Published online 21 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W91–W96***
doi:10.1093/nar/gkn258

RNALogo: a new approach to display structural RNA alignment

Tzu-Hao Chang[1], Jorng-Tzong Horng[1,2] and Hsien-Da Huang[3,4,*]

[1]Department of Computer Science and Information Engineering, [2]Department of Life Science, National Central University, Chung-Li 320, [3]Institute of Bioinformatics and [4]Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan

Received February 19, 2008; Revised April 7, 2008; Accepted April 20, 2008

ABSTRACT

Regulatory RNAs play essential roles in many essential biological processes, ranging from gene regulation to protein synthesis. This work presents a web-based tool, RNALogo, to create a new graphical representation of the patterns in a multiple RNA sequence alignment with a consensus structure. The RNALogo graph can indicate significant features within an RNA sequence alignment and its consensus RNA secondary structure. RNALogo extends Sequence logos, and specifically incorporates RNA secondary structures and mutual information of base-paired regions into the graphical representation. Each RNALogo graph is composed of stacks of letters, with one stack for each position in the consensus RNA secondary structure. RNALogo provides a convenient and high configurable logo generator. An RNALogo graph is generated for each RNA family in Rfam, and these generated logos are accumulated into a gallery of RNALogo. Users can search or browse RNALogo graphs in this gallery to receive additional perspectives of known RNA families. RNALogo is now available at: http://rnalogo.mbc.nctu.edu.tw/.

INTRODUCTION

Regulatory RNA molecules, including iron responsive elements (IRE), riboswitches and microRNAs precursors, play significant roles in many essential biological processes, ranging from gene transcriptional regulation, post-transcriptional regulation to protein synthesis. Sequence logo (1) is a promising tool for graphically representing sequence patterns from a multiple alignment of DNA, RNA and protein sequences. The logos comprise stacks of letters, one stack for each position in the sequence. The overall height of the stack at each position reflects the information contents determined according to Shannon information (2), which is given by $R_{seq} = \log_2 N - (-\sum_{n=1}^{N} p_n \log_2 p_n)$. Here, p_n indicates the observed frequency of the symbol n at a specific sequence position, and N represents the number of distinct symbols for the given sequence type, either four for DNA and RNA or 20 for protein (3). R_{seq} denotes the difference between the maximum possible entropy and that of the observed symbol distribution. WebLogo (3), which is an extended development of Sequence logo, provides a convenient and highly configurable program for generating sequence logos.

The structure logo program (4) includes prior frequencies on the bases, accommodates gaps in the alignments, and points out the mutual information of base-paired regions in RNA according to Sequence logo representation. CorreLogo combines information contents of nucleotides and mutual information of two different positions into 3D representation to determine correlations between bases and potential RNA base pairs (5). enoLOGOS generates sequence logos, which include energy measurements, probabilities matrices and alignment matrices, to represent the mutual information of different positions of the consensus sequence (6).

RNALogo is a web-based tool to create a new graphical representation of the patterns in a multiple RNA sequence alignment with a consensus structure. RNALogo provides a convenient and highly configurable logo generator. This work creates RNALogo graph for each RNA family in Rfam, and collects these generated logos into an RNALogo gallery. Users can search or browse RNALogo graphs in this gallery to receive further perspectives of known noncoding RNA families.

METHODS

Generation of RNALogo graph

The visualizing representation of RNALogo incorporates the representation of the standard Sequence logos, as well

*To whom correspondence should be addressed. Tel: +886 3 5712121 56952; Fax: +886 3 5729288; Email: bryan@mail.nctu.edu.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

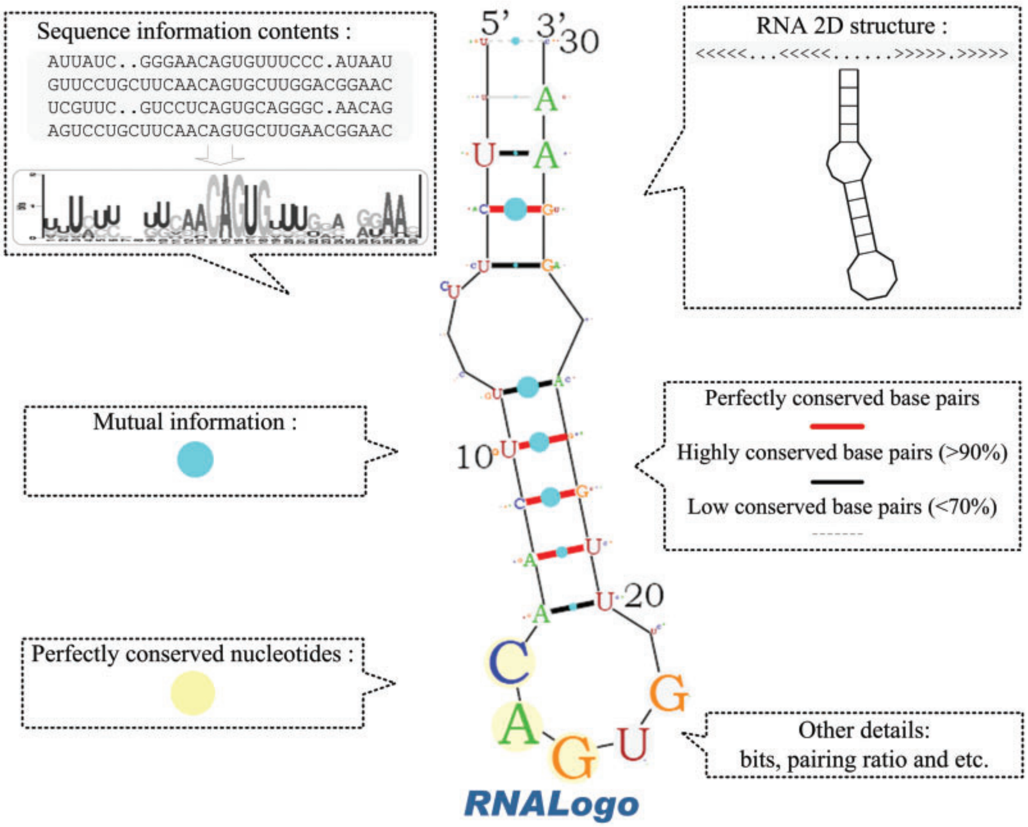
Sequence information contents :
AUUAUC..GGGAACAGUGUUUCCC.AUAAU
GUUCCUGCUUCAACAGUGCUUGGACGGAAC
UCGUUC..GUCCUCAGUGCAGGGC.AACAG
AGUCCUGCUUCAACAGUGCUUGAACGGAAC
RNA 2D structure :
<<<<<...<<<<<......>>>>>.>>>>>
5'
3'
30
10
20
Mutual information :
Perfectly conserved base pairs
Highly conserved base pairs (>90%)
Low conserved base pairs (<70%)
Perfectly conserved nucleotides :
Other details:
bits, pairing ratio and etc.
RNALogo

Figure 1. The RNALogo graph representation.

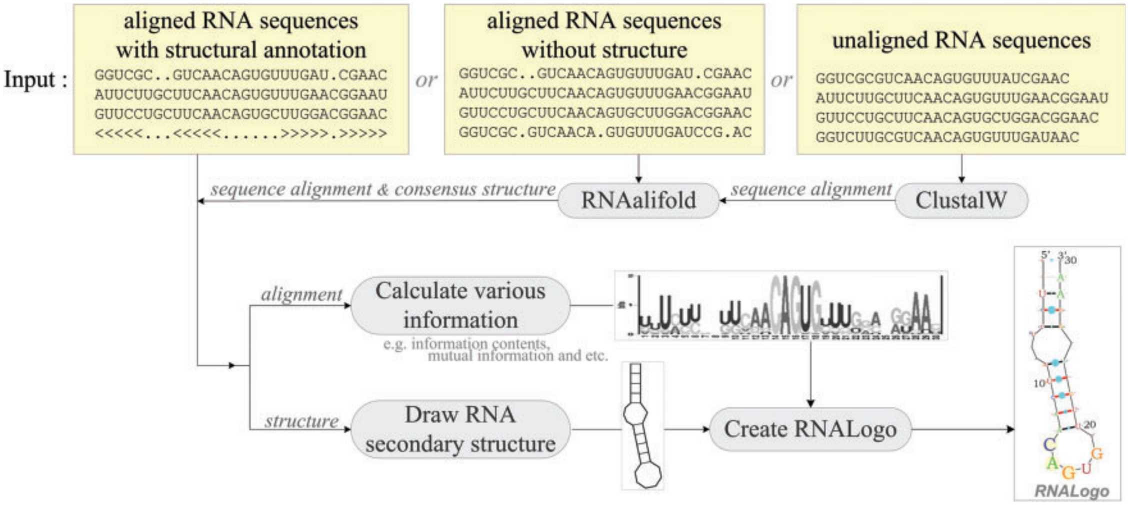
Input :
aligned RNA sequences
with structural annotation
GGUCGC..GUCAACAGUGUUUGAU.CGAAC
AUUCUUGCUUCAACAGUGUUUGAACGGAAU
GUUCCUGCUUCAACAGUGCUUGGACGGAAC
<<<<<...<<<<<......>>>>>.>>>>>
or
aligned RNA sequences
without structure
GGUCGC..GUCAACAGUGUUUGAU.CGAAC
AUUCUUGCUUCAACAGUGUUUGAACGGAAU
GUUCCUGCUUCAACAGUGCUUGGACGGAAC
GGUCGC.GUCAACA.GUGUUUGAUCCG.AC
or
unaligned RNA sequences
GGUCGCGUCAACAGUGUUUAUCGAAC
AUUCUUGCUUCAACAGUGUUUGAACGGAAU
GUUCCUGCUUCAACAGUGCUGGACGGAAC
GGUCUUGCGUCAACAGUGUUUGAUAAC
sequence alignment & consensus structure
RNAalifold
sequence alignment
ClustalW
alignment
Calculate various information
e.g. information contents, mutual information and etc.
structure
Draw RNA secondary structure
Create RNALogo
RNALogo

Figure 2. The flowchart for RNALogo graph generation.

as the RNA secondary structures and the mutual information of base-paired regions. Each RNALogo graph is composed of stacks of letters, one stack for each position in the consensus RNA secondary structure of an RNA alignment. The RNALogo graph can indicate significant features within an RNA sequence alignment and its consensus RNA secondary structure. As demonstrated in Figure 1, users can intuitively observe various significant features, such as the sequence conservation, the mutual information of base-paired regions, the consensus RNA secondary structure and the structural conservation of a RNA family, by the representation of RNALogo graph.

Figure 2 illustrates the flowchart for generating an RNALogo graph. To facilitate the generation of an RNALogo graph, RNALogo allows three types of input: (i) aligned RNA sequences with a consensus structure; (ii) aligned RNA sequences without any consensus structures, and (iii) unaligned RNA sequences. First, to input aligned RNA sequences with a consensus structure, the RNA sequence alignment with a consensus structure can be directly processed to generate an RNALogo graph. Secondly, to input aligned RNA sequences without any consensus structures, the proposed tool incorporates RNAalifold, which is a program in Vienna RNA Package (7) for predicting a consensus secondary structure of a set of aligned sequences, to generate a consensus RNA secondary structure from the input alignment, and then generate a RNALogo graph from the aligned RNA sequences and the corresponding consensus structure. Finally, for unaligned RNA sequences, the web server employs ClustalW (8) to align the input sequences, then utilizes these sequences to predict its consensus structure and generate its RNALogo graph.

The shape of the consensus structure of the input RNA sequences is sketched using RNAplot, which is an RNA secondary structure drawing program in Vienna RNA Package (7). Other information, such as information contents and the mutual information, are drawn on the RNA secondary structure based on their corresponding positions. The RNALogo includes the structure logo program (4) to obtain the information content for each position and the mutual information of base-paired regions of a structural RNA alignment. The mutual information is a convenient measurement of the dependence between two different positions of RNA secondary structure (9). Additionally, RNALogo supports the option of annotating pseudo-knots on the RNALogo graph simply by specifying the location of the interaction regions of the consensus structure.

The default colors for nucleotide symbols in an RNALogo graph are A, U, C and G are green, red, blue and orange, respectively, as shown in Figure 1. Users can select a variety of coloring schemes. The red, black and dotted lines between base pairs in helical regions denote the perfectly conserved, highly conserved and low-conserved base pairs, respectively, and the thickness of lines between base pairs denotes the base-paired ratio between the two corresponding positions. The stack of nucleotides marked by a yellow filled circle indicates that the nucleotide is perfectly conserved in all RNA sequences, as in the example displayed in Figure 1. The mutual information of base-paired regions is represented by the blue filled circle, and a larger circle represents stronger mutual information of the base pair.

The RNALogo graph can be generated in a variety of common image formats, namely JPG, PNG, TIF and PDF, all of which are appropriate for display, printing and editing. Another format, Scalable Vector Graphics (SVG), which is a language for describing two-dimensional graphical application in Extensible Markup Language (XML), is also supported for online editing of RNALogo graphs. Furthermore, RNALogo provides various flexible options for customizing logos, including image size, image title, line width, the font size of nucleotide symbols and the font type of nucleotide symbols. That is the RNALogo graph, the standard sequence logos of the input RNA sequence alignment are also provided. These options are described in detail in the documentation at the RNALogo web server.

Gallery of RNALogo

Rfam is a large collection of multiple sequence alignments and covariance models covering many common noncoding RNA families (10). To improve the representation of these noncoding RNA families, an RNALogo graph was created for each RNA family in Rfam, and these logos were accumulated into an RNALogo gallery. Around 600 noncoding RNA families were obtained from the Rfam database, and categorized into several classes according to their functions, as listed in Figure S1 (Supplementary Data). Tables S1 and S2 present the data statistics of the RNA families (Supplementary Data). Users can search or browse the gallery using Rfam accessions to retrieve the RNALogo graphs of known noncoding RNA families for further investigation.

CASE STUDIES

Several examples of regulatory RNAs were chosen from the RNALogo gallery to demonstrate the capabilities of the RNALogo graph representation.

Case I: IRE

The IRE is a short conserved stem–loop structure in untranslated regions (UTRs) of various mRNA, and can negatively regulate the translation of gene whose products are involved in iron metabolism by binding with iron response protein (IRP). Previous studies suggest the base-paired regions have no sequence-specific requirement (11), and that the mutation in the loop regions reduces binding affinity of IRE for IRPs (12). The RNALogo graph (Figure 3a) of the Rfam IRE family (Rfam accession RF00037) clearly demonstrates that the bases in the loop region are highly conserved, whereas those in the helical region are poorly conserved.

Case II: purine riboswitch

Riboswitches, which are found in the 5′-UTR of mRNAs, act as *cis*-acting genetic regulatory elements composed of a metabolite-responsive aptamer domain in a specific RNA secondary structure (13). The RNALogo graph (Figure 3b) of Rfam purine riboswitch family (Rfam accession RF00167) indicates that the sequences in multibranch loop region and hairpin loop regions are more conserved than those in other regions. This observation is supported by previous studies, which have indicated that the multibranch loop region is the aptamer domain, which is the high-affinity ligand binding regions, and is sensitive to single-point mutations (14). Moreover, the two terminal loops are predicted to form a pseudo-knot, which is an important structure for purine riboswitch (15). The RNALogo graph of the purine riboswitch exhibits this phenomenon via pseudo-knot annotation, which is denoted by the purple lines at positions 41–42 and 66–67.

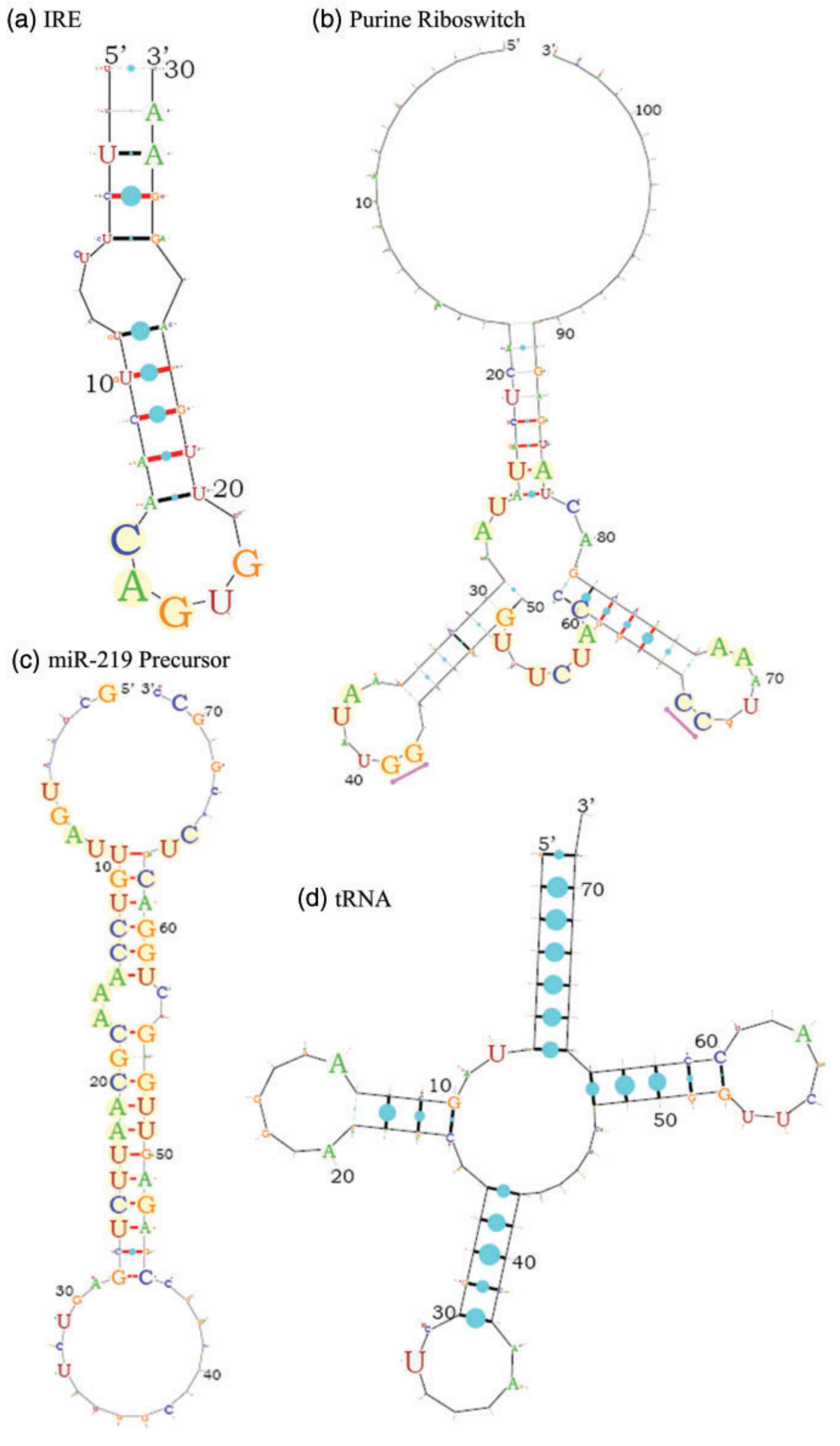

Figure 3. Examples of RNALogo graphs. (**a**) RNALogo graph of IRE family. (**b**) RNALogo graph of purine riboswitch family. The purple lines represent the pseudo-knot interaction regions. (**c**) RNALogo graph of microRNA miR-219 precursor family. (**d**) RNALogo graph of tRNA family.

Case III: microRNA miR-219 precursor

MicroRNAs participate in gene post-transcriptional regulation by suppressing the translation of coding genes (16). In general, mature miRNAs are often identified at both the 5′ and 3′ arms of microRNA precursors (17). The RNALogo graph (Figure 3c) of microRNA mir-219 precursor (Rfam accession RF00251) indicates that the bases in the region of mature miRNA are highly conserved, as would be expected.

Table 1. Comparison of RNALogo with other previously developed tools

Comparing features	RNALogo	Sequence logo (1)	Structure logos (4)	enoLOGOS (6)	CorreLogo (5)
Display information on RNA secondary structure	Yes	—	—	—	—
Display mutual information of base pairs	Yes	—	Yes	Yes	Yes
Support frequency plot drawing	Yes	Yes	—	—	—
Support pseudoknot annotation	Yes	—	—	—	—
Support various input formats	Yes	—	—	Yes	—
Support the input of unaligned sequences	Yes	—	—	—	—
Visualization of alternative interaction	—	—	—	Yes	Yes
Visualization/cutoff by standard deviation	—	—	—	—	Yes
Small sample error correction taken into account	—	—	—	—	Yes

Case IV: transfer RNA

The transfer RNA (tRNA) family is a good example for illustrating the compensatory mutation to retain the RNA secondary structure. The RNALogo graph (Figure 3d) of the Rfam tRNA family (Rfam accession RF00005) indicates low conservation in the tRNA sequences, but high conservation in the tRNA structures. As demonstrated in Figure S2, RNALogo, enoLOGOS and CorreLogo represent the mutual information of base-paired regions in different ways. The RNALogo graph directly displays the mutual information on RNA secondary structure, making it more intuitive and simpler than either enoLOGOS or CorreLogo.

In general, the nucleotides of the functional sites of regulatory RNA molecules are more conserved than other nonfunctional regions. For example, the bases of loop regions within the structure of IRE (11) and purine riboswitch (14) are more conserved than those of helical regions, whereas the bases of helical regions within the structure of microRNA precursors are more conserved than the bases of loop regions (16). These cases indicate that IRE, purine riboswitch, microRNA miR-219 precursor and tRNA have different significant features, which can be intuitively observed in the RNALogo graph.

AVAILABILITY

The web server of RNAlogo will be continuously maintained and updated. The web server is now freely available at: http://rnalogo.mbc.nctu.edu.tw/.

DISCUSSION

The representation of RNALogo graph and the RNALogo program as compared with other previously developed tools, namely Sequence logo (1), Structure logo (4), enoLOGOs (6) and CorreLogo (5), are shown in Table 1. RNALogo graph is the only representation that can intuitively display the information content of each base and the mutual information of base pairs within the RNA secondary structures. In particular, RNALogo supports the pseudo-knot annotation, various input formats and the input of unaligned sequences for the generation of RNALogo graph.

SUMMARY

This work presents a novel graphical representation of the patterns in an aligned RNA sequence with a consensus structure. Several significant features, including sequence conservation, structural conservation and the mutual information of base-paired regions, can be revealed within an RNA sequence alignment and its consensus RNA secondary structure. A web-based tool, RNALogo, is also developed to provide a rapid, effective and highly configurable logo generator. A gallery of RNALogo for Rfam RNA families is also built. Users can browse RNALogo graphs in this gallery to obtain further perspectives of known RNA families. Finally, several examples of regulatory RNAs, namely IRE, Riboswitches, miR-219 precursors and tRNA, are employed to reveal demonstrate the capabilities of the RNALogo graph representation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 96-3112-E-009-002 and NSC 95-2311-B-009-004-MY3. Special thanks for financial support from the National Research Program for Genomic Medicine (NRPGM), Taiwan. This work was also partially supported by MOE ATU. Ted Knoy is appreciated for his editorial assistance. Funding to pay the Open Access publication charges for this article was provided by National Science Council of the Republic of China and MOE ATU.

Conflict of interest statement. None declared.

REFERENCES

1. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
2. Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. John Wiley & Sons, New York.

3. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
4. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.*, **13**, 583–586.
5. Bindewald,E., Schneider,T.D. and Shapiro,B.A. (2006) CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucleic Acids Res.*, **34**, W405–W411.
6. Workman,C.T., Yin,Y., Corcoran,D.L., Ideker,T., Stormo,G.D. and Benos,P.V. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.
7. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures (The Vienna RNA Package). *Monatshefte Chemie*, **125**, 167–188.
8. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
9. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.
10. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
11. Bettany,A.J., Eisenstein,R.S. and Munro,H.N. (1992) Mutagenesis of the iron-regulatory element further defines a role for RNA secondary structure in the regulation of ferritin and transferrin receptor expression. *J. Biol. Chem.*, **267**, 16531–16537.
12. Jaffrey,S.R., Haile,D.J., Klausner,R.D. and Harford,J.B. (1993) The interaction between the iron-responsive element binding protein and its cognate RNA is highly dependent upon both RNA sequence and structure. *Nucleic Acids Res.*, **21**, 4627–4631.
13. Coppins,R.L., Hall,K.B. and Groisman,E.A. (2007) The intricate world of riboswitches. *Curr. Opin. Microbiol.*, **10**, 176–181.
14. Gilbert,S.D., Love,C.E., Edwards,A.L. and Batey,R.T. (2007) Mutational analysis of the purine riboswitch aptamer domain. *Biochemistry*, **46**, 13297–13309.
15. Gilbert,S.D., Stoddard,C.D., Wise,S.J. and Batey,R.T. (2006) Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *J. Mol. Biol.*, **359**, 754–768.
16. Kloosterman,W.P. and Plasterk,R.H. (2006) The diverse functions of microRNAs in animal development and disease. *Dev. Cell*, **11**, 441–450.
17. Griffiths-Jones,S. (2006) miRBase: the microRNA sequence database. *Methods Mol. Biol.*, **342**, 129–138.