PhD in Computational Structural Biology **and Bioinformatics**

Subject: Modeling and scoring of protein-RNA complexes

Supervised by: Pr. Christine Froidevaux, Dr. Jérôme Azé (LRI Bioinformatics group CNRS and AMIB group INRIA, University Paris Sud) and Dr. Julie Bernauer (LIX and AMIB group INRIA)

Location: LRI, building 490, Univ. Paris Sud, Orsay

Contact: Jerome.Aze@lri.fr

Abstract:
The function of RNA molecules depends on their interaction with one or many partners. Upon interaction, RNA molecules often undergo large conformation changes. Understanding how these molecules interact with proteins would allow better targeting for therapeutic studies. In RNA molecules, large conformation changes may occur in the protein-RNA docking step. This PhD project contains two parts: - improving and developing approaches for modeling RNA-protein interactions and RNA conformation changes, - conceiving new machine learning methods able to deal with different representations of an object. These would allow to work on different conformations of a molecule -RNA or protein or complex-, with a limited set of experimentally known conformations.

Context:
The function of RNA molecules depends on their interaction with one or many partners. Upon interaction, RNA molecules often undergo large conformation changes. Understanding how these molecules interact with proteins would allow better targeting for therapeutic studies. The CAPRI (Critical Assessment of PRediction of Interactions) challenge[1] has shown that classical docking procedures largely fail when large conformation change occurs and when RNA is involved [1]. This is especially true for RNA molecules, whose large-scale dynamics remain often unknown. Modeling RNA conformational changes is made hard by the inherent flexible nature of their structure but also by the electrostatics involved. These are hard to model and often lead to computationally expensive simulations. Even if for small RNA molecules, molecular dynamics can be used, such simulations are hard to extend to larger molecules and protein-RNA complexes.

For many diseases, such as cancer and HIV, microRNA molecules play a very important role regulating gene expression by guiding the RISC. Some miRNAs have been shown to suppress tumors [2] and are thus ideal candidates for the development of therapeutic agents. Even if various computational techniques have been developed to predict miRNA targets [3] , none of these consider the structural aspects of the interactions between components of the RISC and miRNA.

Objectives (scientific objectives):
This PhD project contains two objectives. The first objective is related to computational structural biology: developing new approaches for modeling RNA interactions and flexibility . The second is related to bioinformatics: conceiving new machine learning techniques to learn models from very imperfect data using multiple representations of an object. One of the specificity of this project is that it implies multi-scale modeling. We would start from a set of coarse-grained conformations, select the most promising subset of conformations that should be refined and generate new conformations from this subset at higher resolution. In this new approach, machine learning and the conformational generation have to be closely linked.

To get better models for RNA and its interactions, the plan is to improve upon the existing techniques by

1 http://capri.ebi.ac.uk

developing new models and algorithms for the study of RNA binding to proteins and RNA conformational changes. The LRI bioinformatics group and collaborators has experience in protein-protein docking scoring functions and RNA structure evaluation that could be extended to protein-nucleic acid complexes [4-7].
The combination of Voronoi models at a coarse-grained level and powerful machine learning techniques allows the accurate scoring of protein-protein complexes. Our actual machine learning approach uses several machine learning approaches (evolutionary algorithm, decision trees, decision rules, …). By adapting these approaches to protein-RNA complexes, we would have a fast and efficient technique for scoring large protein-RNA complexes where conformational changes are involved.
Working with RNA molecules instead of proteins introduce many major differences in the machine learning approaches. RNA conformations are often smaller than protein conformations. This has an impact on the values of the descriptors used. Also because of the difference in size between RNAs and proteins, it is often more difficult to generate conformations closed to the biological solution -near native solutions- at the modeling stage. The machine learning algorithms therefore need to take into account all theses specificities to be able to learn good predictive models from data that are not very close to the real solution.
Another aspect of the machine learning will concern the ability to interact with the modeling stage in order to gradually focus on the most promising conformations in a multi-scale setting. Multi-scale modeling is emerging in the community [8-9]. The proposed interaction between modeling and prediction using machine learning is new. It implies to be able to develop algorithms that have the ability to select different models and representations depending on the estimated distance to the solution. The closest we are to the solution, the more precise the generated conformations need to be. One way to increase the quality of the conformation may be simply to change representation from coarse-grained to atomic level. This will largely increase the computing time required to generate conformations and it must only be used where the conformations are closed enough to the solution.

Work program:
The first step will be to be able to efficiently model RNA: i.e. generate models that are closed enough to the known biological solution in a relatively computationally inexpensive manner.
The second step will deal with the improvement of the existing machine learning approaches in order to evaluate the efficiency of the modeling stage: can we score efficiently models we generated?
The final step will be to develop new machine learning approaches that may need to interact with the modeling stage in a multi-scale fashion. This would allow to gradually focus on the best model created by the modeling approach.

Prerequisite:
Good skills in computational biology and RNA modeling are required for this PhD. Skills in Machine Learning will also be appreciated.

References
[1] Lensink, M. F. & Wodak, S. J. (2010). *Docking and scoring protein interactions: CAPRI 2009.*, Proteins 78 : 3073-3084.
[2] Hammond, S. M. (2006). *MicroRNA therapeutics: a new niche for antisense nucleic acids.*, Trends Mol Med 12 : 99-101.
[3] Sethupathy, P.; Megraw, M. & Hatzigeorgiou, A. G. (2006). *A guide through present computational approaches for the identification of mammalian microRNA targets.*, Nat Methods 3 : 881-886.
[4] Bernauer, J.; Azé, J.; Janin, J. & Poupon, A. (2007). *A new protein-protein docking scoring function based on interface residue properties.*, Bioinformatics 23 : 555-562.
[5] Bernauer, J.; Bahadur, R. P.; Rodier, F.; Janin, J. & Poupon, A. (2008). *DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.*, Bioinformatics 24 : 652-658.
[6] Bernauer, J.; Huang, X.; Sim, A. Y. L. & Levitt, M. (2011). *Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation.*, RNA 17 : 1066-1075.
[7] Bourquard, T.; Bernauer, J.; Azé, J. & Poupon, A. (2011). *A collaborative filtering approach for protein-*

protein docking scoring functions., PLoS One 6 : e18541.
[8] Ayton, G.S. and Voth G.A. (2010). Multiscale computer simulation of the immature HIV-1 virion. Biophys Journal, 99(9) : 2757-2765
[9] Bernauer, J.; Flores, S.; Huang, X.; Shin, S. & Zhou, R. (2011). Multi-scale modelling of biosystems: From Molecular to Mesoscale - *Session Introduction.*, Pac Symp Biocomput : 177-180.