M2 INTERNSHIP IN STRUCTURAL BIOINFORMATICS AND DEEP LEARNING

Supervisors:
Dr. GALOCHKINA Tatiana (Assistant Professor) and Dr. GHEERAERT Aria (Postdoctoral fellow)
Professional Address: UMRS 1134 - BIGR, 8 Rue Maria Helena Vieira da Silva, 75014 Paris, France

Laboratory:
Institution: UMR_S 1134 Inserm et Université Paris Cité - Biologie intégrée du globule rouge

Description of the team (people with their position, main topics, website):

The Master student will work in the structural bioinformatics group DSIMB (2 Assistant Professors, 2 Associate Professors, 1 Researcher, 1 Research Engineer, 1 PostDoc, 5 PhD students). Our team has extensive expertise in methodological developments for structural bioinformatics problems such as: i) modelling and analysis of protein dynamics; ii) protein structure and dynamics prediction using machine learning approaches; iii) development of databases and specific tools for a range of distinct protein families (among others: membrane proteins, camelid antibodies and small disulfide bridge proteins). DSIMB team is internationally recognized for the development of Protein Blocks (PBs), the most widely used structural alphabet in the world applied to analysis and prediction of local protein conformations. DSIMB has also participated in the international CASP 11 and 13 competitions and finished in top 10 for the difficult target category.

The Master student will work with Dr. Tatiana Galochkina in the framework of the SugarPred project funded by ANR. Dr. Galochkina is a specialist in molecular modelling of complex systems and in deep learning applied to the problem of structural bioinformatics. The student will be co-supervised by Dr. Aria Gheeraert, a PostDoc recruited for the same project.

website: https://www.dsimb.inserm.fr/ , https://sites.google.com/view/tatiana-galochkina/
Tel: +33 (0) 1 81 72 43 30 email: tatiana.galochkina@u-paris.fr

Title: Prediction of protein-carbohydrate binding sites using deep learning methods

Keywords: structural bioinformatics; protein-carbohydrate interactions; deep learning

Project summary:

Protein-carbohydrate (PC) interactions play a key role in various biological processes. However, as compared to protein-protein or drug-protein interactions, PC interactions remain much less studied due to the difficulties related to their experimental investigation. In this context, we recovered all the available experimental structures of protein-carbohydrate complexes and curated a new protein-carbohydrate binding site database containing thorough annotation and clustering of the observed types of different binding sites. The student will participate in the second stage of SugarPred and will implement deep learning models aiming at predicting potential carbohydrate binding sites on a protein surface and their classification according to the proposed annotation. The tools developed during the internship will be made publicly available through scientific publications and open source code. They will also be used throughout the next steps of the SugarPred project.

Description of the project:

Carbohydrates (sugars) along with proteins, nucleic acids and lipids, are one of the four building blocks of life and are abundantly present on the surface of any living cell. Protein-carbohydrate interactions mediate a variety of essential biological processes such as cell adhesion, signalling, migration of tumor cells, interactions between immune cells and microorganisms and recognition between host and pathogens. Most of these mechanisms are enabled by non-covalent protein-carbohydrate (PC) interactions observed for a variety of protein classes including antibodies, lectins, sugar transporters, and enzymes. Despite their crucial role for a range of biological processes, carbohydrate structures and their functional mechanisms remain poorly described as compared to proteins, nucleic acids or lipids. Deciphering the mechanism of sugar recognition is crucial for the development of specific carbohydrate-derived therapeutics as well as protein agents aimed at sugar binding.

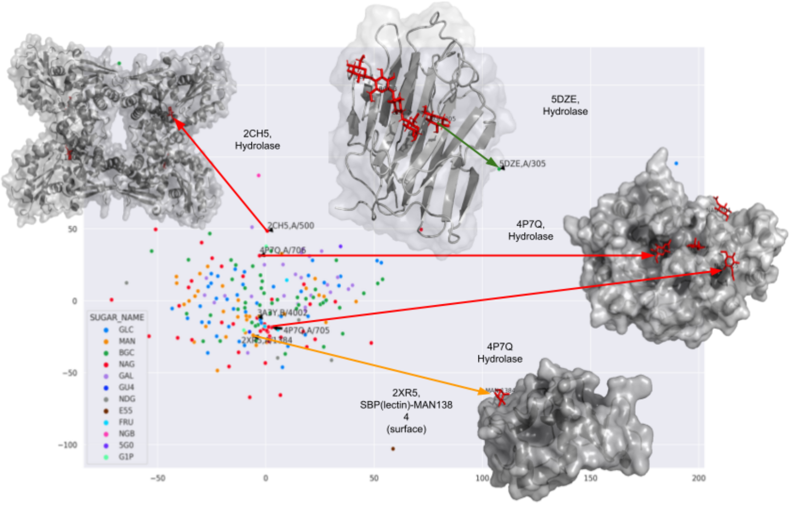

***Fig. 1**. Diversity of sugar binding sites projected in two-dimensional space using multi-dimensional scaling.*

In the framework of ANR SugarPred project, we have developed a curated database, annotated and clustered different binding sites (BS) observed in experimental structures. The Master student will use the obtained annotations to build machine learning models for prediction of protein-carbohydrate BS from protein structure by implementing different approaches. First, the student will focus on extracting the available information on the structural properties of PC complexes with particular attention paid to the flexibility of the considered protein fragments. Each PC binding site will be characterised by physicochemical properties, geometry, interaction types and evolutionary information on the protein sequence fragments. The student will apply unsupervised and semi-supervised machine learning techniques for the extraction of the most explanatory descriptors and their combinations for efficient BS prediction using classical architectures. These results will then be used to develop a more powerful model for sugar binding site prediction using graph representation of a protein. The student will train several state-of-the-art graph neural network models including additional descriptors for graph node encoding in accordance with the feature importance evaluated in the first part of the project.

The internship aims to answer the following questions: what structural descriptors of a protein fragments indicate a susceptibility towards carbohydrate binding? Is there a significant difference between these structural features for different carbohydrates? Can the diversity of carbohydrate binding sites be reduced to several general classes? Are there groups of carbohydrates that have a tendency to bind to the same types of binding site? Thus, the project will contribute to the significant exploration of the protein-carbohydrate interactions and provide a solid background for the further development of the carbohydrate binding site prediction tools. Due to the prevalence of protein-carbohydrate interactions in biological processes, the developed tools have potential for a variety of biological and biomedical applications including drug design of carbohydrate-based compounds for treatment of viral and bacterial diseases.

Five recent publications of the team:

1. SWORD2: Hierarchical analysis of protein 3D structures (2022) Cretin G, Galochkina T, Vander Meersche Y, de Brevern A G, Postic G, Gelly J-C., *Nucleic Acids Re*s, gkac370
2. MEDUSA: Prediction of Protein Flexibility from Sequence (2021) Y Vander Meersche, G Cretin, AG de Brevern, JC Gelly, T Galochkina, *Journal of Molecular Biology* 433 (11), 166882
3. PYTHIA: Deep Learning Approach for Local Protein Conformation Prediction (2021) G Cretin, T Galochkina, AG de Brevern, JC Gelly, *International journal of molecular sciences* 22 (16), 8831
4. New insights into GluT1 mechanics during glucose transfer (2019) T Galochkina, MNF Chong, L Challali, S Abbar, C Etchebest, *Scientific reports* 9 (1), 1-14
5. Conformational dynamics of the single lipopolysaccharide O-antigen in solution (2016) T Galochkina, D Zlenko, A Nesterenko, I Kovalenko, M Strakhovskaya, A Rubin, *ChemPhysChem* 17 (18), 2839-53

Additional questions:

Does this project constitute the first steps of a PhD thesis that will be supported by a PhD fellowship? For the moment, there is no funding for a PhD thesis after the internship but an application for funding will be submitted.

Have you had the opportunity to supervise a master student before? Yes, you can contact my fellow master students from M1BI and M2BI (https://sites.google.com/view/tatiana-galochkina/teaching). The following students made important contribution to the projects during their internship:
Yann Vander Meersche - 1 paper after M1 internship, PhD fellowship
Yani Ren - 1 paper in preparation, 1 submitted
Thomas Bailly - 1 paper and 1 database in preparation

Do you have any special accommodation or fellowship for foreign students? No