Visualizing RNA Secondary Structure Base Pair Binding Probabilities using Nested Concave Hulls

Joris Sansen, Romain Bourqui, Patricia Thebault, Julien Allali, David Auber

Visualizing RNA Secondary Structure Base Pair Binding Probabilities using Nested Concave Hulls

Joris Sansen and Romain Bourqui and Patricia Thebault and Julien Allali and David Auber

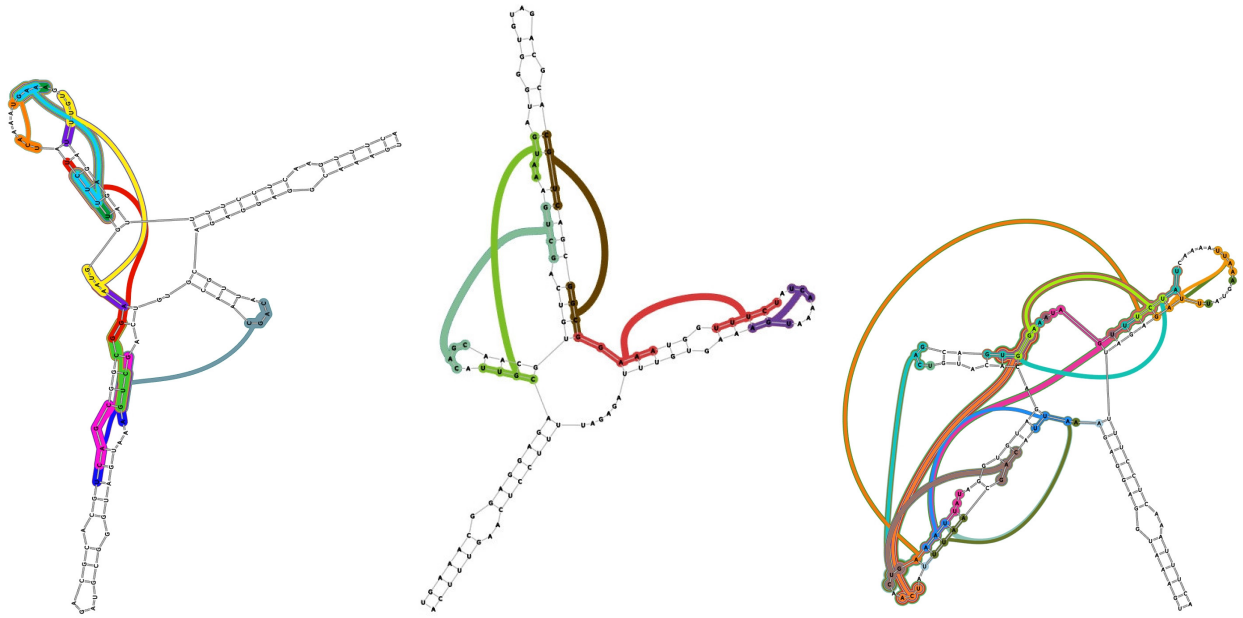

Fig. 1. MFE-structure of ncRNA whith potential pairs binding probabilities. From left to right: human, denisovan and ancestral chimp. The two generation parameters are the same for each sample: minimum number of base-pairs binding = 3 and number of levels =10 but the smallest probabilities have been filtered out for *Human* and *Ancestral chimp*.

Abstract—The challenge 1 of the BIOVIS 2015 design contest consists in designing an intuitive visual depiction of base pairs binding probabilities for secondary structure of ncRNA. Our representation depicts the potential nucleotide pairs binding using nested concave hulls over the computed MFE ncRNA secondary structure. Thus, it allows to identify regions with a high level of uncertainty in the MFE computation and the structures which seem to match to reality.

INTRODUCTION

Among all the discovered RNAs, a major part of them are not transformed into proteins, the non-conding RNAs (ncRNAs). Nevertheless, they are involved in many cellular processes and have a variety of catalytic properties. While primary structure of RNAs is a linear sequence of nucleotides, secondary structure refers to strong binding between pairs of nucleotides. This structure is often considered as characteristic of ncRNAs classes and their biological functions. Being able to predict a ncRNA secondary structure therefore helps to predict its functions. Actually, computational biologists have developed methods to compute the probabilities of base-pairing. These probabilities are then computed to constitute a Minimum Free Energy-structure (MFE) which represents the most likely base pairs binding. These two results are represented within a matrix called a dot-plot. In such matrix, half part represents all the potential base-pairs bindings while the other half contains the results that represent the MFE-structure. Indeed, dot-plots depict the MFE-structure and all potential bindings but it is almost impossible to understand the resulting RNA 2D structure and conclude about its biological functions. Another common technique lies in representing the MFE-structure as a graph. However, the MFE-structure is a prediction that might not depict the reality and that is why the strength of base pairs binding is depicted colors. Nevertheless, such representation does not depict the filtered out potential associations, *i.e.* the other potential pair bindings which might influence the resulting two-dimensional representation. The challenge 1 of 2015 BIOVIS Design Contest consists in designing a visual representation of the probabilities of base pairs binding. We choose to represent the rejected base pairs using nested concave hulls over the graph of the MFE-secondary structure of the ncRNA. In the following, we introduce our depiction, and detail its design. Then we describe the computational process and provide an analysis of the resulting picture. Finally we discuss the strengths and weaknesses of our technique and present conclusions.

- *Joris Sansen is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: jsansen@labri.fr.*
- *Romain Bourqui is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: bourqui@labri.fr.*
- *Patricia Thebault is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: thebault@labri.fr.*
- *Julien Allali is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: allali@labri.fr.*
- *David Auber is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: auber@labri.fr.*

1 Design

Our method (see Fig. 1 and Fig. 2) depicts the major unselected potential bases pairs sequences over the RNA secondary structure. Using the bases pairs computed for the Minimum Free Energy structure, we depict the ncRNA secondary structure as a graph. Potential base pairs binding rejected for this structure are then emphasized with the addition of nested concave hulls which put forward the subsequence. The more a sub-sequence is wrapped into hulls, the more its probability is important. Paired sub-sequences are linked to improve the pairs tracking and the bundling algorithm described in [2] associated to bezier curve bending is used to improve the design and visibility of the base pairs binding. This makes possible to ease the identification of paired sub-sequences by reducing link crossings and clutter. Finally, we ease the identification of paired subsequences providing a different color to each hulls and corresponding links. Thus, while displaying the MFE ncRNA secondary structure, we made possible to depict the potential base pairs binding and their relative values. Furthermore, our depiction eases the identification of stems and loops that could be transformed if different base pairs binding were selected, or on the contrary those who would remain stable.

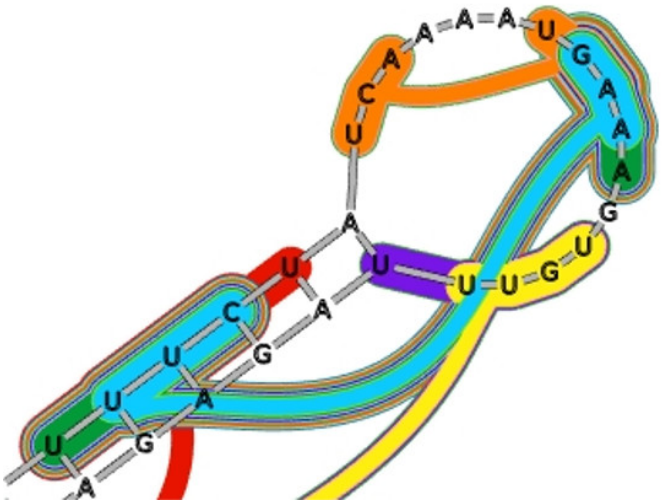

Fig. 2. Zoom on a loop of Human ncRNA secondary structure for design details.

2 Data processing

The first step to generate our depiction consists in generating the representation of the ncRNA secondary structure. Nucleotides are laid out using the primary sequence of nucleotides associated to the MFE-structure base pairs binding. This combination makes possible to draw the MFE-structure using a dedicated algorithm as the one described by Auber *et al.* in [1].

The second step is the search for the alternative pairs bindings. Our algorithm loop through the dot-plot and seek over the pairing probabilities for diagonal sets of minimum 3 matching bases missing in the computed MFE-structure. This algorithm needs two parameters: i) the minimum number of pairs to search in the set and, ii) the number of different levels of probability required. The second is used to define a threshold which determine the minimum base pair probability value authorized in the set. This threshold is simply defined as the range of probabilities ($max-min$) divided by the number of levels required. This process is repeated and the threshold progressively increased to pick smaller and smaller sets with a range of probability more and more restrained. Additionnally, we can tune the threshold in order to filter out of the selectable sets the ones with negligible matching probabilities.

The last step consists in the creation of nested concave hulls. The nucleotides of the selected sets are wrapped into hulls onto the MFE-structure and connected with a link to ease the identification of pairwise elements. The technique used to wrap the sets of nucleotides is the one described by Lambert *et al.* in [3]. This technique makes possible to order and nest hulls by size of sets. Thus, even sets with low pairing probabilities can be depicted and identified while putting forward the most likely sequences.

3 Results

To generate the Fig. 1, we used the previously described generating algorithm with a minimum number of base pairs of 3 and 10 different levels of probabilities. One can easily notice that some regions of the ncRNA present a few other potential base pairs binding indicating a high level of uncertainty, while other regions present just a few potential bindings. Indeed, the two upper stems seem quite uncertain since a few other bases pairing were possible. Moreover, three loops near these stems have an important amount of potential different bindings which could completely change the ncRNA secondary structure. For example, the left stem and its two inner loop could induce the appearance of a larger loop on its extremity by pairing the blue, pink or green sequences. On the contrary, the two bottom stems and loops does not have a lot of other potential nucleotides pairing possibilities. It seems to reveal that the prediction might be correct and correspond to reality.

The generation of this depiction was restrained by the tune of a few parameters which can improve or reduce the readability of the resulting visualization. Indeed, the result of the sequence detection process may vary depending on the two parameters of the generation algorithm: minimum number of base per sequence and value of the minimum probability threshold. Moreover, the bundling process induces variation in the readability of the structure and must be tuned too.

Conclusions

We have presented here an effective and intuitive depiction of a ncRNA secondary structure showing the most important potential base-pairs binding rejected during the MFE-structure computation. Our method puts forward the different possibilities while emphazing the most probable pair bindings and the influence they may have onto the predicted MFE-structure. The use of nested concave hulls produces a pleasant depiction of the potential base pairs binding which reveals the uncertainty in the predicted MFE-structure and the region of the structure that should remain stable with different base-pairs binding.

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. Auber, M. Delest, J.-P. Domenger, and S. Dulucq. Efficient drawing of rna secondary structure. *J. Graph Algorithms Appl.*, 10(2):329–351, 2006.

[2] A. Lambert, R. Bourqui, and D. Auber. Winding roads: Routing edges into bundles. *Computer Graphics Forum*, 29(3):853–862, 2010.

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