Visualizing Ensembles of Predicted RNA Structures and Their Base Pairing Probabilities

Peter Kerpedjiev and Ivo Hofacker

Abstract—In this design contest submission we present an enhanced version of a traditional RNA dot plot containing a multitude of extra features and data, foremost among which is the inclusion of diagrams for the top Zuker sub-optimal RNA secondary structures. This new design facilitates and eases the interpretation of the dot plot by providing the viewer with an immediate representation of which structures the displayed base-pair probabilities belong to.

1 Introduction

The traditional RNA dot plot conveys the probability that a particular base-pair is present in the ensemble of predicted structures. This information is presented as a 2D scatter plot, where the size of the rectangular marks is proportional to the probability of a pairing between nucleotide i (on the x-axis) and nucleotide j (on the y-axis). The upper right triangle of the plot displays this information for the ensemble of predicted secondary structures whereas the bottom left displays only the pairs present in the minimum-free energy structure (MFE). The dot plot is useful in conveying to the viewer that some nucleotides may have a propensity to form differing base-pairs. At first glance, it shows whether there are stems which are consistent across the whole ensemble and which nucleotides they encompass.

Beyond this application, however, it becomes difficult (albeit far from impossible) to extract extra information. The key unanswered question, in our opinion, is which structures correspond to the indicated base-pairs? As previously mentioned, the pairs corresponding to the MFE structure are shown in the lower left hand corner. What does this structure look like, however? What about the other base-pairs in the upper right section? Which structures do those correspond to? How many different structures do they correspond to? Which can be found in the same sub-optimal structure?

With these questions in mind, we set about redesigning the dot plot to include actual secondary structure diagrams in the background. The result, shown in Figure 1, gives the viewer an answer to each of the questions posed above and more. It further provides a platform which can be extended to create an interactive tool to ease the exploration of the data presented in the visualization.

2 Design Considerations

Our design was created to answer some basic questions that researchers might ask about an ensemble of predicted RNA structures, as well as to provide some minor improvements to the way the traditional dot plot is laid out. In each section we describe what we did, why we did it, as well as how we feel it could be improved with an interactive version of our design.

2.1 What does the MFE structure look like?

Description: In the traditional use case, one receives a secondary structure diagram representing the minimum free energy structure in one file and the dot plot in another. We strive to unite these two representations by showing the MFE structure in the background of the dot plot. Such an approach is alluded to in a figure in [3], but we go one step further and arrange the MFE structure along with other sub-optimal structures and scale their size according to their expected population in the Boltzmann ensemble of predicted secondary structures.

Motivation: The give the viewer an immediate representation of the MFE secondary structure.

2.2 Which other structures are predicted?

Description: RNA folding, being a kinetic process, leads to the presence of more than one particular structure in solution. We display a subset of these sub-optimal structures, along with the MFE structure, in the background of the dot plot. Based on the energy of each predicted structure, one can calculate its expected weight within the ensemble and use it to scale the size of its secondary structure using a squarified treemap layout [1]. Only structures which correspond to base-pairs with a probability above a certain threshold (see next section) are displayed.

Motivation: The MFE structure can quickly be compared to the other predicted structures in the ensemble in terms of not only structure, but also energy value.

Potential Improvement: Some structures can appear quite small. An interactive version of the plot can enlarge them when one hovers over a base pair belonging to that structure.

2.3 Which structures do the predicted base pairs correspond to?

Description: The upper right hand corner of the dot plot shows all of the potential predicted base pairs above a certain probability value (0.08 in our case, 0.00001 in the traditional dot plot). We chose a higher cut-off due to the simple fact that a lower cutoff would yield points so small as to be virtually indistinguishable without a magnifying glass. Each of the dots is colored to match the color of the best sub-optimal secondary structure containing that base pair. Recall that these structures are displayed in the background of the dot plot.

Motivation: This encoding helps to link the predicted base pairs with the structures they are expected to appear in.

Potential Improvement: Increasing the size on mouse-over, as suggested in the previous section, should help alleviate this issue. Clicking on a structure could also be employed to highlight/enlarge the base pairs belonging to it.

2.4 Which base pairs in a structure are displayed in the dot plot?

Description: The MFE and sub-optimal structures in the background are generated by finding the lowest energy structure given a base pair constraint. Within those structures, we highlight the pairs which, when constrained to being paired, lead to the prediction of that structure. These also correspond to the base-pairs displayed as dots on the dot plot.

Motivation: By highlighting the base pairs in the secondary structure, one can easily see not only how many, but which pairs in a sub-optimal structure are represented in the dot plot.

Potential Improvement: The identity of the base-pairs could be clarified by drawing lines between the secondary structure and the dots when users hover the mouse over the dots.

2.5 Minor improvements

Nucleotide Numbering: We added the positions of the nucleotides to the margins. To avoid clutter, we only add the numbers for nucleotides

- *Peter Kerpedjiev (pkerp@tbi.univie.ac.at) and Ivo Hofacker (ivo@tbi.univie.ac.at) are both at the University of Vienna.*

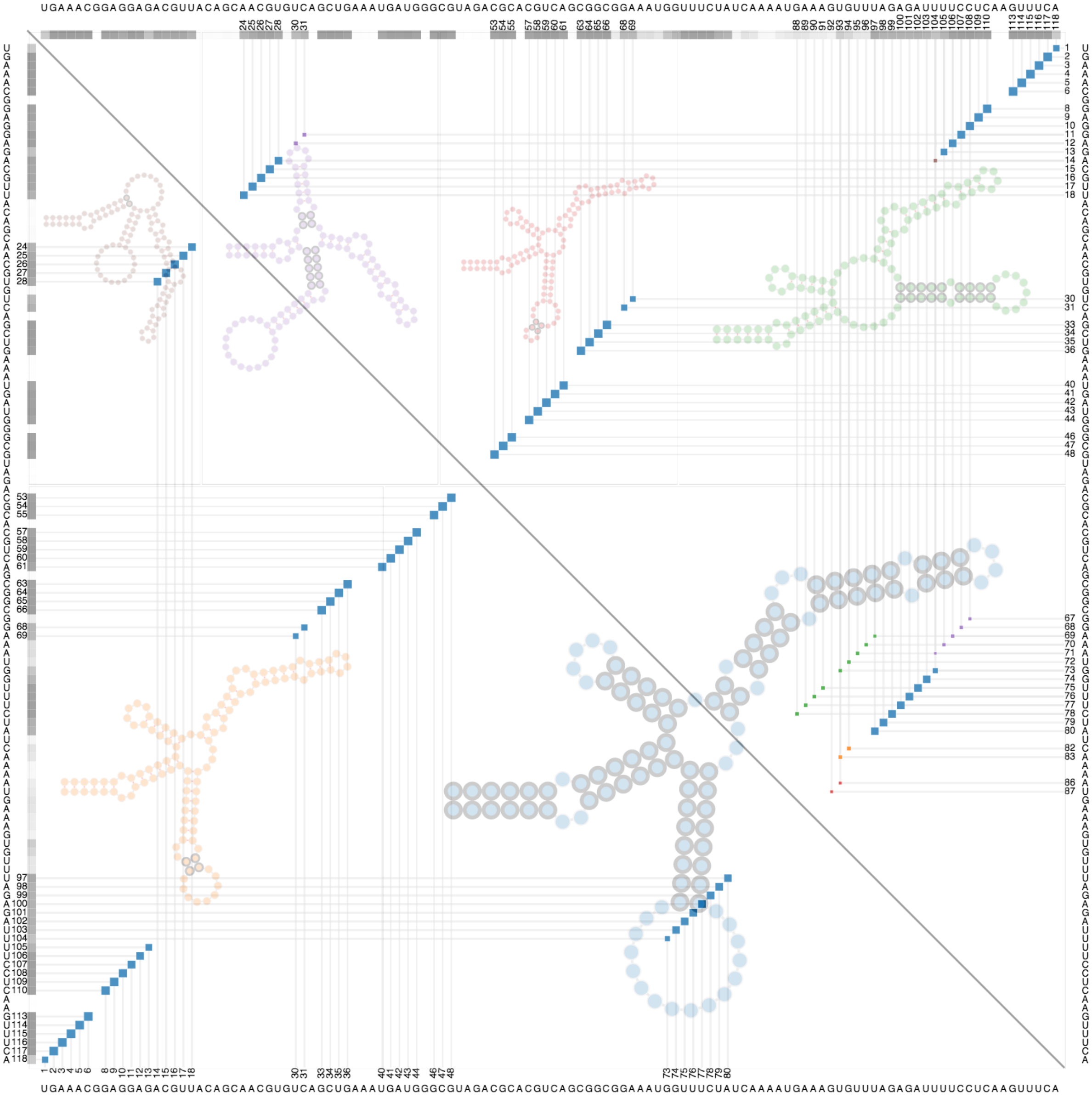

Fig. 1. Sample enhanced dot plot for the Human Highly Accelerated Region 1A provided in the contest data.

that have the potential to be in a base-pair (i.e. have a probability greater than the threshold).

Numbering Guide Lines: To guide the viewer in reading out the identity and position of each paired nucleotide, we have added faint lines from each dot on the dotplot to the numbers of the nucleotides in the margin

Total Pairing Probability: The summed pairing probability for each nucleotide is encoded as colored squares on the upper and left border of the plot. This provides an overview of which nucleotides are likely to be paired in the whole ensemble. It can be used as a comparison with data from probing experiments.

3 Generation

The data for the plot is generated by a python script which makes use of the python binding of the ViennaRNA package. The actual plot is rendered in the browser using the D3.js and fornac.js libraries. Such a format makes it easy to add interactivity to the current design.

4 Availability

The code for creating this visualization is available at:

`https://github.com/pkerpedjiev/dotstruct`

A higher resolution rendering of Figure 1 can be found at:

`http://www.tbi.univie.ac.at/~pkerp/dotplus/`

Acknowledgments

We wish to thank Ronny Lorenz for his help with the ViennaRNA package python interface and Stefan Hammer for his tireless work on creating the secondary structure visualization tool **forna** [2] which paved the way for this submission.

References

[1] M. Bruls, K. Huizing, and J. J. Van Wijk. *Squarified treemaps*. Springer, 2000.

[2] P. Kerpedjiev, S. Hammer, and I. L. Hofacker. forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *submitted*, 2015.

[3] R. B. Lyngsø, J. W. Anderson, E. Sizikova, A. Badugu, T. Hyland, and J. Hein. Frnakenstein: multiple target inverse RNA folding. *BMC bioinformatics*, 13(1):260, 2012.