Accounting for RNA secondary structure allows improved classification and prediction of RNA base triples

D.S. Mikhailova

Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow region, Russia, darya.mikhailova@phystech.edu

V.A. Popova

Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow region, Russia, lerra.popova@mail.ru

E.F. Baulin

Institute of Mathematical Problems of Biology RAS - the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Moscow region, Russia, baulin@lpm.org.ru

M.A. Roytberg

Institute of Mathematical Problems of Biology RAS - the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Moscow region, Russia, mroytberg@lpm.org.ru

Analysis of RNA spatial structures and their motifs remains a challenge of structural bioinformatics. Biological functions of RNA molecules depend on their structures which are stabilized by secondary and tertiary interactions between nucleotides. RNA tertiary structure consists of relatively weak interactions formed on top of the secondary structure forming recurrent tertiary motifs. These motifs often play the role of functional units in RNA either by stabilizing the RNA structure or by recognizing other molecules [1].

RNA base triples are one of the most important RNA tertiary motifs [2] usually formed by canonical Watson-Crick base-pairs and noncanonical base-pairs among three nucleotides, two of which are usually involved in double helical regions. RNA base triples are often found in RNA pseudoknots, for instance forming A-minor motifs within H-knots. There were attempts to classify some types of tertiary motifs [3,4] but there is no common description of tertiary motifs regarding the secondary structure environment of their nucleotides.

In our previous work [5], we proposed a way to generalize widely used Nearest Neighbor Model (NNM, [6]) to apply it to arbitrary secondary structures including pseudoknotted structures. Here we propose a classification of RNA base triples regarding the secondary structure environment of their nucleotides. According to the classification, each nucleotide of a base triple is linked with the respective secondary structure element and each pair of nucleotides is linked with their relative position (from one stem/loop, from adjacent stem and loop or from distant elements).

The proposed classification allowed distinguishing particular motifs with distinct properties. For example, A-minors formed by bases within one internal loop were found to be a part of a kink-turn motif. We found base triples formed by nucleotides from three distant structural elements which is unexpected since it is energetically unfavorable. Furthermore, almost half of A-minors were found in clusters. We annotated all A-minor clusters (A-clusters) and found that the only known cluster A-patch is not the only possible architecture.

Finally, we used the classification of RNA base triples to apply machine learning for prediction of RNA base triples from RNA sequence and secondary structure. The trained machine learning model showed 93% precision and 95% recall on the balanced dataset. Features describing RNA base triple classes gave more than 50% significance according to the calculated feature importance values. It was shown that the model is able to predict base triples in RNA molecule types that were absent in training dataset with 83% precision and 58% recall.

The obtained results demonstrate the advantages of the proposed classification. Moreover the classification is not limited to RNA base triples and is applicable to other types of tertiary motifs.

1. Butcher, S. E., Pyle, A. M. (2011). The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Accounts of chemical research,* **44(12)**:1302-1311.

2. Devi G, Zhou Y, Zhong Z, Toh DF, Chen G. (2015) RNA triplexes: from structural principles to biological and biotech applications. *Wiley Interdisciplinary Reviews: RNA*, **6(1)**:111-28.

3. Almakarem AS, Petrov AI, Stombaugh J, Zirbel CL, Leontis NB. (2011) Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic acids research*. gkr810.

4. Klosterman PS, Hendrix DK, Tamura M, Holbrook SR, Brenner SE. (2004) Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucleic acids research*. **32(8)**:2342-52.

5. Baulin, E., Yacovlev, V., Khachko, D., Spirin, S., & Roytberg, M. (2016). URS DataBase: universe of RNA structures and their motifs. *Database*, **2016**.

6. Turner, D. H., & Mathews, D. H. (2009). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, **38**(suppl_1): D280-D282.