BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 19 2012, pages 2523–2526
doi:10.1093/bioinformatics/bts461

Sequence analysis Advance Access publication July 23, 2012

RILogo: visualizing RNA–RNA interactions

Peter Menzel[1,*], Stefan E. Seemann[2,3] and Jan Gorodkin[2,3]

[1]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen DK-2200, [2]Center for non-coding RNA in Technology and Health and [3]IKVH, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: With the increasing amount of newly discovered non-coding RNAs, the interactions between RNA molecules become an increasingly important aspect for characterizing their functionality. Many computational tools have been developed to predict the formation of duplexes between two RNAs, either based on single sequences or alignments of homologous sequences. Here, we present *RILogo*, a program to visualize inter- and intramolecular base pairing between two RNA molecules. The input for *RILogo* is a pair of structure-annotated sequences or alignments. In the latter case, *RILogo* displays the alignments in the form of sequence logos, including the mutual information of base paired columns. We also introduce two novel mutual information based measures that weigh the covariance information by the evolutionary distances of the aligned sequences. We show that the new measures have an increased accuracy compared with previous mutual information measures.

Availability and implementation: *RILogo* is freely available as a stand-alone program and is accessible via a web server at http://rth.dk/resources/rilogo.

Contact: pmenzel@gmail.com

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 19, 2012; revised on June 16, 2012; accepted on July 18, 2012

1 INTRODUCTION

An increasing amount of functional non-coding RNAs (ncRNAs) has been identified in recent years and computational predictions of possibly functional structured RNAs yield vast amounts of candidate RNA genes (Gorodkin *et al*., 2010). Typically, RNA molecules interact with proteins or with other RNA molecules (including messenger RNAs) by forming intermolecular duplexes. Computational predictions of RNA–RNA interactions usually follow the thermodynamic energy models for the secondary structure prediction of a single RNA sequence. Algorithms for the simultaneous prediction of intra- and intermolecular structure between two given sequences have a higher prediction accuracy, but also have a high time complexity (Bernhart *et al*., 2006; Huang *et al*., 2009). Another class of algorithms only focuses on the prediction of intermolecular duplexes between two RNA sequences, while neglecting possible intramolecular base pairings, and achieves linear time complexity in the best case (Tafer and Hofacker, 2008; Wenzel et al, 2012). When several homologous sequences are available, consensus structure prediction methods can also be applied for predicting intermolecular base pairs. For example, `PETcofold` (Seemann *et al*., 2011a) first predicts intra-molecular pairing and afterwards intermolecular pairing in the previously unpaired regions. Another approach also based on multiple sequence alignments is `ripalign` (Li *et al*., 2011), which computes inter- and intramolecular base pairs simultaneously using a partition function algorithm.

The sequence pattern of a given multiple DNA or protein alignment is easily visualized by sequence logos (Schneider and Stephens, 1990). Each alignment column is represented by a stack of letters from the DNA/protein alphabets, where the overall height of the stack denotes the Shannon information content of that position and the size of a letter denotes its frequency in that column. The most frequent letter is placed on top of the stack. Several programs were developed to automatically create sequence logos from a given input alignment (Beitz, 2000; Crooks *et al*., 2004).

When studying RNA alignments, the secondary structure of the RNAs needs to be considered. Structure logos (Gorodkin *et al*., 1997) include the annotation of base pairing and the mutual information (MI) of two paired alignment positions in the logo itself. Single or consensus RNA structures are usually plotted in a circular planar graph (Auber *et al*., 2006), e.g. `VARNA` (Darty *et al*., 2009) or `R2R` (Weinberg and Breaker, 2011). `RNALogo` (Chang *et al*., 2008) combines a planar plot of the consensus secondary structure with stacks of letters for each base and displays the MI of base pairs. 3D logos are plotted by `Correlogo` (Bindewald *et al*., 2006). However, these programs do not include RNA–RNA interactions. RNA structures and interactions can be visualized by arcs or lines connecting pairing bases in their primary sequences, e.g. in Huang *et al*. (2009).

Here, we introduce *RILogo*, which combines sequence logos with an arc representation of the secondary structure of both intra- and intermolecular base pairs of two interacting RNA genes. Additionally, the MI of base pairs is calculated and included in the logo by colouring the arcs or adding an ‘M’ character on top of the letter stacks, with a height corresponding to the MI. It has been suggested that measures including phylogenetic information are the most powerful (Akmaev *et al*., 2000). However, the widely used measures ignore information from the phylogenetic relationships between the sequences (Lindgreen *et al*., 2006). Thus, we introduce two novel mutual information based measures, which use pairwise evolutionary distances between sequences in order to give different weights to the

*To whom correspondence should be addressed.

covariance information contained in individual sequences from the alignment.

An early version of *RILogo* was previously used to display structure predictions in the PETcofold web server (Seemann *et al.*, 2011b). Here, we make the enhanced version of *RILogo* available both as a stand-alone program and via a web server. It now includes our new evolutionary mutual information measures, customization options and is freely available under the GNU LGPL.

2 IMPLEMENTATION AND RESULTS

2.1 RNA interaction logos

The input to *RILogo* is a pair of multiple structural RNA alignments or single sequences, including a consensus secondary structure annotation, which is allowed to contain pseudoknots. From the input alignments, two sequence logos are calculated and plotted one above the other, with the second logo displayed in 3′- to 5′-orientation (Fig. 1). Intramolecular base pairs are displayed by arcs in each logo. The arc representation of an RNA secondary structure is formally a graph with bases represented as nodes and base pairings as edges, where the nodes are displayed alongside each other while the edges are drawn as connecting arcs. Pseudoknots can easily be spotted as crossing arcs. Each arc is coloured by a gradient that denotes the MI of the underlying base pair. Intermolecular base pairing is displayed by straight lines between the respective columns of the two sequence logos. Both logos are horizontally arranged in a way, so that the average distance between interacting bases is minimized. *RILogo*'s native output is an SVG file. The web server additionally provides other vector and bitmap output formats. Figure 1 shows an example for the interaction of the bacterial sRNA *OxyS* and its target site in the *fhlA* mRNA.

2.2 Evolutionary-based MI measure

RILogo implements two novel measures for covariance of base pairs. The standard measure, an adaptation of the Kullback–Leibler divergence to base pairs, is defined for two alignment columns i and j as

$$\mathrm{MI}_{ij} = O_{ij} \log_2 \frac{O_{ij}}{E_{ij}} + (1 - O_{ij}) \log_2 \frac{(1 - O_{ij})}{(1 - E_{ij})} \tag{1}$$

where O_{ij} is the frequency of observed base pairs and E_{ij} of expected base pairs (Gorodkin *et al.*, 1997). The frequencies of canonical base pairs at columns i and j are counted as

$$O_{ij} = \frac{1}{N} \sum_{s \in S} C_s \theta(s_i, s_j); \quad E_{ij} = \frac{1}{N^2} \sum_{s \in S} C_s \theta(s_i) \cdot \sum_{s \in S} C_s \theta(s_j) \tag{2}$$

where S is the set of all sequences, $N = |S|$ and s_i is the ith base in sequence s. The term C_s is the phylogenetic distance to the other sequences described below in equation (3). $\theta(s_i, s_j)$ is 1 if (s_i, s_j) is a canonical base pair and 0 otherwise. $\theta(s_i)$ is 1 if s_i is equal to any t_i that forms a canonical base pair (t_i, t_j) where $t \in S$ and 0 otherwise. We define canonical base pairs as G:C, C:G, G:U, U:G, A:U and U:A. Note that MI_{ij} strictly speaking is not a mutual information measure, but a distance between the quantities O_{ij} and E_{ij} measuring observed and expected covariance of base changes between columns i and j. Here, we do not strictly distinguish between 'mutual information' and 'distance'. Lindgreen *et al.* (2006) compared MI with other alternative measures and showed that $\mathrm{MI}^{W\diamond P}$ is the best measure that does not include the structural context of base pairs. $\mathrm{MI}^{W\diamond P}$ uses only canonical base pairs and includes gap penalties by replacing the second summand in equation (1) with the term $-N_{ij}^{\mathrm{G}} \cdot \beta$ (see Supplementary material). Other types of gap corrections have been applied as well (Hertz *et al.*, 1990). By adding this *ad hoc* gap correction, the MI_{ij} can be <0 when columns i and j contain many gaps. Thus, it is not a strict information measure in the classical sense, and we refer to it as an approximate mutual information measure or approximate Kullback–Leibler distance. However, the gap correction has empirically been shown to increase the accuracy of the MI_{ij} (Lindgreen *et al.*, 2006) and in practical applications, as also done in *RILogo*, a negative MI_{ij} will be set to zero. Without gap correction and in the special case with equiprobable background frequencies for all base pairs, MI_{ij} reduces to a classical Shannon style information content. A known drawback of MI and $\mathrm{MI}^{W\diamond P}$ is that structurally neutral substitutions of one base within a base pair (e.g. G:C to G:U) do not contribute to the mutual information. Additionally, a phylogenetic bias in the sequence alignment can distort the individual contributions from each sequence. For example, compensatory base pair changes between sequences of high evolutionary distance are likely to happen. In contrast, compensatory substitutions in closely related sequences are less likely and, thus, contribute more information.

Hence, we introduce two new measures, treeMI and $\mathrm{treeMI}^{W\diamond P}$, which weigh the base pair frequencies using averaged pairwise phylogenetic distances. While $C_s = 1$ for all s in MI and $\mathrm{MI}^{W\diamond P}$ in equation (2), we define C_s in treeMI and $\mathrm{treeMI}^{W\diamond P}$ as

$$C_s = 1 - d_{\mathrm{avg}}(s)/N_d \tag{3}$$

where $d_{\mathrm{avg}}(s)$ is the average pairwise distance from sequence s to all other sequences in the tree and N_d is defined as $\sum_s d_{\mathrm{avg}}(s)$. This gives a higher weight to sequences that are phylogenetically proximal to all other sequences and less weight to more distant sequences. The normalization by the number of sequences N in equation (2) results in treeMI>0 also for fully conserved columns, which would not be the case if $N = \sum_s C_s$. Both $\mathrm{MI}^{W\diamond P}$ and $\mathrm{treeMI}^{W\diamond P}$ give values in the interval $\{-1.0, 2.0\}$, with -1 when each base pair for a given column contains one gap. The maximum of 2.0 bit is achieved when exactly the six possible base pairs are present. The MI is displayed by an additional letter M (with height $\mathrm{MI}_{ij}/2$) in the sequence logos, as in Gorodkin *et al.* (1997). The total height of a letter stack (y-axis) can exceed 2 bit, since it sums up two measures, Shannon Information and (approximate) mutual information. An alternative (and more rigorous way of) display could be a 3D plot with the measures on respective axes.

For evaluation, we compared treeMI and $\mathrm{treeMI}^{W\diamond P}$ with MI and $\mathrm{MI}^{W\diamond P}$ based on how well they discriminate between true and false base pairs. For this comparison, we selected a set of 961 structure-annotated RNA alignments from Rfam 10.1 (Gardner *et al.*, 2011), by keeping only those with at least five sequences, average pairwise sequence identity (PID) $\leq 90\%$, and removed

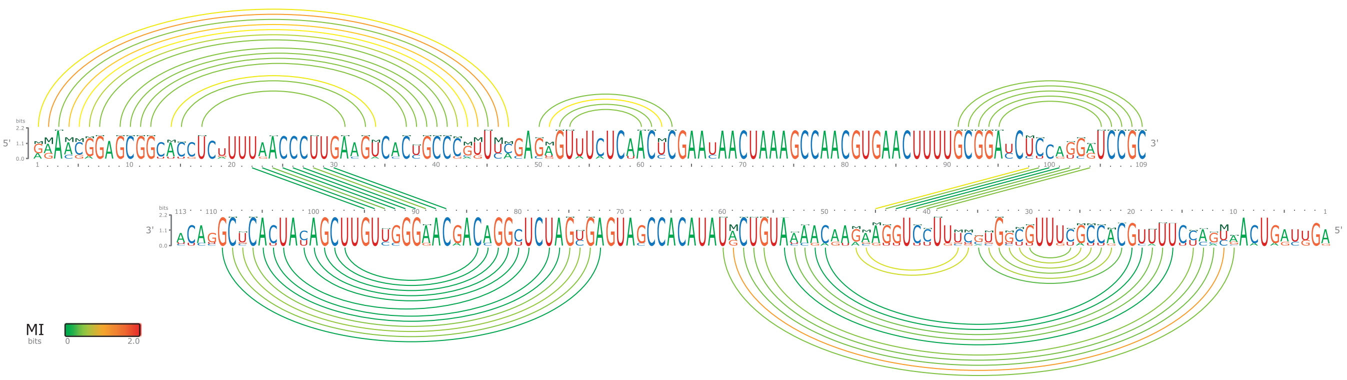

Fig. 1. Output of *RILogo* showing a sequence logo for bacterial RNA orthologs of the small RNA *OxyS* and its target *fhlA* [alignments from Huang *et al*. (2009)]. Intramolecular (arcs) and intermolecular (lines) base pairs are coloured by their mutual information using treeMI$^{W\diamond P}$ in log-scale. The *y*-axes show the *summed information content* (in bits) of the conservation for each position (the relative entropy) and its contribution to the (approximate) mutual information

redundant sequences with $\geq$ 99% similarity. Phylogenetic trees were inferred from the alignments with `FastTree2` (Price *et al*., 2010), which were then used to calculate d_{avg} in treeMI and treeMI$^{W\diamond P}$. Like Lindgreen *et al*. (2006), we calculate the best Matthews Correlation Coefficient (MCC; Matthews, 1975) for each alignment, where the best MCC is derived from the MI score that differentiates most accurate between annotated base pairs and unpaired bases.

Both treeMI and treeMI$^{W\diamond P}$ identified annotated base pairs better than MI and MI$^{W\diamond P}$, with an average MCC of 0.238 for treeMI and 0.241 for treeMI$^{W\diamond P}$, compared to an average MCC of 0.181 for MI and 0.217 for MI$^{W\diamond P}$. Figure 2a shows that treeMI captures the covariation within most alignments better than MI which results in better predictions (higher MCCs) of the annotated base pairs. By excluding those alignments with very low base pair prediction accuracy (MCC <0.1), the information gained by phylogeny even increases: average MCC (MI) = 0.315, average MCC (treeMI) = 0.395. Additionally, we observed that treeMI gives better results than treeMI$^{W\diamond P}$ for MCC >0.3, while treeMI$^{W\diamond P}$ is better than treeMI for lower MCC (Fig. 2b), due to the little covariation in these alignments. In contrast, MI$^{W\diamond P}$ gives higher MCCs than MI, as already observed in Lindgreen *et al*. (2006) (see Supplementary Fig. 1).

The average PID is negatively correlated to the alignment covariance, and thus, to the information that can be gained by the MI measure. By taking weights for evolutionary distances into account, however, treeMI$^{W\diamond P}$ improves MI$^{W\diamond P}$ in alignments with higher average PID, where less covariance is available (see Supplementary Table 1 and Fig. 1). In contrast, alignments with highly divergent sequences carry more covariance information, so that there is less improvement by including evolutionary distances. Supplementary Table 1 shows that treeMI performs better than treeMI$^{W\diamond P}$ for alignments with low average PID, because it also includes non-canonical base pairing in the mutual information measure.

3 CONCLUSION

RILogo visualizes two RNA sequence alignments by sequence logos, their intra- and intermolecular base pairing and the

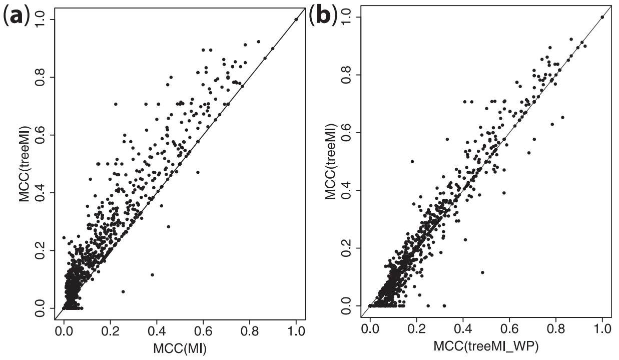

Fig. 2. Pairwise comparison of the MCCs for each alignment for the measures **(a)** MI and treeMI and **(b)** treeMI$^{W\diamond P}$ and treeMI

mutual information of base pairs. It also introduces two novel mutual information based measures: treeMI and treeMI$^{W\diamond P}$, that perform better than previous measures by including additional information from evolutionary distances. *RILogo* is accessible via a web server and its output can easily be customized. We publish the stand-alone version of *RILogo* under the GNU LGPL license, i.e. it is allowed to be explicitly integrated in other software/servers to display RNA–RNA interactions and sequence logos.

Funding: P.M. is funded by the 7th framework programme for research and technological development of the European Union (Project #265933 *Hotzyme*). This work was supported by the Danish Council for Independent Research (Technology and Production Sciences), The Danish Council for Strategic Research (Programme Commission on Strategic Growth Technologies), as well as the Danish Center for Scientific Computing.

Conflict of Interest: none declared.

REFERENCES

Akmaev,V. *et al*. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, **16**, 501–512.

Auber,D. *et al*. (2006) Efficient drawing of RNA secondary structure. *J. Graph Algorithms Appl*., **10**, 329–351.

Beitz,E. (2000) TEXshade: shading and labeling of multiple sequence alignments using LATEX2 epsilon. *Bioinformatics*, **16**, 135–139.

Bernhart,S. *et al*. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol*., **1**, 3.

Bindewald,E. *et al*. (2006) CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucleic Acids Res*., **34**, W405–W411.

Chang,T. *et al*. (2008) RNALogo: a new approach to display structural RNA alignment. *Nucleic Acids Res*., **36**, W91–W96.

Crooks,G. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.

Darty,K. *et al*. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Gardner,P. *et al*. (2011) Rfam: Wikipedia, clans and the 'decimal' release. *Nucleic Acids Res*., **39**, D141–D145.

Gorodkin,J. *et al*. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci*., **13**, 583–586.

Gorodkin,J. *et al*. (2010) De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol*., **28**, 9–19.

Hertz,G.Z. *et al*. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *CABIOS*, **6**, 81–92.

Huang,F. *et al*. (2009) Partition function and base pairing probabilities for RNA–RNA interaction prediction. *Bioinformatics*, **25**, 2646–2654.

Li,A. *et al*. (2011) RNA–RNA interaction prediction based on multiple sequence alignments. *Bioinformatics*, **27**, 456–463.

Lindgreen,S. *et al*. (2006) Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics*, **22**, 2988–2995.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem. Biophys. Acta*, **405**, 442–451.

Price,M.N. *et al*. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Schneider,T. and Stephens,R. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.

Seemann,S. *et al*. (2011a) PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, **27**, 211–219.

Seemann,S. *et al*. (2011b) The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. *Nucleic Acids Res*., **39**, W107–W111.

Tafer,H. and Hofacker,I. (2008) RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

Weinberg,Z. and Breaker,R. (2011) R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, **12**, 3.

Wenzel *et al*. (2012) RIsearch: Fast RNA-RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics*, doi: 10.1093/bioinformatics/bts519.