M2 INTENSHIP

Supervisors :

- **Massinissa HAMIDI** (IBISC, Univ. Évry Paris-Saclay) **Fariza TAHI** (IBISC, Univ. Évry Paris-Saclay) (contact: massinissa.hamidi@univ-evry.fr, fariza.tahi@univ-evry.fr)

Title: “Exploring Gradient-Based Metalearning for RNA 3D Structure Prediction”

Determining the 3D structure of ribonucleic acid (RNA) chains is essential to understanding their function and role in the various stages of living organisms and viruses. Due to the high cost of experimental methods (NMR, cristallography, etc.), computational methods could be very helpful. Although methods have been proposed in the literature for several years, the task remains open. For proteins, this problem has witnessed tremendous advances in recent years: DeepMind’s AlphaFold2 [1] made a giant leap in solving the 3D structure prediction problem for many types of single-chain protein structures using deep learning. Unfortunately, RNA still remains challenging [2]: unlike for proteins, (i) data of known 3D RNA structures are not available in large quantities; (ii) RNA are not stable and thus may have different 3D conformations; (iii) RNA sequences can vary from a few nucleotides to several tens of thousands of nucleotides.

It is suggested in the literature that the only way to address these challenges is for the quantity of RNA structures or sequence alignments to catch up with the amount of protein data that is currently accessible for models like AlphaFold [2]. We think the solution does not solely lie in the quantity of data but in finding suitable search biases and principled ways to incorporate domain knowledge into the learning process. The metalearning paradigm can provide answers to these challenges. This paradigm aims to improve a learning model’s generalization capabilities by leveraging prior knowledge from a family of tasks and accumulating past experience in a meaningful way [5, 6, 3]. Gradient-based metalearning approaches are examples of this paradigm, where the goal is to learn a model that knows how to adapt to new tasks or domains using limited quantities of data [5, 13, 4]. These approaches made tremendous breakthroughs in many applications where adaptation to new tasks required only a few learning examples.

Recently, only a very few studies in the literature [7, 8, 9, 10] have started to address the use of metalearning in computational biology. Works like [7] and [8] have been limited to the prediction of non-coding RNA using metalearning, leaving the structural level unexplored. In this internship, we want to investigate metalearning for the problem of 3D structure prediction of RNA chains. In particular, the ability to leverage the 3D conformations coming from multiple known species and know how to adapt rapidly to new ones under few available samples.

Furthermore, we want to investigate how prior knowledge can be leveraged to guide the adaptation process further [6]. For example, we can exploit the knowledge base of prominent

RNA structural patterns provided in the CaRNAval dataset maintained by the LISN laboratory at UPSaclay [11]. Concretely, the prior knowledge in the form of RNA structural patterns can, for example, be used to devise better parameterizations for the optimization landscapes induced by the initial learning problem.

We will use the RNANet [12] database, developed by our research team, that integrates various information on RNA, including sequences, families (e.g., MSA multiple sequence alignments), secondary structures, 3D structures, etc.

The steps of the internship will, first, consist of the study of the state-of-the-art on RNA 3D structure prediction and gradient-based metalearning approaches. Second, frame the problem of RNA 3D prediction in the metalearning setting and build a first metalearning-based architecture for RNA 3D structure prediction. Third, study the prominent RNA structural patterns included in the CaRNAval knowledge base and propose a way to leverage such structural patterns to devise better parameterizations for the learning process. Finally, benchmark with RNANet dataset and possibly other datasets.

Position statement on the chosen research topic

From a methodological point of view, we want to develop new metalearning approaches that can effectively deal with limited data in predicting the 3D structure of RNA and incorporate prior knowledge into the learning process.

From a bioinformatic perspective, we would like to propose an efficient tool for predicting RNA 3D structures, a tool that could be used in a personalized medicine project we are involved in, the IHU Prometheus project (2024-2034) on Sepsis, where RNAs can be potential biomarkers and/or therapeutic targets.

The developed tool will be available through our EvryRNA bioinformatics platform (http://evryrna.ibisc.univ-evry.fr), a platform providing the scientific community with several tools developed under the team for the analysis and prediction of non-coding RNAs.

The internship can lead to a Ph.D. thesis to further deepen the use of metalearning for the prediction of RNA 3D structures.

References

[1] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

[2] Schneider, Bohdan, et al. "When will RNA get its AlphaFold moment?." Nucleic Acids Research 51.18 (2023): 9522-9532.

[3] Hospedales, Timothy, et al. "Meta-learning in neural networks: A survey." IEEE transactions on pattern analysis and machine intelligence 44.9 (2021): 5149-5169.

[4] Nichol, Alex, and John Schulman. "Reptile: a scalable metalearning algorithm." arXiv preprint arXiv:1803.02999 2.3 (2018): 4.

[5] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." International conference on machine learning. PMLR, 2017.

[6] Hamidi, Massinissa. Metalearning guided by domain knowledge. Diss. Université Paris-Nord-Paris XIII, 2022.

[7] Li, Zhongshen, et al. "CoraL: interpretable contrastive meta-learning for the prediction of cancer-associated ncRNA-encoded small peptides." Briefings in Bioinformatics 24.6 (2023): bbad352.

[8] Bonidia, Robson P., et al. "BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria." Briefings in Bioinformatics 23.4 (2022): bbac218.

[9] Wu, Xue, et al. "Meta-learning shows great potential in plant disease recognition under few available samples." The Plant Journal (2023).

[10] Rodrigues, Vânia, and Sérgio Deusdado. "Metalearning approach for leukemia informative genes prioritization." Journal of Integrative Bioinformatics 17.1 (2020): 20190069.

[11] Reinharz, Vladimir, et al. "Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families." Nucleic acids research 46.8 (2018): 3841-3851.

[12] Becquey, Louis, Eric Angel, and Fariza Tahi. "RNANet: an automatically built dual-source dataset integrating homologous sequences and RNA structures." Bioinformatics 37.9 (2021): 1218-1224.

[13] Raghu, Aniruddh, et al. "Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML." International Conference on Learning Representations. 2019.