Published online 19 March 2012

Nucleic Acids Research, 2012, Vol. 40, No. 12 ***e95***
doi:10.1093/nar/gks241

R-CHIE: a web server and R package for visualizing RNA secondary structures

Daniel Lai[1,2], Jeff R. Proctor[1,2], Jing Yun A. Zhu[1,2] and Irmtraud M. Meyer[1,2,*]

[1]Department of Computer Science and [2]Department of Medical Genetics, Centre for High-Throughput Biology, University of British Columbia, Vancouver V6T 1Z4, Canada

Received December 15, 2011; Revised February 15, 2012; Accepted March 1, 2012

ABSTRACT

Visually examining RNA structures can greatly aid in understanding their potential functional roles and in evaluating the performance of structure prediction algorithms. As many functional roles of RNA structures can already be studied given the secondary structure of the RNA, various methods have been devised for visualizing RNA secondary structures. Most of these methods depict a given RNA secondary structure as a planar graph consisting of base-paired stems interconnected by roundish loops. In this article, we present an alternative method of depicting RNA secondary structure as arc diagrams. This is well suited for structures that are difficult or impossible to represent as planar stem-loop diagrams. Arc diagrams can intuitively display pseudo-knotted structures, as well as transient and alternative structural features. In addition, they facilitate the comparison of known and predicted RNA secondary structures. An added benefit is that structure information can be displayed in conjunction with a corresponding multiple sequence alignments, thereby highlighting structure and primary sequence conservation and variation. We have implemented the visualization algorithm as a web server R-CHIE as well as a corresponding R package called R4RNA, which allows users to run the software locally and across a range of common operating systems.

INTRODUCTION

The tertiary or three-dimensional structure of many RNAs plays a key role in defining the biological function of the molecule. Both *in vivo* and *in silico*, the proper formation of tertiary structure depends on the correct secondary structure, i.e. the pairs of nucleotide positions in the RNA sequence that form complementary base pairs. For both experimentally and computationally derived RNA secondary structures, diagrammatic visualization has been key to evaluating and gleaning biological insight, thereby prompting the development of new computational methods for RNA secondary structure visualization.

A handful of computationally generated representations of RNA secondary structures exist, the most prominent being the planar format which consists of base pairs stems and unpaired loops (1). One is the circular version where the sequence is drawn as a circle and base pairs as chords (2). Another one is the linear version of the circle that shows the sequence as horizontal line and the base pairs as arcs, i.e. semicircles connecting the two respective base-pairing positions in the sequence. Other formats include energy dot plots (3) which are coloured, square base-pairing matrices (4), the dot-bracket or Vienna format (5) and the closely related mountain plot figures (6). Finally, trees (7) and graphs (8) in the strict mathematical sense have also been used to display the topology of RNA secondary structures.

For any experimentally determined or theoretically predicted RNA structure, a critical method of evaluation is to analyze the degree of structure conservation between homologous sequences (9). Strong evidence for RNA structure conservation are so-called pairs of compensatory mutations that retain the base-pairing ability, but change the base-pairing nucleotides (co-variation). A quick method of visually surveying the quality of a given multiple sequence alignment and a corresponding secondary structure prediction is to highlight co-varying pairs of alignment columns (10).

During the development and evaluation of our comparative helix prediction method TRANSAT (11), we had to develop a new method for RNA secondary structure visualization that was able to show (i) conflicting base pairs (i.e. base pairs involving the same sequence position), (ii) the degree of helix conservation and corresponding co-variation within the multiple sequence alignment and (iii) allows the comparison of two different structures simultaneously (e.g. from different sources). We also wanted

*To whom correspondence should be addressed. Tel: +21 604 827 4232; Fax: +21 604 822 5485; Email: irmtraud.meyer@cantab.net

the figures to be (iv) visually pleasing and intuitive to grasp. Our requirement to visualize conflicting base pairs rules out all formats except circle, linear and dot plots. Finally, the additional need to simultaneously show several structures in conjunction with a corresponding multiple sequence alignment, led us to choose a linear format.

While various powerful visualization programs already exist, only a few actively supported tools visualize RNA secondary structure in a linear fashion (12,13). These tools, however, lacked the features that we required, as they were not designed to handle conflicting base pairs nor display multiple sequence alignments simultaneously with a structure. Finally, the need to create such diagrams in a high-throughput and scripted manner, rather than restricted by a graphical user interface, made their adoption and modification difficult (12,13).

In the following, we present a highly modified and new method employing a linear format that we call 'arc diagrams' (14) to fulfill our above requirements (i–iv). In addition, we provide a web server R-CHIE that accepts four common secondary structure formats and secondary structure for the quick visualization of data with our method to generate publication quality figures. For further customization and local use, we also make a corresponding R package (15) called R4RNA available that leverages the graphical and computational framework of the interactive and easily scriptable language.

MATERIALS AND METHODS

The R-CHIE web server

Located at http://www.e-rna.org/r-chie/, the R-CHIE web server provides a simple interface for generating six different types of arc diagrams instantaneously with instructions and examples, accessible by all major browsers. Descriptions and usage of the six different types of diagrams are as follows:

Single structure

This is the most basic type of arc diagram, essentially identical to the typical linear diagrams observed in other publications, with the exception of much more powerful graphical options. This arc diagram shows the RNA sequence of interest drawn as a horizontal line from 5′ to 3′, left to right, with 'arcs', drawn above the horizontal line. Each arc depicts a base pair of the RNA structure and connects the respective sequence positions involved in that base pair.

For predicted structures, it is not uncommon for individual structural features such as helices or base pairs to be assigned individual scores such as energetic contributions (16) or statistical significance (11). In order to retain and visualize this valuable information, our method can assign different colours to individual arcs according to their corresponding scores, using palettes obtained from ColorBrewer (17), or those specified by the user. Alternatively, colouring arcs can also be done manually and independently of value, e.g. when certain base pairs or structure features such as pseudoknots are to be especially highlighted. In addition, base pairs can also be filtered by their scores and a lower or upper threshold value can be imposed.

Double structure

A double structure arc diagram is obtained by starting with a single structure diagram, and drawing a second structure below the horizontal sequence line. Any colouring and filtering options can be applied to the top and the bottom structure jointly or separately (Figure 1).

This type of arc diagram is especially useful when comparing two—perhaps radically different—alternative structures for the same sequence. It is also useful for comparing two similar structures, e.g. derived from two different structure prediction methods or comparing a predicted to an experimentally validated structure.

Overlapping structure

The type of figures seen in the TRANSAT paper (11) allow to quickly and entirely visually evaluate the performance of a structure prediction method by comparing the predicted to the known structure. In order to do this best, we aim to simultaneously visualize the sensitivity and the positive predictive value, i.e. specificity, of the prediction method (Figure 2).

Similar to the double structure plot, arcs are seen both above and below the horizontal sequence line. Instead of showing all arcs corresponding to one structure above the horizontal line and all arcs corresponding to the other structure below, however, the first is interpreted as a 'predicted' structure, the second as reference i.e. 'known' structure, and the two structures are overlapped. For this, the algorithm identifies all predicted base pairs that overlap with those of the known structure (i.e. a 'true positive' in the performance evaluation), and draws corresponding arcs above the line, coloured by the score of the base pair in the predicted structure. Any predicted base pair that is not part of the known structure (i.e. a 'false positive') is drawn below the sequence line, also coloured by its score. Any known base pair that was not predicted (i.e. a 'false negative') is drawn above the sequence line in black. Base pairs that are neither part of the known nor the predicted structure (i.e. 'false negatives') are not shown at all.

With a single glance, this type of diagram shows both the sensitivity and specificity of the structure prediction and readily highlights new base pairs that are not part of the known structure and may warrant further investigation.

Creating two overlapping structures from predictions of two different algorithms against the same known structure and juxtaposing the resulting diagrams is also an interesting method of comparing and highlighting the differences, as was done extensively in the TRANSAT paper (11).

Single structure co-variation

Adding a multiple sequence alignment beneath a single structure arc diagram provides a powerful means of displaying both the secondary structure and corresponding evidence for base pair conservation and co-variation. As the state of art in RNA secondary structure prediction in terms of prediction performance is currently provided

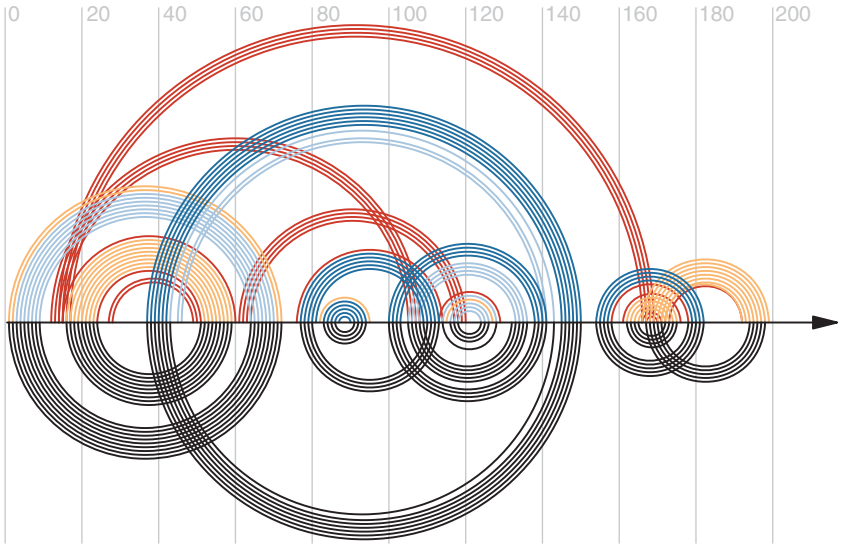

Figure 1. An example of a 'double structure arc diagram', showing the Cripavirus Internal Ribosomal Entry Site [family RF00458 from the RFAM database (20)]. The RNA secondary structure shown above the horizontal sequence line has been predicted by TRANSAT (11). Every arc corresponds to one base pair whose colour indicates its *P*-value, where dark blue is ≤1e-06, light blue is ≤1e-05, orange is ≤1e-04 and red is ≤1e-03 (*P*-value threshold). The RNA structure shown below the horizontal sequence line shows the consensus RNA structure from RFAM.

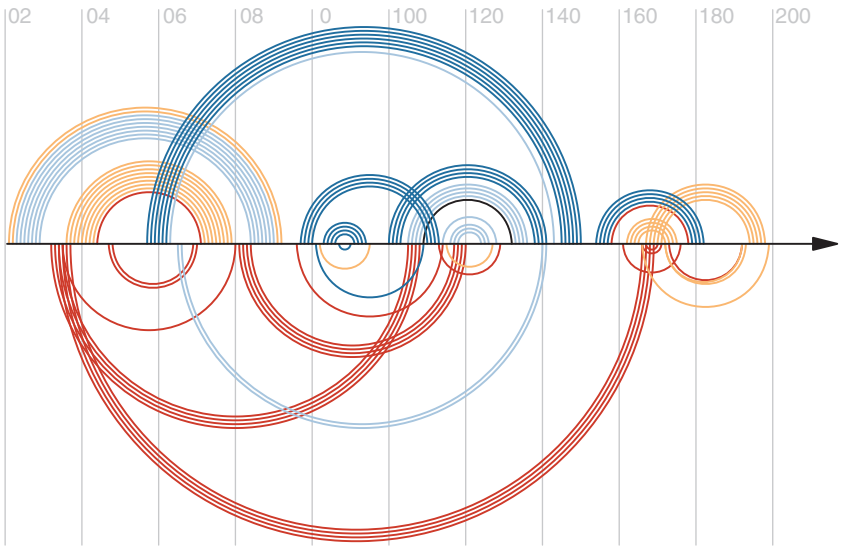

Figure 2. An example of a 'overlapping structure arc diagram', overlapping the TRANSAT predicted structure and the RFAM consensus structure of family RF00458 presented in Figure 1. The structure shown above the horizontal sequence is the known structure in black, coloured by *P*-value if correctly predicted by TRANSAT (best in blue and worst in red, see Figure 1 for exact *P*-value colours). The arcs below the line represent novel base pairs predicted by TRANSAT not found in the known RFAM structure. Such a diagram can give} a qualitative description of a predicted structure's performance, where high *sensitivity* would result in a high proportion of top helices being coloured, and high specificity would result in a majority of helices above the line. On the other hand, the novel base pairs observed on the bottom half, may indicate alternative structural elements not yet experimentally verified, but worth investigating, especially in light of strong evolutionary evidence (Figure 3).

by comparative methods that typically take a fixed alignment as input, this type of arc diagram is especially useful for evaluating both structures given an alignment and *vice versa*.

For this type of arc diagram, the arcs are drawn on top of the sequence line as usual while the multiple sequence alignment is shown below as a block of parallel black lines, each representing one sequence of nucleotides (with gray for gaps) from the multiple sequence alignment. Two alignment columns at the base of a single arc represent the two columns of base-pairing nucleotides. The corresponding nucleotides are shown in green or blue if they represent a valid canonical base pair, and red if they are not. The most frequent base pairs are coloured in green. Blue indicates a compensatory mutation relative to the green pairs (dark blue for a double-sided mutation, light blue for a single-sided mutation). This style of colouring is similar to existing programs (10,18,19). In the case where different types of base pairs occurs at the same frequency, green is assigned to the base pair more commonly observed according to Structure Statistics from the Comparative RNA Web Site (21).

Given a structure and a corresponding multiple sequence alignment, the web server automatically applies this colouring, allowing for a quick evaluation of how well (or poorly) the different structural features are supported by co-variation and gap patterns.

Given two different ways of aligning the same set of sequences, two arc diagrams of this type can also be used to highlight the effect that the alignment quality has on the corresponding structure prediction.

One small caveat is the technical inability of the co-variation colouring to be displayed simultaneously for conflicting helices. When faced with conflicting helices, our algorithm makes a greedy decision to select and colour the first helix that it observes in the input. A user can therefore simply rank conflicting helices or base pairs in the input file to ensure that the most dominant features are being coloured.

Additional web server appearance options that users may adjust include displaying sequences as blocks instead of lines, including the nucleotide base on the block, and including the sequence descriptions left of each sequence. The specific colours used to highlight base pairs in the multiple sequence alignment can also be customized, and one may even completely ignore base pairs and colour the alignment based on nucleotides.

While the structure conservation for one base pair can be summarized in a single numerical value as done for some figures in the RFAM database (20), a coloured multiple sequence alignment as in these arc diagrams retains more detailed information. If desired however, we offer the ability to colour arcs according to structure conservation, co-variation and percent canonical base pair.

Double structure co-variation

Two multiple sequence alignment blocks, one for each structure, are inserted between the top and bottom arcs of a double arc diagram, highlighting the conservation and co-variation for each structure.

An embellishment of the double structure arc diagram, the double structure co-variation diagrams show not only the differences between two structures, but also allow the evaluation of the different base pairs in light of evolutionary evidence. The input multiple sequence alignment is duplicated into two identical blocks. The top alignment and its co-variation annotation refer to the top RNA structure and the bottom alignment to the bottom RNA structure.

Overlapping structure co-variation

A natural extension of the arc diagrams mentioned so far is to combine overlapping structure arc diagrams with co-variation plots. These are similar to double structure co-variation plots, but are drawn according to the same rules as overlapping structure arc plots (Figure 3).

This type of diagram is of great use when evaluating new helices (drawn below the sequence line in overlapping structure arc plots) by providing evolutionary evidence (or lack of) for their existence. The same rules to resolving conflicting helices exist as outlined above for single structure co-variation plots, as do the rules for ensuring consist co-variation plot colours.

Input

There are two main types of text inputs to the web server; those specifying secondary structures and those specifying multiple sequence alignments.

For RNA secondary structure and to cater for the wider RNA community, our method accepts most common output formats: dot bracket or Vienna format (22), MFOLD's connect format (23) and Gutell's bpseq format (21). Additionally, we also accept a variant of Shapiro's original region table (1) that we refer to as the 'helix format'. This describes each helix as one line which contains the following fields: start position of outer base pair, end position of outer base pair, helix length (in terms of number of base pairs), helix score. Differing from Shapiro's definition, we add a header line which includes the length of the sequence, along with any other comments one may want to retain, e.g. the primary sequence. This helix format provides an extremely compact means of representing complex RNA structures, and also allows for unambiguous specification of conflicting, pseudo-knotted and overlapping base pairs.

For multiple sequence alignment, our program accepts standard FASTA format (24).

To control the appearance of the resulting figure showing the desired type of arc diagram, a standard options panel is available to automate and fine-tune colouring and filtering of base pairs on the fly.

Output

After generating the figure, which typically takes on the order of seconds, a static web page rendering the figure is displayed from which the figure can be downloaded as .png or .pdf format.

The R4RNA R package

The plotting functionality of the web server is driven by an R script built on top of an R package called R4RNA which we make available for offline and local use and which can be downloaded from http://www.e-rna.org/r-chie/ release for public use under the GPLv3 license.

Written in R (15) (which is freely downloadable at http://www.r-project.org/ for all major operating systems), the package is capable of producing the same plots as the server with a few interactive function calls, and allows for even more fine-tune control, automation, and customized diagrams, and output formats. This is especially convenient for Bioinformatics research groups that can call functions of our R package within existing programs and analysis pipelines. Our software is well documented and includes a comprehensive manual and instructional vignettes as well as examples.

RESULTS AND DISCUSSION

We here present a new computational method for visualizing RNA secondary structures in conjunction with corresponding multiple sequence alignments which

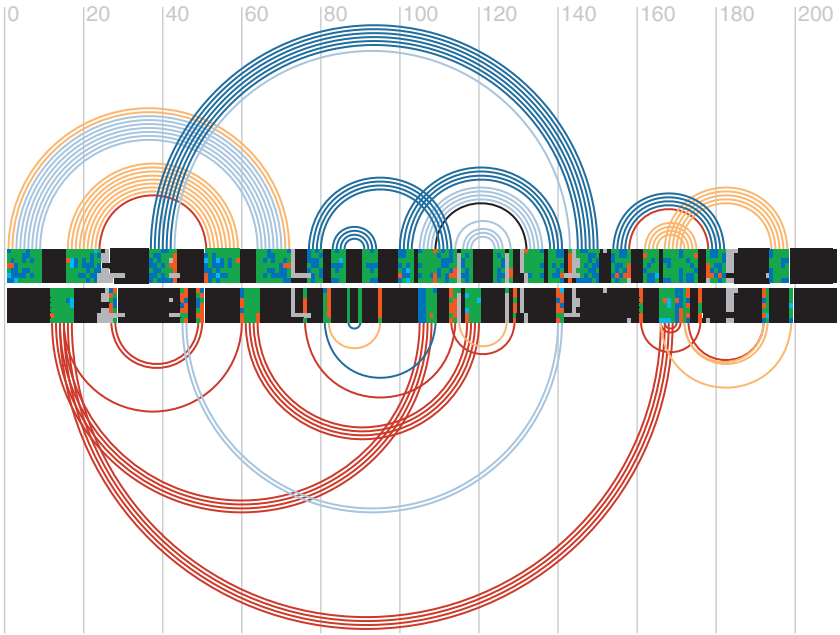

Figure 3. An example of an 'overlapping covariance arc diagram', of the TRANSAT predicted structure, overlapping the RFAM consensus structure of family RF00458 (see Figure 2 for a detailed explanation of the known top and novel bottom arcs). The colouring of arcs indicates their *P*-value (detailed in Figure 1) from best (blue) to worst (red) and gray for conflicting base pairs. Between the arcs are two covariance blocks representing the same seven sequences in a gapped multiple sequence alignment from RFAM. Unpaired nucleotides are in black and gaps are in gray. For columns at the ends of a single non-conflicting arc, bases are assigned green if they are valid base pairs (G:C, A:U, G:U and the reversed pairings), or else they are red. For the green valid base pairs, if the base pair varies from the most commonly observed base pair in the column, then it is coloured blue to signify *co-variation*, or compensatory mutations to retain the base-pairing potential of the positions (dark blue for two sided, and light blue for one sided). The colouring of base pairs for arcs gives a qualitative representation of how well the base pair is conserved, and can be used to infer the validity of predicted base pairs, or the quality of an alignment given a known structure. For example, while most of the novel base pairs predicted to exist are simply extensions or slight shifts of known helices, the three largest red helices in the bottom show very high sequence conservation.

can either be used via a web server R-CHIE or offline and locally via a corresponding R package called R4RNA. Our method readily creates six different types of arc diagrams which cover numerous useful applications. These range from visualizing the evolutionary evidence for a given RNA secondary structure to comparisons of two RNA secondary structure and performance evaluations of RNA structure prediction methods. The key feature of all six types of arc diagrams is that details that are typically lost in a numerical evaluation are highlighted and can be visually interpreted in a straightforward and intuitive way.

Our method makes several major improvements with respect to existing methods that depict RNA secondary structures in a linear way. These include the colouring of structural features according to their score (e.g. free energy, *P*-value, log-likelihood), the joint display of an RNA structure with a corresponding multiple sequence alignment which highlights the evolutionary patterns that support the different structural features, and comparison plots which allow the quick visual inspection of sensitivity and specificity of a predicted structure with respect to a reference structure. In addition, all types of arc plots can display structural features that are mutually exclusive or would render the overall RNA structure pseudo knotted.

Our R-CHIE web server and the corresponding R4RNA R package can be freely accessed and downloaded from http://www.e-rna.org/r-chie/. In addition to several examples, we have also generated single structure co-variation arc diagrams of all seed alignments in the RFAM database (20) which can be downloaded from our web page.

FUTURE DIRECTIONS

We intend to add additional features to the R4RNA R package in the future which will likely include functions related to RNA structural Bioinformatics analysis beyond RNA structure visualization. We also welcome suggestions for improvements on the R-CHIE web server or the R4RNA R package from the research community. Finally, to ensure long-term support and development of the package for RNA analysis, we aim to officially submit the R4RNA package to the BIOCONDUCTOR repository (25) once the package is sufficiently developed.

ACKNOWLEDGEMENTS

The authors would like to thank our former group members Nick Wiebe and Casper Shyr for designing a prototype of the visualization software in Perl and Asymptote which motivated the complete re-implementation in R and as a web server. We also thank Adi Steif from our group for testing the web server; and Dave Brent and Renee Stephen from the

Department of Computer Science at the University of British Columbia for providing technical assistance with the web server.

FUNDING

Canadian Institutes of Health Research/Michael Smith Foundation for Health Research Strategic Training Program in Bioinformatics at the University of British Columbia; Natural Sciences and Engineering Research Council of Canada (NSERC) Alexander Graham Bell Canada Graduate Scholarships (to D.L. and J.R.P.); NSERC Undergraduate Student Research Awards Collaborative Research Experiences for Undergraduates - Canada (to J.Y.A.Z.); NSERC Discovery Grant (to I.M.M.). Funding for open access charge: NSERC Discovery Grant (to I.M.M.).

Conflict of interest statement. None declared.

REFERENCES

1. Shapiro,B.A., Lipkin,L.E. and Maizel,J. (1982) An interactive technique for the display of nucleic acid secondary structure. *Nucleic Acids Res*., **10**, 7041–7052.
2. Nussinov,R., Pieczenik,G., Griggs,J.R. and Kleitman,D.J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math*., **35**, 68.
3. Jacobson,A.B. and Zuker,M. (1993) Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol*., **233**, 261–269.
4. Tinoco,I., Uhlenbeck,O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.
5. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem*., **125**, 167–188.
6. Hofacker,I.L. and Stadler,P.F. (1999) Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput. Chem*., **23**, 401–414.
7. Le,S.Y., Nussinov,R. and Maizel,J.V. (1989) Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res*., **22**, 461–473.
8. Gan,H.H. (2003) Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res*., **31**, 2926–2943.
9. Pace,N., Thomas,B. and Woese,C. (1999) Probing RNA structure, function, and history by comparative analysis. In: Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World*, 2nd edn. Cold Spring Habor Press, Cold Spring Harbor, NY, pp. 113–142.
10. Griffiths-Jones,S. (2005) RALEE–RNA ALignment editor in Emacs. *Bioinformatics*, **21**, 257–259.
11. Wiebe,N.J.P. and Meyer,I.M. (2010) TRANSAT– method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. *PLoS Comput. Biol*., **6**, e1000823.
12. Wiese,K., Glen,E. and Vasudevan,A. (2005) jViz. Rna-A Java tool for RNA secondary structure visualization. *IEEE Trans. Nanobiosci*., **4**, 212–218.
13. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.
14. Wattenberg,M. (2002) Arc diagrams: visualizing structure in strings. *Proceedings of the IEEE Symposium on Information Visualization*, Boston, MA, pp. 110–116.
15. R Development Core Team (2011) *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
16. Zuker,M. and Jacobson,A.B. (1998) Using reliability information to annotate RNA secondary structures. *RNA*, **4**, 669–679.
17. Harrower,M. and Brewer,C.A. (2003) ColorBrewer.org: an online tool for selecting colour schemes for maps. *Cartogr. J*., **40**, 27–37.
18. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol*., **2**, e33.
19. Bendaña,Y.R. and Holmes,I.H. (2008) Colorstock, SScolor, Ratón: RNA alignment visualization tools. *Bioinformatics*, **24**, 579–580.
20. Griffiths-Jones,S. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.
21. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Müller,K.M. *et al*. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinf*., **3**, 2.
22. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.
23. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.
24. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *PNAS*, **85**, 2444–2448.
25. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.