Identification of potential conserved RNA secondary structure throughout influenza A coding regions

WALTER N. MOSS,[1] SALVATORE F. PRIORE,[1] and DOUGLAS H. TURNER
Department of Chemistry and Center for RNA Biology, University of Rochester, Rochester, New York 14627-0216, USA

ABSTRACT

Influenza A is a negative sense RNA virus of significant public health concern. While much is understood about the life cycle of the virus, knowledge of RNA secondary structure in influenza A virus is sparse. Predictions of RNA secondary structure can focus experimental efforts. The present study analyzes coding regions of the eight viral genome segments in both the (+) and (–) sense RNA for conserved secondary structure. The predictions are based on identifying regions of unusual thermodynamic stabilities and are correlated with studies of suppression of synonymous codon usage (SSCU). The results indicate that secondary structure is favored in the (+) sense influenza RNA. Twenty regions with putative conserved RNA structure have been identified, including two previously described structured regions. Of these predictions, eight have high thermodynamic stability and SSCU, with five of these corresponding to current annotations (e.g., splice sites), while the remaining 12 are predicted by the thermodynamics alone. Secondary structures with high conservation of base-pairing are proposed within the five regions having known function. A combination of thermodynamics, amino acid and nucleotide sequence comparisons along with SSCU was essential for revealing potential secondary structures.

Keywords: influenza; RNA; secondary structure; structure prediction; codon suppression

INTRODUCTION

Influenza A virus is a significant public health threat causing more than an estimated 200,000 severe infections and 41,400 deaths each year in the United States (Dushoff et al. 2006). The influenza A virus is a negative (−) sense RNA virus composed of eight discrete genomic segments. Each segment is packaged as a ribonucleoprotein (RNP) complex that contains multiple structural NP proteins and the heterotrimeric polymerase consisting of the PB1, PB2, and PA protein subunits (Compans 1972; Noda et al. 2006; Ye et al. 2006). Each genome segment serves as a template for the synthesis of two distinct positive (+) sense RNA molecules within the nucleus of infected cells. The (+)RNAs serve protein coding (mRNA) and genomic replication (cRNA) functions and are generated by distinct mechanisms in vivo (Bouloy et al. 1978; Plotch et al. 1981; Shapiro and Krug 1988).

In general, RNA secondary structure is important for viral viability. For example, internal ribosome entry sites (IRES), which allow mRNA to internally initiate translation and bypass canonical ribosomal scanning, are heavily structured and found in many viruses (Kieft 2008). Other examples are the Hepatitis Delta Virus (HDV) ribozyme that is used for maturation of the viral RNA (Kuo et al. 1988), the tRNA-like structures found in the 3′ untranslated region (UTR) of many plant viruses (Weiner and Maizels 1987; Dreher 2009), the frameshifting signals that allow some viruses to encode overlapping open reading frames (ORFs) (Jacks et al. 1988; Dam et al. 1990), viral packaging signals (Clever et al. 1995), and many more. Recently, widespread secondary structure of an HIV-1 genome was deduced using free energy minimization coupled with chemical probing (Watts et al. 2009).

The de novo discovery of structured regions in long RNA strands, such as viral RNAs, has been approached with a variety of techniques (Washietl et al. 2005a; Schroeder 2009; Mathews et al. 2010). One method is based on searching for regions predicted to be unusually stable thermodynamically (Washietl et al. 2005b; Uzilov et al. 2006). When the search occurs in coding regions, it is possible to also consider the effect of RNA structure on codon evolution (Pedersen et al. 2004). In particular, RNA structural constraints lead to suppression of variation in the third (wobble) position of amino acid codons. Suppression of synonymous codon

[1]These authors contributed equally to this work.

Reprint requests to: Douglas H. Turner, Department of Chemistry and Center for RNA Biology, RC Box 270216, University of Rochester, Rochester, NY 14627-0216, USA; e-mail: turner@chem.rochester.edu; fax: (585) 276-0205.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2619511.

usage (SSCU) has been used to identify structured RNA elements in viral genomic RNAs (Simmonds and Smith 1999; Tuplin et al. 2004). Here, we use a combination of thermodynamics, SSCU, amino acid and RNA sequence comparison to reveal potential secondary structures in influenza.

The influenza virus is an interesting target for structural analysis because it uses RNA exclusively throughout its life cycle; no DNA intermediate is involved. Each (−)RNA carries conserved 5′ and 3′ sequences that can base-pair to circularize the molecule (Hsu et al. 1987). This pairing provides a binding site for the heterotrimeric polymerase that carries out the synthesis of both mRNA and cRNA molecules (Hagen et al. 1994). Additionally, two studies have described structures in the segment 8 (+)RNA that encode the nonstructural (NS1) and nuclear export protein (NEP, formerly NS2) (Gultyaev et al. 2007; Ilyinskii et al. 2009). Segment 8 (+)RNA has been the most extensively studied with regard to secondary structure. RNA secondary and tertiary structures have also been proposed to be important in the splicing of segment 8 mRNA (Plotch and Krug 1986; Nemeroff et al. 1992) and in viral packaging of (−)RNA (Muramoto et al. 2006; Marsh et al. 2007, 2008; Hutchinson et al. 2008; Liang et al. 2008). Outside of segment 8, very little is known about RNA secondary structure.

Improved knowledge of secondary structure in influenza may shed light on important aspects of influenza biology and lead to new therapeutic targets. RNA structural motifs may be targeted with oligonucleotides (Childs et al. 2002, 2003; Disney et al. 2004) or small molecules (Mei et al. 1998; Sucheck and Wong 2000; Wilson and Li 2000; Gallego and Varani 2001; Childs-Disney et al. 2007; Disney et al. 2008; Lee et al. 2009; Pushechnikov et al. 2009) to disrupt viral function.

MATERIALS AND METHODS

Sequences

Sequence data were obtained from the National Center for Biotechnology Information (NCBI) Influenza Virus Resource page (Bao et al. 2008). For prediction of conserved structural RNA, six full influenza A genome sets were used from human, avian, and swine strains, H5N1 and H1N1 (Taxonomy IDs: 755298 [Human H5N1], 279728 [Avian H5N1], 287864 [Swine H5N1], 865618 [Human H1N1], 768723 [Avian H1N1], 762299 [Swine H1N1]). For analysis of synonymous codon suppression, sets of sequences were obtained for each separate segment by downloading all nonredundant influenza A sequences.

Prediction of conserved structural regions

The six influenza A genome sets were divided by segment. Coding regions were translated in silico with BioEdit (Hall 2001), and protein sequences were aligned with ClustalW (Larkin et al. 2007) using the default protein parameters. The aligned sequences were converted back into nucleotides, now aligned based on the protein sequence, and submitted to RNAz 2.0 (Gruber et al. 2010) for prediction of potentially conserved structures. It was important to use protein-based alignments, as the quality was much improved over nucleotide alignments. The amino acid alignments also allowed analysis of how RNA structure may influence codon evolution.

RNAz predictions were run in both strand orientations (+/−) using a 120-nt window size, 10-nt step size, and with the program's default filtering parameters. RNAz uses a support vector machine (SVM) to make predictions based on the following five criteria: minimum predicted free energy (MFE) from single sequence structure calculations, *Z*-score, structure conservation index (SCI), average pairwise sequence identity (APSI), and number of sequences in the alignment. The *Z*-score measures the "excess" predicted minimum free energy of folding for a native sequence versus random sequence. A dinucleotide shuffling model for calculating *Z*-scores was used that reduces background over the default mononucleotide model setting. The SCI is the consensus structure free energy divided by the average of the individual sequence free energies in the alignment and is a measure of how well represented the consensus structure is in individual sequence folds. Based on the above criteria, RNAz assigns a classification value, herein referred to as p-class, to indicate the probability that a given region contains structure. For this study, RNAz predictions with a p-class of >0.5 are considered structured. RNAz was trained on representative sequences including rRNAs, spliceosomal RNAs, tRNAs, miRNAs, small nucleolar RNAs, nuclear RNase P, and SRP RNA (Gruber et al. 2010).

Structural predictions in both the (+)RNA and (−)RNA strand orientation can arise for the same window because base-pairing is also possible in the reverse complement sequence of the strand that contains the true or functionally conserved structure (Reiche and Stadler 2007). To check predictions for these structural "echoes," fragments corresponding to overlapping RNAz hits were concatenated and submitted to the program RNAstrand (Reiche and Stadler 2007), which uses an SVM to predict strand bias. RNAstrand uses four criteria for identifying asymmetries between structure in the (+)RNA and (−)RNA strand orientations: average folding free energy of individual sequences in the alignment, folding free energy of the consensus secondary structure as calculated by RNAalifold (Bernhart et al. 2008), mean free energy *Z*-score of the individual sequences in the alignment, and SCI. The resulting "*P*-value" ranges from 0 to 1, where 1 implies high likelihood for structure in the given strand orientation. The RNAstrand SVM was trained on a similar set of structures as RNAz (Reiche and Stadler 2007).

Analysis of suppression of synonymous codon usage

All nonredundant sequences for each segment were aligned using MAFFT (Katoh et al. 2002, 2005) with the FFT-NS-1 strategy optimized for very large alignments. A Perl script was written to randomly select sequences from the alignment while simultaneously restricting the APSI to <95% to avoid selecting very similar sequences. For segments 4 and 6, the two segments encoding the antigenic proteins, 300 sequences were selected, while 100 sequences were used for all other segments. In each case, the larger sequence sets were appended to the alignments used in the RNAz analysis. Gaps were removed from the sequences, which were then translated in silico and submitted to ClustalW as above.

The resulting amino acid alignments were converted back into RNA sequences (now aligned with respect to encoded amino

acids) and submitted to the Simmonics package (Simmonds and Smith 1999) for analysis of the suppression of synonymous codon usage (SSCU). Sequence scans were run analyzing the synonymous sites using mean pairwise distance measurements. The SSCU was calculated for windows of 15 nt, with a 3-nt step size. In segments 7 and 8, the calculation of SSCU was done separately for each alternatively spliced product. This analysis was repeated with resampled alignments to check for sampling bias, and results were not significantly different.

Secondary structure modeling

Preliminary structural models were built when RNAz or strong SSCU predictions correlated with functional annotations. Initial models were built with RNAalifold (Bernhart et al. 2008), the same algorithm used to calculate the free energies in RNAz and RNAstrand. RNAalifold was run with the default program parameters. The alignments submitted to RNAalifold contained the six genome sequences used for the RNAz analysis or the alignments used for the SSCU analysis. The RNAalifold secondary structure predictions were compared to predictions from Dynalign (Mathews and Turner 2002; Mathews 2004; Harmanci et al. 2007), which simultaneously optimizes sequence alignment and consensus structure for two sequences. The Dynalign calculation input consisted of the two most distant sequences (by APSI) in the alignment. Sequence comparison for each secondary structure model was carried out with all available unique sequences. Free energies for non-pseudoknot structures were predicted using nearest-neighbor free energy parameters (Xia et al. 1998; Mathews et al. 2004).

To scan for potential pseudoknots, which are forbidden in the RNAz and Dynalign prediction algorithms, the program DotKnot was used (Sperschneider and Datta 2010). DotKnot is a heuristic algorithm that searches for stems in long RNAs from a secondary structure prediction dotplot and then assembles candidate pseudoknots (Sperschneider and Datta 2010). When RNAs were >1000 nt in length, the maximum length allowed in DotKnot, they were cut into overlapping windows (~100 nt overlap) and submitted to the program. Free energies for pseudoknots were predicted using published parameters optimized either by comparison of predictions to known structures (Dirks and Pierce 2003) or to results from a diamond lattice model (Cao and Chen 2006, 2009).

In vitro folding of RNA

Representative cluster a (5′-GGGUGAUGCCCCAUUCCUUGAUC GGCUUCGCCGAGA UCAGAAGUCCCUAAGAGGAAGAGGCAG CACUC-3′) and cluster b (5′-GGGUGAUGCUC CCUUUGAUGAC AGACUCAGAAGAGAUCAAAAGGCAUUAAAGGGAAGAGGCAG CACUC-3′) sequences for segment 8 region 81–148 were synthesized from deoxyoligonucleotide templates with an Ambion MEGAscript T7 transcription kit. Purified RNA was 5′-end-labeled with [γ-^{32}P]ATP and purified with an Ambion NucAway spin column. Labeled RNA was renatured in 10 mM Tris-HCl (pH 7.0) and 100 mM KCl by heating for 2 min to 90°C and then slow-cooling to 37°C. $MgCl_2$ was added to a final concentration of 5, 10, or 15 mM, and the RNA was incubated for 20 min at 37°C. Folded RNAs were fractionated on a nondenaturing 8% polyacrylamide gel. The dried gel was exposed to a phosphorscreen, and bands were detected with a Bio-Rad Personal Molecular Imager.

RESULTS

Evidence for conserved RNA secondary structure in influenza virus genome segments

Segment 8 (NS1/NEP)

Segment 8 is the smallest influenza RNA and one of the two (+)RNA segments that undergoes splicing. The alignment of six sequences used in the RNAz analysis has a length of 838 nt and APSI of 86.4%. Of all the influenza segments, segment 8 has the most widespread distribution of RNAz predictions of conserved secondary structure, which is consistent with its having the most negative average *Z*-score across the alignment [−0.95 in the (+)RNA sense] (Table 1). It also has the highest average RNAz p-class of any segment [0.30 in the (+)RNA sense]. Thus, segment 8 has the strongest bias toward predicted structure in the (+)RNA (Tables 1, 2).

Three regions are predicted to contain conserved secondary structure in segment 8 (+)RNA (Fig. 1; Table 2). Region 371–690 has the most favorable single window *Z*-score, p-class, and SSCU of any predicted region in any segment (−3.70, 1.00, and 0.00, respectively). This region also contains the 3′ splice site at position 487. After 487, the codon use in the NS1 ORF is severely suppressed because of overlap with the frameshifted NS2 ORF. Codon suppression in the NS2 ORF starts to abate toward the end of the structural region at 371–690 (Fig. 1).

Regions 21–180 and 181–300 of segment 8 have moderate values for *Z*-score, SCI, p-class, and SSCU (Table 2). Region 21–180, however, has individual windows with values that strongly predict structure. The fifth most favorable local *Z*-score occurred in 21–180. Strong SSCU occurs near positions 60, 150, and 550. The 5′ splice site of segment 8 occurs at position 30.

Segment 7 (M1/M2)

Segment 7 is the second smallest viral segment and is also spliced to produce two protein products. The alignment

TABLE 1. Average values of metrics used for structured RNA prediction

	(−)RNA			(+)RNA			
	Z-score	SCI	p-class	*Z*-score	SCI	p-class	SSCU
SEG8	−0.33	0.59	0.12	−0.95	0.65	0.30	0.33
SEG7	−0.22	0.80	0.16	−0.55	0.78	0.15	0.39
SEG6	−0.03	0.66	0.01	−0.22	0.71	0.04	0.55
SEG5	−0.35	0.64	0.03	−0.44	0.64	0.07	0.61
SEG4	−0.24	0.21	0.00	−0.29	0.22	0.01	0.57
SEG3	0.17	0.79	0.06	0.18	0.78	0.03	0.56
SEG2	−0.06	0.67	0.02	−0.10	0.70	0.02	0.49
SEG1	−0.18	0.65	0.03	−0.16	0.62	0.02	0.52
Average	*−0.16*	*0.63*	*0.05*	*−0.32*	*0.64*	*0.08*	0.50
StdDev	*0.17*	*0.18*	*0.06*	*0.34*	*0.18*	*0.10*	0.10

TABLE 2. Summary of RNAz and SSCU scans for predicted structural region

Segment	Strand[a]	Region	*Z*-score[b,c]	SCI[b,c]	p-class[b,c]	SSCU[b]
8	+/− (0.00)	21–180	−1.84 (−2.39)	0.78 (0.87)	0.69 (0.92)	0.26 (0.10)
8	+ (1.00)	181–300	−1.41	0.62	0.63	0.42 (0.33)
8	+ (1.00)	371–690	−2.27 (−3.70)	0.77 (0.94)	0.74 (1.00)	0.21 (0.00)
7	+ (0.91)	71–330	−1.70 (−2.81)	0.84 (0.92)	0.60 (0.99)	0.40 (0.02)
7	+ (1.00)	841–990	−1.58 (−1.79)	0.86 (0.89)	0.56 (0.78)	0.20 (0.03)
6	+ (1.00)	531–670	−1.31 (−2.05)	0.90 (0.92)	0.58 (0.94)	0.52 (0.39)
5	+ (1.00)	1–160	−1.98 (−3.01)	0.83 (0.88)	0.81 (0.97)	0.56 (0.23)
5	+/− (0.22)	441–560	−0.36	0.67	0.00	0.68 (0.54)
			−2.05	*0.78*	*0.89*	
5	+ (0.52)	1031–1250	−1.32 (−2.85)	0.65 (0.74)	0.20 (0.81)	0.59 (0.43)
5	+ (0.98)	1381–1494	−1.64 (−1.83)	0.91 (0.93)	0.86 (0.95)	0.19 (0.05)
4	+/− (0.00)	961–1080	−2.30	0.28	0.54	0.62 (0.53)
3	+/− (0.15)	41–290	−0.28 (−1.07)	0.73 (0.88)	0.00 (0.02)	0.55 (0.34)
			−1.22 (−2.11)	*0.93 (1.00)*	*0.63 (0.99)*	
3	+/− (0.00)	1161–1280	−0.91	0.64	0.01	0.57 (0.43)
			−1.32	*0.84*	*0.59*	
3	+ (1.00)	1611–1860	−1.06	0.87	0.29	0.63 (0.44)
			−1.50	*0.84*	*0.53*	
3	+ (1.00)	1941–2120	−1.23 (−1.88)	0.97 (1.00)	0.69 (0.96)	0.33 (0.09)
2	+/− (0.31)	51–170	−2.25	0.92	0.96	0.33 (0.15)
			−1.44	*0.89*	*0.58*	
2	+/− (0.00)	491–610	−0.75	0.68	0.00	0.52 (0.38)
			−2.38	*0.85*	*0.97*	
1	+/− (0.06)	831–960	−1.45 (−1.49)	0.73 (0.73)	0.22 (0.25)	0.50 (0.33)
			−1.81 (−1.83)	*0.83 (0.87)*	*0.83 (0.90)*	
1	+/− (0.00)	1041–1160	−0.46	0.59	0.00	0.51 (0.32)
			−1.36	*0.84*	*0.52*	
1	+ (1.00)	2101–2220	−1.68	0.88	0.71	0.35 (0.07)

[a]Here +/− indicates regions where predicted conserved structure favors the (−)RNA but with low RNAstrand (Reiche and Stadler 2007) probability. [RNAstrand *P*-values for (+) and (+/−) regions are in parentheses.]
[b]Average values across overlapping windows are shown. The most favorable value for a window in a region, when not equal to the average, is shown within parentheses.
[c]Italicized values are for the (−)RNA calculations and appear beneath the (+)RNA results.

used in the RNAz analysis has a length of 982 nt and an APSI of 92.5%. As shown for (+)RNA in Table 1, segment 7 has the second most favorable average *Z*-score (−0.55) and the second strongest average SSCU and RNAz p-class (0.39 and 0.15), indicating a high overall probability of structure in the (+)RNA (Table 1).

RNAz predicts a high probability of structure in two regions of segment 7 (Fig. 1; Table 2). Regions 71–330 and 841–990 have a high RNAstrand probability for structure in the (+)RNA. Region 841–990 is notable for having, respectively, the second and third strongest average and single window SSCU with respect to all other segments. The SSCU that overlaps with this region continues upstream and also overlaps with the nearby 3′ splice site at position 715 (Fig. 1).

Region 71–330 has moderate values for *Z*-score, SCI, p-class, and SSCU, but single window values are very strong (Table 2). This region has the second strongest single window p-class and SSCU value of any region. This predicted structural region occurs just downstream from the 5′ RNA splice site at position 26, which falls within the region of suppressed synonymous codon usage (Fig. 1).

Segment 6 (NA)

The alignment for segment 6 is 1460 nt in length and has an APSI of 86.7%. As shown in Table 1, segment 6 has the third highest average SCI for the (+)RNA, but the *Z*-score of −0.22 averaged over the entire coding region predicts a roughly average amount of structure for segment 6 when compared to the other segments.

RNAz predicts a moderate probability of structure in region 531–670, but the single window 551–670 has favorable values for *Z*-score, SCI, and p-class (Fig. 2; Table 2). The RNAz predictions are for structure in the (+)RNA, consistent with RNAstrand predictions (Fig. 2; Table 2). Synonymous codon usage appears to be highly suppressed only at positions 100–250, which does not overlap with the RNAz predicted conserved structural regions (Fig. 2).

Segment 5 (NP)

The alignment for segment 5 is 1494 nt in length and has an APSI of 88.1%. Average *Z*-score across the alignment is −0.35 and −0.44 for the (−)RNAs and (+)RNAs, respectively, suggesting an above average amount of

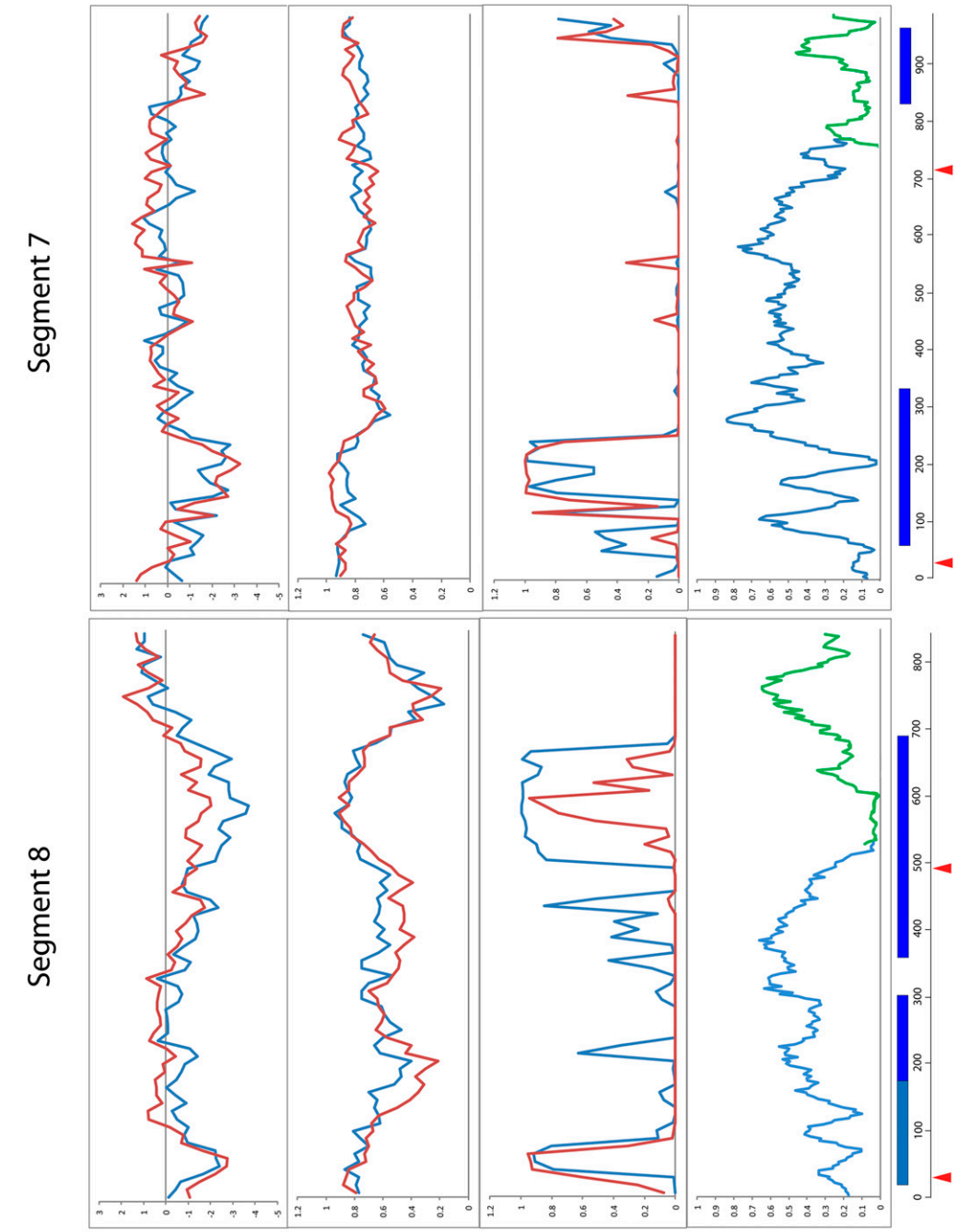


FIGURE 1. Results of RNAz calculations and suppression of synonymous codon usage (SSCU) studies for coding regions of segments 8 and 7. (Red lines) (−)RNA; (blue lines) (+)RNA. The first plot gives the calculated *Z*-score, which is a measure of the "excess" free energy of folding a native RNA sequence vs. random sequence. The second plot gives the structure conservation index (SCI), which indicates how well represented the consensus structure is in predictions of individual sequence secondary structures (*y*-axis indicates the fraction of conservation). The third plot shows the RNAz probability (p-class) of the presence of conserved structural RNA. The *bottom* panel shows the results for the SSCU calculations, which measure the variation at the third codon position (*y*-axis gives the distance at the third codon position). Here, low distance/variation indicates strong SSCU. Results for the larger ORF (blue) and the smaller (green). (Below the *bottom* panel) The common *x*-axis indicates the input alignment position in nucleotides. (Dark blue bars) Overlapping RNAz predictions clearly in the (+) sense; (light blue bars) RNAz predictions with ambiguous strand bias; (red arrows) the splice sites.

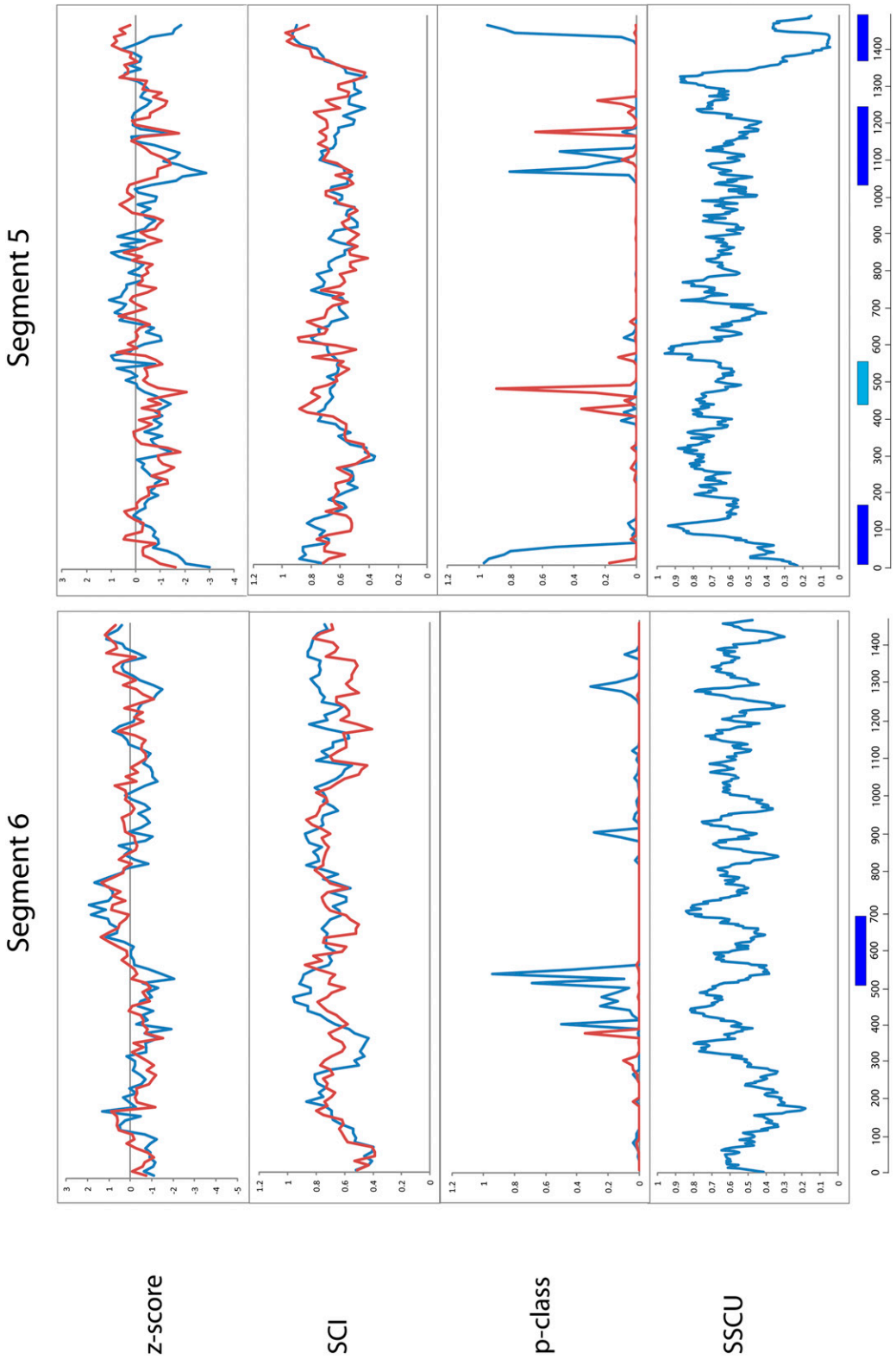


FIGURE 2. Results of RNAz calculations and suppression of synonymous codon usage (SSCU) studies for segments 6 and 5. Figure annotations are as in Figure 1.

structure in segment 5 compared to the other segments (Fig. 2; Table 1).

In segment 5, RNAz predicts a high probability of structure solely in (+)RNA at regions 1–160, 1031–1250, and 1381–1494, and of ambiguous RNAstrand predicted strand bias at position 441–560 (Fig. 2; Table 2). Region 1–160 has the fourth most favorable average *Z*-score of any (+)RNA region (Table 2).

Toward the ends of segment 5, there are two regions with strong SSCU that match up with the two high-probability RNAz predicted regions of structure in the (+)RNA. The region of SSCU that overlaps with region 1381–1494 is, on average, the strongest for any region (Table 2).

Segment 4 (HA)

The alignment for segment 4 is 1770 nt in length and has an APSI of 73.4%. On average, segment 4 has the lowest SCI of any of the influenza segments in both strands (0.21 and 0.22) (Fig. 3; Table 1). Average *Z*-scores of −0.24 and −0.29 in the (−)RNA and (+)RNA, respectively, suggest a slightly above average amount of structure in segment 4 compared to the other segments (Table 1).

In segment 4, RNAz predicts a single conserved structural region at 961–1080 in the (+)RNA (Fig. 3), but RNAstrand favors structure in the (−)RNA with very low confidence (Table 2).

There is below average SSCU corresponding to the RNAz prediction at 961–1080 (Fig. 3). There is also moderate SSCU toward the 5′ end of the RNA, which does not correspond to any RNAz predictions (Fig. 3).

Segment 3 (PA)

The alignment for segment 3 is 2151 nt in length and has an APSI of 91.9%. The positive *Z*-scores of 0.17 and 0.18 for the (−)RNA and (+)RNA, respectively, suggest a lack of overall conserved structure in both strands (Table 1).

In segment 3, however, RNAz predicts four regions with potentially conserved structure (Fig. 3). Two regions are predicted by RNAstrand to have bias for structure in the (+)RNA, 1611–1860, and 1941–2120, while regions 41–290 and 1161–1280 are predicted to have ambiguous strand bias (Table 2).

Region 1941–2120 has the highest average SCI of any (+)RNA region, a high average and single window p-class, and strong SSCU (Table 2). Another region, ∼500–800, has strong SSCU but does not occur near any RNAz predictions of structure or any known features of the viral RNA.

Segment 2 (PB1/PB1-F2)

The alignment for segment 2 is 2151 nt in length and has an APSI of 89.3%. The average *Z*-score for segment 2 in the (+)RNA is −0.10 (Table 1), suggesting a below average amount of conserved structure compared to the other segments. Nevertheless, RNAz predicts structure in two regions, 51–170 and 491–610: RNAstrand predictions are ambiguous in both regions (Fig. 4; Table 2).

Region 51–170 has strong RNAz scores in the (+)RNA, while region 491–610 has strong RNAz scores in the (−)RNA. Segment 2 has two regions with strong SSCU (Fig. 4). One region spans positions 50–150, which corresponds to both the RNAz-predicted region of structure at 51–170 and the start of the internal ORF for PB1-F2 at position 91 (Fig. 4). The second region with SSCU is at the 3′ end of the RNA and does not overlap with known or predicted features of the RNA.

Segment 1 (PB2)

The alignment for segment 1 is 2280 nt in length and has an APSI of 88.8%. The average *Z*-scores of −0.18 and −0.16 for the (−)RNA and (+)RNA, respectively, suggest the presence of an average amount of conserved structure in segment 1 compared to the other segments (Table 1).

In segment 1, RNAz predicts three regions with potentially conserved secondary structure: 831–960, 1041–1160, and 2101–2220 (Fig. 4). Only region 2101–2220 has unambiguous strand bias, as predicted by RNAstrand (Table 2). This region is also the only one to overlap with an area of strong SSCU (Fig. 4; Table 2). Toward the 5′ end of the RNA is another region that shows SSCU (Fig. 4).

Comparisons to previous predictions of RNA structures

Two of the RNAz-predicted regions of conserved structure in segment 8 contain previously predicted secondary structures (Gultyaev et al. 2007; Ilyinskii et al. 2009). Region 21–180 contains a fragment that has been proposed to fold into a structure that influences NS protein expression (Ilyinskii et al. 2009). Using the RNAz alignment, RNAalifold predicts a consensus structure for this region similar to the published one (Fig. 5). Using the SSCU alignment for RNAalifold, however, results in a different predicted secondary structure (Fig. 6). The two models are identical at the basal stem, but the nucleotides between positions 90 and 139 are folded differently (Figs. 5, 6). In the alternative model of Figure 6, the multi-branch loop from nucleotides 100–130 (Fig. 5) is folded into a tetraloop hairpin. The tetraloop model has less mutational support, a slightly lower (81.3% vs. 83.6%) conservation of canonical base-pairing, and much less favorable predicted free energy (−9.6 vs. −19.7 kcal/mol at 37°C) than the multi-branch loop in Figure 5. A similar tetraloop structure has been proposed for clade B influenza strains (Gultyaev et al. 2010).

To test whether a subset of sequences prefers the tetraloop structure, sequences were clustered using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal and Michener 1958). All unique sequences clustered into

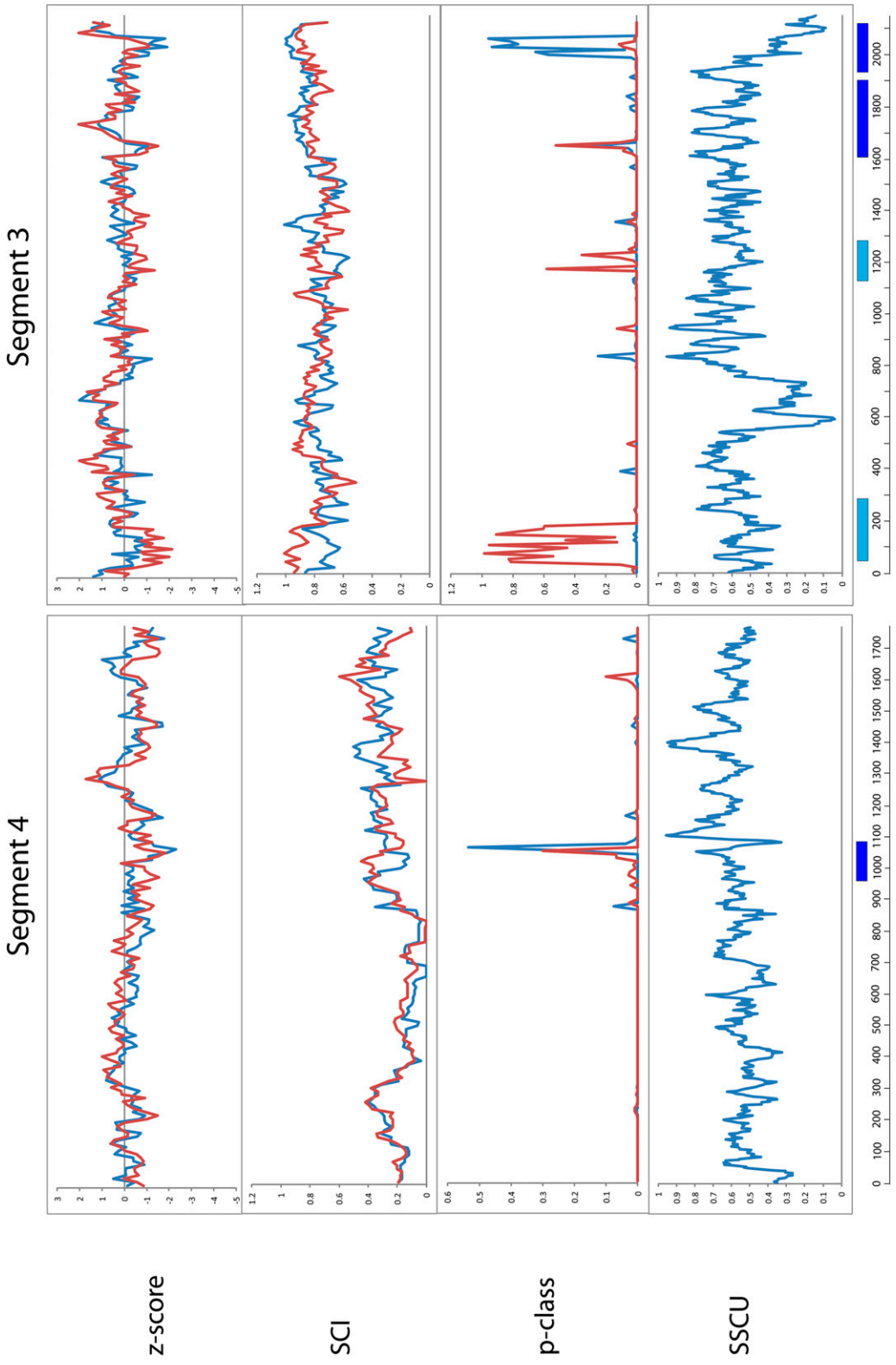


FIGURE 3. Results of RNAz calculations and suppression of synonymous codon usage (SSCU) studies for segments 4 and 3. Figure annotations are as in Figure 1.

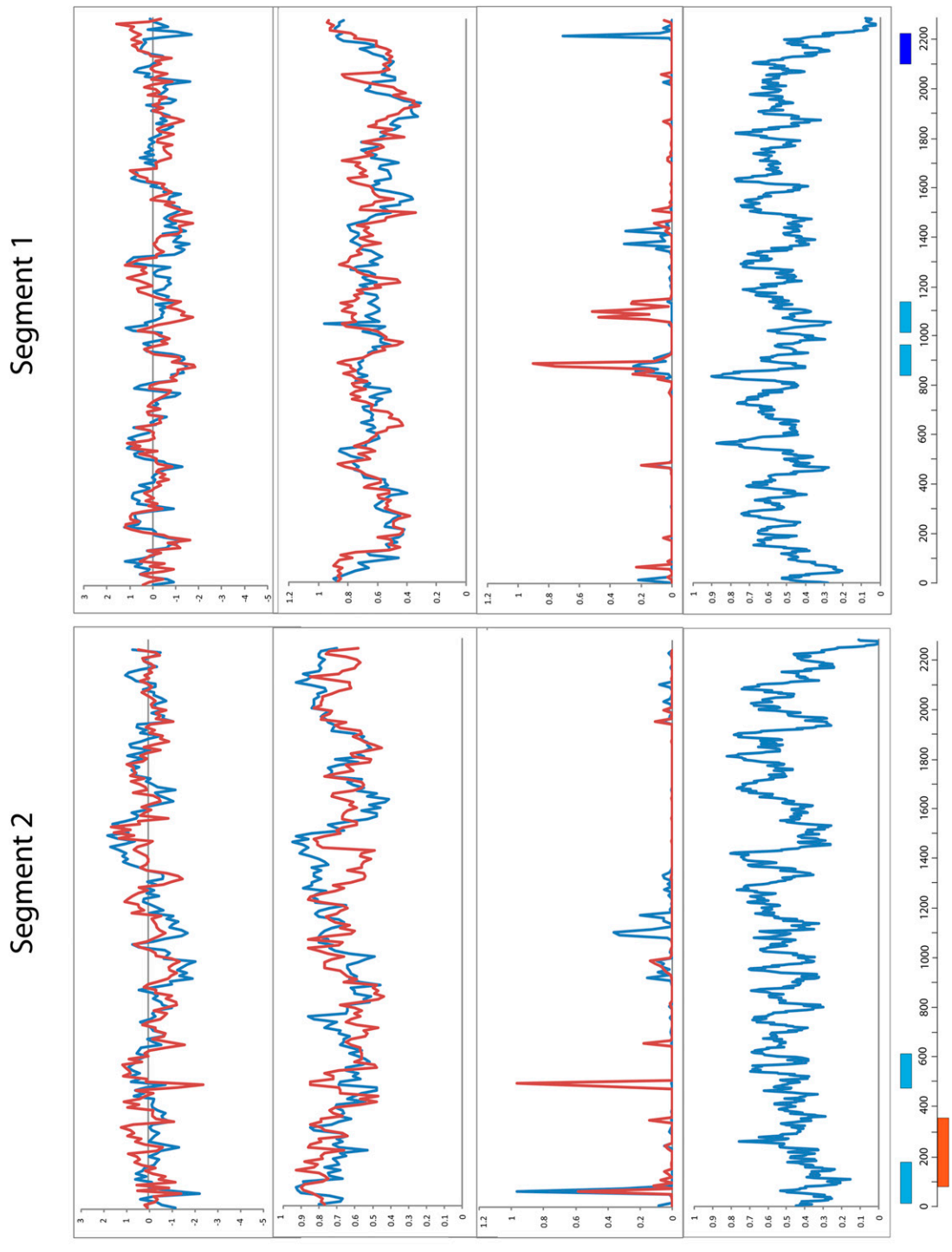
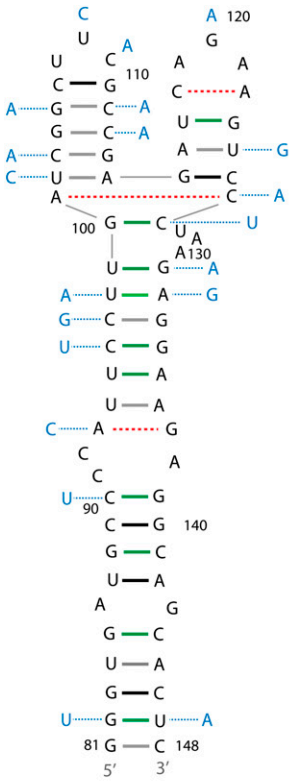


FIGURE 4. Results of RNAz calculations and suppression of synonymous codon usage (SSCU) studies for segments 2 and 1. Figure annotations are as in Figure 1; (orange bar) the internal ORF for the PB1-F2 product.



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can	%can b
81	148	792(110)	0	0	1	2	0	12	0	0	0	0	1	87	47	0	0	84.4	100.0
82	147	136	0	92	110(110)	590	0	8	0	0	0	2	0	4	0	0	1	98.4	100.0
83	146	936(110)	0	0	0	2	0	2	0	0	0	0	0	2	1	0	0	99.5	100.0
84	145	0	0	0	839(109)	0	6	3	0	5	0	1	0	0	1	87(1)	0	89.7	99.1
85	144	878(110)	0	0	0	54	0	9	0	0	1	0	0	0	0	0	0	98.9	100.0
86	143	0	0	30(1)	0	0	0	1	674(104)	236(5)	0	0	0	1	1	0	0	3.2	0.9
87	142	0	0	0	923(110)	0	4	0	0	0	0	2	0	0	1	13	0	98.3	100.0
88	141	876(107)	0	0	0	56(3)	0	5	0	0	1	0	0	2	3	0	0	98.8	100.0
89	140	0	937(110)	0	0	0	4	0	0	0	0	0	2	0	0	0	0	99.8	100.0
90	139	0	735(9)	0	27(16)	0	103(82)	0	8(2)	2	0	1(1)	0	0	0	66	1	91.7	97.3
93	137	0	153(109)	1	0	0	12(1)	0	739	1	35	0	0	0	0	2	0	17.6	100.0
94	136	0	0	0	923(106)	0	3	1	0	0	0	0	0	0	12(2)	4(2)	0	98.2	96.4
95	135	0	0	0	869(100)	0	71(9)	0	0	0	0	1(1)	0	0	0	2	0	99.7	99.1
96	134	0	679	0	0	0	262(110)	0	0	0	0	0	0	0	0	1	0	99.9	100.0
97	133	0	808	0	0	0	4	1(1)	13(4)	0	107(105)	0	0	0	0	10	0	86.1	0.0
98	132	0	3	0	552	0	271(1)	1	109(107)	1(1)	2(1)	1	0	0	0	3	0	87.6	0.9
99	131	0	78	0	305(109)	0	436(1)	2	78	32	4	4	0	0	0	4	0	86.9	100.0
100	127	743(1)	0	0	0	195(109)	0	4	0	0	0	0	0	0	1	0	0	99.5	100.0
101	126	2	0	183	0	0	0	0	19(2)	133(108)	0	0	1	602	3	0	0	19.6	0.0
102	114	0	67(2)	0	437(48)	0	20(2)	2	0	0	0	1	1	0	0	402(58)	13	55.6	47.3
103	113	0	803	0	0	0	1	0	136(109)	1(1)	0	0	1	0	0	0	0	85.7	0.0
104	112	807(1)	0	0	0	1	0	131(108)	0	0	2(1)	0	0	0	2	0	0	83.2	0.9
105	111	765	0	1	0	19	0	38(36)	12(12)	58(58)	4(4)	0	2	39	4	0	1	87.4	0.0
106	110	0	906(107)	0	0	0	7	0	21(1)	1(1)	4	0	1	0	0	3(1)	0	13	97.3
115	125	937(110)	0	0	0	0	0	0	0	0	1	0	2	1	1	0	0	85.4	100.0
116	124	0	0	823	0	1	0	0	115(110)	1	0	0	0	3	0	0	0	96.8	0.0
117	123	0	133(38)	0	145(10)	0	638(59)	0	1	1	0	0	0	0	0	25(3)	0	99.5	97.3
118	122	0	123(14)	0	0	0	0	0	1	3	0	0	0	0	0	816(96)	0	97.1	12.7

Canonical pair (>98%).
Canonical pair, but most frequent change is to non-canonical pair.
Non-canonical pair, but most frequent change is to a canonical pair.
Canonical pair with most frequent change to a consistent pair.
Canonical pair with most frequent change to a compensatory pair.

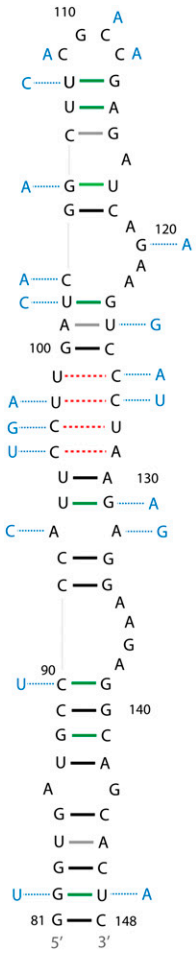
Average: 83.6% (68.5%)

AA	Leu	Gly			Asp			Ala			Pro			Phe			Leu			Asp			Arg			Leu			Arg			Arg		
codon	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii
nt	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114
A	87	96	2	5	0	941	0	2	0	10	1	0	741	0	0	0	13	110	110	0	937	0	137	0	110	22	0	12	84	4	96	131	1	841
G	806	734	940	3	942	1	0	938	0	0	0	0	35	1	0	0	108	3	6	942	2	2	0	941	826	4	1	10	0	938	16	2	941	87
C	1	1	0	87	0	0	13	0	939	802	942	943	155	4	2	680	818	6	82	0	1	483	804	0	3	910	1	262	854	1	810	809	1	1
U	48	112	1	847	0	1	930	3	4	131	0	0	12	938	941	262	4	824	745	1	3	458	1	2	4	7	940	659	5	0	21	1	0	14
MBL Str	(	(	(	(	(	.	(	(	(	(	.	.	.	(	(	(	(	(	(	(	.	(	(	(	(	(	.	.	.	)	)	)	)	)
TL STR	(	(	(	(	(	.	(	(	(	(	(	(	.	(	(	.	.	.	.	(	(	(	.	(	(	(	(	(	.	.	.	.	)	)
%cons	0.86	0.78	1.00	0.90	1.00	1.00	0.99	0.99	1.00	0.85	1.00	1.00	0.79	0.99	1.00	0.72	0.87	0.87	0.79	1.00	0.99	0.51	0.85	1.00	0.88	0.97	1.00	0.70	0.91	0.99	0.86	0.86	1.00	0.89

AA	Asp			Glu			Lys			Try			Leu			Arg			Gly			Arg			Gly			Ser			Thr			Leu
codon	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i
nt	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148
A	1	942	2	4	933	309	942	819	171	1	0	133	4	0	800	930	343	557	11	1	871	928	3	866	95	0	5	936	237	9	935	2	118	13
G	938	1	0	0	8	629	0	124	772	115	1	19	0	0	139	11	596	385	932	942	71	3	939	76	846	941	1	4	674	1	6	0	0	0
C	2	0	158	939	0	2	1	0	0	3	942	608	744	2	1	2	0	0	0	0	0	12	0	0	0	2	881	1	2	879	1	939	140	928
U	1	0	783	0	1	3	0	0	0	824	0	183	195	941	3	0	4	1	0	0	1	0	1	1	2	0	56	2	30	54	1	2	685	2
MBL Str	(	(	(	.	.	.	.	.	)	)	)	.	)	.	.	.	)	)	)	)	)	)	.	.	)	)	)	)	.	)	)	)	)	)
TL STR	)	.	)	)	.	.	.	.	)	)	)	.	.	.	.	)	)	.	)	)	.	.	.	.	)	)	)	)	.	)	)	)	)	)
%cons	1.00	1.00	0.83	1.00	0.99	0.67	1.00	0.87	0.82	0.87	1.00	0.64	0.79	1.00	0.85	0.99	0.63	0.59	0.99	1.00	0.92	0.98	1.00	0.92	0.90	1.00	0.93	0.99	0.71	0.93	0.99	1.00	0.73	0.98

FIGURE 5. RNAalifold predicted secondary structure from the RNAz alignment for fragment of 5′ predicted secondary structure region from segment 8 (+)RNA. This structure was also predicted by Ilyinskii et al. (2009). Base pairs are color annotated with information from base pair counts (tabulated to the *right* of the structure) from an alignment of all available unique sequences. The color annotation key is given below the table. Pairing type is given at the *top* of the table; canonical pairs to the *left*, and noncanonical to the *right*. The "%can" column gives the percentage of canonical pairs found in those aligned positions. Italicized alignment positions (*i–j*) are for symmetric internal loop bases. The average percent conservation of the whole structure is given *below* the table. Base pair counts for all unique sequences and cluster b sequences are given without and with parenthesis, respectively. Cluster b consensus sequence is indicated by light blue nucleotides. The predicted free energies of folding, $\Delta G_{37}°$ (Mathews et al. 2004), for the consensus sequence of all unique sequences is −19.7 kcal/mol and for cluster b sequences is −8.6 kcal/mol. Nucleotide composition by alignment position is summarized at the *bottom* of the figure. The structure is notated in bracket notion, codon position is indicated by roman numerals, and consensus amino acid sequence is notated at the *top* of the table. The percent conservation for each position is also given.

two broad groups consistent with the previously described phylogeny for segment 8 (Kawaoka et al. 1998; Basler et al. 2001). As shown in the table in Figure 6, a cluster of 110 sequences (cluster b) increases the average canonical pairing of the tetraloop structure to 98.8%, much higher than the 68.5% for cluster b in the multi-branch loop structure (see table in Fig. 5). Specifically, cluster b has canonical base pairs at positions 93–132, 96–129, 97–128, 98–127, and 99–126. For cluster b, the predicted free energies for the multi-branch loop versus the tetraloop structure are −8.6 and −23.8 kcal/mol at 37°C, respectively. The results provide strong support for the tetraloop structure in cluster b sequences. The limited understanding of all factors affecting in vivo folding of RNA, however, leaves both the multi-branch and tetraloop



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can	%can b
81	**148**	792(110)	0	0	1	2	0	12	0	0	0	0	1	87	47	0	0	84.4	100.0
82	**147**	136	0	92	110(110)	590	0	8	0	0	0	2	0	4	0	0	1	98.4	100.0
83	**146**	936(110)	0	0	0	2	0	2	0	0	0	0	0	2	1	0	0	99.5	100.0
84	145	0	0	0	839(109)	0	6	3	0	5	0	1	0	0	1	87(1)	0	89.7	99.1
85	**144**	878(110)	0	0	0	54	0	9	0	0	1	0	0	0	0	0	0	98.9	100.0
86	143	0	0	30(1)	0	0	0	1	674(104)	236(5)	0	0	0	1	1	0	0	3.2	0.9
87	**142**	0	0	0	923(110)	0	4	0	0	0	0	2	0	0	1	13	0	98.3	100.0
88	**141**	876(107)	0	0	0	569(3)	0	5	0	0	1	0	0	2	3	0	0	98.8	100.0
89	**140**	0	937(110)	0	0	0	4	0	0	0	0	0	2	0	0	0	0	99.8	100.0
90	**139**	0	735(9)	0	27(16)	0	103(82)	0	8(2)	2	0	1(1)	0	0	0	66	1	91.7	97.3
91	**134**	0	941(109)	0	0	0	0	0	1(1)	0	0	0	0	0	0	1	0	99.8	99.1
92	**133**	0	932(109)	0	0	0	0	0	0	0	0	0	0	0	0	11(1)	0	98.8	99.1
93	132	0	144(108)	1	8	0	4(1)	24	226	514	11	0	0	0	0	11(1)	0	16.6	99.1
94	**131**	0	0	0	339(107)	0	595(1)	0	0	0	1	4	0	0	0	4(2)	0	99.0	98.2
95	**130**	0	0	0	928(110)	0	11	0	0	0	0	0	0	0	2	2	0	99.6	100.0
96	129	0	74	0	197(107)	0	65(3)	0	0	0	0	0	1	0	0	602	3	35.7	100.0
97	128	0	0	13(4)	0	108(106)	0	0	0	0	0	4	2	0	0	0	816	12.8	100.0
98	127	2	0	107(107)	4	1(1)	0	0	0	0	0	87(1)	6	3(1)	733	0	0	12.1	98.2
99	126	1	1	9	116(108)	4	18(2)	1	0	13	0	151	59	88	460	3	19	15.8	100.0
100	**125**	941(110)	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	99.8	100.0
101	124	0	0	818	0	2	0	0	115(110)	1	0	3	0	3	0	0	1	87.0	0.0
102	**123**	0	374(54)	0	61(7)	0	397(43)	1	0	0	1	0	0	0	0	109	0	88.2	94.5
104	**118**	937(110)	0	0	0	0	0	4	0	0	0	0	0	0	2	0	0	99.4	100.0
105	**117**	140(31)	0	93(60)	0	684(9)	0	2	0	0	0	4	1	17(10)	0	0	2	97.2	90.9
106	115	0	905(108)	0	0	0	7	0	22(2)	0	4	0	2	0	0	1	1	96.8	98.2
107	**114**	0	0	0	838(106)	0	87(4)	1	0	0	0	14	0	0	1	1	0	98.2	100.0
108	**113**	0	261(101)	0	0	0	658(6)	0	12(2)	0	10	0	0	0	1	1(1)	0	97.5	97.3

Average: 81.3% (98.8%)

FIGURE 6. Alternative secondary structure for the region shown in Figure 5. This was predicted by RNAalifold using the SSCU alignment of segment 8 sequences. Figure annotations and base pair counts are as described in Figure 5. Base pair counts for all unique sequences and cluster b sequences are given without and with parentheses, respectively. Cluster b consensus sequence is indicated by light blue nucleotides. The predicted free energies of folding, $\Delta G_{37}°$ (Mathews et al. 2004) for the consensus sequence of all unique sequences is −9.6 kcal/mol and for cluster b sequences is −23.8 kcal/mol.

structures as viable alternatives for all influenza sequences. To test whether cluster a and b sequences adopt different conformations, in vitro–synthesized RNA was folded under different Mg^{++} concentrations at 37°C. As shown in Figure 7, cluster b sequences migrate distinctly from cluster a and appear to adopt two different folds.

The other previously described structure in segment 8 occurs in the 440–690 region, which encompasses the 3′ splice site. The RNAalifold predicted consensus structure for this region contains a hairpin identical to previous reports (Fig. 8, top; Gultyaev et al. 2007). The predicted secondary structure of this fragment did not change with the SSCU alignment and is also predicted by Dynalign. The overall conservation of canonical base-pairing in this hairpin is 92.8%, and the predicted free energy at 37°C is −18.9 kcal/mol. An alternative pseudoknot conformation has been proposed for this region with average canonical pairing of 95.3% (Fig. 8, bottom; Gultyaev et al. 2007). The predicted free energy for the pseudoknot is −9 kcal/mol at 37°C, using parameters from Dirks and Pierce (2003) or Cao and Chen (2009). Current knowledge, however, does not allow

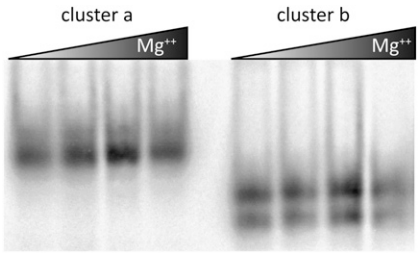
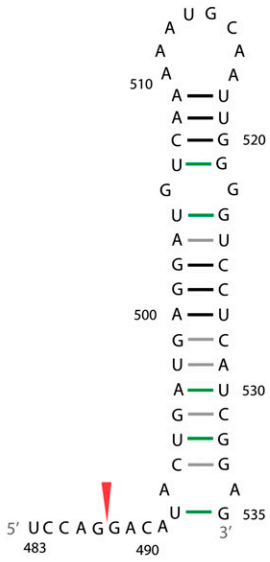
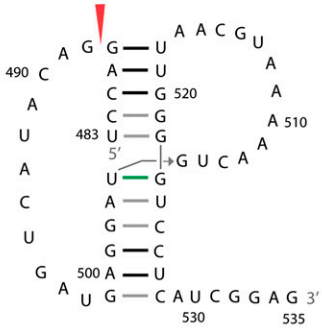


FIGURE 7. Nondenaturing 8% polyacrylamide gel of in vitro folded cluster a and cluster b sequences from Figures 5 and 6 (see Materials and Methods). Final Mg^{++} concentrations are 0, 5, 10, and 15 mM. Two bands are apparent in the clade b samples: slower and faster migrating products that account for 57% and 43% of the integrated band intensity, respectively.



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can
492	535	0	5	0	2	0	367	0	0	0	0	0	0	0	0	0	0	100.0
493	*534*	0	0	4	4	0	0	3	47	275	0	35	0	1	4	0	0	2.1
494	533	0	331	0	1	0	5	0	8	0	3	0	0	0	0	24	1	90.3
495	532	0	0	0	34	0	340	0	0	0	0	0	0	0	0	0	0	100.0
496	531	184	0	2	0	4	0	0	0	0	0	1	4	148	30	0	0	50.9
497	530	1	0	249	0	59	0	0	0	0	0	28	1	2	1	0	33	82.6
498	529	0	0	0	292	0	6	4	2	50	0	0	0	0	1	19	0	79.7
499	528	351	0	0	0	4	0	1	0	0	0	0	2	15	0	0	0	95.2
500	**527**	0	0	371	0	1	0	0	0	1	0	0	0	1	0	0	0	99.5
501	**526**	370	0	0	0	1	0	1	0	0	0	0	0	2	0	0	0	99.2
502	**525**	347	0	0	0	8	0	8	0	0	0	0	0	10	0	0	0	95.2
503	524	0	0	360	0	1	0	0	1	2	0	0	0	10	0	0	0	96.5
504	523	0	0	0	89	0	280	0	0	0	0	2	0	0	3	0	0	98.7
505	*522*	57	0	0	0	5	1	70	3	0	236	0	0	1	0	0	0	16.9
506	521	0	0	0	22	0	351	0	1	0	0	0	0	0	0	0	0	99.7
507	**520**	0	369	0	0	0	0	0	0	0	0	0	1	0	0	4	0	98.7
508	**519**	0	0	365	0	2	0	0	1	0	0	1	0	2	0	0	3	98.1
509	**518**	0	0	374	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0

Average: 92.8%



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can
483	522	0	6	0	32	0	232	35	1	3	1	5	0	0	58	0	0	72.4
484	521	0	341	0	1	0	11	0	0	0	0	0	0	0	0	21	0	94.4
485	**520**	0	367	0	0	0	2	0	0	0	0	0	1	0	0	4	0	98.7
486	**519**	0	0	369	0	0	0	0	1	0	0	0	0	2	0	0	2	98.7
487	**518**	0	0	1	0	372	0	0	0	0	0	0	0	0	0	0	0	100.0
499	528	351	0	0	0	4	0	1	0	0	0	0	2	15	0	0	0	95.2
500	**527**	0	0	371	0	1	0	0	0	1	0	0	0	1	0	0	0	99.5
501	**526**	370	0	0	0	1	0	1	0	0	0	0	0	2	0	0	0	99.2
502	525	347	0	0	0	8	0	8	0	0	0	0	0	10	0	0	0	95.2
503	524	0	0	360	0	1	0	0	1	2	0	0	0	10	0	0	0	96.5
504	**523**	0	0	0	89	0	280	0	0	0	0	2	0	0	3	0	0	98.7

Average: 95.3%

AA	Phe			Gln			Asp			Ile			Leu			Met			Arg			Met			Ser			Lys			Met			Glu			Leu			Gly			Ser			Ser			Ser			Glu		
codon	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii
nt	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535
A	4	0	0	372	1	4	363	0	371	0	328	8	0	150	251	52	15	373	2	10	373	0	4	1	0	368	374	369	348	358	0	2	0	350	299	0	0	4	22	70	89	2	8	1	1	1	365	0	0	34	25	282	2
G	36	0	0	0	372	369	10	0	1	0	3	3	0	188	60	4	356	1	372	363	1	0	369	0	0	2	0	4	22	4	0	371	1	22	71	0	1	369	352	240	280	1	0	0	0	0	8	0	0	340	347	47	372
C	6	362	372	2	0	0	1	368	1	5	0	357	0	4	34	19	2	0	0	0	0	0	0	0	374	3	0	1	4	1	0	0	372	1	1	0	2	1	0	58	3	10	358	372	1	369	1	5	367	0	0	5	0
U	328	12	2	0	0	1	0	6	1	369	43	6	374	31	29	299	0	0	0	0	0	374	1	373	0	1	0	0	0	11	374	1	1	1	3	374	371	0	0	5	2	361	8	1	372	4	0	369	7	0	1	39	0
HP Str	.	.	.	.	.	.	.	.	(	.	(	(	(	(	(	(	(	(	(	.	(	(	(	.	.	.	.	.	.	.	.	.	)	)	)	)	.	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)
PK Str	(	(	(	(	(	.	.	.	.	.	.	.	.	.	.	.	[	[	[	[	[	[	.	.	.	.	.	.	.	.	.	.	)	)	)	)	)	]	]	]	]	]	]	.	.	.	.	.	.	.	.	.
%cons	0.88	0.97	0.99	0.99	1.00	0.99	0.97	0.98	0.99	0.99	0.88	0.95	1.00	0.50	0.67	0.80	0.95	1.00	0.99	0.97	1.00	1.00	0.99	1.00	1.00	0.98	1.00	0.99	0.93	0.96	1.00	0.99	0.99	0.94	0.80	1.00	0.99	0.99	0.94	0.64	0.75	0.97	0.96	0.99	0.99	0.99	0.98	0.99	0.98	0.91	0.93	0.76	0.99

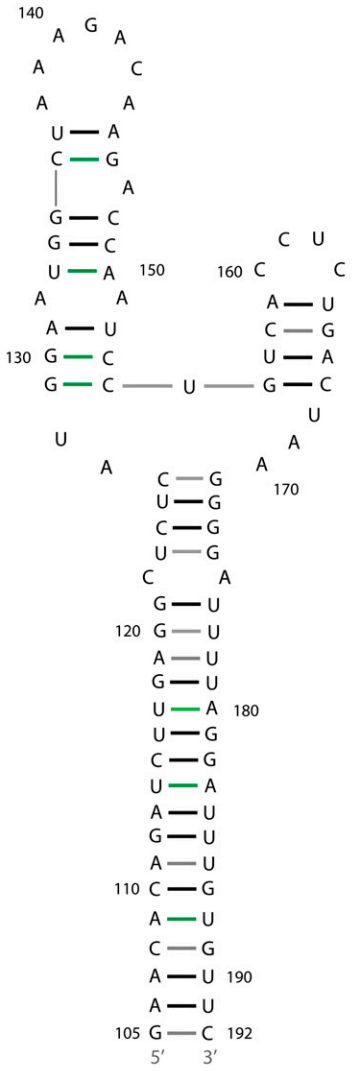
FIGURE 8. Secondary structure models for fragment of 3′ predicted secondary structure region from segment 8 (+)RNA. The *top* structure is for the hairpin predicted by RNAalifold on the SSCU alignment, while the alternative pseudoknot conformation is shown *below*. These structures were also predicted by Gultyaev et al. (2007). Figure annotations and base pair counts are as described in Figure 5. The predicted free energy of folding, $\Delta G_{37}°$ (Mathews et al. 2004), for the consensus hairpin is −18.9 kcal/mol. Predicted free energies for the pseudoknot were calculated as −9 kcal/mol (Dirks and Pierce 2003; Cao and Chen 2009).

inclusion of tertiary interactions in these predictions. Such interaction can be very favorable (Theimer et al. 1998; Liu et al. 2009). Gultyaev et al. (2007) provided evidence for equilibrium between hairpin and pseudoknot structures in some strains of influenza.

Structure prediction in regions of known function

RNAalifold calculations of the 5′ region of segment 7 using the six sequence RNAz alignment and the SSCU alignment generate a structure with a stem topped by a multi-branch loop and having 91.2% overall base-pair conservation (Fig. 9). Dynalign calculations of the two most distant sequences predict the same structure with a predicted free energy of −30.0 kcal/mol at 37°C (Fig. 9). The size of this structure is 88 nt compared to the 68-nt structure in segment 8 (Fig. 5). The segment 7 and 8 putative structures are, respectively, 79 and 51 nt downstream from 5′ splice sites.

To assess whether the region surrounding the 3′ splice site of segment 7 could form a pseudoknot structure similar to the one proposed by Gultyaev et al. (2007) for segment 8 (Fig. 8), the program DotKnot (Sperschneider and Datta



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can
105	192	563	0	20	0	105	0	0	0	0	0	0	0	221	1	0	0	75.6
106	**191**	0	0	910	0	3	0	0	0	1	0	0	0	0	0	0	1	99.8
107	**190**	0	0	900	0	11	0	0	0	1	0	0	0	0	0	0	3	99.6
108	189	0	657	0	9	0	56	1	1	0	0	2	6	0	0	164	19	78.9
109	**188**	0	0	878	0	26	0	0	0	1	0	7	0	0	0	0	3	98.8
110	**187**	0	891	0	0	0	7	0	0	0	0	0	0	0	0	17	0	98.1
111	186	2	1	407	0	3	0	0	0	2	0	65	29	44	4	1	357	45.1
112	**185**	0	0	1	0	914	0	0	0	0	0	0	0	0	0	0	0	100.0
113	**184**	0	0	906	0	0	0	0	2	0	0	0	0	6	0	0	1	99.0
114	**183**	0	40	0	519	0	271	0	0	0	0	21	0	0	6	55	2	90.8
115	**182**	0	909	0	0	0	4	0	1	0	1	0	0	0	0	0	0	99.8
116	**181**	0	1	0	1	0	913	0	0	0	0	0	0	0	0	0	0	100.0
117	**180**	0	109	0	483	0	180	3	7	18	1	0	0	0	0	113	0	84.5
118	**179**	0	0	0	0	913	0	0	0	0	1	0	0	0	0	0	1	99.8
119	178	0	0	876	0	2	0	0	0	0	0	0	0	36	0	0	0	96.1
120	177	32	0	44	0	764	1	3	4	0	59	5	0	0	0	0	1	92.1
121	**176**	0	0	0	0	911	0	0	0	0	0	2	0	0	0	0	1	99.7
122	*175*	0	19	0	14	0	2	0	0	0	0	0	0	0	0	877	0	3.8
123	174	0	22	0	114	0	508	4	203	16	32	10	0	0	2	2	0	70.5
124	**173**	0	906	0	0	0	2	0	5	0	0	0	1	0	0	0	0	99.3
125	**172**	0	1	0	0	0	913	0	0	0	0	1	0	0	0	0	0	99.9
126	171	0	708	0	5	0	21	0	11	1	2	0	0	0	0	166	1	80.2
129	**154**	789	0	0	0	121	0	3	0	0	1	0	0	0	1	0	0	99.5
130	**153**	858	0	0	0	50	0	7	0	0	0	0	0	0	0	0	0	99.2
131	**152**	0	0	906	0	3	0	0	0	5	0	0	0	1	0	0	0	99.3
132	*151*	0	0	0	1	0	0	247	1	665	0	0	0	1	0	0	0	0.1
133	**150**	0	0	0	845	0	60	1	0	1	0	6	0	0	2	0	0	98.9
134	**149**	914	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	100.0
135	**148**	912	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	100.0
136	**146**	0	846	0	0	0	48	0	15	0	3	0	1	0	0	1	0	97.8
137	**145**	0	0	0	912	0	2	0	0	0	0	0	0	0	1	0	0	99.9
156	**167**	902	0	0	0	4	0	1	0	0	0	0	0	5	1	0	0	99.2
157	**166**	0	0	0	908	3	1	1	0	1	0	1	0	0	0	2	0	99.5
158	165	0	892	0	0	0	0	0	1	0	0	0	0	0	0	19	3	97.5
159	**164**	0	0	908	0	4	0	0	0	0	0	1	0	1	0	0	0	99.8

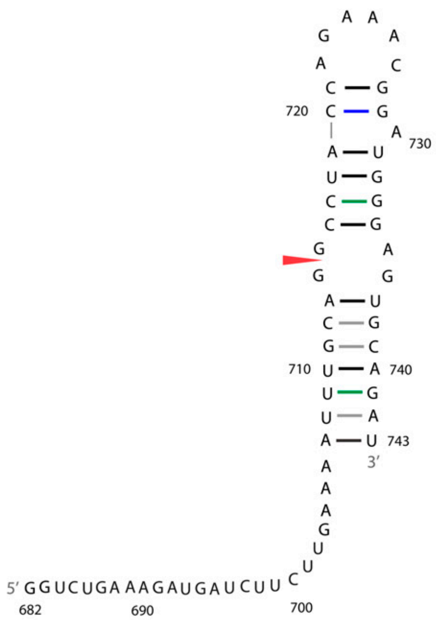
Average: 91.2%

AA	Lys	Asn			Thr			Asp			Leu			Glu			Ala			Leu			Met			Glu			Trp			Leu			Lys			Ile			Arg		
codon	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i
nt	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148
A	241	911	901	1	879	0	453	1	914	0	1	0	25	0	913	48	0	0	219	5	0	12	912	0	0	0	912	667	1	0	0	15	0	889	914	912	3	913	1	911	912	1	899	0
G	670	3	11	1	26	0	5	914	0	0	1	0	4	914	2	860	911	0	36	0	0	2	0	0	914	915	3	247	1	915	915	3	0	26	1	2	912	2	0	1	2	913	16	0
C	0	1	3	846	3	908	388	0	1	97	909	1	222	1	0	1	1	898	24	907	1	875	2	5	0	0	0	0	0	0	0	848	0	0	0	1	0	0	914	1	1	1	0	912
U	1	0	0	67	7	7	69	0	0	818	4	914	664	0	0	6	2	16	635	2	914	26	1	910	1	0	0	1	913	0	0	48	915	0	0	0	0	0	0	2	0	0	0	3
Str	(	(	(	(	(	(	(	(	(	(	(	(	(	(	(	(	(	.	(	(	(	(	.	.	(	(	(	.	(	(	(	(	.	.	.	.	.	.	.	)	)	.	)
%cons	0.73	1.00	0.98	0.92	0.96	0.99	0.50	1.00	1.00	0.89	0.99	1.00	0.73	1.00	1.00	0.94	1.00	0.98	0.69	0.99	1.00	0.96	1.00	0.99	1.00	1.00	1.00	0.73	1.00	1.00	1.00	0.93	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00

AA	Pro		Ile			Leu			Ser			Pro			Leu			Thr			Lys			Gly			Ile			Leu			Gly			Phe			Val			Phe		
codon	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**	i	ii	**iii**
nt	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192
A	0	847	913	5	7	3	0	5	1	1	910	0	2	8	0	0	19	912	1	9	914	912	172	0	0	137	892	0	3	0	0	617	1	0	574	0	0	3	17	1	174	1	1	0
G	0	60	1	0	0	1	2	908	1	0	4	0	1	0	0	0	893	1	0	2	0	3	742	914	914	765	21	0	64	0	1	297	914	915	311	2	0	1	898	0	714	0	0	0
C	914	2	1	1	858	790	0	0	2	914	0	899	911	3	887	1	0	0	909	58	1	0	0	0	1	2	0	0	32	36	0	0	0	0	6	6	0	79	0	0	6	0	0	788
U	1	6	0	909	50	121	913	1	911	0	1	16	1	904	27	913	3	1	4	845	0	0	1	1	0	10	0	915	814	878	914	0	0	0	23	907	915	832	0	914	21	914	914	125
Str	)	)	.	)	)	)	.	(	(	(	(	.	.	.	.	)	)	)	)	.	.	.	)	)	)	)	.	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)
%cons	1.00	0.93	1.00	0.99	0.94	0.86	1.00	0.99	1.00	1.00	0.99	0.98	1.00	0.99	0.97	1.00	0.98	1.00	0.99	0.92	1.00	1.00	0.81	1.00	1.00	0.84	0.98	1.00	0.89	0.96	1.00	0.68	1.00	1.00	0.63	0.99	1.00	0.91	0.98	1.00	0.78	1.00	1.00	0.86

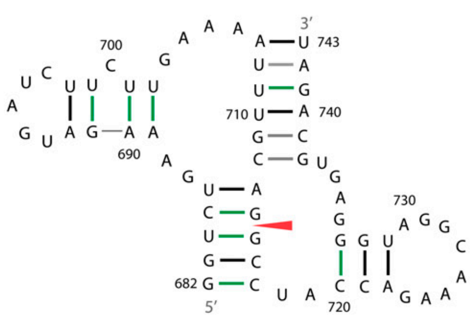
FIGURE 9. RNAalifold predicted secondary structure for fragment of 5′ predicted secondary structure region from the RNAz and the SSCU alignments of segment 7 (+)RNA. Figure annotations and base pair counts are as described in Figure 5. The predicted free energy of folding, $\Delta G_{37}°$ (Mathews et al. 2004), for the consensus sequence is −30.0 kcal/mol.

2010) was used to scan the entire segment 7 (+)RNA coding sequence from genome set 755298. A potential pseudoknot structure was identified that incorporates the 3′ splice site (Fig. 10, bottom). Additional base pairs were manually added in the loop regions between nucleotides 690 and 701. As an alternative folding, RNAalifold calculations on alignments



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can
707	743	0	0	533	0	0	0	0	0	0	0	2	0	2	0	0	1	99.07
708	742	0	0	0	492	0	6	0	0	3	0	2	0	0	14	20	0	92.74
709	741	0	2	0	207	0	317	0	0	0	1	1	0	0	0	10	0	97.77
710	740	0	0	0	537	0	0	0	0	0	0	0	0	0	1	0	0	99.81
711	739	449	0	0	0	0	0	0	0	0	0	0	0	87	0	0	0	83.77
712	738	0	524	0	0	0	0	0	0	0	0	0	1	0	0	13	0	97.4
713	737	0	0	537	0	0	0	0	0	0	0	0	0	1	0	0	0	99.81
714	*736*	0	0	0	0	18	0	4	1	0	515	0	0	0	0	0	0	3.346
715	*735*	0	0	1	0	3	0	362	37	19	115	0	0	0	0	1	0	0.743
716	734	0	531	0	0	0	5	0	0	0	1	0	0	0	0	0	1	99.63
717	733	0	454	0	0	0	43	0	40	0	0	0	0	0	0	0	0	92.55
718	732	0	1	0	0	0	536	0	0	0	0	1	0	0	0	0	0	99.81
719	731	0	0	531	0	0	0	0	1	3	0	3	0	0	0	0	0	98.7
720	729	0	252	0	128	0	43	0	0	0	0	0	1	0	0	112	1	78.77
721	728	0	536	0	0	0	0	0	0	0	0	0	1	0	0	1	0	99.63

Average: 95.3 %



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can
682	717	449	0	0	0	43	0	40	0	0	0	0	4	1	0	0	0	91.6
683	716	531	0	0	0	5	0	0	0	0	1	0	0	1	0	0	0	99.6
684	715	0	22	0	54	0	450	0	5	1	3	0	0	0	1	2	0	97.8
685	714	0	482	0	0	0	51	0	1	0	1	0	0	0	0	1	0	99.4
686	713	0	0	0	536	0	0	0	0	2	0	0	0	0	0	0	0	99.6
689	702	2	0	410	0	118	0	0	0	0	0	0	0	8	0	0	0	98.5
690	701	0	0	497	0	41	0	0	0	0	0	0	0	0	0	0	0	100.0
691	699	30	0	17	0	462	0	0	0	0	2	0	0	27	0	0	0	94.61
692	698	0	0	535	0	2	0	0	0	1	0	0	0	0	0	0	0	99.81
707	743	0	0	533	0	0	0	0	0	0	0	2	0	2	0	0	1	99.1
708	742	0	0	0	492	0	6	0	0	3	0	2	0	0	14	20	0	92.7
709	741	0	2	0	207	0	317	0	0	0	1	1	0	0	0	10	0	97.8
710	740	0	0	0	537	0	0	0	0	0	0	0	0	0	1	0	0	99.8
711	739	449	0	0	0	0	0	0	0	0	0	0	0	87	0	0	0	83.8
712	738	0	524	0	0	0	0	0	0	0	0	0	1	0	0	13	0	97.4
720	733	0	366	0	0	0	171	0	0	0	0	0	0	0	0	0	0	100.0
721	732	0	537	0	0	0	0	0	0	0	0	0	0	0	0	0	1	99.8
722	731	0	0	533	0	0	0	0	1	3	0	1	0	0	0	0	0	99.1

Average: 97.2 %

AA	Gly			Leu			Lys			Asp			Asp			Leu			Leu			Glu			Asn			Leu			Gln
codon	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i
nt	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712
A	1	1	6	1	2	240	538	418	497	44	536	2	64	538	0	0	1	0	68	0	0	1	530	526	536	535	3	0	0	87	0
G	533	537	3	1	0	293	0	120	41	494	2	0	474	0	0	0	0	2	0	0	0	537	6	0	1	0	0	1	0	449	0
C	4	0	24	483	0	1	0	0	0	0	0	27	0	0	76	537	0	57	469	0	10	0	0	3	0	1	20	12	0	0	538
U	0	0	505	51	536	4	0	0	0	0	0	509	0	0	461	1	537	479	1	538	528	0	2	9	1	2	514	525	538	0	0
HP Str	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	(	(	(	(	(	(
PK Str	{	{	{	{	{	.	.	(	(	(	(	.	.	.	.	.	)	)	.	)	)	.	.	.	.	[	[	[	[	[	[
%cons	0.99	1.00	0.94	0.90	1.00	0.54	1.00	0.78	0.92	0.92	1.00	0.95	0.88	1.00	0.86	1.00	1.00	0.89	0.87	1.00	0.98	1.00	0.99	0.98	1.00	0.99	0.96	0.98	1.00	0.84	1.00

AA	Gln		Ala			Tyr			Gln			Lys			Arg			Met			Gly			Ala			Gln			Met	
codon	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii	iii	i	ii
nt	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743
A	538	1	57	0	40	0	535	0	0	537	43	537	531	442	17	1	240	536	3	0	0	0	382	4	0	13	0	537	217	516	0
G	0	537	480	1	0	0	0	0	0	0	489	0	6	88	0	536	295	0	1	537	538	537	152	516	0	524	0	0	320	6	0
C	0	0	1	532	454	1	0	366	538	0	3	0	1	6	520	1	1	1	0	0	0	0	0	0	1	1	538	1	0	14	2
U	0	0	0	5	43	537	3	171	0	1	0	0	0	1	0	0	1	1	534	1	0	1	4	18	537	0	0	0	1	2	536
HP Str	(	.	.	(	(	(	(	(	(	.	.	.	.	.	.	)	)	.	)	)	)	)	.	.	)	)	)	)	)	)	)
PK Str	}	}	}	}	}	.	.	(	(	(	.	.	.	.	.	.	.	.	)	)	)	.	.	.	.	]	]	]	]	]	]
%cons	1.00	1.00	0.89	0.99	0.85	1.00	0.99	0.68	1.00	1.00	0.91	1.00	0.99	0.82	0.97	1.00	0.55	1.00	0.99	1.00	1.00	1.00	0.71	0.96	1.00	0.97	1.00	1.00	0.59	0.96	1.00

FIGURE 10. Secondary structures predicted for a fragment of the 3′ region of segment 7 (+)RNA. The *top* structure is for the hairpin predicted by RNAalifold on the SSCU alignment, while the alternative pseudoknot conformation is shown *below*. Figure annotations and base pair counts are as described in Figure 5. The predicted free energy of folding, $\Delta G_{37}°$ (Mathews et al. 2004), for the consensus hairpin is −14.3 kcal/mol. Predicted free energies for the pseudoknot were calculated as −7 (DP) or −4 (CC) kcal/mol, depending on the parameters used (Dirks and Pierce 2003; Cao and Chen 2006). A slipped helix with G691–C700 and A692–U699 pairs results in a more favorable predicted free energy range of −12 (DP) to −9 (CC), but less favorable percentage of canonical pairing of 79.4% and 89.2%, respectively, for these base pairs.

from this region generated a consensus hairpin where two additional conserved base pairs could be manually added at positions 720–729 and 721–728 (Fig. 10, top). Dynalign (Mathews and Turner 2002) calculations of the two most distant sequences also predict the hairpin with a predicted free energy of −14.3 kcal/mol at 37°C. The pseudoknot has slightly higher pair conservation than the hairpin (97.2% vs. 95.3%, respectively) and has more consistent mutations (Fig. 10). Predicted free energies for the pseudoknot range from −7 to −4 kcal/mol at 37°C, using parameters from Dirks and Pierce (2003) and Cao and Chen (2006), respectively. If the bulged C700 is slipped to pair with G691 and an A692–U699 pair forms, then the predicted free energies range from −12 to −9 kcal/mol at 37°C using parameters from Dirks and Pierce (2003) and Cao and Chen (2006), respectively. This pairing scheme, however, reduces the pair

conservation to 79.4% and 89.2%, respectively, for these two base pairs. Co-transcriptional folding would be expected to generate preferentially the pseudoknot.

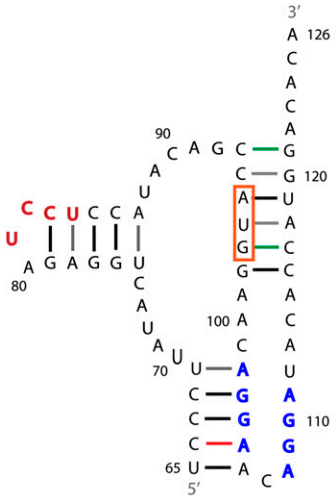
DotKnot (Sperschneider and Datta 2010) was also used to scan the whole segment 2 (+)RNA sequence from genome set 755298. Three potential pseudoknot folds are predicted for the region encompassing positions 65–121. All three models incorporate the start codon of the PB1-F2 ORF into a helix with base-pair conservation of 98.0% (Fig. 11, nucleotides 93–98 and 116–121). The other helix of the pseudoknot is formed by bases upstream of and downstream from the PB1-F2 start codon. The most conserved helix predicted is shown in Figure 11. The other possibilities are represented by the blue-shaded nucleotides, which can base-pair with the red-shaded nucleotides to form two alternative helices. Predicted free energies for the most conserved pseudoknot range from −14 kcal/mol (Dirks and Pierce 2003) to −8 kcal/mol (Cao and Chen 2009) at 37°C. Interestingly, the nucleotides at positions 66 and 105 always form canonical or CA pairs (Fig. 11).

To determine if structurally distinct subsets of sequences exist for Figures 8–11, an UPGMA analysis was conducted as described for segment 8. No obvious pattern for these structures was detected between the different sequence clusters.

Base-pair and codon conservation in secondary structure models

Table 3 summarizes the average base-pair composition of the structures in Figures 5–6 and 8–11. When noncanonical pairing is observed, it is predominantly AC followed by AG. Occurrence of AA, UU, GG, and CC is almost always <1%. The one apparent exception is the 5′ tetraloop model for segment 8 (Fig. 6), but virtually all of the noncanonical pairs are confined to the cluster a sequences, which better support the multi-branch loop structure rather than the tetraloop structure. If cluster a sequences are eliminated from the tetraloop percentages, then the pattern for noncanonical pairing follows the average pattern.

In Table 4, nucleotide conservation is analyzed in the context of codon position between predicted double-stranded and single-stranded regions. On average, the predicted double-stranded regions are more constrained at all three codon positions. This is statistically significant for the second and third codon positions (statistical *P*-values of 0.05). The only cases in which predicted single-stranded regions are significantly more constrained than predicted double-stranded regions are for the 3′ HP/PK model structures in segments 7 and 8. In segment 8 this occurs mainly where a nucleotide is predicted to be single-stranded



i	j	GC	CG	AU	UA	GU	UG	GA	AG	AA	GG	UU	CC	AC	UC	CA	CU	%can
65	**106**	0	0	0	466	0	1	0	0	0	0	0	0	0	0	0	0	100.0
66	**105**	0	127	0	66	0	9	0	0	0	0	0	0	0	0	265	0	43.3
67	**104**	0	465	0	0	0	0	0	1	0	0	0	0	0	0	1	0	99.6
68	**103**	0	466	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0
69	102	1	18	4	272	0	17	19	57	34	10	5	0	0	6	22	1	67.0
75	87	0	0	0	324	1	13	3	0	11	0	64	0	0	22	26	3	72.4
76	**86**	466	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	100.0
77	**85**	467	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0
78	84	3	0	363	0	10	0	0	5	4	0	25	0	54	2	0	1	80.5
79	**83**	463	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0	99.4
93	**121**	0	411	0	0	0	56	0	0	0	0	0	0	0	0	0	0	100.0
94	120	0	450	0	0	0	0	0	0	0	0	0	0	0	0	14	3	96.4
95	**119**	0	0	460	0	3	0	0	0	1	0	0	0	2	0	0	1	99.1
96	118	0	0	0	446	0	1	0	0	1	0	1	0	0	4	14	0	95.7
97	**117**	412	0	0	0	39	0	14	0	0	2	0	0	0	0	0	0	96.6
98	**116**	467	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0

Average: 90.6 %

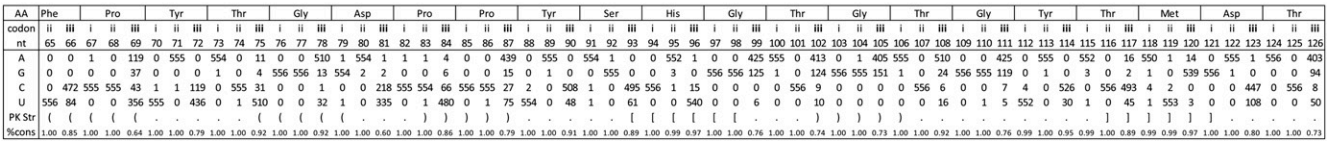


FIGURE 11. DotKnot (Sperschneider and Datta 2010) secondary structure model for 5′ region 65–126 from segment 2 of genome set 755298. Other predictions by DotKnot for this region are represented by blue shaded nucleotides, which can base-pair with the red shaded nucleotides to form two alternate 5′ helices leaving nucleotides 65–80 unpaired. Base pair counts from an alignment of all unique segment 2 sequences are shown to the *right*. The nucleotides boxed in orange in the 3′ pseudoknot helix are the start codon for the internal ORF for the PB1-F2 product. This ORF is shifted +1 compared to the PB1 coding region. Figure annotations and base pair counts are as described in Figure 5. Predicted free energies for the pseudoknot were calculated as −14 (DP) or −8 (CC) kcal/mol, depending on the parameters used (Dirks and Pierce 2003; Cao and Chen 2009).

TABLE 3. Average base-pair percentage across model structures

Structure[a]	GC/CG	AU/UA	GU/UG	AG/AG	AC/CA	UC/CU	AA	UU	GG	CC
SEG8 5′ MBL	48.3 (40.5)	24.5 (31.0)	10.9 (16.3)	5.8 (4.8)	3.2 (7.0)	0.4 (0.1)	0.9 (0.1)	0.1 (0.1)	0.6 (0.0)	0.0 (0.0)
SEG8 5′ TL	46.4 (50.8)	20.0 (38.6)	14.7 (9.5)	1.5 (0.3)	6.2 (0.8)	8.9 (0.0)	2.3 (0.0)	1.2 (0.1)	0.1 (0.0)	0.3 (0.0)
SEG8 3′ HP	32.7	36.1	23.9	0.5	4.0	1.2	0.9	0.5	0.1	0.1
SEG8 3′ PK	43.4	29.8	22.2	1.2	1.6	1.5	0.1	0.2	0.0	0.1
SEG7 5′ MBL	36.7	35.2	22.6	1.0	2.3	1.3	0.1	0.4	0.2	0.1
SEG7 3′ HP	30.4	46.6	14.1	0.6	7.7	0.3	0.1	0.2	0.0	0.0
SEG7 3′ PK	39.2	44.2	12.4	1.2	2.5	0.2	0.1	0.1	0.1	0.1
SEG2 PK	56.4	32.1	2.0	1.3	5.3	1.4	0.7	1.3	0.2	0.0
Average	40.6 (40.2)	33.5 (36.7)	15.8 (15.8)	2.1 (1.9)	4.5 (4.3)	1.8 (0.6)	0.6 (0.2)	0.4 (0.3)	0.2 (0.1)	0.1 (0.1)
StdDev	6.9 (7.0)	9.1 (6.2)	6.6 (6.7)	2.2 (2.0)	2.6 (3.0)	2.9 (0.6)	0.8 (0.3)	0.4 (0.2)	0.2 (0.1)	0.1 (0.1)

Values in parentheses represent average base-pair conservation for cluster b sequences.
[a]MBL represents the structures with multi-branch loops (Figs. 5, 9). TL represents the tetraloop structure in Figure 6. HP represents the hairpin structures in Figures 8 and 10. PK represents the pseudoknot structures in Figures 8, 10, and 11.

in one structure and double-stranded in the other. Many positions where consistent and compensatory mutations are observed show lower conservation than average. When compared to the amino acid sequence, almost all of these changes maintain the coding potential of protein sequence and predicted RNA secondary structure. For example, there are several third codon positions that have fourfold degeneracy, yet observed mutations are usually restricted to two nucleotides that preserve predicted base-pairing (Figs. 5, 6, 8–11).

DISCUSSION

The bioinformatics analysis of influenza coding regions provides evidence for many areas with potentially conserved RNA secondary structure and supports previous SSCU studies of influenza A (Gog et al. 2007). With the exception of segment 3, *Z*-scores are negative in both strand orientations across the eight coding regions. Results for the coding regions predict more conservation of RNA secondary structure in the (+)RNA over the (−)RNA. No predicted structured region strongly favors the (−)RNA (Table 2). This is a reasonable result as the influenza (−)RNA is tightly associated with multiple copies of the NP protein to form viral ribonucleoproteins (vRNPs) that are packaged into viral particles. NP protein binding melts RNA secondary structure (Baudin et al. 1994). Except for interactions between the (−)RNA UTRs, which are important for associating with viral polymerase (Fodor et al. 1994), it is unlikely that most of the (−)RNA possesses extensive conserved secondary structure in the areas analogous to the (+)RNA coding region.

Of the 20 structured regions predicted in influenza RNA, 11 are clearly predicted to be in the (+)RNA, while RNAstrand predictions for the remaining nine are ambiguous (Table 2). These ambiguous regions occur mainly where RNAz predicts structure solely in the (−)RNA, but RNAstrand predicts strand bias toward the (−)RNA with extremely low probability.

Additional evidence for the predicted structured regions comes from the observed overlap of RNAz predictions with areas of strong SSCU and with biologically significant sections of influenza RNA. In all but segments 2 and 3 (Figs. 3, 4), areas with high SSCU overlapped regions with predicted conserved RNA secondary structure. The constraint on synonymous codon site variation may be a manifestation of the preservation of RNA secondary structure (Simmonds and Smith 1999; Pedersen et al. 2004; Tuplin et al. 2004). This is supported by the observed constraint on double-stranded versus single-stranded RNA at third codon positions (Table 4).

Especially compelling are the instances in which predictions of conserved RNA secondary structures overlap with both SSCU and with known biologically important regions. In segment 8, there are two regions where the predicted conserved secondary structure overlaps with sites

TABLE 4. Percent nucleotide conservation by codon position and RNA secondary structure

Structure	i (ssRNA)	i (dsRNA)	ii (ssRNA)	ii (dsRNA)	iii (ssRNA)	iii (dsRNA)
SEG8 5′ MBL	81.4	93.4	85.6	96.0	65.2	82.3
SEG8 5′ TL	89.4	96.4	93.7	96.9	77.5	83.2
SEG8 3′ HP	97.6	95.3	86.4	97.0	91.7	90.3
SEG8 3′ PK	97.0	96.4	99.4	98.1	79.0	88.5
SEG7 5′ MBL	88.0	98.3	87.0	99.8	82.4	85.2
SEG7 3′ HP	96.3	99.0	99.3	99.6	86.0	80.5
SEG7 3′ PK	98.0	95.5	99.5	98.0	79.0	88.5
SEG2 PK	99.7	99.8	99.9	99.8	80.3	84.7
Average	93.4	96.8	93.8	98.1	80.1	85.4
StdDev	6.4	2.2	6.5	1.5	7.6	3.4

Codon positions are indicated by lowercase roman numerals i, ii, and iii.

of SSCU and with viral splice sites (Fig. 1). Others have also described structured RNA elements within these regions of predicted conserved secondary structure (Gultyaev et al. 2007; Ilyinskii et al. 2009). Using the RNAz alignment, RNAalifold predicted a multi-branch loop for the 5′ region of segment 8 (Fig. 5) that is similar to the one proposed previously (Ilyinskii et al. 2009). Additionally, a tetraloop structure was predicted using the SSCU alignment. This structure is not as well supported by all unique sequences for this area. However, the cluster b sequences support this structure quite well, raising the possibility that some strains may adopt a tetraloop structure in this region. Perhaps the RNAalifold calculation revealed the tetraloop structure because the SSCU alignment was biased toward cluster b strains. A recent review article also suggests a tetraloop structure as an alternative fold for clade B influenza strains (Gultyaev et al. 2010). It is likely that the cluster b described in this study corresponds to clade B influenza phylogeny (Kawaoka et al. 1998; Basler et al. 2001). In vitro folding of representative cluster a and b sequences have distinct migration on native gels, thus supporting the prediction that these groups adopt different conformations in nature (Fig. 7).

Interestingly, a reported mutant (NS1mut3841) to the 5′ region of segment 8 was shown to inhibit NS1 protein expression (Ilyinskii et al. 2009). It was proposed that this inhibition occurred due to disruption of RNA secondary structure. The expanded sequence data set, however, provides an alternative possibility. A nucleotide count on an alignment of all unique sequences for this region shows that four of the five substitutions in NS1mut3841 occur naturally (Fig. 12). Alignment position 122, however, is never observed to mutate naturally to the C residue in NS1mut3841. This apparently critical position is predicted to be single-stranded in both structural models (Figs. 5, 6). Therefore, it is less likely that RNA secondary structure is responsible for the inhibition of protein production.

An alternative explanation for the observed inhibition of protein expression is in the mutant amino acid sequence. The mutations in NS1mut3841 lead to a protein product with two new alanine residues (only one of which is observed in the alignment of all unique wild-type sequences) (Fig. 12). These alanines are separated by two intervening residues (AXXA). Previous studies have found that AXXA is able to substitute for RXXL in the destruction box motif (Yamano et al. 1996) and signal the cell's proteolytic machinery to degrade the protein product. The reported effect of the NS1mut3841 mutation (Ilyinskii et al. 2009) may be due to AXXA marked protein destruction in the eukaryotic 293 human embryo kidney cell (HEK) expression system used.

Near the 3′ splice site of segment 8, Gultyaev et al. (2007) described an RNA secondary structure in equilibrium between a hairpin and pseudoknot. RNAalifold calculations

Pos.	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125
AA		D			R			L			R			R			D			Q			K			
nt	G	A	C	C	G	G	C	U	U	C	G	C	C	G	A	G	A	U	C	A	G	A	A	G	U	C
HP Str.	(	(	(	.	(	(	(	(	(	.	.	.	.	)	)	)	.	)	)	.	.	.	.	)	)	)
ML Str.	(	.	(	(	(	(	(	.	.	.	)	)	)	)	)	(	(	(	.	.	.	.	.	)	)	)
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
A	0	937	0	137	0	110	22	0	12	84	4	96	131	1	841	1	942	2	4	933	309	**942**	819	171	1	0
G	942	2	2	0	941	826	4	1	10	0	938	16	**2**	941	**87**	938	1	0	0	8	629	0	124	**772**	115	1
C	0	1	483	804	0	3	910	1	262	854	1	810	809	**1**	1	2	0	158	939	0	2	1	0	0	3	942
U	1	3	458	1	2	4	7	940	659	5	0	21	1	0	14	1	0	783	0	1	3	0	0	0	824	0
A		1			0			0			0			**1**			0			0			0			
C		0			0			0			4			0			0			0			0			
D		936			0			0			0			0			935			0			0			
E		0			0			0			0			0			2			0			0			
F		0			0			5			0			0			0			0			0			
G		2			0			0			0			1			1			0			0			
H		0			0			0			2			0			2			4			0			
I		0			0			22			0			0			0			0			0			
K		0			0			0			2			1			0			4			819			
L		0			0			909			0			0			0			1			0			
M		0			2			0			0			0			0			0			0			
N		0			0			0			0			0			1			0			0			
P		0			0			1			0			0			0			0			0			
Q		0			0			0			0			0			0			925			0			
R		0			939			1			932			939			0			8			124			
S		0			0			0			3			0			0			0			0			
V		3			0			4			0			0			0			0			0			
W		0			1			0			0			0			0			0			0			
Y		1			0			0			0			0			1			0			0			

FIGURE 12. Segment 8 sequence from nucleotides 100–125 compared to amino acid coding for region with mutations reported by Ilyinskii et al. (2009). The *top* row has the alignment positions. *Below* are given the consensus amino acid sequence and primary nucleotide sequence. Position 102 is a C to represent the consensus sequence for this region, but a U is used at this position in Figure 5 to more accurately represent the canonical pairing in the structural model. Natural occurrences of the NS1mut3841 mutations made by Ilyinskii et al. (2009) are underlined, while the mutations never observed naturally are boxed.

on this fragment also predict the hairpin (Fig. 8). As verified by native gel electrophoresis, Gultyaev et al. (2007) demonstrated that more evolutionarily recent sequences of H5N1 influenza A favor the hairpin over the pseudoknot. This is consistent with free energy predictions for the hairpin (−18.9 kcal/mol) versus the pseudoknot (−9 kcal/mol). However, nucleotides 493 and 496–498 at the bottom of the hairpin model (Fig. 8, top) exhibit lower conservation than the rest of the structure and are predicted to be single-stranded in the pseudoknot (Fig. 8, bottom). Thus, it appears that some mutations in these sequences favor the pseudoknot. However, the UPGMA clustering in this analysis did not reveal any clear pattern for sequences that prefer one structure over the other.

The proximity of the predicted structured region shown in Figures 5 and 6 to the 5′ splice site, coupled with the structured region at the 3′ splice site in Figure 8, suggests that these structures may be involved in segment 8 splicing. Others have noted that RNA secondary structure may play an important role in splicing of influenza virus segment 8 (Plotch and Krug 1986; Nemeroff et al. 1992). Interestingly, segment 7 RNA displays a similar pattern of predicted secondary structure surrounding the 5′ and 3′ splice sites (Figs. 9, 10). Both structures have similar sizes and spacing in relation to the 5′ and 3′ splice sites as their segment 8 counterparts. The 3′ splice sites contain a putative hairpin with a possible alternative pseudoknot conformation in both segments 7 and 8. The putative structures at the 5′ and 3′ splice sites in segments 7 and 8 suggest experiments to test the proposed structural models.

Segment 2 provides another correspondence between a predicted structured region and a known feature of influenza RNA. The predicted structured region at 51–170 (Fig. 4) has moderate SSCU and includes the start site of the internal ORF for PB1-F2, a small protein product thought to increase virulence through a pro-apoptotic mechanism (Chen et al. 2001; Gibbs et al. 2003; Chanturiya et al. 2004; Zamarin et al. 2005, 2006). DotKnot predicts a pseudoknot in this region that encompasses the PB1-F2 start codon in a 3′ helix that displays 98.0% base-pair conservation and consistent mutations. Other possibilities predicted by DotKnot for the 5′ helix are described in the Figure 11 caption. AC is the most common noncanonical pair at position 66–105, which is intriguing. In fact, AC is the most common noncanonical pair for almost all structural models (Table 3).

There are many examples where noncanonical base pairs play important roles in RNA secondary structure (Nagaswamy et al. 2002). In particular, AC pairs can induce little perturbation to the A-form helix and can play a role in protein recognition (Jang et al. 1998; Lima et al. 2002). AC is the only noncanonical base pair that can preserve a canonical C1′–C1′ distance in the helix. This has been observed primarily when the N1 position of the adenosine is protonated (Leontis et al. 2002), but recent NMR spectra reveal an almost identical C1′–C1′ distance even in the absence of protonation (Y Lerman, SD Kennedy, N Shankar, M Parisien, F Major, and DH Turner, unpubl.). In segment 2, nucleotides at positions 99–101 and 113–115 (Fig. 11) offer the possibility of two additional AC base pairs capped with a sheared purine–purine base pair, similar to the motif found in helix 8 of the signal recognition particle RNA (Ataide et al. 2011), which would maintain helicity between the predicted pseudoknot helixes.

Another interesting possibility for this pseudoknot (Fig. 11) is suggested by the sequences at positions 89–92 and 122–126, which are identical to the *Tetrahymena thermophila* group I ribozyme sequences that facilitate a 180° turn in the 3D structure (Cate et al. 1996; Guo et al. 2004). Thus, it is possible that a similar turn orients the helix between nucleotides 75 and 87. The mechanism for expression of PB1-F2 product is not definitively known but is presumed to occur via ribosomal scanning (Chen et al. 2001). RNA structure in this region may facilitate the initiation of translation at the PB1-F2 ORF as seen in other viruses with expressed internal ORFs (Ryabova and Hohn 2000; Pooggin et al. 2006).

There are three instances in which predicted structured regions overlap with strong SSCU but do not correspond to known annotations: at the 5′ and 3′ ends of segments 5 (Fig. 2) and in the 3′ end of segments 3 and 1 (Figs. 3 and 4, respectively). These are intriguing regions of the influenza (+)RNA. There is apparent conserved structure that may be constraining codon variation within regions with no current annotation. These predictions suggest RNA structures with novel function in the influenza virus.

The remaining regions with predicted conserved structure do not overlap with strong SSCU or current annotations. These may be false positives, or the RNA structure conservation may not be constraining enough to have noticeable SSCU.

There are also two regions where strong SSCU does not overlap with predicted structural regions: positions 500–800 in segment 3 and toward the 3′ end of segment 2 (Figs. 3, 4). These constrained sites may be due to other factors acting on the evolution of the sequence (e.g., RNA–protein interactions), or they could be due to RNA structure not predicted by RNAz. Long-range base pairs would be missed in the structure prediction, which divides sequences into windows.

Although only a subset of sequences was used for SSCU studies and secondary structure predictions, structural models are supported by sequence data from all unique sequences for each region. Examining the conservation of each codon position in the context of the proposed secondary structure models adds additional support. Due to the constraint to maintain amino acid coding potential, few compensatory mutations were observed. However, Table 4 shows that when all unique sequences are considered for each structured region, predicted double-stranded nucleotides are more constrained than predicted single-stranded nucleotides. When nucleotides did vary more than average, it was at tolerant codon positions that preserve the amino acid identity and

often resulted in consistent and compensatory mutations in the structural model (Figs. 5, 6, 8–11). This conservation recapitulates the global SSCU findings and favors RNA secondary structure as the source of the suppression.

This analysis paves the way for a focused approach to experimental secondary structure determination in influenza A virus. Most importantly, the secondary structures predicted here surround previously defined functional annotations, therefore maximizing the possibility that they are functionally important.

ACKNOWLEDGMENTS

S.F.P. is a trainee in the Medical Scientist Training Program funded by NIH T32 GM07356. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or NIH. During the course of this work S.F.P. was supported by NIH T32 GM068411 from an Institutional Ruth L. Kirschstein National Research Service Award. This work was also supported by NIH RO1 GM22939. We thank Prof. S.-J. Chen and Dr. S. Cao for help with their pseudoknot parameters.

Received January 6, 2011; accepted April 1, 2011.

REFERENCES

Ataide SF, Schmitz N, Shen K, Ke A, Shan SO, Doudna JA, Ban N. 2011. The crystal structure of the signal recognition particle in complex with its receptor. *Science* **331:** 881–886.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. 2008. The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82:** 596–601.

Basler CF, Reid AH, Dybing JK, Janczewski TA, Fanning TG, Zheng H, Salvatore M, Perdue ML, Swayne DE, García-Sastre A. 2001. Sequence of the 1918 pandemic influenza virus nonstructural gene (NS) segment and characterization of recombinant viruses bearing the 1918 NS genes. *Proc Natl Acad Sci* **98:** 2746–2751.

Baudin F, Bach C, Cusack S, Ruigrok RW. 1994. Structure of influenza virus RNP. I. Influenza virus nucleoprotein melts secondary structure in panhandle RNA and exposes the bases to the solvent. *EMBO J* **13:** 3158–3165.

Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. 2008. RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9:** 474–486.

Bouloy M, Plotch SJ, Krug RM. 1978. Globin mRNAs are primers for the transcription of influenza viral RNA in vitro. *Proc Natl Acad Sci* **75:** 4886–4890.

Cao S, Chen SJ. 2006. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res* **34:** 2634–2652.

Cao S, Chen SJ. 2009. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* **15:** 696–706.

Cate JH, Gooding AR, Podell ER, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* **273:** 1678–1685.

Chanturiya AN, Basanez G, Schubert U, Henklein P, Yewdell JW, Zimmerberg J. 2004. PB1-F2, an influenza A virus-encoded proapoptotic mitochondrial protein, creates variably sized pores in planar lipid membranes. *J Virol* **78:** 6304–6312.

Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O'Neill R, Schickli J, Palese P. 2001. A novel influenza A virus mitochondrial protein that induces cell death. *Nat Med* **7:** 1306–1312.

Childs JL, Disney MD, Turner DH. 2002. Oligonucleotide directed misfolding of RNA inhibits *Candida albicans* group I intron splicing. *Proc Natl Acad Sci* **99:** 11091–11096.

Childs JL, Poole AW, Turner DH. 2003. Inhibition of *Escherichia coli* RNase P by oligonucleotide directed misfolding of RNA. *RNA* **9:** 1437–1445.

Childs-Disney JL, Wu M, Pushechnikov A, Aminova O, Disney MD. 2007. A small molecule microarray platform to select RNA internal loop–ligand interactions. *ACS Chem Biol* **2:** 745–754.

Clever J, Sassetti C, Parslow TG. 1995. RNA secondary structure and binding sites for gag gene products in the 5′ packaging signal of human immunodeficiency virus type 1. *J Virol* **69:** 2101–2109.

Compans RW. 1972. Structure of the ribonucleoprotein of influenza virus. *J Virol* **10:** 795–800.

Dam EBT, Pleij CWA, Bosch L. 1990. RNA pseudoknots and translational frameshifting on retroviral, coronaviral and luteoviral RNAs. *Virus Genes* **4:** 121–136.

Dirks RM, Pierce NA. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem* **24:** 1664–1677.

Disney MD, Childs JL, Turner DH. 2004. New approaches to targeting RNA with oligonucleotides: Inhibition of group I intron self-splicing. *Biopolymers* **73:** 151–161.

Disney MD, Labuda LP, Paul DJ, Poplawski SG, Pushechnikov A, Tran T, Velagapudi SP, Wu M, Childs-Disney JL. 2008. Two-dimensional combinatorial screening identifies specific aminoglycoside-RNA internal loop partners. *J Am Chem Soc* **130:** 11185–11194.

Dreher TW. 2009. Role of tRNA-like structures in controlling plant virus replication. *Virus Res* **139:** 217–229.

Dushoff J, Plotkin JB, Viboud C, Earn DJD, Simonsen L. 2006. Mortality due to influenza in the United States—an annualized regression approach using multiple-cause mortality data. *Am J Epidemiol* **163:** 181–187.

Fodor E, Pritlove DC, Brownlee GG. 1994. The influenza virus panhandle is involved in the initiation of transcription. *J Virol* **68:** 4092–4096.

Gallego J, Varani G. 2001. Targeting RNA with small-molecule drugs: Therapeutic promise and chemical challenges. *Acc Chem Res* **34:** 836–843.

Gibbs JS, Malide D, Hornung F, Bennink JR, Yewdell JW. 2003. The influenza A virus PB1-F2 protein targets the inner mitochondrial membrane via a predicted basic amphipathic helix that disrupts mitochondrial function. *J Virol* **77:** 7214–7224.

Gog JR, Afonso EDS, Dalton RM, Leclercq I, Tiley L, Elton D, Von Kirchbach JC, Naffakh N, Escriou N, Digard P. 2007. Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Res* **35:** 1897–1907.

Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. 2010. RNAz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput* **15:** 69–79.

Gultyaev AP, Heus HA, Olsthoorn RCL. 2007. An RNA conformational shift in recent H5N1 influenza A viruses. *Bioinformatics* **23:** 272–276.

Gultyaev AP, Fouchier RAM, Olsthoorn RCL. 2010. Influenza virus RNA structure: Unique and common features. *Int Rev Immunol* **29:** 533–556.

Guo F, Gooding AR, Cech TR. 2004. Structure of the *Tetrahymena* ribozyme: Base triple sandwich and metal ion at the active site. *Mol Cell* **16:** 351–362.

Hagen M, Chung TD, Butcher JA, Krystal M. 1994. Recombinant influenza virus polymerase: Requirement of both 5′ and 3′ viral ends for endonuclease activity. *J Virol* **68:** 1509–1515.

Hall TA. 2001. *BioEdit: A user-friendly biological sequence alignment editor and analysis, version 5.09.* Department of Microbiology, North Carolina State University, Raleigh, NC.

Harmanci AO, Sharma G, Mathews DH. 2007. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics* **8:** 130–150.

Hsu MT, Parvin JD, Gupta S, Krystal M, Palese P. 1987. Genomic RNAs of influenza viruses are held in a circular conformation in virions and in infected cells by a terminal panhandle. *Proc Natl Acad Sci* **84:** 8140–8144.

Hutchinson EC, Curran MD, Read EK, Gog JR, Digard P. 2008. Mutational analysis of *cis*-acting RNA signals in segment 7 of influenza A virus. *J Virol* **82:** 11869–11879.

Ilyinskii PO, Schmidt T, Lukashev D, Meriin AB, Thoidis G, Frishman D, Shneider AM. 2009. Importance of mRNA secondary structural elements for the expression of influenza virus genes. *OMICS* **13:** 421–430.

Jacks T, Madhani HD, Masiarz FR, Varmus HE. 1988. Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. *Cell* **55:** 447–458.

Jang SB, Hung LW, Chi YI, Holbrook EL, Carter RJ, Holbrook SR. 1998. Structure of an RNA internal loop consisting of tandem C-A+ base pairs. *Biochemistry* **37:** 11726–11731.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059–3066.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33:** 511–518.

Kawaoka Y, Gorman OT, Ito T, Wells K, Donis RO, Castrucci MR, Donatelli I, Webster RG. 1998. Influence of host species on the evolution of the nonstructural (NS) gene of influenza A viruses. *Virus Res* **55:** 143–156.

Kieft JS. 2008. Viral IRES RNA structures and ribosome interactions. *Trends Biochem Sci* **33:** 274–283.

Kuo MY, Sharmeen L, Dinter-Gottlieb G, Taylor J. 1988. Characterization of self-cleaving RNA sequences on the genome and antigenome of human hepatitis delta virus. *J Virol* **62:** 4439–4444.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–2948.

Lee MM, Pushechnikov A, Disney MD. 2009. Rational and modular design of potent ligands targeting the RNA that causes myotonic dystrophy 2. *ACS Chem Biol* **4:** 345–355.

Leontis NB, Stombaugh J, Westhof E. 2002. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* **30:** 3497–3531.

Liang Y, Huang T, Ly H, Parslow TG. 2008. Mutational analyses of packaging signals in influenza virus PA, PB1, and PB2 genomic RNA segments. *J Virol* **82:** 229–236.

Lima S, Hildenbrand J, Korostelev A, Hattman S, Li H. 2002. Crystal structure of an RNA helix recognized by a zinc-finger protein: An 18-bp duplex at 1.6 Å resolution. *RNA* **8:** 924–932.

Liu B, Shankar N, Turner DH. 2009. Fluorescence competition assay measurements of free energy changes for RNA pseudoknots. *Biochemistry* **49:** 623–634.

Marsh GA, Hatami R, Palese P. 2007. Specific residues of the influenza A virus hemagglutinin viral RNA are important for efficient packaging into budding virions. *J Virol* **81:** 9727–9736.

Marsh GA, Rabadan R, Levine AJ, Palese P. 2008. Highly conserved regions of influenza A virus polymerase gene segments are critical for efficient viral RNA packaging. *J Virol* **82:** 2295–2304.

Mathews D. 2004. Predicting the secondary structure common to two RNA sequences with Dynalign. *Curr Protoc Bioinformatics* 12.4.1–12.4.11.

Mathews DH, Turner DH. 2002. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* **317:** 191–203.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292.

Mathews DH, Moss WN, Turner DH. 2010. Folding and finding RNA secondary structure. *Cold Spring Harb Perspect Biol* doi: 10.1101/cshperspect.a003665.

Mei HY, Cui M, Heldsinger A, Lemrow SM, Loo JA, Sannes-Lowery KA, Sharmeen L, Czarnik AW. 1998. Inhibitors of protein–RNA complexation that target the RNA: Specific recognition of human immunodeficiency virus type 1 TAR RNA by small organic molecules. *Biochemistry* **37:** 14204–14212.

Muramoto Y, Takada A, Fujii K, Noda T, Iwatsuki-Horimoto K, Watanabe S, Horimoto T, Kida H, Kawaoka Y. 2006. Hierarchy among viral RNA (vRNA) segments in their role in vRNA incorporation into influenza A virions. *J Virol* **80:** 2318–2325.

Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, Fox GE. 2002. NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res* **30:** 395–397.

Nemeroff ME, Utans U, Kramer A, Krug RM. 1992. Identification of *cis*-acting intron and exon regions in influenza virus NS1 mRNA that inhibit splicing and cause the formation of aberrantly sedimenting presplicing complexes. *Mol Cell Biol* **12:** 962–970.

Noda T, Sagara H, Yen A, Takada A, Kida H, Cheng RH, Kawaoka Y. 2006. Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature* **439:** 490–492.

Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. 2004. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res* **32:** 4925–4936.

Plotch SJ, Krug RM. 1986. In vitro splicing of influenza viral NS1 mRNA and NS1-beta-globin chimeras: Possible mechanisms for the control of viral mRNA splicing. *Proc Natl Acad Sci* **83:** 5444–5448.

Plotch SJ, Bouloy M, Ulmanen I, Krug RM. 1981. A unique cap (m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. *Cell* **23:** 847–858.

Pooggin MM, Ryabova LA, He X, Fütterer J, Hohn T. 2006. Mechanism of ribosome shunting in rice tungro bacilliform pararetrovirus. *RNA* **12:** 841–850.

Pushechnikov A, Lee MM, Childs-Disney JL, Sobczak K, French JM, Thornton CA, Disney MD. 2009. Rational design of ligands targeting triplet repeating transcripts that cause RNA dominant disease: Application to myotonic muscular dystrophy type 1 and spinocerebellar ataxia type 3. *J Am Chem Soc* **131:** 9767–9779.

Reiche K, Stadler PF. 2007. RNAstrand: reading direction of structured RNAs in multiple sequence alignments. *Algorithms Mol Biol* **2:** 6–15.

Ryabova LA, Hohn T. 2000. Ribosome shunting in the cauliflower mosaic virus 35S RNA leader is a special case of reinitiation of translation functioning in plant and animal systems. *Genes Dev* **14:** 817–829.

Schroeder S. 2009. Advances in RNA structure prediction from sequence: New tools for generating hypotheses about viral RNA structure-function relationships. *J Virol* **83:** 6326–6334.

Shapiro GI, Krug RM. 1988. Influenza virus RNA replication in vitro: Synthesis of viral template RNAs and virion RNAs in the absence of an added primer. *J Virol* **62:** 2285–2290.

Simmonds P, Smith DB. 1999. Structural constraints on RNA virus evolution. *J Virol* **73:** 5787–5794.

Sokal R, Michener C. 1958. A statistical method for evaluating systematic relationships. *U Sci Pap Univ Kansas Nat Hist Mus* **38:** 1409–1438.

Sperschneider J, Datta A. 2010. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Res* **38:** e103. doi: 10.1093/nar/gkq021.

Sucheck SJ, Wong CH. 2000. RNA as a target for small molecules. *Curr Opin Chem Biol* **4:** 678–686.

Theimer CA, Wang Y, Hoffman DW, Krisch HM, Giedroc DP. 1998. Non-nearest neighbor effects on the thermodynamics of unfolding of a model mRNA pseudoknot. *J Mol Biol* **279:** 545–564.

Tuplin A, Evans DJ, Simmonds P. 2004. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. *J Gen Virol* **85:** 3037–3047.

Uzilov AV, Keegan JM, Mathews DH. 2006. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics* **7:** 173–202.

Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. 2005a. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* **23:** 1383–1390.

Washietl S, Hofacker IL, Stadler PF. 2005b. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci* **102:** 2454–2459.

Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460:** 711–716.

Weiner AM, Maizels N. 1987. tRNA-like structures tag the 3′ ends of genomic RNA molecules for replication: Implications for the origin of protein synthesis. *Proc Natl Acad Sci* **84:** 7383–7387.

Wilson WD, Li K. 2000. Targeting RNA with small molecules. *Curr Med Chem* **7:** 73–98.

Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37:** 14719–14735.

Yamano H, Gannon J, Hunt T. 1996. The role of proteolysis in cell cycle progression in *Schizosaccharomyces pombe*. *EMBO J* **15:** 5268–5279.

Ye Q, Krug RM, Tao YJ. 2006. The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. *Nature* **444:** 1078–1082.

Zamarin D, García-Sastre A, Xiao X, Wang R, Palese P. 2005. Influenza virus PB1-F2 protein induces cell death through mitochondrial ANT3 and VDAC1. *PLoS Pathog* **1:** e4. doi: 10.1371/journal.ppat.0010004.

Zamarin D, Ortigoza MB, Palese P. 2006. Influenza A virus PB1-F2 protein contributes to viral pathogenesis in mice. *J Virol* **80:** 7976–7983.