iDotter – an interactive dot plot viewer

Daniel Gerighausen[1,2]
daniel@informatik.uni-leipzig.de

Alrik Hausdorf[1]
hausdorf@informatik.uni-leipzig.de

Sebastian Zänker[1]
sebastianz541@gmail.com

Dirk Zeckzer[1]
zeckzer@informatik.uni-leipzig.de

[1]Image and Signal Processing Group, Leipzig University

[2]Bioinformatics, Leipzig University

ABSTRACT

Bioinformaticians judge the likelihood of the overall RNA secondary structure based on comparing its base pair probabilities. These probabilities can be calculated by various tools and are frequently displayed using dot plots for further analysis. However, most tools produce only static dot plot images which restricts possible interactions to the capabilities of the respective viewers (mostly PostScript-viewers). Moreover, this approach does not scale well with larger RNAs since most PostScript viewers are not designed to show a huge number of elements and have only legacy support for PostScript. Therefore, we developed iDotter, an interactive tool for analyzing RNA secondary structures. iDotter overcomes the previously described limitations providing multiple interaction mechanisms facilitating the interactive analysis of the displayed data. According to the biologists and bioinformaticians that regularly use out interactive dot plot viewer, iDotter is superior to all previous approaches with respect to facilitating dot plot based analysis of RNA secondary structures.

Keywords

Bioinformatics Visualization, Tabular Data, User Interfaces, Dot Plots

1 INTRODUCTION

In bioinformatics, one frequent task is judging the likelihood of the overall RNA secondary structure. This judgment is based on comparing the base pair probabilities of RNA secondary structures. Therefore, the probabilities for two nucleotides of an RNA sequence forming such base pairs are calculated. Dot plots are used for displaying probabilities or similarity measures between a row and a column of a matrix. Hence, dot plots are frequently used for RNA secondary analysis displaying the probability of a row and a column nucleotide forming a base pair.

Most currently available tools produce dot plots in postscript (ps) format (e.g., [6,8]). These ps-images are then viewed using suitable postscript viewers. However, postscript itself is no longer actively developed and was replaced by the portable document format (pdf). Moreover, the images are static and possible interactions are restricted to standard *viewing* interactions like zooming and panning the image. Further, the scalability of this approach is low as during zoom-in the nucleotide sequence that is displayed at the border of the image might not be visible any more.

Therefore, we developed iDotter, an interactive tool for analyzing RNA secondary structures that overcomes these limitations. Concretely, the contributions of this paper are:

- Sophisticated zooming and panning methods
- Presenting details for each dot on demand
- Highlighting of elements in the dot plot
- Recoloring of the dot plot
- Export of parts or the whole dot plot for further analysis
- A powerful API for using iDotter within analysis pipelines
- A sharing function for collaborative analyses

iDotter provides an interactive web interface that is implemented using the current state of the art web-programming languages HTML5, PHP, and JavaScript.

It proved superior to all previous approaches, and is already regularly used by biologists and bioinformaticians.

2 BACKGROUND AND RELATED WORK

Dot plots were introduced by Gibbs and Mcintyre [5]. Originally, dot plots were used to visualize alignments of two nucleotide sequences or proteins. A dot plot is a two dimensional matrix where the sequences 'A' and 'B' that are compared are visualized on the x- and y-axis, respectively. A dot in a cell means that Sequence 'A' is similar to Sequence 'B' at this nucleotide/amino acid (position). Both color and size of a dot represent the similarity of the sequences calculated using application dependent measurements. With the aid of dot plots, identifying highly similar regions between two sequences is easily possible. These regions are the diagonal lines in the matrix. An example for an interactive dot plot viewer for alignments was introduced by Sonnhammer and Durbin [12]. We, however, focus on RNA folding structures that can not be handled by their program. Moreover, we allow additional interactions like highlighting, semantic zoom, and export of (sub-)sequences not provided by their tool.

While the nucleotide sequence (RNA primary structure) is important for the analysis of RNA sequences, the folded structure of the RNA (RNA secondary structure) provides additional vital information. With the emergence of RNA folding tools [8–10], visualizing RNA secondary structure became more and more important to foster its analysis. Tools like Varna [4] or the NAVIEW algorithm [2] generate graph-based, node-link visualizations of RNA secondary structure showing *one* possible folding of the RNA, only. Further, dot plots were adapted to visualize the predicted base pair probabilities within a single RNA sequence. Thus, they support analyzing the changes of an RNA sequence between different species. Usually, the size of a dot describes the probability of a base pair between the corresponding nucleotides.

Static dot plots can be calculated with R using the R package R-CHIE [3]. The ViennaRNA package [8] can generate one dot plot in postscript format for each RNA secondary structure prediction (an example being shown in Figure 1a). Moreover, the ViennaRNA Web Services [6] provide the functionality of the ViennaRNA package without the necessity to compile the package. Therefore, it can be used platform independently. We use the ViennaRNA package for generating the initial dot plots, importing the data from them, and providing additional interactive visualizations for analyzing them. While the original dot plots of Gibbs and Mcintyre [5] for alignments show the same information in the upper and the lower triangle, dot plots generated by RNA folding software contain *two different* folding predictions, e.g., the energetically best solution and all possible base pair probabilities in the upper and lower triangles, respectively. As iDotter is based on the latter, it supports comparing *two different* folding probabilities predicted by the respective folding algorithms.

An alternative for visualizing RNA secondary predictions is the arc diagram introduced by Wattenberg [14] and later implemented as arc plot in R [7]. The RNA sequence is plotted as a linear sequence and an arc between two nucleotides describes a base pair while the color of an arc might encode the probability of the pair. Besides the fact, that this approach has limited scalability, the arcs produce a lot of clutter and it is hard to determine the corresponding base pair.

Arc diagrams and dot plots can be used for character sequence comparison (alignment) in general. For arc diagrams this was already introduced in the original paper [14]. Abdul-Rahman et al. [1] use dot plots to visualize text alignments between different documents.

3 PROBLEM, SOLUTION, AND METHOD

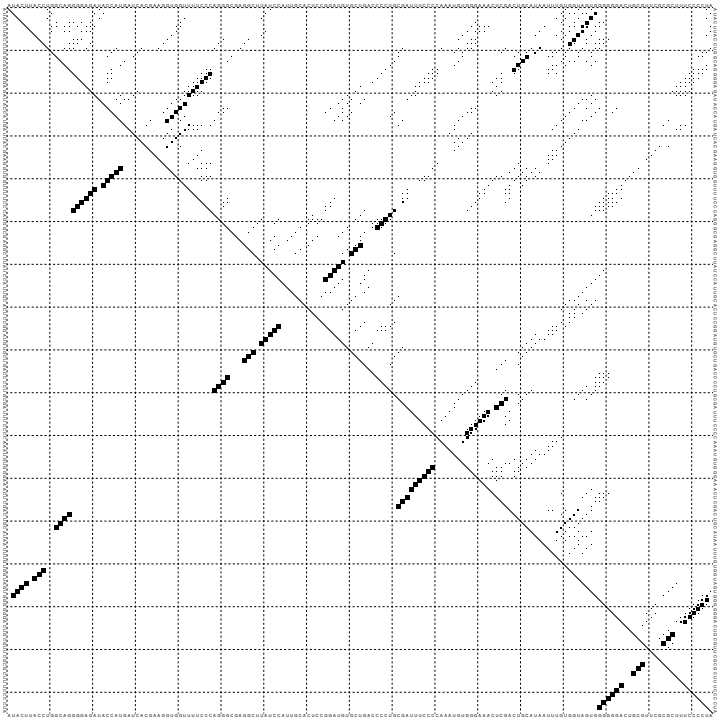
3.1 Problem Statement and Proposed Solution

Current state and Issues A dot plot fulfills the standard design goals taken from the information visualization literature [13]. Dot plots

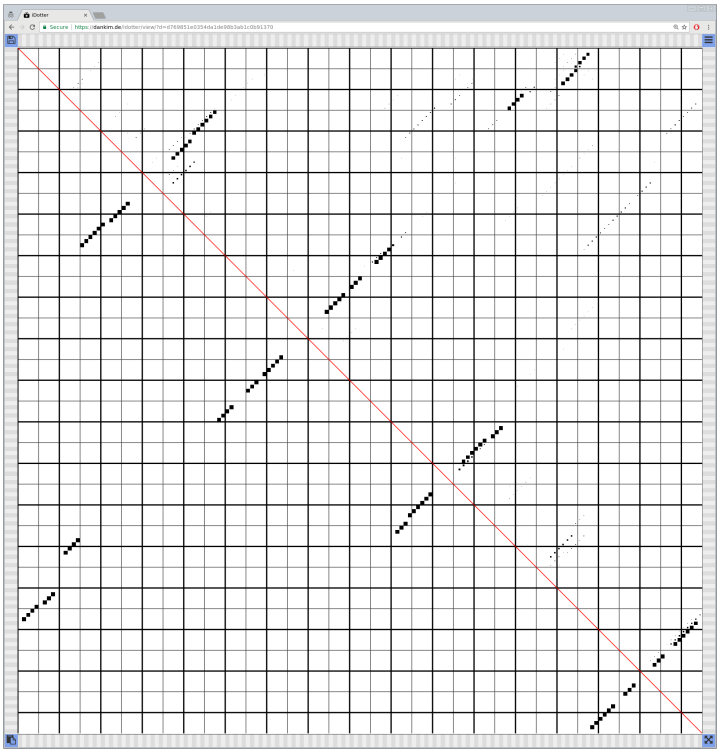
1. are flexible (can be used for different tasks and application areas)
2. are space efficient
3. provide a good overview of the data
4. ease the identification of pattern in the data
5. are fast to create

However, dot plots are static images without any interaction provided. Figure 1 shows a relatively small RNA having a length 165nt. As can be seen in the ps version (Figure 1a), the nucleotide names are no longer readable. Moreover, it is difficult to impossible to spot small base pair probabilities. Zooming into a part of the ps view is possible (Figure 1a). Then, all dots become larger and small base pair probabilities are more easily spotted. However, due to the limitations of the ps-viewers, the nucleotide sequence related to the zoomed-in area might no longer be visible.

Solution To overcome these limitations of existing dot plot generators, we propose iDotter, a fully interactive web-interface that supports experts in analyzing RNA secondary structure. iDotter is based on the dot plots generated by existing folding tools and provides

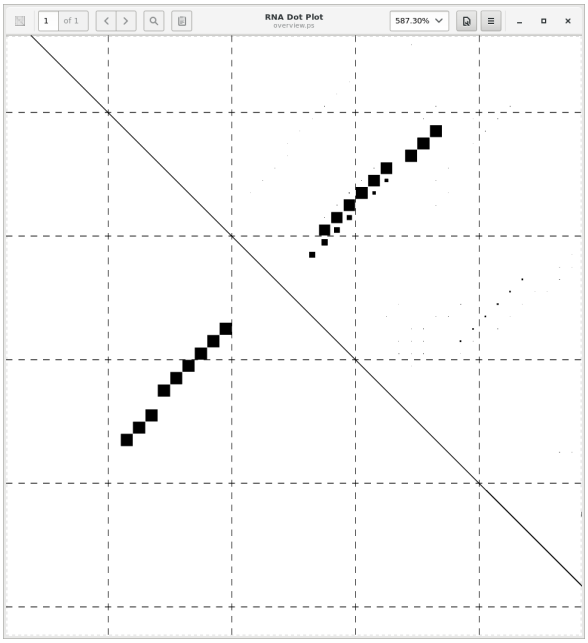


(a) Overview of the postscript dot plot generated by ViennaRNA [8]. The nucleotide sequence is always shown at the borders.

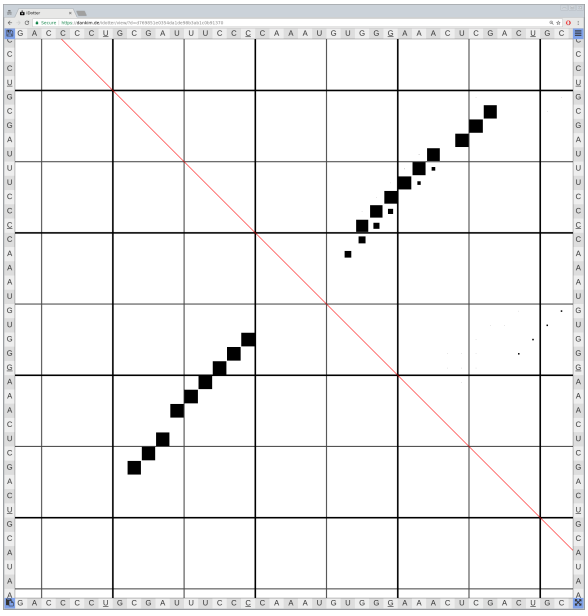


(b) Overview of the iDotter interface showing the same dot plot as Figure 1a. The nucleotide sequence is only shown at the borders, if the nucleotides are readable in the current zoom level.

Figure 1: Complete dot plot showing base pair probabilities of an RNA. Black squares are used for showing the possibility of two nucleotides forming a base pair. The probability for forming a base pair is encoded in the size of these squares. Large squares imply a high probability, while small squares imply a low probability. The diagonal is used as a landmark only. It divides the upper-right triangle showing the centroid probabilities from the lower-left triangle showing the probabilities according to the energetically best solution.



(a) PostScript-View, zoom-in of Figure 1a



(b) iDotter view, zoom-in of Figure 1b

Figure 2: Comparison of the dot plot interfaces after zooming into a sub-sequence. In the postscript view, the nucleotide sequence is no longer visible, while in iDotter it stays visible at all borders easing the analysis of sub-sequences.

additional interactions. After importing the data (Section 3.2.1), the dot plot is shown in the web browser (Section 3.2.2). Then, the expert can zoom in and out as well as pan the view (Section 3.2.3). Moreover, the expert can mark rectangular regions of dots in the dot plot as well as single columns and single rows (Section 3.2.3). Finally, the highlighted part of the dot plot or the complete dot plot can be exported in postscript-format (Section 3.2.4). A web-based API provides a connection with dot plot generating services (Section 3.2.5).

3.2 Methods

3.2.1 Data Import

After starting iDotter, the original ps-file generated by the ViennaRNA package [8] is transformed into a JSON file by iDotter, if the JSON file does not already exist. To do so, the RNA sequence, as well as the ubox and lbox containers are extracted from the ps-file and stored in a JSON array representing the box plot. Each ubox and lbox container comprises an x- and a y-coordinate designating the cell in the dot plot matrix, the size of the dot, and the color of the dot. The color information is optional. By transforming the input file into a generic JSON file iDotter can easily be extended for other input types by implementing a corresponding import routine.

3.2.2 Dot Plot View

The JSON file is imported by iDotter and the complete dot plot (zoom out) is displayed in the browser (Figure 1b). This follows the Shneiderman Mantra, presenting an “overview first” [11]. On each border, the nucleotide sequence is displayed. For convenience, the diagonal showing the same nucleotide on both the x- and the y-axis is shown in red. At the same time, this diagonal separates the upper from the lower triangle of the matrix. In the upper and lower triangle, either the same or two different base pair probabilities (encoded as size in the input file) are shown. The size of each dot is relatively encoded depending on the zoom level so that the expert can compare the probabilities easily on each zooming level. As default, we show the centroid probabilities in the upper and the energetically best solution probabilities in the lower triangle of the matrix, respectively. The probability of a dot is mapped to its size. The color can be used to represent, e.g., the conservation of the sequence between species. An adaptive background grid is displayed to enable an easy counting of the base pairs. Matching the zoom-level of the dot plot, the different grid levels can be shown or faded out. This view corresponds to the zoomed out standard dot plots, except that the nucleotide sequence is not shown, if the text becomes unreadable.

3.2.3 *Dot Plot Interaction*

The second step in Shneiderman's Mantra is "zoom and filter" [11]. The expert can use the semantic zoom to more closely analyze a sub-sequence (Figure 2b). The expert benefits from the sequence labels staying visible at all borders of the dot plot all the time. This is an improvement over the state of the art (Figure 2a) where the nucleotide sequence might disappear during zoom. This improves the scalability with respect to the size of the data that can be analyzed conveniently. Moreover, the individual nucleotides of the nucleotide sequence are only shown, if the zoom level allows displaying them in a readable manner. Otherwise, they are hidden (Figure 1b). The semantic zoom is triggered by mouse wheel motion. Moreover, the expert can pan the viewport by holding the left mouse button and moving the mouse.

Filtering is not provided for the original data. It would not be useful in this context. However, parts of the dot plot can be selected and this selection can then be exported (see below). This corresponds to a filtering step while its primary use is for reporting and collaborating.

The third step in Shneiderman's Mantra is "details on demand" [11]. While working with the dot plot, information about individual dots can be displayed on demand as a tool tip by mouse over. All available information is shown (Figure 3). Thus, the user can get exact information about the nucleotides (names and positions) involved in a base pair even though the respective names are not longer visible at the corners because they would be too small to read. Moreover, the values for the size (here: representing the base pair probability) and the color are shown.

Following the taxonomy of Yi et al. [15], selecting dots is provided by iDotter. The expert can mark a dot by left clicking on it (Figure 4a). Then, the selected dot is highlighted with the dot marker color. Moreover, the expert can select multiple dots by left clicking into the viewing area and dragging the mouse while holding the 'Shift' key pressed. This creates a rectangular region. Within this region, all columns and rows that contain dots are highlighted with the dot marker color (Figure 6). Additionally, the selected dots are highlighted in a different color (currently yellow) in both cases. Deselecting a region of dots is achieved by pressing the 'Ctrl' key while using the mouse. Besides marking dots, the expert can mark single columns by left clicking on them (Figure 4b). Then, the selected column is highlighted with the line marker color (see Figure 6). In the same way, the expert left clicks on a row to select it (Figure 4b). Both—marking dots as well as marking columns and rows—can be combined (Figure 5) to mark those parts of the dot plot that are of interest to the expert. Finally, the highlighting can be reset by pressing the 'Remove Marker' button in the settings view

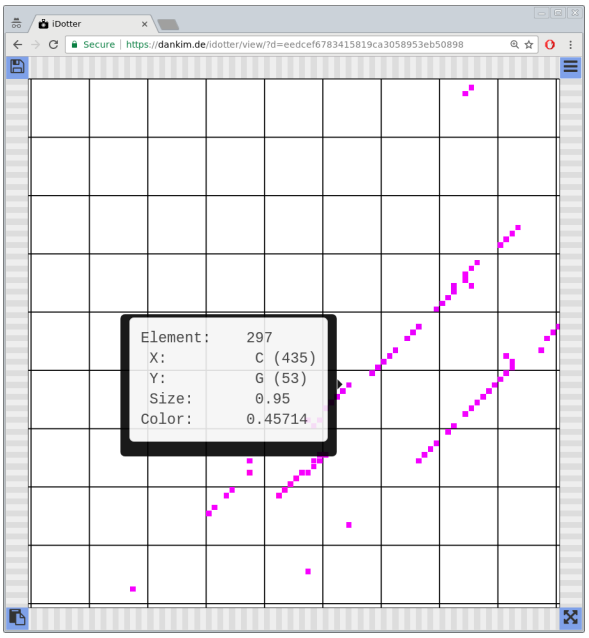


Figure 3: Each dot provides details on demand by mouse-over showing a tool tip: element ID, X shows the nucleotide of the column and its position, and Y shows the nucleotide of the row and its position. The (biological) attributes mapped onto 'Size' and 'Color' are application dependent.

which is invoked by left clicking on the 'three horizontal bars' icon in the upper right corner of the dot plot.

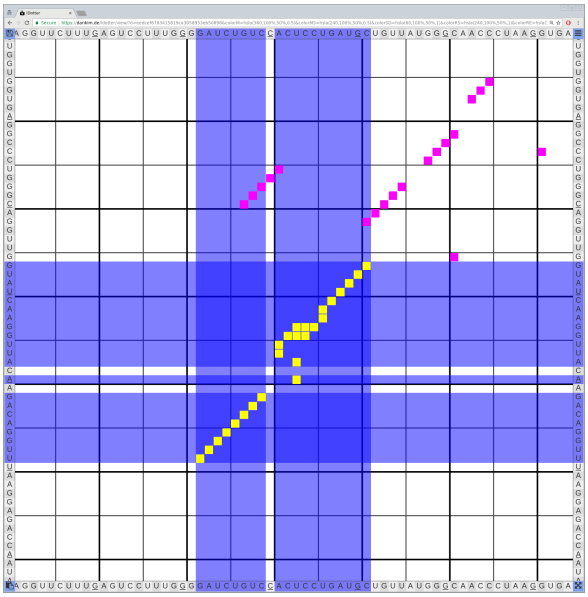
The color coding (corresponding to 'encode' [15]) can be adapted using the settings view. The two colors, the color gradient is generated from (Figure 6) can be changed. This directly influences the colors of the dots. Moreover, the colors of the marked dots (Dotmarker Color) and of the marked lines (Linemarker Color) can be chosen.

3.2.4 *Data Export*

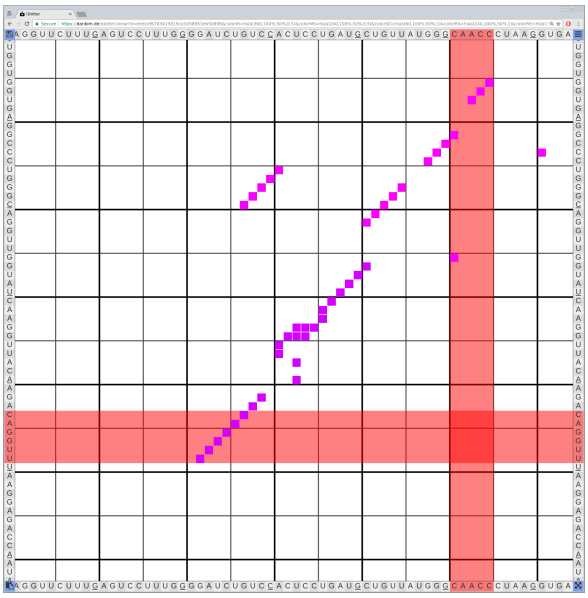
After working with the data, the expert can export the *highlighted parts* of the dot plot into a new ps-file for publication or other purposes by pressing the disc icon in the *upper left* corner and selecting the menu item 'export only selection'. Further, the expert can export the *complete* dot plot into a new ps-file by pressing the disc icon and selecting the menu item 'export all'. This is useful, if the API functionality of iDotter is used.

3.2.5 *API*

We designed a web-based API that provides a connection with dot plot generating services like the ViennaRNA Web Services [6]. This API supports direct import of ps-files into the view, pre-selecting highlighted regions, and exporting the highlighted regions for automatic workflows. The API is controlled by URL parameters. This type of control provides iDotter with additional possibilities for collaboration between users. The



(a) The 'mark dot' interaction allows selecting single dots by clicking on them. In this case, the selected dot is highlighted with the dot marker color (see Figure 6). Moreover, multiple dots can be selected by marking a rectangular region. (Clicking into the viewing area and dragging the mouse while holding the 'Shift' key pressed. For deselection, the 'Ctrl' key should be pressed instead.) All columns and rows that contain dots in the selected region are highlighted with the dot marker color (Figure 6). All selected dots are highlighted in a different color (currently yellow).



(b) The 'mark row' interaction allows selecting single rows by clicking on them. In this case, the selected row will be highlighted with the line marker color (see Figure 6). In the same way, columns can be selected.

Figure 4: Highlighting dots (a), and rows and columns (b).

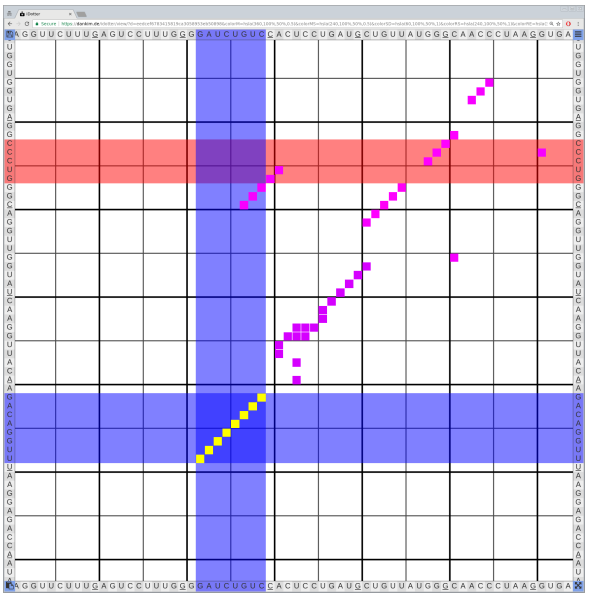


Figure 5: Marking dots and regions of dots (Figure 4a) and marking rows and columns (Figure 4b) can be combined.

expert can export her current settings, like zoom level, position, and color settings, and share these with her collaborators or save them for documentation purposes. The URL export is triggered by pressing the clipboard icon in the *lower left* corner. The URL contains all necessary parameters and is copied into the clipboard of the operating system. The expert can copy it afterwards to any application.

3.3 Interaction Properties

iDotter supports all interaction mechanisms required by the scientists for the analysis of dot plots. All useful steps of Shneiderman's Mantra [11] are supported. Moreover, the interactions 'select' and 'encode' proposed by Yi et al. [15] are supported.

4 EVALUATION

Dot plots are one of the default visualizations for the analysis of secondary RNA structure predictions. Therefore, the requirement was to enhance and extend this visualization for state of the art interaction techniques. In our case study our biological collaborator used iDotter for analyzing the evolution of so called long non coding RNAs (lncRNA). Since these RNAs are longer than 200nt, it is challenging to analyze the generated dot plots in ps-format due the lack of interactivity. Furthermore, it is hard to compare specific regions between different dot plots. For that reason, the expert used the interactivity features for selecting regions of interests. By exporting these regions with the API from all investigated RNA samples, it was possible to detect evolutionary changes between several species.

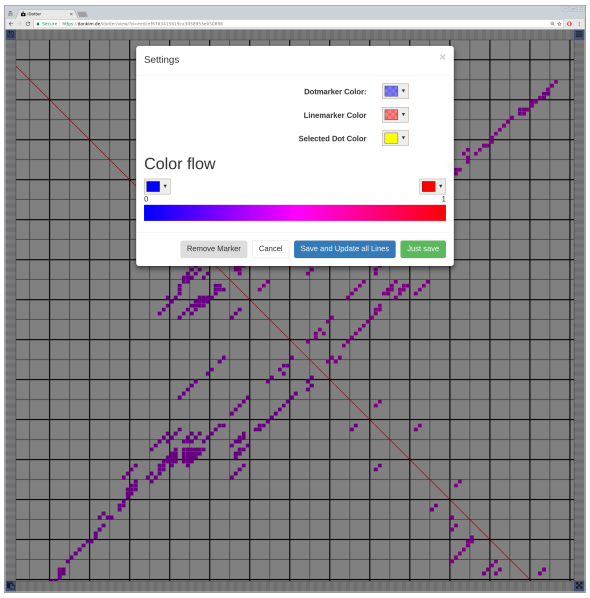


Figure 6: In contrast to the postscript visualization, iDotter provides choosing the color gradient. Additionally, choosing the highlighting colors for dots (Dotmarker Color, Figure 4a) and columns/rows (Linemarker Color, Figure 4b) is possible. Moreover, it is possible to reset highlighting in the dot plot by pressing the 'Remove Marker' button.

5 DISSEMINATION AND FUTURE WORK

The iDotter project is available under the GNU GPL v3 on `https://git.gurkware.de/biovis/idotter.git`. In the future, a close integration with the ViennaRNA Web Services [6] using the already existing API (Section 3.2.5) will be provided.

With respect to interaction, it is planned to add a small inset that provides an overview which part of the RNA sequence is currently zoomed-in. All interactions of the Shneiderman Mantra [11] and the interactions "select", "encode", "abstract/elaborate" (details on demand) and "filter" from the taxonomy proposed by Yi et al. [15] are already provided. Regarding the remaining three interactions from latter taxonomy, "explore" requires a closer integration with the folding tools using the already existing API. The "reconfigure" and "connect" interactions, however, are beyond the scope of the analysis task.

Further, it is planned to extend iDotter for analyzing data from other application areas similar to, e.g., alignments or text similarity [1]. Adapting iDotter for the different input file formats is straight forward. For this, a new data wrapper has to be created in iDotter that transforms the input data into a valid JSON input file.

6 CONCLUSION

We introduced iDotter, an interactive dot plot viewer for RNA secondary predictions. According to the biologists and bioinformaticians that regularly use out interactive dot plot viewer, iDotter is outperforms previous approaches with respect to facilitating dot plot based analysis of RNA secondary structures. By using the different interaction methods the experts were able to generate new insights and new hypotheses for their further work. The API enables the automated usage of iDotter in analysis pipelines or common RNA folding web services. The collaboration functionality allows the expert sharing her focus with her collaborators and documenting her insights.

7 ACKNOWLEDGMENTS

We thank all our colleagues from the BSV and Bioinformatics research groups for fruitful discussions on earlier versions of the project. This work was partially funded by the German Federal Ministry of Education and Research (BMBF) within the project Competence Center for Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig (BMBF grant 01IS14014B).

8 REFERENCES

[1] A. Abdul-Rahman, G. Roe, M. Olsen, C. Gladstone, R. Whaling, N. Cronk, R. Morrissey, and M. Chen. Constructive Visual Analytics for Text Similarity Detection. *Computer Graphics Forum*, 36(1):237–248, 2016. doi: 10.1111/cgf.12798

[2] 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.

[3] D. Charif and J. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H. Roman, and M. Vendruscolo, eds., *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pp. 207–232. Springer Verlag, New York, 2007.

[4] K. Darty, A. Denise, and Y. Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–5, 2009.

[5] A. J. Gibbs and G. A. Mcintyre. The Diagram, a Method for Comparing Sequences. *European Journal of Biochemistry*, 16(1):1–11, 1970. doi: 10.1111/j.1432-1033.1970.tb01046.x

[6] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker. The vienna RNA websuite. *Nucleic acids research*, 36(suppl 2):W70–W74, 2008.

[7] D. Lai, J. R. Proctor, J. Y. A. Zhu, and I. M. Meyer. R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic acids research*, p. gks241, 2012.

[8] R. Lorenz, S. H. Bernhart, C. H. Z. Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[9] N. R. Markham and M. Zuker. UNAFold. *Bioinformatics: Structure, Function and Applications*, pp. 3–31, 2008.

[10] J. S. Reuter and D. H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):129, 2010.

[11] B. Shneiderman. *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations*. VL '96. IEEE Computer Society, Washington, DC, USA, 1996.

[12] E. L. Sonnhammer and R. Durbin. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, 167(1):GC1–GC10, 1995.

[13] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 ed., 2004.

[14] M. Wattenberg. Arc diagrams: visualizing structure in strings. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pp. 110–116, 2002. doi: 10.1109/INFVIS.2002.1173155

[15] J. S. Yi, Y. ah Kang, J. Stasko, and J. Jacko. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, Nov 2007. doi: 10.1109/TVCG.2007.70515