Visualization of sRNA-mRNA interaction predictions

Joris Sansen, Patricia Thébault, Isabelle Dutour and Romain Bourqui
Université de Bordeaux
{jsansen, thebault, dutour, bourqui}@labri.fr

Abstract

The central dogma in molecular biology postulated that 'DNA makes RNA makes protein', however this dogma has been recently extended to integrate new biological activities involving small non-coding RNAs, called sRNA*s. In particular, it has been shown that* sRNA*s regulate the production of proteins by interacting on* mRNA*s to regulate positively or negatively their translations. That regulation of the* mRNA *translation is done by forming a base-pairing between the RNAs sequences of bases.* In silico *methods have been proposed by the bioinformatics community to provide a list of putative interactions to be experimentally validated. However, such approaches suffers from a poor specificity and therefore produce a large number of false predictions. In this paper, we present a new visualization technique for sRNA-mRNA interactions emphasizing the involved regions on the sRNA secondary structure drawing. Our approach also supports interactive exploration as the user can select and highlight interactions. We demonstrate the usefulness of our approach by a case study on* E. coli *bacteria performed by domain experts.*

1 Introduction

For many decades, *DNA makes RNA makes proteins* have been considered as the central dogma in molecular biology. In that model, the *transcription* process synthetizes a *mRNA* (*messenger* RNA) from one specific region of the DNA molecule, called *gene*. The *translation* of that mRNA then enables the organism to synthetize a protein corresponding to the expressed gene. While that dogma provides an overview of the overall protein synthesis process, it has been recently extended to integrate new biological activities [17]. In particular, it has been shown that a family of small non-coding RNAs, called *sRNA*, allows the regulation of the organism at different scales [16, 20]. New sequencing technologies (*NGS*) together with specialized *mRNA* enrichment and tiling array techniques [22] have revealed the existence of a plethora of small regulatory RNAs in bacteria. The identification of the regulatory role of these *sRNA*s functions implies to carry out time-consuming and expensive biological experiments. *In silico* methods (*i.e.* computational methods) have therefore been developed (*e.g.* [1]) to prioritize gene candidates before designing an experimental protocol. However, these bioinformatics approaches are often poorly efficient in term of specificity and the number of false interactions can be large [21].

In this paper, we focus on the sRNA regulation that operates onto the translation of mRNAs into proteins. Such a regulation involves the formation of a base-pairing between the two molecules [23]. Such folding modifies the structure and stability of the *mRNA* to positively or negatively regulate its translation into protein or even its stability. In order to provide to biologists a reduced list of putative interactions to be experimentally validated, our visualization helps to filter out false positive predictions. To do so, an important information is the region of the sRNA, *i.e.* a sequence of contiguous bases, where putative interactions have been predicted. Indeed, a large number of predicted interactions on a region of the molecule may indicate that this region has been constrained during evolution. It therefore provides arguments for further investigation. The region information may not be sufficient as the self-folding of the RNA sequence creates constraints on region accessibility. To address that issue, our visualization method displays RNA secondary structure (*i.e.* self-folding of the RNA sequence in 2D) to help the expert to identify reachable regions of the molecule.

To the best of our knowledge, there is very few related work dedicated to the visualization of RNA-RNA interactions. *rNAV* software [7] and *CopraRNA* web service [26] focus on the analysis of sRNA-mRNA interactions at the genome scale. In these tools, interacting regions are displayed using histogram representation where the abscissa represents the sequence of RNA bases (called *RNA primary sequence*) while the ordinate represents the number of interaction predicted for each base of the molecule. Using such representation, one can easily identify highly interacting regions whereas no information is given about the accessibility of these regions.

Considering that the representation of the secondary structure is provided, the problem addressed in this paper can also be seen as an overlapping clusters visualization problem together with the automatic positioning of these clusters labels. In this context, each interacting region cor-

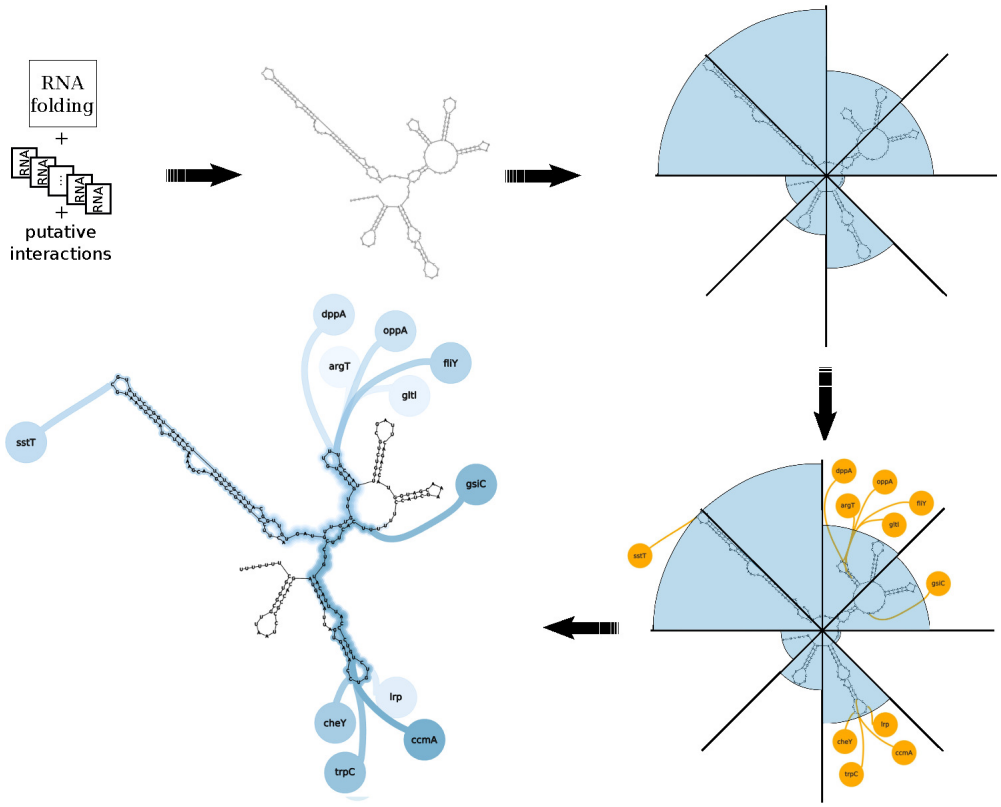

Figure 1: Overview of our method. First step consists in drawing the sRNA secondary structures. Then angular sectors are computed to approximate that drawing. Next mRNA targets are next laid out around the secondary structure and linked to the corresponding region. Finally interactions are rendered using concave hulls [11].

responds to a cluster that overlaps and as regions overlap then the corresponding clusters overlap as well. And, a cluster label (region label) corresponds to the putative interacting mRNA. Again there is, to the best of our knowledge, no related work adressing these two problems together. On one hand, some methods have been designed for the visualization of overlapping clusters in a fixed layout (*e.g.* [6, 11, 15]). The basics of these methods are usually to compute hulls or envelops surroundding each cluster. However, none of these methods allows the automatic positioning of these cluster labels. For a recent survey on the visualization of clusters in graphs, the reader can refer to [24]. On the other hand, one can find in the litterature several approaches for automatic positioning of labels (*e.g.* [3, 13, 14]) but very few refers to clusters labels (*e.g.* [13]) while the other refer to positioning of element labels. In [13], the proposed approach supports the automatic postionning of elements labels. As mentionned in the paper, the method could be extended to elements of arbitrary shape (a cluster could here be considered as an element). However the complex shapes of the clusters considered here would increase drastically the computational cost of such method.

In this paper, we present a new visualization pipeline allowing to show sRNA-mRNA interactions by emphasizing the involved regions on a RNA secondary structure drawing. Our approach also supports interactive exploration as the user can select and highlight interactions, *i.e.* a specific interaction or the set of interactions involving a specific base. The remainder of this paper is structured as follows. In section 2, we give some definitions as well as the terminology used in this paper. We next describe in detail each step of that pipeline in section 3. We then present some results and show the efficiency of our method on a case study in section 4. Finally, we draw a conclusion and give directions for future work.

2 sRNA-mRNA interactions : challenges

A small RNA chain (molecule) can be represented as a linear sequence of letters that symbolize the 4 different bases (A, C, G and U) and are connected by inflexible covalent bonds. *In vivo*, the linear molecule folds into its 3D structure according to the flexible hydrogen bonds that can be formed by some base association (*e.g.* A with U).

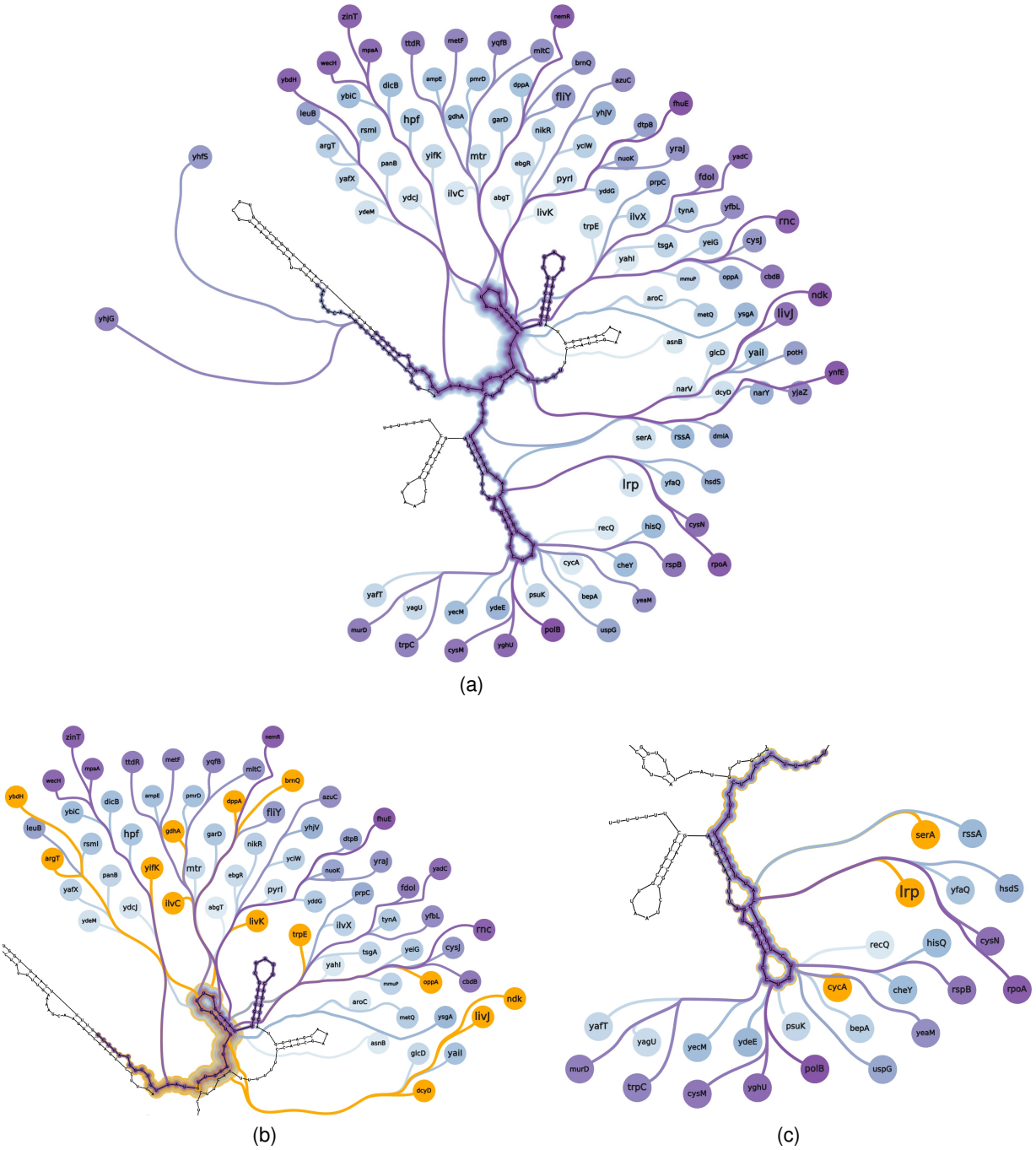

Figure 2: (a) Visualization of the *gcvB* sRNA secondary structure with the 100 best predicted mRNA targets. Hulls have been colored according to the prediction scores. (b) and (c) Zoom on the two highly interacting regions of *gcvB* together with the corresponding mRNA targets. Experimental validated targets have been higlighted in orange.

A simplification of the real structure is given by the secondary structure that focus on the hydrogen bonds that can be represented as a planar graph. Typically, a secondary structure will show up the successive hairpin loops that can be formed by the nucleotide chain. A hairpin brings closer two parts of the small RNA molecule which is crucial for the biological functionality of the molecule. The interaction between a *sRNA* and its *mRNA* targets is first initiated

with few accessible bases (for instance, the bases that are not base pairing other bases of the *sRNA*) before being extended with less accessible bases. As mentionned above, prediction tools perform poorly in term of specificity, visualization techniques are therefore needed to filter out false positive interactions. Several biological features can be considered to improve the candidate confidence level:

- A small RNA may interact with several mRNAs. To maintain the capacity of interacting with different mRNA sequences, a particular region of the sRNA may be affected by an evolutive selective pressure. In other word, a region that can interact with several mRNAs is an interesting candidate,
- The accessibility of the interacting region may be another positive criterion for evaluating the candidates.

The main challenge in this context is to design a visualization technique that emphasizes both the regions involved in putative interactions and their accessibility.

3 Method

Figure 1 shows an overview of our method to build an interactive representation of sRNA-mRNA interactions from a focused sRNA and a list of putative mRNA targets. To do so, we propose a 4-steps pipeline. The first step of the pipeline consists in drawing the secondary structure of the sRNA. In the second step, we compute angular sectors. Such sectors are used to approximate the drawing of the secondary structure and allows to define areas where the mRNA targets can be positionned without overlap. During the third step, we assign each mRNA to an angular sector and compute their positions. Interacting mRNAs are then linked to their corresponding regions of the sRNA. Finally concave hulls are computed to emphasize all the putative interactions as well as the corresponding regions of the sRNA.

3.1 Drawing RNA secondary structure

As mentionned above, the first step of our pipeline is to draw the secondary structure of the sRNA. There exists many drawing algorithms in the literrature (*e.g.* [2,4,8,9]). In this work, we used the algorithm proposed by Auber *et al.* [2] to compute the secondary structure drawing. This algorithm produces layouts similar to manually drawn ones. To do so, it contains three main steps: (i) modeling the secondary structure with a tree; (ii) drawing the tree of (i); and (iii) computing base positions according to the coordinates computed in (ii).

3.2 Sector computation

The second step of our pipeline consists in computing angular sectors together with an associated distance. To compute them, we first determine the center of the drawing. To do so, several options are possible, for instance, the center of the bounding box (or the bounding circle) of the drawing or the barycenter of the base positions. We use the tree of [2] for modeling the secondary structure. More precisely, the center of the drawing is set to the center of the central loop of the secondary structure. In that tree, such loop can be retrieved by traversing the tree from the root node and stopping at the first node of out-degree more than 2. Once the center of the drawing has been determined, the plane is split into k angular sectors. For each sector, we compute the maximal distance between the center of the drawing and each base of the structure falling in that sector.

3.3 Assigning sectors and positionning interacting RNAs

During the third step, mRNAs are assigned to an angular sector and then positionned within that angular sector.

Each putative interaction involves a specific region of the sRNA sequence. For each interacting mRNA, the barycenter of the bases of that region is computed and the mRNA is assigned to the sector in which the barycenter falls. For each angular sector, the interacting mRNAs are laid out on concentric layers. These mRNAs are therefore assigned to a layer and then ordered within each layer to reduce clutter in the final representation. We can easily compute, from the distance and the angle of the sector, the number of layers as well as the number of mRNAs (considering their size as given) per layer. To assign the mRNAs to each layer, we sort them according to the scores provided by the prediction tool. Interacting mRNAs are then assigned to layers starting from the closest to the center to the farest. Using that strategy emphasizes predictions with the highest levels of confidence. To order the mRNAs of each layer, our method is similar to algorithm for minimizing the number of edge crossings in hierarchical drawing of DAG. The goal here is not to minimize the number of edge crossings but rather to minimize the average distance between a mRNA and its region of interaction. This is achieved by computing, for each mRNA, the barycenter of the bases of its region. These barycenters are then projected on the circular arc defined by the angular sector and the corresponding distance. Interacting RNAs of the layer are finally ordered according the positions of their corresponding projections.

3.4 Building concave hulls

To highlight the regions of interaction and the corresponding mRNAs, we use the technique of Lambert *et al.* [11]. This method allows to emphasize the clusters of an overlapping decomposition of a graph by building concave hulls around each cluster. To use that technique, we need to link the mRNAs to their corresponding region of interaction. To minimize the clutter in the representation, each interacting RNA is linked to the closest base of its corresponding region. In addition, our method bundles these

links with *Winding Road* algorithm [10]. The method of Lambert *et al.* [11] is then applied to compute hulls around each mRNA and the corresponding region. In order to emphasize the scores associated to each putative interaction (according to the prediction tool), a color mapping is finally used for coloring the hulls.

The main advantage of the method of Lambert *et al.* [11] is to highlight regions of the secondary structure where the number of predictions is high (see figure 2). In addition, this method also supports relevant interaction tools. First it supports the selection of a single hull to ease the identification of a specific interaction and the corresponding region. Second, it also supports the selection of all hulls which contain specific bases. The last allows to identify all putative interactions involving these bases.

4 Case study

To illustrate the efficiency of our approach, we present in this section a case study based on the *E. coli* sRNA, called *gcvB*. Of particular interest, the *E. coli gcvB* is an homolog of the well characterized *Salmonella gcvB* sRNA for which more than 10 targets have been experimentally validated in the *Salmonella* genome [19]. Exploiting the validated data known for *Salmonella*, we investigated the *E. coli gcvB* with the objective to confirm the regulation pattern (involving the same region in the sRNA and mRNAs) and to identify new putative candidates. The small evolutive distance between both bacteria may allow to transfer the knowledge known from *salmonella* to *E. coli*.

The input data of the *E. coli* case study have been generated with: RNAfold [12] to compute the *gcvB* secondary structure ; intaRNA software [1] to predict the 100 best mRNA target candidates of *gcvB*.

Figure 2 shows the resulting visualization of the prediction with in the central part the *gcvB* secondary structure, and in its periphery the putative targets (labeled with the name of the target gene). The interaction of these targets are represented onto the *gcvB* sequence with envelops of different thicknesses (see figure 2). By exploiting the visual information given by the thickness of these envelops, two regions with an high number of predictions can be observed and having accessible hairpin-loops. Selecting bases of these hairpin-loops allows then to highlight these two regions as well as the corresponding putative mRNA targets (see figures 2.(b) and (c)). Thanks to this resulting visualization, users can expertize the predicted targets according to their sRNA interacting region. Of particular interest, they can both exploit the information of the sRNA secondary structure and the conservative region criteria. A group of mRNAs interacting with a same *gcvB* region that may contain accessible bases is a relevant argument to improve the confidence level of their predictions. This observation is confirmed by most of the experimental validated targets extracted from the sRNAtarbase database [25].

Moreover, for the Salmonella *gcvB* sRNA, a secondary structure has been proposed [18], composed of five hairpins and two single strand regions (namely R1 and R2). For the R1 region, experimental works have shown the *gcvB* regulatory role with mRNAs involved in the peptide transport or/and the acid stress response (for example see [18]). Concerning the R2 region, three targets have been validated so far [19]. As displayed in figures 2.(b) and (c),the two regions of interest according to the visualization approach presents closed similarity with the *salmonella* validated data in terms of accessibility and multi targeting region. The candidate targets are interacting with one of the two highlighted regions with an interaction that could be initiated with accessible bases. Moreover, other targets following the same region constraints can be prioritize and give candidate of interest for designing an experimental protocol.

Conclusion

In this paper, we presented a new technique that allows to visualize sRNA-mRNA interactions. Our approach allows to visualize the regions where these interactions occur on a drawing of the sRNA secondary structure. While emphasizing regions with an high number of predictions provides some arguments for further investigation, the secondary structure drawing allows to filter out interactions that are unlikely to happen as involving an unaccessible part of the molecule. In addition, interaction tools supporting the highlighting of specific putative interactions and of interactions involving specific bases of the sRNA are also provided. Finally, we showed the efficiency of our approach with a case study on *E. coli* bacteria.

As future work, we plan to integrate the mRNA/gene biological activity information to gain a global view of the involved biological features. Indeed, as suggested by Beisel and Storz [5], the multiple mRNAs targeted by one sRNA may be related to the same biological process.

References

[1] Busch A, Richter AS, and Backofen R. Intarna: efficient prediction of bacterial srna targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, Dec 2008.

[2] D. Auber, M. Delest, S. Dulucq, and J.-P. Domenger. Efficient drawing and comparison of RNA secondary structure. *Journal of Graph Algorithms and Applications*, 10:329–351, 2006.

[3] E. Bertini, M. Rigamonti, and D. Lalanne. Extended excentric labeling. In *Computer Graphics Forum*, volume 28, pages 927–934. Wiley Online Library, 2009.

[4] R.E. Bruccoleri and G. Heinrich. An improved algorithm for nucleic acid secondary structure display. *Computer applications in the biosciences: CABIOS*, 4(1):167–173, 1988.

[5] Beisel CL and Storz G. Base pairing small rnas and their roles in global regulatory networks. *FEMS Microbiol Rev*, 34(5):866–882, Sep 2010.

[6] C. Collins, G. Penn, M. Sheelagh, and T. Carpendale. Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations. *IEEE Trans. Vis. Comput. Graph.*, 15(6):1009–1016, 2009.

[7] J. Dubois, A. Ghozlane, P. Thebault, I. Dutour, and R. Bourqui. Genome-wide detection of sRNA targets with rNAV. In *Symposium on Biological Data Visualization*, pages 81 – 88, United States, Oct 2013.

[8] K. Han, D. Kim, and H. J. Kim. A vector-based method for drawing RNA secondary structure. *Bioinformatics*, 15(4):286–297, 1999.

[9] K. Han and Byun Y. PSEUDOVIEWER2: Visualization of RNA pseudoknots of any type. *Nucleic Acids Res*, 31(13):3432–40, Jul 2003.

[10] A. Lambert, R. Bourqui, and D. Auber. Winding roads: Routing edges into bundles. In *Computer Graphics Forum*, volume 29, pages 853–862. Wiley Online Library, 2010.

[11] A. Lambert, F. Queyroi, and R. Bourqui. Visualizing patterns in node-link diagrams. In *16th Int. Conference on Information Visualisation 2012*, pages 48–53. IEEE, 2012.

[12] R. Lorenz, S.H. Bernhart, C.H. Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, and I.L. Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):1, 2011.

[13] M. Luboschik, H. Schumann, and H. Cords. Particle-based labeling: Fast point-feature labeling without obscuring other visual features. *IEEE Trans. on Visualization and Computer Graphics,*, 14(6):1237–1244, 2008.

[14] Y. Meng, H. Zhang, M. Liu, and S. Liu. Clutter-aware label layout. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific*, pages 207–214. IEEE, 2015.

[15] W. Meulemans, N. Henry Riche, B. Speckmann, B. Alper, and T. Dwyer. Kelpfusion: A hybrid set visualization technique. *IEEE Trans. on Visualization and Computer Graphics,*, 19(11):1846–1858, 2013.

[16] P Romby and EGH Wagner. Exploring the complex world of rna regulation. *Biol Cell*, 100(1):e1–e3, 2008.

[17] J.A. Shapiro. Revisiting the central dogma in the 21st century. *Ann N Y Acad Sci*, 1178:6–28, Oct 2009.

[18] C.M. Sharma, F. Darfeuille, T.H. Plantinga, and J. Vogel. A small rna regulates multiple abc transporter mrnas by targeting c/a-rich elements inside and upstream of ribosome-binding sites. *Genes Dev*, 21(21):2804–2817, Nov 2007.

[19] CM Sharma, K Papenfort, SR Pernitzsch, HJ Mollenkopf, JCD Hinton, and J Vogel. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the hfq-dependent gcvb small rna. *Mol Microbiol*, 81(5):1144–1165, Sep 2011.

[20] G. Storz, J. Vogel, and K.M. Wassarman. Regulation by small rnas in bacteria: Expanding frontiers. *Mol Cell*, 43(6):880–891, Sep 2011.

[21] P Thébault, R Bourqui, W Benchimol, C Gaspin, P Sirand-Pugnet, R Uricaru, and I Dutour. Advantages of mixing bioinformatics and visualization approaches for analyzing srna-mediated regulatory bacterial networks. *Brief Bioinform*, 16(5):795–805, 2015.

[22] A Toledo-Arana, O Dussurget, G Nikitas, N Sesto, H Guet-Revillet, D Balestrino, E Loh, J Gripenland, T Tiensuu, K Vaitkevicius, M Barthelemy, M Vergassola, M-A Nahori, G Soubigou, B Régnault, J-Y Coppée, M Lecuit, J Johansson, and P Cossart. The listeria transcriptional landscape from saprophytism to virulence. *Nature*, 459(7249):950–956, Jun 2009.

[23] A. Toledo-Arana, F. Repoila, and P. Cossart. Small noncoding rnas controlling pathogenesis. *Curr Opin Microbiol*, 10(2):182–188, Apr 2007.

[24] C. Vehlow, F. Beck, and D. Weiskopf. The state of the art in visualizing group structures in graphs. In *Eurographics Conference on Visualization (EuroVis)-STARs*, pages 21–40, 2015.

[25] J. Wang, T. Liu, B. Zhao, Q. Lu, Z. Wang, Y. Cao, and W. Li. srnatarbase 3.0: an updated database for srna-target interactions in bacteria. *Nucleic Acids Res*, 44(D1):D248–D253, Jan 2016.

[26] P.R. Wright, J. Georg, M. Mann, D.A. Sorescu, A.S. Richter, S. Lott, R. Kleinkauf, W.R. Hess, and R. Backofen. CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Res.*, 42(W119-23), Jul 2014.