Implementation of statistical tools for the structural investigation of Intrinsically Disordered Proteins

Project supervisors:
Juan CORTÉS
LAAS-CNRS, Toulouse
juan.cortes@laas.fr
Pierre NEUVIAL
IMT, Toulouse
pierre.neuvial@math.univ-toulouse.fr
Javier GONZÁLEZ-DELGADO
IMT, LAAS-CNRS, Toulouse
javier.gonzalez-delgado@math.univ-toulouse.fr

Summary:
Understanding the relationships between protein sequence, structure and function is a fundamental problem in biology, and is essential for numerous applications in biomedicine and biotechnology. Up to now, this problem has mostly been addressed for proteins having a well-defined three-dimensional form, which represent around 60% of the human proteome. The remaining proportion corresponds to proteins that do not present a stable structure, the so-called Intrinsically Disordered Proteins (IDPs), but which are nevertheless fully functional and perform essential tasks. Due to their flexibility, the structural investigation of IDPs is extremely challenging, and requires a tight combination of experimental and computational methods. Reliable and accurate methods to characterize and compare conformational ensemble models is a key component in this context. For instance, they are required to quantify structural effects of mutations, or to assess the quality of molecular dynamics simulations.

Due to the intrinsic probabilistic nature of IDPs, novel methods to analyze their structural properties have to be conceived from a statistical point of view. Researchers at LAAS-CNRS and at IMT are currently collaborating on this topic. More precisely, they are developing a statistical comparison tool, based on recent advances in **Optimal Transport** Theory, as well as **unsupervised learning** methods for data supported on non-euclidean spaces. A first prototype has been successfully implemented in R. However, the required computations being expensive, a more efficient and user-friendly implementation is needed in order to make this tool available to the Structural Biology community. The aim of this project is to properly implement such methods in **Python**, creating a ready-to-use tool adapted to the format and structure of protein data which is used in this community. We expect that this project will lead to a scientific publication.

Expected skills:
Strong background in statistics is mandatory, as well as good programming skills (Python, R).
Background in structural biology is not necessary, but it would be a plus.

Possibility of funding:
The student will be provided with a monthly stipend of around 550 euros during up to six months.

Applications:
Please send an email containing your CV to the three project supervisors, indicating in the subject "Candidate IDPs stats project".