Master " Sciences, Technology, Health

Mention M2 -BI

In Silico Drug Design

INTERNSHIP SUBJECT version 3-4pages
(after validation)

Academic Year 2020/2021

Name of the Laboratory Manager: Claudia Huber Administrative Staff (Computational Pharmacy)**Laboratory:** Department of pharmaceutical sciences
Precise address of the Laboratory: Klingelbergstrasse 50 , 4056 Basel, Schweiz
Name of the host team leader and name of the team: Dr. Lill , Computational Pharmacy

---

Name of Internship Leader(s): Manuel Sellner
Phone number : +41612076598
E-Mail : manuel.sellner@unibas.ch
Specialty of the internship: Research ~~/ private, national /~~ international
Is this topic a first step towards a thesis: ~~yes~~/no

Indicate by a few key words, the scientific orientation of the subject:
deep learning, computational pharmacology, compound-protein interaction

---

Title: *Implementation of a multi objective neural network for the prediction of protein-ligand interactions*

Project Summary (15 lines):

Machine learning approaches provide a set of tools that can improve discovery and decision making for well-specified questions with abundant, high-quality data for instance compound-protein interaction understanding can greatly facilitate drug development. (https://doi.org/10.1093/bioinformatics/btaa524) . To overcome the current limitations of structure and ligand-based methods; structure free models are developed based on simplified molecular-input line-entry systems and primary sequences of proteins.

In bioinformatics, machine learning-based methods that predict compound-protein interaction (CPIs) play an important role in the virtual screening for drug discovery. End-to-end representation learning for discrete symbolic data (e.g. words in natural language processing) using deep neural networks has demonstrated excellent performance on various difficult problems. For the CPI problem, data are provided as discrete symbolic data, i.e. compounds are represented as graphs where the vertices are atoms, the edges are chemical bonds, and proteins are sequences in which the characters are amino acids. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences doi: 10.1093/bioinformatics/bty535 .

The project includes the implementation and improvement of a published neural network. (https://doi.org/10.1016/j.cels.2020.03.002) with the goal of predicting qualitative protein-ligand interactions as well as binding affinities. The neural network will then be integrated into a platform for the prediction of adverse effects of small molecules.

Master " Sciences, Technology, Health

MORE DETAILED PRESENTATION OF THE SUBJECT (2 TO 3 PAGES)
THE PURPOSE OF THIS MORE DETAILED VERSION OF YOU M2 SUBJECT IS TO PREPARE YOUR INTERNSHIP (BY DISCUSSING WITH YOUR FUTURE SUPERVISORS AND BY INTEGRATING A BASIC BIBLIOGRAPHY ON YOUR SUBJECT)

Scientific context

Indicate the existing state of the art on your subject globally (bibliography) then **what already exists on the subject** at the laboratory and team level. **Specify what already exists**, the **limits** that explain your internship topic and what will be your work during the internship

CPI's, compound-protein interaction understanding can greatly facilitate drug development. Identifying compound–protein interaction (CPI) is a crucial task in drug discovery and chemogenomic studies, and proteins without three-dimensional structure account for a large part of potential biological targets, which requires developing methods using only protein sequence information to predict CPI. However, sequence-based CPI models may face some specific pitfalls, including using inappropriate datasets, hidden ligand bias and splitting datasets inappropriately, resulting an overestimation of their prediction performance. (doi: 10.1093/bioinformatics/btaa524)

Conventional methods, such as structure-based virtual screening and ligand-based virtual screening, have been studied for decades and gained great success in drug discovery, they can often be time-consuming and biased because of the atomic resolution structures of protein-ligand complexes and are limited in numbers of solved structures in addition, some cases are not suitable to apply conventional screening methods. (https://doi.org/10.1038/s41573-019-0024-5)
To overcome the current limitations of structure and ligand-based methods; structure free models are developed based on simplified molecular-input line-entry systems and primary sequences of proteins.

For this master thesis project, we aimed to develop a neural network for the prediction of both protein-ligand non-covalent interactions between the atoms of a compound and the residues of its protein partner and predict binding affinities like Ki, Kd, IC50, based on molecular graph and primary sequence of the protein.(https://doi.org/10.1016/j.cels.2020.03.002), thus I will have to learn state-of-the-art deep learning techniques, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Graph Convolution Networks – Graph Wrap Unit and Neural Attentions, through this master thesis to accomplish this goal.

This project is part of Manuel Sellner's thesis "Development of an in-silico platform for early detection of side effects", which is already implemented.

Questions and objectives:

What will be your questions and objectives (adapted for 5 to 6 months of research)

- Learn state-of-the-art deep learning techniques
- Learn how to use Pytorch and scikit-learn
- Reproduce and improve a published neural network
- **Improvements will most likely be on the side of the model architecture**
- Integrate the neural network in a platform for the prediction of adverse effects of small molecules
- Communicates and present works

Anne.Camproux@univ-paris-diderot.fr

Master " Sciences, Technology, Health

Data, tools and methods:

Data are the core foundation of deep learning models, and in a way what a model learns mainly depends on the datasets it is fed, and inappropriate datasets make the model easily deviate from the goal.
Firstly, data will be downloaded from MONN's GitHub page, to have an idea of the kind of data that can be used, then if times allows it, a new dataset will be created following the dataset construction protocol available on the GitHub page. The protocol enables to generate the dataset file containing : input information of protein-ligands and their affinity values, the non-covalent interactions between proteins and ligands, extract the interaction information from PLIP output and the pocket positions from PDBbind and sequence alignment between the sequence from the complex structures and the UniProt sequences.

Tools (Python):

- Pytorch : https://pytorch.org
- DeepLearning with Pytorch eli stevens, luca antiga, and thomas viehmann https://pytorch.org/assets/deep-learning/Deep-Learning-with-PyTorch.pdf
- Sickit learn https://scikit-learn.org/stable/

Bibliography (5 publications min)

- MONN A Multi-Objective Neural Network for Predicting Compound-Protein Interactions and Affinities https://doi.org/10.1016/j.cels.2020.03.002
- Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning https://doi.org/10.1038/nbt.3300
- Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences: https://doi.org/10.1093/bioinformatics/bty535
- TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments https://doi.org/10.1093/bioinformatics/btaa524
- DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks (https://doi.org/10.1093/bioinformatics/btz111)
- A review on compound-protein interaction prediction methods: Data, format, representation and model: https://doi.org/10.1016/j.csbj.2021.03.004

Schedule on 5.5 months

Month 1: Literature search and training
Month 2 & 3: Implementation and validation of MONN model
Month 4 & 5: Improvement of model
21st of June : master thesis presentation

Anne.Camproux@univ-paris-diderot.fr

Master " Sciences, Technology, Health

Return by e-mail: anne-claude.camproux@univ-paris-diderot.fr,