Campagne 2025 Contrats Doctoraux Instituts/Initiatives

Projet de recherche doctorale SCAI:
Mathematical and molecular modeling for single-stranded nucleic acids design

Scientific context and objectives

Single-stranded nucleic acids (ssNAs) play an important role for cells thriving, since they are involved in structural, functional and regulatory functions and in protein synthesis. In addition, synthetic ssNAs can be exploited as biosensors or as therapeutic and diagnostic tools [1]. Indeed, thanks to the specific conformations they can adopt, they can recognize a plethora of molecular targets, spanning from small molecules to whole cells

SsNAs function highly depends on their secondary (i.e. their base pairing pattern) and tertiary (i.e. their 3D organization) structures. Therefore, when designing a ssNA able to mimic a natural one or to bind to a given molecular target, it is primordial to consider the folding the ssNA should have. In addition, ssNAs are characterized by a high level of structural flexibility, which is relevant for the exercise of their function since it allows the ssNAs to increase the interaction with their molecular target.

Because of the relevance of ssNAs' secondary and tertiary structures, much effort has been paid to try to predict and compare [2] the 2D and 3D folding of this kind of molecules. However, so far, none of the available tools, including AlphaFold3 (AF3), takes into account the intrinsic ssNAs' flexibility, since just a single or a few conformations are provided as output.

In addition, ssNAs rational design requires not only the prediction of the most probable conformations for a given sequence (i.e. solving the folding problem), but also to retrieve all the sequences having the desired conformation among their most probable ones (i.e. solving the inverse folding problem). Indeed, it often happens that a ssNA with a required function and, therefore, a searched structure, is already known. Several algorithms are available at this scope, based for example on thermodynamic models or on evolutionary models (see for instance [3] for a recent review on the subject). Nevertheless, they usually provide a limited number of solutions and they do not consider the ssNAs flexibility.

Within the Num4Lyme team-project funded by the Institute des Sciences du Calcul et des Données (SU), aimed to define a new reliable diagnostic of Lyme disease, we worked on developing ssNAs capable of recognizing a surface protein (CspZ) of the bacteria causing the Lyme disease. This has been initially done experimentally by a SELEX procedure, a succession of rounds of selection and amplification of a random initial DNA or RNA library incubated with the target, allowing to retrieve a few ssNAs, called aptamers, capable to strongly recognize the molecular target. After 12 rounds of SELEX we could retrieve the sequences of a subset of a sample (∼100 000 sequences/round) of the ssNAs binding to CspZ and characterize them in terms of abundance and dissociation constant [4]. These data allowed us to tune a **Restricted Boltzmann Machine (RBM) model** which is **able to generate new sequences** which are rather different from those experimentally determined, but close in terms of log-likelihood and cosine similarity of sparse vectors derived from a k-mers representation of the sequences and of dense vectors directly derived from the average of the hidden variables generated by the RBM model. The closest sequences to the best performing experimentally selected aptamers have been synthesized and tested and preliminary results show that some of them are able to recognize CspZ (**article in preparation** [5]). However, this model takes only indirectly into account the ssNAs structures. This might lead to the generation of sequences with a different folding as compared to the required one. Therefore, **within this PhD thesis project we want to address the inverse folding problem, by directly focusing on the two levels of structures of ssNAs**. **The first step will consist in the generation of ssNA sequences satisfying a desired secondary structure**. This will be done by tuning a chosen **machine learning model**, among a Variational Autoencoder, a Graph Neural Network, or a Recursive Neural Network, which, in contrast to RBM, are able to deal with input values (ssNA sequences and secondary structures) of different lengths, and have established their usefulness in biological applications [6]. The training dataset will be built from public databases (e.g. RFAM or RNA STRAND 2.0), allowing

to dispose of a large number of ssNA sequences together with their experimental secondary structures. The possibility of having sequence constraints (using a IUPAC nomenclature) will be also included, in order to allow to block positions that are relevant for the ssNAs function.

The sequences generated by the model with the desired folding will be clustered in terms of sequence diversity and/or their characteristics (e.g. the log-likelihood), and **the most representative sequences will be 3D modeled**, at first **with AF3** to provide an initial static 3D structure **and**, subsequently, **by enhanced sampling molecular dynamics techniques**, such as Gaussian Accelerated Molecular Dynamics. This will allow to extensively **explore the conformational space of ssNAs**, thus taking into account their intrinsic flexibility, which might impact their function. Additionally, if a precise molecular target is defined, as in the case of the Num4Lyme team-project, the complexes ssNAs-molecular target will be predicted and modeled in the same way.

Coherence with the institute/initiative

The project is coherent with SCAI and the PostGenAI@Paris project: i) it is **interdisciplinary**, involving both mathematical and molecular modeling; ii) it focuses on one of the main areas of interest of PostGenAI@Paris, namely the **health area**, since it will allow to generate ssNAs sequences and structures exploitable for therepeutical and diagnostic applications; iii) it is dependent on **large-scale simulations**, and last, but not least, iv) it makes uses of **artificial intelligence** to reach its goal, through the mathematical model that will be developed and the use of AF3.

PhD supervisors

Ghislaine GAYRAUD (the PI) is a Professor at the Université de Technologie de Compiègne (UTC) and is affiliated to the Laboratoire de Mathématiques Appliquées de Compiègne (LMAC). She develops her work on both theoretical and applied statistics including stochastic modeling for dependent data.

Miraine DÁVILA FELIPE is a *Maître de Conférences* at UTC, affiliated to the LMAC. She works on stochastic models and their applications to biology, in particular to molecular evolution and epidemics spreading. Therefore, G.G. and M.D.F. will supervise the mathematical modeling aspects of the project.

Irene MAFFUCCI is a *Maître de Conférences* at UTC and is affiliated to the CNRS UMR7025 Génie Enzymatique et Cellulaire laboratory. **She will defend her HDR in June 2025**. She has a strong experience developing computational procedures to design bio-inspired molecules for protein binding. This includes investigating the conformational preferences of the designed molecules by means of enhanced sampling molecular dynamics simulations. I.M. will supervise the molecular modeling part of the project.

PhD candidate profile

The PhD candidate should have a bioinformatics / biostatistics background. Knowledge of machine learning models is strongly recommended. Programming skills are required.

References

[1] Kumar Kulabhusan, P., Hussain, B., Yüce, M.: Current Perspectives on Aptamers as Diagnostic Tools and Therapeutic Agents. Pharmaceutics **12**(7), 646 (2020). doi:10.3390/pharmaceutics12070646

[2] Binet, T., Avalle, B., **Dávila-Felipe**, M., **Maffucci**, I.: AptaMat: a matrix-based algorithm to compare single-stranded oligonucleotides secondary structures. Bioinformatics **39**, 2022–0504490414 (2023). doi:10.1093/bioinformatics/btac752

[3] Wirecki, T.K., Lach, G., Badepally, N.G., Moafinejad, S.N., Jaryani, F., Klaudel, G., Nec, K., Baulin, E.F., Bujnicki, J.M.: DesiRNA: structure-based design of RNA sequences with a replica exchange Monte Carlo approach. Nucleic Acids Research **53**(2) (2025). doi:10.1093/nar/gkae1306

[4] Guérin, M., Vandevenne, M., Matagne, A., Aucher, W., Verdon, J., Paoli, E., Ducrotoy, J., Octave, S., Avalle, B., **Maffucci**, I., Padiolleau-Lefèvre, S.: Selection and characterization of DNA aptamers targeting the surface borrelial protein CspZ with high-throughput cross-over SELEX (2025). doi:10.1101/2025.01.13.632687

[5] Issouani, E.-M., Guerin, M., Padiolleau-Lefèvre, S., **Maffucci**, I., **Dávila-Felipe**, M., **Gayraud**, G.: DNA aptamer design through RBM and k-mer representations. In preparation (2025)

[6] Zeng, X., Wang, F., Luo, Y., Kang, S.-g., Tang, J., Lightstone, F.C., Fang, E.F., Cornell, W., Nussinov, R., Cheng, F.: Deep generative molecular design reshapes drug discovery. Cell Reports Medicine **3**(12), 100794 (2022). doi:10.1016/j.xcrm.2022.100794