Université de Paris

M2 internship (6 months)

Optimizing a physical RNA force-field via Machine Learning

The importance of the study of RNA molecules has been highlighted by the recent pandemic, with the SARS-CoV-2 virus featuring an RNA-based genome and a replication mechanism controlled by non-coding RNA. The ncRNAs of viral genomes are often more conserved because they perform fundamental functions related to the ability of the genome to be translated, making them interesting targets for drug design. However, the flexible and dynamic nature of RNAs, where multiple structures can be adopted by the same sequence, represents one of the main challenges associated with targeting them with small molecules. Computational modeling using dedicated force-fields can provide a coherent view of the molecule, which can follow this dynamical behavior and include the effect of the environment.
Physical modeling consists in describing a system as a set of particles interacting via some energetic potentials that are responsible for the forces acting among the particles. The set of interactions and the equations describing them in mathematical terms are commonly referred to as a "force field" which can in turn be used to carry out simulations such as molecular dynamics or Monte Carlo. In our group, we develop a coarse-grained RNA model and force-field (HiRE-RNA). Among the currently existing models it is the one retaining the highest resolution and it has been designed with energy potentials chosen empirically to reproduce the main structural behavior of RNAs. The model force-field was originally optimized using standard techniques (genetic algorithm maximizing the energy difference between a native structure and decoys) and limited experimental data. The recent advent of machine learning (ML) provides an array of tools with the potential to dramatically improve both the functional form and parameter optimization.

The main goal of this project is the optimization of our RNA model through ML to obtain a cutting-edge RNA force field to facilitate building functional three-dimensional structures for RNA molecules. We will employ machine learning to optimize the model exploiting extensively the structural data available in databanks and the sparse thermodynamic and dynamic data available from experiments. This approach will allow our model to give much more accurate and reliable structural predictions and to be deployed on systems of more complex architectures than currently possible.
Our aim here is to anchor our force field model deep into the corresponding physics by adapting recent and promising Symbolic Regression algorithms to our data format and selecting the possible improvements in the functional form of the force field uncovered by this technique, based on sound physical principles.
The M2 internship will be the first step of a larger project where we propose to first use the existing functional form of the force field and train its 100+ coefficients and then to then build upon the ML pipeline developed in the first step to learn additional terms of the force field. The first step will serve

two purposes: i) improving the existing, physics-based force field and ii) establish an accuracy baseline for further improvement.

The work will be divided in 4 phases:
1. Set up the global optimization scheme coupling Pytorch to the coarse-grained force-field code.
2. Generate an appropriate training set of RNA structures.
3. Run the optimization on the training set.
4. Run the optimized force-fields on a set of benchmark systems.

We will test the performance of the new force field by running enhanced sampling simulations such as Replica Exchange Molecular Dynamics and Basin Hopping Monte Carlo to generate the view of the energy landscape and of the thermodynamic behavior of a few molecules with which we have extensive experience simulating with both the coarse-grained model in its previous version and with atomistic simulations, and for which there is abundant experimental evidence to compare our results, as directly studied by our experimental team.

The work will be co-supervised by Pr. Samuela Pasquali, main developer of the HiRE-RNA model, and by Dr. Frederic Lechenault, expert in machine learning, at Ecole Normal Superieure in Paris. Computing facilities at LPENS at BFA laboratories, as well access to the French national computing cluster, will be available to the student.

The internship will last 6 months, and the student will receive an allowance of approximately 500 euros/month. It will have to take place starting in the first quarter of 2022.

If interested, please send a motivation letter and a CV to:
samuela.pasquali@u-paris.fr

Samuela Pasquali
Physics professor

Université de Paris, Faculté des Sciences Pharmaceutiques et Biologiques
Laboratoire Cibles Thérapeutiques et Conception de Médicaments, CNRS UMR 8038
and
Laboratoire Biologie Fonctionnelle et Adaptative, CNRS UMR 8251
Equipe Modélisation Computationnelle des Interactions Protéine-Ligand
Bâtiment Lâmarck A, bureau 418
35 rue Hélène Brion, 75205, Paris Cedex 13
tel: 0157278279