Visualizing Uncertainty of RNA Sequence Base Pairing Variants

Fleur Jeanquartier, Claire Jean-Quartier and Andreas Holzinger

Abstract—This work describes a design oriented approach to visualizing uncertainty of RNA secondary structure probabilities. We address the challenge of finding an intuitive visual representation of encoding uncertainty in RNA secondary structures. We highlight certain limitations and present three different but not exclusive approaches for tackling this challenge.

1 Introduction

In molecular biology researchers have to deal with a decreasing certainty when predicting secondary structures of RNA sequences. Practical testing is limited, computational methods fill the gap in the data with predicted and hence uncertain data. Computational biologists have developed methods to predict the secondary structures (2D folding views of RNA) from a primary sequence of RNA. The outputs of this calculation includes the minimum free energy structure (MFE), the thermodynamically favored and most likely structure, and equilibrium base pairing probabilities. These outputs are typically visualized as a "dot plot", where a box on a square grid of nxn (n is the sequence's length) encodes the base pair binding probability in its area on a logarithmic scale. In addition, the predicted MFE structure is often represented as a secondary structure graph.

2 Background

Dot plots (base pair probability matrices) are a common way for visualizing secondary structure calculations. The squares in the plot area represent a pair (x,y), while either color, transparency, blur effects or size of a dot is used to indicate the probability of a base pair [13]. For today, conservation consensus dot plots can even be interactively controlled to some extent: For example, Sorescu et al. [12] describes a mechanism to specify a threshold probability for dynamic visualization adaptation. However, dot plot representations for base pair probabilities are also said to be confusing when complexity rises, and therefore alternative representations exist too. Base pairings visualization can also be found as linear and circular representations. Alberts et al. [1] introduced so called "RNAbow" diagrams. Hofacker [6] described a software package for analyzing secondary structures and rendering structures as mountain plot and other representations.
When speaking of uncertainty, uncertain data sets may have diverse sources, including data acquisition (signal-to-noise ratio), data mapping (pre-processing and post-processing) and the visualization method itself. Uncertainty can be described as a composite of different concepts, such as errors, accuracy, and subjectivity [4]. Visualizing uncertainty is a difficult problem in all kinds of scientific domains too [5, 11, 2, 8]. Potter et al. [10] already identified uncertainty representations commonly used in visualization and presented a taxonomy of visualization approaches.
None of the mentioned research already dealt with visually encoding uncertainty of the complete set of folding possibilities into one single visualization.
Therefore, we submit this entry to the BioVis 2015 Design Contest [3], that addresses the challenge of visualizing uncertainty of RNA secondary structures. In the following, we describe our visual approaches to the challenge of visualizing uncertainty.

- *Fleur Jeanquartier, Claire Jean-Quartier and Andreas Holzinger are with the Research Unit HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz. E-mail:* {f.jeanquartier,c.jeanquartier,a.holzinger}@hci-kdd.org

3 Visual Approach to Challenge 1

We address the first contest's challenge, namely visualizing uncertainty. The problem is defined as follows:

3.1 Problem:

Design an intuitive visual representation of RNA secondary structure to encode the uncertainty within all the possible base pairing possibilities. The top-right triangle of a dot plot encodes base pairing probabilities and the bottom-left triangle represents the MFE structure. The RNA sequence of n nucleotides is shown on the edge of the nxn square grid. The MFE secondary structure is visualized as a graph, where the color of each nucleotides depicts the strength of base pairing. The challenge is to design a structural representation that is in line with the uncertainty.
To deal with this challenge, however, using the right visualization technique is a question of scaling: An unanswered question remains: What is the limit of possible base pairing probability matrices that can be visualized within one single visualization? Since the number ob potential secondary structures is exponential to the rna sequence's length n [9]. Therefore, we present the following three different approaches for (interactive) visual analysis of rna base pair configurations:

3.2 Approach 1:

One possible interactive visualization approach is sketched in Fig. 1:

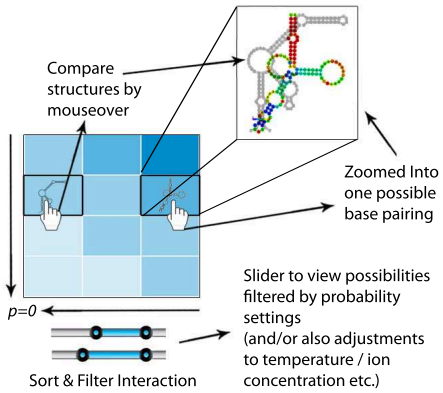

Fig. 1. Visualizing encoded uncertainty of RNA secondary structure possibilities as interactive heatmap including detail view

Holzhuter et al. [7] have shown that particularly heat maps can be dangerous as they can be over-plotted. It is possible, up to a certain amount, to visualize the ensemble organized in a heatmap. But, as common to information visualization, there will be the necessity to integrate interactive exploration features for zoom and filter. We also sketched such interaction integrations. The slider filter at the bottom supports viewing only those rectangles that are related to the most probable configurations but also allows for highlighting the unusual ones. Different perspectives support the interactive visual analysis approach. Additional interactions should be taken into account, like a

slider for filtering specific temperature areas and/or ion concentration settings and adding a switch for sorting not only by probability but also other data variables (i.e. number of base pairs, hairpins, free energy).

3.3 Approach 2:

To overcome some of the heatmap's limitations, another additional or alternative approach is visualizing the complete set of dot plot representations as interactive visual analysis approach making use of the "Rolodex"-art metaphor (also known from window manager in operating systems, apple's time machine or windows exposé), illustrated in Fig. 2. All possible structures are visualized as matrices one after another, while the most probable, the MFE, is the first one on top and behind lay the less probable ones. Interaction allows for toggling through all the possible structures seamlessly while clicking on upper right part of the dot plot all secondary structures are shown in a details view the following manner: All the possible configurations are shown at once, while the most probable is on top. Below all other configurations are shown but with increased transparency values. The most likely is therefore 100% opaque, while the less likely ones are more translucently renderer.
Additionally, Eterna's animation metaphor can be used: Single bases and base pairs within the details view can be animated insofar, as the base pairs movement in pixel per second is related to the structure's folding stability and probability.

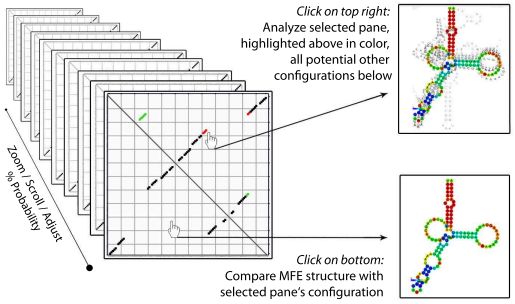

Fig. 2. Visualizing encoded uncertainty of RNA secondary structure base pairings by exploring complete set/ensemble at once

3.4 Approach 3:

Last but not least, another possible approach could be visualizing all possibilities not as box but as part of a network graph, sketched in Fig. 3. The graph is composed by the complete ensemble of structures as follows: Each node represents one possible folding structure, each edge stands for a user defined number x similar base pairs between two structures, while the whole graph integrates the complete "picture". Thereby, similar base pair areas can be marked with another color (compare sketched red area in Fig. 3)
The nodes' transparency (or color/contrast variance) depicts the probability of the particular structure. The node that stands for the MFE is highlighted (in darkest contrast or special color) as the root or center of the graph as the most probable base pairing combination. If the MFE is not the most probable configuration, the visualization can be adapted to distinguish between root, as most probable one, and MFE, as a node somewhere else within the graph highlighted by another color.
According to the dynamic programming algorithm for all subsequences (i, j) of a dot plot, the less probable folding possibilities can be traced back too. Less probable configurations are marked in a translucent manner: The more like configurations are represented by nodes with higher opacity while the more unlikely ones are rendered with less opacity.
Regarding the interaction: By adjusting x certain isles are highlighted, where the configurations represented by the nodes within an isle are more similar to each other. Additional network analysis approaches may further suite the rna analysis process.

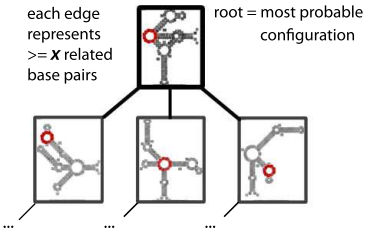

Fig. 3. Visualizing encoded uncertainty of RNA secondary structure by putting focus on the configurations' related base pairs as network graph

4 Material And Methods

Due to the fact, that the submission should be no more than 2 pages we include only a few figures into it. We also recommend watching a short animation, that depicts some details about the three different visualization approaches and the structural representation that is in line with the uncertainty: `http://youtu.be/PZp5GNpNZX4`.

5 Terms and Conditions

By submitting this entry, we give the BioVis 2015 organizers permission to publish it in conference-related materials. Any usage or reference to any submission will include full credit to its authors.

Acknowledgments

We gratefully acknowledge the dataset provided by Maria Beatriz Walter Costa, Henrike Indrischek, Katja Nowick and Christian Hner zu Siederdissen at The University of Leipzig for the purposes of the BioVis 2015 Contest.

References

[1] D. P. Aalberts and W. K. Jannen. Visualizing rna base-pairing probabilities with rnabow diagrams. *RNA*, 19(4):475–478, 2013.
[2] M. Albrecht, A. Kerren, K. Klein, O. Kohlbacher, P. Mutzel, W. Paul, F. Schreiber, and M. Wybrow. On open problems in biological network visualization. In *Graph Drawing*, pages 256–267. Springer, 2010.
[3] BioVis. Design contest, 2015.
[4] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *SimVis*, pages 143–156, 2006.
[5] C. D. Hansen, M. Chen, C. R. Johnson, A. E. Kaufman, and H. Hagen. *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*. Springer, 2014.
[6] I. L. Hofacker. Rna secondary structure analysis using the vienna rna package. *Current protocols in bioinformatics*, pages 12–2, 2009.
[7] C. Holzhüter, A. Lex, D. Schmalstieg, H.-J. Schulz, H. Schumann, and M. Streit. Visualizing uncertainty in biological expression data. In *IS&T/SPIE Electronic Imaging*, pages 82940O–82940O. International Society for Optics and Photonics, 2012.
[8] A. Holzinger, M. Schwarz, B. Ofner, F. Jeanquartier, A. Calero-Valdez, C. Roecker, and M. Ziefle. Towards interactive visualization of longitudinal data to support knowledge discovery on multi-touch tablet computers. In *Availability, Reliability, and Security in Information Systems*, pages 124–137. Springer, 2014.
[9] G. Pavesi, G. Mauri, M. Stefani, and G. Pesole. Rnaprofile: an algorithm for finding conserved secondary structure motifs in unaligned rna sequences. *Nucleic acids research*, 32(10):3258–3269, 2004.
[10] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, 2012.
[11] J. Smith, D. Retchless, C. Kinkeldey, and A. Klippel. Beyond the surface: current issues and future directions in uncertainty visualization research. In *Proceedigs of the 26th International Cartographic Conference*, pages 1–10, 2013.
[12] D. A. Sorescu, M. Möhl, M. Mann, R. Backofen, and S. Will. Carnaalignment of rna structure ensembles. *Nucleic acids research*, page gks491, 2012.
[13] A. Wilm, K. Linnenbrink, and G. Steger. Construct: improved construction of rna consensus structures. *BMC bioinformatics*, 9(1):219, 2008.