Cardioviral RNA structure logo analysis: entropy, correlations, and prediction

Xiao-Zhou Chen · Huai Cao · Wen Zhang · Ci-Quan Liu

Received: 6 March 2008 / Accepted: 14 April 2009 /
Published online: 29 August 2009
© Springer Science + Business Media B.V. 2009

Abstract In recent years, there has been an increased number of sequenced RNAs leading to the development of new RNA databases. Thus, predicting RNA structure from multiple alignments is an important issue to understand its function. Since RNA secondary structures are often conserved in evolution, developing methods to identify covariate sites in an alignment can be essential for discovering structural elements. Structure Logo is a technique established on the basis of entropy and mutual information measured to analyze RNA sequences from an alignment. We proposed an efficient Structure Logo approach to analyze conservations and correlations in a set of Cardioviral RNA sequences. The entropy and mutual information content were measured to examine the conservations and correlations, respectively. The conserved secondary structure motifs were predicted on the basis of the conservation and correlation analyses. Our predictive motifs were similar to the ones observed in the viral RNA structure database, and the correlations between bases also corresponded to the secondary structure in the database.

This work was partly supported by NSFC (Natural Science Foundation of China, Nos. 90303018 and 90208018). We thank the referees for help.

X.-Z. Chen · H. Cao · C.-Q. Liu
Modern Biology Research Center, Yunnan University, Kunming 650091, China

X.-Z. Chen
School of Mathematics and Computer Science, Yunnan Nationalities University,
Kunming 650031, China
e-mail: ch_xiaozhou@yahoo.com.cn

W. Zhang
Department of Cell Biology and Genetics, Kunming Medical College, Kunming 650031, China

C.-Q. Liu
Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

C.-Q. Liu (✉)
Theoretical Biology Research Center, Peking University, Beijing 100871, China
e-mail: liucq@ynu.edu.cn

Keywords RNA structure logo • Entropy • Correlations • Predictive motif

1 Introduction

The prediction of secondary and tertiary structures of RNA molecules is difficult. A single RNA strand can form hairpin structures, which is the common motif of the RNA secondary structure. Although the RNA secondary structure does not determine the function of the RNA molecule, it is integral for the function of the RNA [1, 2]. It has been observed that genetically distant groups of RNA viruses exhibit little or no detectable sequence homology due to the high rate of mutations, thus implying a rapid evolution at the sequence level. Functional secondary structures evolve much slower than the underlying sequences. The analysis of RNA structures can yield insights into the evolution of RNAs. In addition, the use of RNA viruses as a foundation for advances towards functional genome analysis is beneficial. The prediction of the secondary structural elements for RNA molecules provides a means for functional genome analysis towards understanding viral phylogeny [3]. With the availability of increasing numbers of sequenced RNAs and the establishment of new RNA databases, such as the ribosomal RNA database [4], the Comparative RNA Web Site [5], and Rfam [6], there is a need for accurately and automatically predicting RNA structures from multiple alignments. Structural prediction from multiple sequence alignments has proven to be a powerful tool [7–9]. Various methods such as RNAalifold [10], ConStruct [11], and COVE [12] have been proposed for automating this process.

One of the methods for identifying conserved secondary structures from multiple alignments of RNA sequences is the mutual information measure [13]. This measure can be used to detect covarying sites [14, 15]. A sequence logo was invented to display patterns in sequence conservation to assist in discovering and analyzing those patterns [16] and to analyze splice sites [17, 18]. It can also provide indications of motifs such as base-paired sites in folded RNA structures [18, 19]. Structure logos [20, 21] can cope with any prior nucleotide distribution as well as allowing for gaps in the alignments and can indicate mutual information of base-paired positions in RNA sequences.

In this paper, we used an entropy and mutual information measure to extend structure logos for RNA structure prediction from a multiple alignment. We applied this method to perform a correlation analysis and check the sequences for primary structure similarity. From this analysis, we developed a predictive motif for detecting nucleotide sequences.

2 Dataset

We searched for functionally important structures in viral genomes and viral messenger RNAs using structure-based alignments, which are conserved among a group of closely related viruses. These structures were then verified for the presence of compensatory mutations. RNAs are ideal candidates for developing novel approaches toward functional genome analysis. We proposed to select genomic RNAs of *Cardiovirus* for our investigation, which belongs to the family Picornaviridae (sRNA virus). This virion has no envelope and possesses an icosahedral symmetry of around 30 nm in diameter. Its RNA viral genome is 8,400 nucleotides, and it is positive-sensed and single-stranded. The 3' region encodes a polyA, while the 5' encodes a genome-linked protein. The *Cardiovirus* is comprised of two

Table 1 Genus *Cardiovirus*

ID	Acc. no.	Length	Virus	Strain
MNGPOLY	L22089	7761	EMCV	Mengo
EMCPOLYP	M81861	7835	EMCV	Ruckert
EMCBCG	M22457	7825	EMCV	B
EMC5ES	K01410	149 (Sfrag)	EMCV	Russian
TMEPP	M16020	8098	TMEV	BeAn 8386
TMEGDVCG	M20562	8105	TMEV	GDVII
TMECG	M20301	8093	TMEV	DA

species, *Encephalomyocarditis virus* (EMCV) and *Theilovirus*. It infects vertebrates and is transmitted from rodents to other animals. Currently, there are no effective treatments. Thus, studying the conserved secondary structural elements and biological function of its viral RNA molecules would help to understand *Cardiovirus* taxonomy and classification, properties, and prevention and control.

The aligned sequences of genus *Cardiovirus* were taken from a publicly available database, the Viral RNA Structure Database (<http://rna.tbi.univie.ac.at/cgi-bin/virusdb.cgi>), which provides both primary sequence similarity and the relationships between sites within the sequences. The sequences were initially aligned based on their primary structure similarity. The secondary structure interactions were then identified and used to improve the alignment. The *Cardiovirus* sequence dataset is shown in Table 1.

Among the eight sequences, the lengths range from 149 to 8105 nt. Varying alignments were used for different regions of the genome by CLUSTAL W. Regions of the 5'-NTR, the coding, and the 3'-NTR of the genome are listed for alignment. Multiple sequence alignments that are allowed are shown in Table 2.

Correlation analysis was applied for all aligned sequences identified in the RNA *Cardiovirus* Genome. A partial list is presented in Tables 3, 4, 5, and 6. These parts are 589 to 613, 649 to 689, 765 to 847, and 6649 to 6686 nt, with the lengths ranging from 25 to 83 nt. The parentheses indicate base pairing. Tables 3, 4, 5, and 6 show the four subsequences along with the base-pairing interactions that were used to align them. As can be seen, gaps have been inserted in the sequences in order to align them. These positions were used in the analysis of the RNA sequences. Multiple sequence alignment of the four ranges is shown in Tables 3, 4, 5, and 6, respectively.

3 Methods

3.1 Entropy

The primary-aligned sequences for each position i was measured in terms of Shannon entropy, H_{ik} , defined as

$$H_{ik} = - \sum_{k \in \Lambda} p_{ik} \log_2 p_{ik} \quad (3.1)$$

where $\Lambda \in \{A, C, G, U, -\}$ is the set of bases; p_{ik} is the observed frequency of base k at position i . It also measures the degree of variation among the different nucleotides at each

Table 3 Multiple-sequence alignment of the four ranges

TMEPP	UCCACAGUGUUCUAU-ACUGUGGA
TMEGDVCG	UUCACAGUGUUCUAU-GCUGUGAA
TMECG	UCCACAGUGCUCUAU-ACUGUGGA
EMCPOLYP	CAUAUUGCCGUCUUUUGGCAAUGUG
EMCBCG	CACAUCACCGUCUUUUGGUGGUGUG
MNGPOLY	UACAUUGUCGUCUGU-GACGAUGUA
FOLDALIGN has at least 8 base pairs in the stem	(((((((.....)))))))))

Table 4 Multiple-sequence alignment of the four ranges

TMEPP	ACGUGCGCGCGGUCUUU-CCGUCUCUCGACAAGCGCGCGU
TMEGDVCG	ACACGCGUGGCGGUUUUUCCGUCUCUCGACAAGCGCGGUGU
TMECG	ACGUGCGCGCGCGGUCUUU-CCGUCUCUCGACAAGCGCGCGU
EMCPOLYP	GCAUUCUAGGGGUCUUU-CCCCUCUCGCCAAAGGAAUGC
EMCBCG	GUAUUCUAGGGGUCUUU-CCCCUCUCGACAAAGGAAUAC
MNGPOLY	GUAUUCUAGGGGUCUUU-CCCCUCUCGACAAAGGAAUAC
FOLDALIGN has at least 11 base pairs in the stem	(((((((.(((.....)))......)))))))))

Table 5 Multiple-sequence alignment of the four ranges

TMEPP	CACACAAAGGCAGCGGAACCCCCCUCUGGUAACAGGAGCC
TMEGDVCG	CACACAAAGGCAGCGGAACCCCCCUCUGGUAACAGGAGCC
TMECG	CACACAAAGGCAGCGGAACCCCCCUCUGGUAACAGGAGCC
EMCPOLYP	CCUUUGCAGGCAGCGGAACCCCCCACCUGGCGACAGGUGCC
EMCBCG	CCUUUGCAGGCAGCGGAUUUUUUCCCCCACCUGGUAACAGGUGCC
MNGPOLY	CCUUUGCAGGCAGCGGAUUUUUUCCCCCACCUGGUGACAGGUGCC
	(((((((((((.(((......((((....))))))))))))))))))
TMEPP	UCUGCGGCCAAAAGCCACGUGGAUAAGAUAUCCACCUUUGUGUG
TMEGDVCG	UCUGCGGCCAAAAGCCACGUGGAUAAGAUAUCCACCUUUGUGUG
TMECG	UCUGCGGCCAAAAGCCACGUGGAUAAGAUAUCCACCUUUGUGUG
EMCPOLYP	UCUGCGGCCAAAAGCCACGUGUAUAAGAUAUACACCUGCAAAGG
EMCBCG	UCUGCGGCCAAAAGCCACGUGUAUAAGAUAUACACCUGCAAAGG
MNGPOLY	UCUGCGGCCGAAAGCCACGUGUAAGACACACCUGCAAAGG
	..)))(....))..(((((((....))))))))))))))

FOLDALIGN has at least 30 base pairs in the stem

Table 6 Multiple-sequence alignment of the four ranges

TMEPP	GCCGCUACCAUCAUCACCAAGGAAUUGAUUGAAGCAGC
TMEGDVCG	GCCGCUACCAUCAUCACAGAGAGUUGAUUGAAGCAGC
TMECG	GCCGCUACCAUCAUCACAGAGAGUUGAUUGAAGCAGC
EMCPOLYP	GCCGCCUCGAUUGUGUCACAGGAGAUAUUCGGGCGGU
EMCBCG	GCCGCCUCGAUUGUGUCACAAGAAAUGAUCUGUGCGGU
MNGPOLY	GCCGCGUCGAUAAUUUCACAAGAAAUGAUCGAUGCGGU
	(((((....((((.....))))....))))))

FOLDALIGN has at least 9 base pairs in the stem

position. For aligned sequences, the appropriate information measures some position i as the information content, I_i , defined as

$$I_i = \sum_{k \in \Lambda} I_{ik} = \sum_{k \in \Lambda} q_{ik} \log_2 \frac{q_{ik}}{p_k} \quad (3.2)$$

where $p_k = 0.25$ is a prior distribution of the bases and q_{ik} is the fraction of base k at position i . We set $p_- = 1$, since $I_{i-} = q_{i-} \log_2 q_{i-} = 0$, when $q_{i-} = 1$ or $q_{i-} = 0$. Information content increases with an increase in the number of the same nucleotides that occupy the position. It can reach a maximum of 2 bits if the same nucleotide occupies the particular position for all sequences in the alignment. The information content calculation was performed for each position in the sequence alignment including gaps from the calculation [22].

3.2 Sequence conservation

The sequence conservation R_{seq} [16] at position i was defined as

$$R_{\text{seq}} = \log_2 N - H_{ik} \quad (3.3)$$

where $N = 4$ denotes the number of distinct bases. The more conserved the sequence, the higher the value R_{seq} . Consequently, the complete sequence conservation per site reaches the maximum possible entropy $\log_2 4 = 2$ bits. For binding sites, it has been found that this total entropy is approximately equal to the amount of information [22].

3.3 Mutual information

Mutual information can be employed to detect compensatory mutations in an alignment [23, 24]. To compute the mutual information M_{ij} of columns i and j in an alignment, the following statistics were required: the frequency p_k of base X in column k , for $k = i, j$, where X is A, C, G, U , or $-$; and the joint frequency p_{ij} of complementary bases $X \cdot Y$, where $X \cdot Y$ is $G \cdot C, C \cdot G, A \cdot U, U \cdot A, G \cdot U$, or $U \cdot G$. The mutual information is defined as

$$M_{ij} = \sum_{i,j} p_{ij}(X \cdot Y) \log_2 \frac{p_{ij}(X \cdot Y)}{p_i(X) p_j(Y)}. \quad (3.4)$$

Note that the mutual information content achieves its lower bound of zero ($M_{ij} = 0$), if the two positions of i and j are independent of one another [$p_{ij}(X \cdot Y) = p_i(X) p_j(Y)$]. The mutual information content reaches its upper bound of value ($M_{ij} = 2$) if the variables are perfectly correlated [$p_{ij}(X \cdot Y) = p_i(X) = p_j(Y)$]. The higher the mutual information calculation is, the more likely that the two positions are correlated. The nucleotide at one position can be estimated with a high likelihood due to the presence of another nucleotide at a separate position [25].

The entropy value was calculated for the primary sequence conservation. Each position was ranked according to its entropy calculation. The mutual information content between all pairs of positions in the domain of the sequences was performed to analyze the correlations. Based on the above analyses, a predictive motif was constructed to represent the aligned RNA sequences. Each RNA sequence was tested against the motif, and the number of mismatches were computed and used as a measurement to score the ability of the motif to represent the sequences.

4 Results

4.1 Primary sequence analysis

In Fig. 1, we plot structural logos of four aligned subsequences in the regions 589 to 613 nt, 649 to 689 nt, 765 to 847 nt, and 6649 to 6686 nt, respectively.

Fig. 1 The logos for the data. The symbols for sequence alignment are *A*, *U*, *G*, *C* and *–*. The letter involving structure is *M*. The sequence-structure logos are **a** for the range 589 to 613 nt; **b** for the range 649 to 689 nt; **c** for the range 765 to 847 nt; **d** for the range 6649 to 6686 nt. The letters *a*, *b*, *c*, *d*, *e*, and *f* indicate base pairing for respective regions

Table 7 Primary sequence analysis

Pos	Entropy	Rank	Pos	Entropy	Rank	Pos	Entropy	Rank
Range: 589 to 613			Range: 649 to 689					
589	1.081	5	649	1.207	6	670	2.000	12
590	0.540	3	650	1.263	8	671	1.500	11
591	1.349	6	651	1.263	8	672	1.000	3
592	2.000	7	652	1.412	10	673	2.000	12
593	1.000	4	653	1.500	11	674	2.000	12
594	0.540	3	654	2.000	12	675	2.000	12
595	1.349	6	655	1.500	11	676	2.000	12
596	1.081	5	656	1.213	7	677	2.000	12
597	1.000	4	657	1.000	3	678	2.000	12
598	0.540	3	658	2.000	12	679	1.350	9
599	2.000	7	659	1.500	11	680	2.000	12
600	2.000	7	659	1.500	11	680	2.000	12
601	2.000	7	659	1.500	11	680	2.000	12
602	0.540	3	660	2.000	12	681	2.000	12
603	2.000	7	661	2.000	12	682	2.000	12
604	-0.251	1	662	2.000	12	683	1.131	5
605	0.000	2	663	1.350	9	684	1.500	11
606	1.000	4	664	2.000	12	685	2.000	12
607	1.349	6	665	2.000	12	686	1.500	11
608	0.540	3	666	2.000	12	687	1.062	4
609	1.081	5	667	-0.310	1	688	1.263	8
610	2.000	7	668	0.000	2	689	1.263	8
611	2.000	7	668	0.000	2	689	1.263	8
612	0.540	3	669	2.000	12			
613	1.081	5				Range: 6649 to 6686		
Range: 765 to 847								
765	2.000	5	807	2.000	5	6649	2.000	7
766	1.000	2	808	2.000	5	6650	2.000	7
767	1.000	2	809	2.000	5	6651	2.000	7
768	1.000	2	810	2.000	5	6652	2.000	7
769	1.000	2	811	2.000	5	6653	2.000	7
770	1.000	2	812	2.000	5	6654	0.715	3
771	1.000	2	813	2.000	5	6655	1.000	5
772	2.000	5	814	2.000	5	6656	2.000	7
773	2.000	1	815	1.349	4	6657	1.000	5
774	2.000	5	816	2.000	5	6658	2.000	7
775	2.000	5	817	2.000	5	6659	2.000	7
776	2.000	5	818	2.000	5	6660	0.540	1
777	2.000	5	819	2.000	6	6661	1.081	6
778	2.000	5	820	2.000	5	6662	2.000	7
779	2.000	5	821	2.000	5	6663	0.540	1
780	2.000	5	822	2.000	5	6664	1.000	5
781	2.000	5	823	2.000	5	6665	2.000	7
782	2.000	5	824	2.000	5	6666	1.000	5
783	0.748	1	825	2.000	5	6667	1.000	5
784	1.349	4	826	2.000	5	6668	1.081	6
785	2.000	5	827	1.000	2	6669	1.081	6
786	2.000	5	828	1.349	4	6670	2.000	7
787	2.000	5	829	2.000	5	6671	2.000	7
788	2.000	5	830	2.000	5	6672	1.000	5

Table 7 (continued)

Pos	Entropy	Rank	Pos	Entropy	Rank	Pos	Entropy	Rank
789	1.000	2	831	2.000	5	6673	1.000	5
790	2.000	5	832	2.000	5	6674	2.000	7
791	2.000	5	833	2.000	5	6675	2.000	7
792	2.000	5	834	2.000	5	6676	2.000	7
793	2.000	5	835	2.000	1	6677	2.000	7
794	2.000	5	836	2.000	1	6678	1.081	6
795	1.349	4	837	2.000	1	6679	0.748	4
796	1.081	3	838	2.000	1	6680	1.081	6
797	2.000	5	839	2.000	1	6681	0.590	2
798	2.000	5	840	2.000	5	6682	2.000	7
799	2.000	5	841	1.000	2	6683	2.000	7
800	2.000	5	842	1.000	2	6684	1.000	5
801	2.000	5	843	1.000	2	6685	2.000	7
802	1.000	2	844	1.000	2	6686	1.000	5
803	2.000	5	845	1.000	2			
804	2.000	5	846	1.000	2			
805	2.000	5	847	2.000	5			
806	2.000	5						

The structural logos contain mutual information from the base pairs. The M symbols indicate that there are significant constraints on the sequences. For some paired positions, such as positions 7, 11, 22, 35, and 37, there are highly conserved structures as well as significant preferences for particular bases (Fig. 1b). This is due to the value of the sequences' entropy and mutual information that were nearly as great as 2 bits. The logos showed that some positions are completely conserved so that the total information is 2 bits, such as positions 4, 11, 12, 13, 15, 22, and 23 in Fig. 1a; 6, 10, 12–14, 16–18, 21, 24–29, 31–33, 36 in Fig. 1b; 1, 8–18, 21–24, 26–30, 33–37, 39–50, 52–62, 65–74, 83 in Fig. 1c; and 1–5, 8, 10–11, 14, 17, 22–23, 26–29, 34–35, 37 in Fig. 1d.

The results of these analyzed primary sequences are shown in Table 7. For each position in the aligned sequences, the entropy shows the conservation of the primary sequence. Each position was ranked according to its entropy calculation. Positions with a low degree of primary sequence conservation were ranked lower while those with a high degree of conservation were ranked higher in the range.

As seen from Table 7, the positions 592, 599–601, 603, 610, 611, 654, 658, 660–662, 664–666, 669, 670, 673–678, 680–682, 685, 765, 772–782, 785–788, 790–794, 797–801, 803–814, 816–826, 829–840, 847; 6649–6653, 6656, 6658, 6659, 6662, 6665, 6670, 6671, 6674–6677, 6682, 6683, and 6685 are completely conserved ($E = 2.00$). The positions 589, 591, 593, 595–597, 606, 607, 609, 613; 649–653, 655–657, 659, 663, 671, 672, 679, 683, 684, 686–689; 765–771, 784, 789, 795, 796, 802, 815, 827, 828, 841–846; 6655, 6657, 6661, 6664, 6667–6669, 6672–6673, 6678, 6680, 6684, and 6686 ($E \geq 1.00$) have some degree of primary sequence conservation. Over 65% of the positions in the sequence alignment were found to be conserved ($E \geq 1.00$). This suggests that the RNA sequences may have the potential to maintain their higher-order structures, despite some degree of variability in their primary sequence. In addition, this may explain the conserved relationships between the pairs of positions in the sequence alignment. Further discussion is provided below regarding the analysis of the mutual information content of the RNA sequences.

4.2 Correlation analysis

In Fig. 2, we show mutual information plots of three aligned subsequences ranging from 589 to 613 nt, 649 to 689 nt, 765 to 847 nt, and 6649 to 6686 nt, respectively. The plot depicts alignments contained along the principal covarying regions, in particular, the main stem that has several base pairs between positions. The mutual information between pairs of positions in the sequence alignment is illustrated by the image. Purple intensities in the image indicate position pairs that have a high degree of correlation with one another. The correlated pairs of positions and the calculated mutual information for them are listed in Table 8 along with their postulated relationships between the position pairs. The relative frequencies of the nucleotides for each correlated position were examined and used to determine the relationship for each pair.

Fig. 2 The mutual information content of an RNA alignment of the *Cardiovirus* sequences. A combined plot of sequence information and gap frequencies is displayed along the edges. The mutual information is indicated by the *color scale*. The sequence information on the logo profiles is indicated in *black*, the gap frequencies in *gray*. The plots are **a** for the range 589 to 613 nt; **b** for the range 649 to 689 nt; **c** for the range 765 to 847 nt; and **d** for the range 6649 to 6686 nt

Table 8 High correlations among pairs of positions

Position 1	Position 2	Type of interaction	Mutual information
Range: 589 to 613			
589 (U)	613 (A)	U–A base pairing	0.363
590 (A)	612 (U)	A–U base pairing	1.170
591 (C)	594 (A)	Any base pairing	0.126
593 (C)	609 (G)	C–G base pairing	0.263
594 (A)	608 (U)	A–U base pairing	1.082
595 (G)	607 (C)	G–C base pairing	0.216
596 (U)	606(A)	U–A base pairing	0.263
597 (G)	605 (G)	3-way interaction	0.208
598 (G)	604 (–)	Any base pairing	0.119
598 (G)	602 (A)	Any base pairing	0.183
Range: 649 to 689			
649 (A)	689 (U)	A–U base pairing	0.415
650 (C)	688 (G)	C–G base pairing	0.363
651 (A)	687 (U)	A–U base pairing	0.363
652 (U)	686 (G)	U–G base pairing	0.126
653 (G)	685 (C)	G–C base pairing	1.000
655 (G)	683 (C)	G–C base pairing	1.000
656 (U)	682 (G)	U–G base pairing	0.263
Range: 765 to 847			
766 (A)	846 (U)	A–U base pairing	1.000
767 (C)	845 (G)	C–G base pairing	0.415
768 (A)	844 (U)	A–U base pairing	1.000
769 (C)	843 (G)	C–G base pairing	0.415
770 (A)	842 (U)	A–U base pairing	0.415
771 (A)	841 (U)	A–U base pairing	1.000
789 (U)	802 (A)	U–A base pairing	1.000
827 (G)	835 (C)	G–C base pairing	1.000
828 (A)	834 (U)	A–U base pairing	0.125
Range: 6649 to 6686			
6654(U)	6681(A)	U–A base pairing	0.350
6655(A)	6679(G)	3-way interaction	0.208
6655(A)	6681(A)	3-way interaction	0.350
6657(C)	6679(G)	C–G base pairing	0.350
6657(C)	6680(A)	3-way interaction	0.263
6660(C)	6678(U)	Any base pairing	0.170
6661(A)	6666(C)	3-way interaction	0.263
6663(C)	6679(G)	C–G base pairing	0.848
6664(A)	6673(U)	A–U base pairing	1.000
6667(A)	6670(A)	3-way interaction	0.263
6667(A)	6668(A)	3-way interaction	0.263
6672(A)	6678(U)	A–U base pairing	0.119

Some distinct areas have a high degree of mutual information, such as position pairs (590, 612), (594, 608), (653, 685), (655, 683); (766, 846), (768, 844), (771, 841), (789, 802), (827, 835), (6664, 6673). These regions of high mutual information appear to interact through U–A and C–G base pairing and form the secondary structure of the RNA. In addition, we observed U–G base pairing interactions in our analysis, such as (652, 686) and (656, 682). Other position pairs, such as (597, 605), (663, 685), (6655, 6681),

Table 9 Motifs for the RNA sequences

Range 589 to 613		Range 649 to 689		Range 765 to 847				Range 6649 to 6686	
Pos	Base	Pos	Base	Pos	Base	Pos	Base	Pos	Base
589	U	649	A	765	C	807	C	6649	G
590	A	650	C	766	A	808	U	6650	C
591	C	651	A	767	C	809	G	6651	C
592	A	652	U	768	A	810	G	6652	G
593	C	653	G	769	C	811	G	6653	C
594	A	654	C	770	A	812	G	6654	U
595	G	655	G	771	A	813	C	6655	A
596	U	656	U	772	A	814	C	6656	C
597	G	657	G	773	G	815	A	6657	C
598	G	658	G	774	G	816	A	6658	A
599	U	659	C	775	C	817	A	6659	U
600	C	660	G	776	A	818	A	6660	C
601	U	661	G	777	G	819	G	6661	A
602	A	662	U	778	C	820	C	6662	U
603	U	663	C	779	G	821	C	6663	C
604	–	664	U	780	G	822	A	6664	A
605	G	665	U	781	A	823	C	6665	C
606	A	666	U	782	A	824	G	6666	C
607	C	667	–	783	C	825	U	6667	A
608	U	668	C	784	C	826	G	6668	A
609	G	669	C	855	C	827	G	6669	A
610	U	670	G	786	C	828	A	6670	G
611	G	671	U	787	C	829	U	6671	A
612	U	672	C	788	C	830	A	6672	A
613	A	673	U	789	U	831	A	6673	U
		674	C	790	C	832	G	6674	U
		675	U	791	C	833	A	6675	G
		676	C	792	U	834	U	6676	A
		677	G	793	G	835	C	6677	U
		678	A	794	G	836	C	6678	U
		679	C	795	U	837	A	6679	G
		680	A	796	A	838	C	6680	A
		681	A	797	A	839	C	6681	A
		682	G	798	C	840	U	6682	G
		683	C	799	A	841	U	6683	C
		684	G	800	G	842	U	6684	A
		685	C	801	G	843	G	6685	G
		686	G	802	A	844	U	6686	C
		687	U	803	G	845	G		
		688	G	804	C	846	U		
		689	U	805	C	847	G		
				806	U				

(6657, 6680), (6661, 6666), (6667, 6668), and (6667, 6670) may be involved in three-way interactions.

Some of the base-pairing interactions have a certain degree of variability ($M = 0.1$ or so). These were found at position pairs (591, 594), (598, 602), (598, 604), (828, 834), (652, 686), and (6660, 6678). Any base-pairing interactions may occur between these positions

with any of the four nucleotides. Therefore, there is a certain degree of primary sequence variability. However, as can be seen in the mutual information calculations, these positions are highly correlated with one another through base-pairing interactions. For example, if some position has an adenine (A), it is highly likely that the position pair will have a uracil (U). Thus, although the RNA sequences can vary greatly in their primary structure, their structures and functions will be preserved as long as their base-pairing interactions are maintained.

The U–A (A–U) and C–G (G–C) interactions observed in our analyses were consistent with those predicted by RNAfold in the Viral RNA Structure Database. In addition, there are other base-pairing interactions, such as the G–U pair, the three-way interactions, and any base pairing which may be important factors for determining the tertiary structure of RNA.

4.3 Predictive motif

RNA molecules can fold back upon themselves to form stem regions and unpaired regions, consisting of hairpin loops, bulges, internal loops, multi-branched loops, start sequences, and external sequences. The interactions between the secondary structural elements form the tertiary structures. Currently, predicting the three-dimensional structure of RNA molecules is difficult with existing knowledge and methods. Successful prediction of the secondary structure is necessary to understand the tertiary structures of RNA. Due to the complexity of the RNA structures, two main approaches used to predict the secondary structure, namely minimum energy [26–28] and comparative sequence analysis [29, 30], have been developed. The first method looks for the lowest free energy out of the possible folds and forms a stable structure. The latter takes into consideration the similarity of the primary sequence and the relationships between sites within the sequences.

We introduced the concepts of entropy and mutual information from information theory. Entropy value measures the degree of sequence conservation in multiple sequence alignments. Mutual information content measures correlation for a pair of nucleotide sites. From sequence conservation and base-pairing interaction analysis, we can correctly predict the conserved secondary structural elements of a set of aligned *Cardiovirus* RNA sequences. The motif is listed by position in Table 9. All symbols follow the aligned DOT-plots as seen in the Viral RNA Structure Database. The predictive motif may describe the set for the nucleotides themselves.

5 Conclusion

We present a method in which entropy and mutual information are used to identify elements for RNA structural prediction from multiple alignments. We used RNA Structure Logo to perform primary sequence and correlative analysis on a set of aligned Cardioviral RNA sequences. We found that the primary sequences display some degree of variability but had conserved base-pairing interactions among distinct sites within the alignment. These relationships helped determined the secondary and tertiary structures of the RNA molecules and may affect their functions. From our analysis, we developed a predictive motif to describe the set of RNA. The RNA sequences used in our study were from the genus *Cardiovirus*. Based on our analysis, we demonstrated that the generated motifs are similar to the ones observed from the Viral RNA Structure Database, and the correlations between

the bases were similar to the ones corresponding to the secondary structures in the database. Future work will involve checking correlations between bases and to predicting all motifs from other viruses in the viral RNA structure database.

Acknowledgements This work was partly supported by NSFC (Natural Science Foundation of China, No. 90303018, and No. 90208018). The authors thank the reviewers for their comments in the improvement of the manuscript.

References

1. Kim, S.H., Suddath, G.J., Quigley, G.J., et al.: Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185**, 435–439 (1974)
2. Shi, H., Moore, P.B.: The crystal structure of yeast phenylalanine tRNA at 1.92 Å resolution: a classic structure revisited. *RNA* **6**, 1091–1105 (2000)
3. Witwer, C., Rauscher, S., Hofacker, I., et al.: Conserved RNA secondary structures in Picornaviridae genomes. *Nucleic Acids Res.* **29**, 5079–5089 (2001)
4. Wuyts, J., Perriere, G., Van De Peer, Y.: The European ribosomal RNA database. *Nucleic Acids Res.* **32**, 101D–103D (2004)
5. Cannone, J.J., Subramanian, S., Schnare, M.N., et al.: The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**, 2 (2002)
6. Griffiths-Jones, S., Bateman, A., Marshall, M., et al.: Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–444 (2003)
7. Rosenblad, M.A., Gorodkin, J., Knudsen, B., et al.: SRPDB: Signal recognition particle database. *Nucleic Acids Res.* **31**, 363–364 (2003)
8. De los Monteros, E.A.: Models of the primary and secondary structure for the 12 S rRNA of birds: a guideline for sequence alignment. *DNA Seq.* **14**, 241–256 (2003)
9. Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., et al.: Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA* **9**, 1084–1097 (2003)
10. Hofacker, I.L., Fekete, M., Stadler, P.F.: Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066 (2002)
11. Luck, R., Graf, S., Steger, G.: ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.* **27**, 4208–4217 (1999)
12. Eddy, S.R., Durbin, R.: RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22**, 2079–2088 (1994)
13. Andrew, K.C., Wong, D., Chiu, K.Y.: An event-covering method for effective probabilistic inference. *Pattern Recogn.* **20**, 245–255 (1987)
14. Chiu, D.K., Kolodziejczak, T.: Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.* **7**, 347–352 (1991)
15. Gutell, R.R., Power, A., Hertz, G.Z., et al.: Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.* **20**, 5785–5795 (1992)
16. Schneider, T.D., Stephens, R.M.: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990)
17. Emmert, S., Schneider, T.D., Khan, S.G., Kraemer, K.H.: The human XPG gene: Gene architecture, alternative splicing and single nucleotide polymorphisms. *Nucleic Acids Res.* **29**, 1443–1452 (2001)
18. Schneider, T.D.: Information content of individual genetic sequences. *J. Theor. Biol.* **189**, 427–441 (1997)
19. Stormao, G.D.: Information content and free energy in DNA–protein interactions. *J. Theor. Biol.* **195**, 135–137 (1998)
20. Gorodkin, J., Heyer, L.J., Stormo, G.D.: Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.* **25**, 3724–3732 (1997)
21. Gorodkin, J., Heyer, L.J., Brunak, S., et al.: Displaying the information contents of structural RNA alignments: the structure logos. *CABIOS* **13**, 583–586 (1997)
22. Schneider, T.D., Stormao, G.D., Gold, L., Ehrenfeuch, A.: Information content of binding sites. *J. Mol. Biol.* **188**, 415–431 (1986)

23. Chiu, D.K., Kolodziejczak, T.: Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.* **7**, 347–352 (1991)
24. Gutell, R.R., Power, A., Hertz, G.Z., et al.: Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.* **20**, 5785–5795 (1992)
25. Gorodkin, J., Stærfeldt, H.H., Lund, O., Brunak, S.: MatrixPlot: visualizing sequence constraints. *Bioinformatics* **15**, 769–770 (1999)
26. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9**, 133–148 (1981)
27. Mathews, D.H., Sabina, J., Zuker, M., et al.: Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999)
28. Zuker, M.: On finding all suboptimal foldings of an RNA molecule. *Science* **244**, 48–52 (1989)
29. Eddy, S.R., Durbin, R.: RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22**, 2079–2088 (1994)
30. Knudsen, B., Hein, J.: RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **15**, 446–454 (1999)