A multi-objective algorithm for interactive prediction of RNA complexes

Mandy Ibéné, Audrey Legendre, Eric Angel and Fariza Tahi

Université Paris-Saclay, Univ Evry, IBISC, 91020, Evry, France.

Abstract. RNAs are less stable than DNA, and their structure can easily change. This allows them to interact with other molecules present in their environment such as ions, proteins or other RNAs, to form complexes. The prediction of the structure of these complexes is an important issue for the determination of their biological roles. We are interested here by RNA complexes, composed of several RNAs interacting with each other, for which the prediction of the global secondary structure is a difficult task. We show how the use of available knowledge and probing data on the considered RNAs can help the prediction.

Keywords: RNA bioinformatics; RNA structure prediction; RNA complexes; secondary structure; SHAPE data; multi-objective optimization; graphs; cliques

1 Introduction

Several methods have been proposed in the literature for RNA secondary structure prediction as well as for RNA-RNA interaction prediction (interaction between two RNAs). However, very few have been proposed for the prediction of the structure of RNA complexes composed of more than two RNAs. RNAs can indeed bind and form complexes with catalytic functions. The prediction of the secondary structure of those complexes is an important issue to determine their functions. There are currently three available tools designed for this purpose: MultiRNAFold (Andronescu *et al.*, 2005), NUPACK (Zadeh *et al.*, 2011) and RCPred (Legendre *et al.*, 2019).

Biologists have sometimes some knowledge that can help the prediction. This can be sporadic information on the structure: pairings, particular motifs, etc. They can also be in possession of experimental data like SHAPE (Bindewald *et al.*, 2011), which gives probing information of the considered RNA. The integration of these user knowledge and probing data into the prediction of secondary structures exists for RNAs alone but to our knowledge not for RNA complexes.

We propose a new method and tool, called C-RCPred, to predict in an interactive way secondary structures of RNA complexes. C-RCPred is based on our previously developed method RCPred (Legendre *et al.*, 2019). Like its predecessor, it takes as input a set of predicted secondary structures per RNA sequence and a set of predicted RNA-RNA interactions per RNA pair. However, C-RCPred stands out through the integration of probing data and user knowledge as well. The tool aims to find the best combinations of these inputs, i.e., the set of RNA complexes structures optimizing simultaneously the free energy, the agreement with user constraints (user knowledge) and the agreement with probing data. For this purpose, C-RCPred solves a multi-objective version of an optimization problem, the *Maximum Clique Problem* (MCP).

2 Methods

We define a weighted graph G (V, E) formed by the input secondary structures and interactions. V is the set of vertices representing secondary structures and interactions. Each vertex $v \in V$ has three weights representing three objectives: (i) minimizing the free energy, (ii) maximizing the agreement with user constraints and (iii) maximizing the agreement with probing data. E is the set of edges representing compatibilities between vertices. An edge exists between two vertices if and only if these two vertices are compatible, i.e. when they represent structures or interactions that can belong to the same complex (meaning that a nucleotide is not involved in different pairings).

Here we are looking for the cliques optimizing the three objectives defined above. Since we are in a multi-objective context, we are looking for a set of cliques forming the best possible trade-off, the so-called Pareto set representing the optimal solutions of possible complexes formed by the RNAs given in input. To perform the search of those cliques in the graph, we use the Breakout Local Search (BLS) heuristic (Benlic and Hao, 2013) which we adapted for the multi-objective approach.

3 Results

To evaluate C-RCPred, we used a dataset composed of 90 RNA complexes, taken from the RNA STRAND database (Andronescu *et al*., 2008).

As mentioned above, tools for predicting (probable) secondary structures of each RNA and for predicting (probable) interactions between each couple of RNAs are used upstream of C-RCPred. In the two cases, we chose to use tools able to return several solutions. We used among others Biokop (Legendre *et al.*, 2018) for secondary structure prediction and IntaRNA (Busch *et al.*, 2008) for RNA-RNA prediction, as well as RNAsubopt (Lorenz *et al.*, 2011) for both.

It is very difficult nowadays to find SHAPE data for RNAs belonging to complexes of more than two RNAs. We therefore integrated in our framework another tool called Shaker (Mautner *et al*., 2019), allowing to generate artificial SHAPE data. Based on a machine learning approach, Shaker predicts SHAPE values on each nucleotide of a given RNA sequence.

We compared C-RCPred with the other existing tools for RNA complexes structure prediction, i.e. MultiRNAFold, NUPACK and RCPred (cited above). C-RCPred allows to return solutions that are on average the closest to the reference structure. As we can see on Figure 1, the average MCC of C-RCPred on all the dataset complexes is more than 0.8, while the one of RCPred is around 0.65, the one of MultiRNAfold is around 0.58 and the one of NUPACK is close to 0.5.

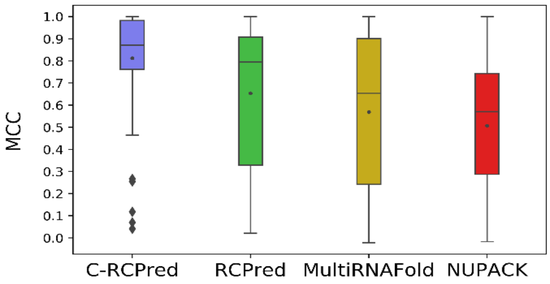

Fig. 1. Distribution of the maximum MCC results obtained on our benchmark dataset by C-RCPred compared to RCPred, NUPACK and MultiRNAFold. For each complex, C-RCPred and RCPred are run 100 times and the maximum MCC is kept, then the average is computed. The maximum MCC is also kept for NUPACK which also returns several solutions. For all the tools, at most 20 solutions are considered in each run. The MCC is calculated as: $MCC = (TP \times TN - FP \times FN)/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$ where *TP* is the number of true positive pairings, *TN* is the number of true negative pairings, *FP* is the number of false positive pairings, and *FN* is the number of false negative pairings.

CRCPred is implemented as an interactive tool. It is available as a web server on the EvryRNA platform (http://EvryRNA.ibisc.univ-evry.fr).

References

1. Andronescu, M., Zhang, Z. C., and Condon, A. (2005). Secondary structure prediction of interacting RNA molecules. *Journal of Molecular Biology*, 345(5), 987–1001.
2. Andronescu, M., Bereg, V., Hoos, H. H., and Condon, A. (2008). RNA STRAND: the RNA secondary structure and statistical analysis database. BMC *Bioinformatics*, 9(1), 340.
3. Benlic, U. and Hao, J.-K. (2013). Breakout local search for maximum clique problems. *Computers & Operations Research*, 40(1), 192–206.
4. Bindewald, E., Wendeler, M., Legiewicz, M., Bona, M. K., Wang, Y., Pritt, M. J., Le Grice, S. F., and Shapiro, B. A. (2011). Correlating SHAPE signatures with three-dimensional RNA structures. *RNA*, 17(9), 1688–1696.
5. Busch, A., Richter, A. S., and Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24), 2849–2856.
6. Legendre, A., Angel, E., and Tahi, F. (2018). Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. *BMC bioinformatics*, 19(1), 13.
7. Legendre, A., Angel, E., and Tahi, F. (2019). RCPred: RNA complex prediction as a constrained maximum weight clique problem. *BMC Bioinformatics*, 20(3), 128.
8. Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1), 1.
9. Mautner, S., Montaseri, S., Miladi, M., Raden, M., Costa, F., and Backofen, R. (2019). ShaKer: RNA SHAPE prediction using graph kernel. *Bioinformatics*, 35(14), i354–i359.
10. Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (2011). NUPACK: analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1), 170–173.