Reg. No: 2020/39/O/ST6/01488; Principal Investigator: prof. dr hab. inż. Marta Xymena Szachniuk

Predicting 3D RNA structures using Generative Adversarial Networks

Marta Szachniuk

Institute of Computing Science, Poznan University of Technology

The subject of the project is embedded in bioinformatics, a field that is a fusion of computing and life sciences. The research aims to create a system for sequence-based prediction of reliable three-dimensional RNA structures with high precision and resolution. We intend to achieve this goal by combining data science and artificial intelligence with the concept of structure generation *in silico*, by recombining structural elements.

Computer prediction of the 3D shape, which the molecules take while folding, is becoming an increasingly common approach to obtaining general knowledge of protein or RNA structures. It is a result of a rapid increase in the size of data stored in databases of amino acid and nucleotide sequences, which are most often not accompanied by information about the spatial molecular structure. Having the sequence, one can find structurally similar molecules and, based on the homology, draw conclusions about the functions of the analyzed molecule and outline its generalized shape (e.g., by predicting it with a computer algorithm). However, it is not enough to answer questions that require precise knowledge of the 3D structure. For example, when identifying molecular binding sites - a key issue explored when designing molecularly targeted drugs. That is why many laboratories around the world carry out experiments - including X-ray spectroscopy, nuclear magnetic resonance, cryoelectron microscopy - which allow for molecule imaging and determining their structures with atomic accuracy. Experimental data allow to improve prediction methods and to refine the generated models.

The first algorithms to predict 3D structures appeared in the 1970s and were dedicated to proteins. Their development intensified, i.a., due to the launch of the CASP (Critical Assessment of protein Structure Prediction) initiative in 1994 - the world championship in protein structure prediction. In the recently concluded CASP14, the AlphaFold 2 artificial intelligence algorithm won spectacularly in all categories of the tournament, and the CASP organizers announced the beginning of a new era in protein prediction (Callaway 2020). A similar achievement with RNA structures is a matter of time, but it requires access to reliable structural data, on which Artificial Intelligence algorithms could be trained. There is not much of this data yet: the Protein Data Bank has collected structures of 1 516 experimentally defined RNA molecules compared to 150 694 protein structures (data as of early December 2020). It is one of the reasons why progress in predicting 3D RNA structures is not keeping pace with progress in protein prediction.

In this project, we plan to develop the world's first 3D RNA structure prediction algorithm based on artificial intelligence. The algorithm will build a molecule by recombining three-dimensional structural elements, which will be modeled by a module applying Generative Adversarial Networks. The networks will be trained on a set of small 3D motifs obtained from both experimental structures and *de novo* simulators. Although *de novo* prediction methods, due to the complexity of computing, are not suitable for modeling large structures, they can generate accurate and reliable models of small molecules or their fragments. Structural element selection will be based on the consensus search algorithm (RNAtive). The overall fold will be predicted by applying existing prediction methods, including the RNAComposer system developed in our laboratory (Popenda et al. 2012), and using knowledge of the geometry of structural motifs that determine the course of the RNA chain, such as the n-way junctions. This knowledge will be obtained from structures collected in databases, including the RNA FRABASE (Popenda et al. 2008) and RNAloops (Wiedemann et al. 2020) developed in our lab, using analysis and data engineering methods.

The developed method will contribute to the generation of 3D RNA structures whose precision and resolution will correspond to those obtained by laboratory experimental techniques. It will make it possible to acquire - in a relatively short time - information, which currently requires time-consuming experiments. It will intensify the research on the relationship between biomolecule structure and function. It will speed up progress in biomedicine and bioengineering.

Callaway E (2020) ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures, *Nature*.

Popenda M, Blazewicz M, Szachniuk M, Adamiak RW (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures, *Nucleic Acids Res* 36(1):D386-D391.

Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW (2012) Automated 3D structure composition for large RNAs, *Nucleic Acids Res* 40:e112.

Wiedemann J, Kaczor J, Antczak M, Milostan M, Zok T, Szachniuk M (2020) RNAloops - a database of RNA multiloops, *submitted for publication*.