Klavs Jermakovs Frida Matiyevskaya

Estimating bias in PPI datasets

Task 1

Link to repository: https://github.com/TheKursta/PPI-bias

Task 2

Background

Proteins are the main workhorses of the cell. They carry out billions of reactions by interacting with other molecules, and more oftenly with other proteins. Protein-protein interactions (PPIs) are essential in understanding the key cellular mechanisms and could reveal new ways to tackle diseases. Yet the main principles behind those interactions are still not fully understood. Above all, many scientists struggle to explain what features are important in PPIs and how these interactions are regulated. The field of PPIs therefore is important to investigate and opens a huge scope for drug development design. Having machine learning models at a hand biologists can speed up their research by identifying the potential drug targets through computer simulations rather than long and expensive experiments.

There is always a possibility that trained models used in research might be biased to a particular set of proteins, not allowing to expand on perhaps much more effective protein targets, therefore resulting in suboptimal research end result or even worse - failure of reaching the goal. These models are trained on pre-existing experimentally validated data that might contain hidden biases that machine learning models when trained upon would happily exploit to maximize their accuracy.

Research goals

Identifying hidden bias patterns and raising awareness about issues in the PPI datasets used in these ML models. Providing a hypothetical solution to benefit future PPI studies through ML methods, ultimately speeding up R&D time of drugs and cell machinery research as a whole.

Research success criteria

Identification of at least one bias in pre-existing PPI dataset and illustrating its impact on machine learning models would render exploratory data-analysis as a success. More distant success criteria is implementation of knowledge and proof that de-biased data leads to bias free fitted ML models.

Inventory of resources

As with any inter-discipline data-analysis project there has to be not just technical expertise on how to work with the data but also domain knowledge and understanding of the data. Luckily our project team has both - student Frida who is an expert on biological systems and underlying biological cellular mechanisms; student Klavs who covers technical aspects of data pre-processing, pattern discovery and fitting ML models. Additionally, the team has numerous connections around the university, both

biologists to help form hypotheses and data science researchers to share the knowledge on how to validate hypotheses through data. Hardware involves both cloud (Google Collab) and on-site (rather powerful PC) solutions to store data and perform data-analysis.

Requirements, assumptions, and constraints.

The data availability is high and it is open-source, available from previous research articles provided that it is properly cited. The work has to be properly formatted with understandable graphs, if there will be any, clear presentation of the problem and results of data-analysis. The major constraint in this project is time. Project execution time-frame is rather strict, the bias analysis and result presentation has to be completed by December 17.

Risks and contingencies

Many subsets of the dataset can lead to getting lost in it, therefore, data will be properly named,stored and documentation will be written along. Another risk is to lose sight of milestones to be accomplished for achieving success, therefore, a calendar with reminders is created. Technical data wrangling troubles may arise when implementing hypotheses, supervisors might give hints to overcome such issues if any.

Terminology

PPIs - protein-protein interactions - interaction between two proteins through physical touch that allows the propagation of a signal.
Interacting pair - two proteins that form an interaction.
Non-interacting pair - two proteins that do not form an interaction.
Protein-hub - protein that has many impossible interactions with many other proteins.
Amino Acids - organic compounds that are represented by specific letters.
Protein sequence - sequence of amino acids.

Costs and benefits

The implementations of gained knowledge in PPI ML models would result in more effective research of disease mechanisms, cell functions and drug-target discovery. Decrease of just 1 year in 3 year of 1.8$ billion drug target discovery would result in multi-million gains for the drug-industry. The cost of research is insignificant and will be executed through free-student labour.

Data-mining goals

Bias in the data has to be explained and presented with means of graphs, simple Naive models or other means.
Data has to be formatted and ready for fitting ML models, and results have to be presented with unambiguous performance scores that are explained. Fit ML models on biased data and assess impact of bias presented or speculated.

Data-mining success criteria

The Naive models fit on bias present better than random results. Data containing protein sequences is converted into suitable format for ML models for fitting. The bias is found in data and explained both - in terms of data science and terms of biology.

TASK 3

Data Understanding

Outline data requirements

PPI ML models are trained on data that contains pairs of proteins and indications whenever they interact or not. Additionally, data should contain protein sequences that will be used as a feature for ML models. Data has to be composed of experimentally validated PPI interactions, and non-interactions in the dataset have to be constructed in a manner that makes biological sense. Ideally, data has to contain more than 1000 entries of interacting and non-interacting proteins in order to even have an attempt to fit an ML model.

Verify data availability

As PPI research in the field of bioinformatics with ML models is well-established and growing, numerous repositories of already assembled PPI datasets are available for several cellular organisms through research articles. In case more specific dataset has to be constructed or more meta-data (features) besides protein sequence is needed, then it will be mined from the extensive UniProtKB protein database, which is the source from which researchers most often create PPI datasets.

Define selection criteria

Data will be selected from already existing published research articles regarding the topic of ML in PPI which are available in NCBI, PubMed, Elsevier, BMC Bioinformatics. The quality of data is also somewhat guaranteed to be validated due to the peer-review process of those published articles. Using such sources provides baseline results of what other researchers have achieved with the data. Already one dataset that is investigated in this project is presented in [1]. In case of bias findings there is possibility to investigate these findings in other suitable PPI datasets [2] and [3].

Describing data

Dataset that will be researched is sourced from [4]. Dataset contains two non-redundant parts in format of tab separated text files (.tsv). One part contains the ID and sequence of a protein, the other part contains pairs of IDs and an indicator of whenever the pair interacts. The dataset contains all the information needed to explore possible bias in the data.

Exploring data

By examining data closer following information was found:

- There are 10’370 proteins with unique IDs and sequences.
- Lengths of sequences are between 24 and 33’432 amino-acids with mean of 613.
- There are 73’108 protein pairs, each with specified interaction.
- Interaction or non-interaction of pair is described with 1’s or 0’s.
- There are ~36’500 interacting and ~36’500 non-interacting protein pairs.

The data seems well-balanced in terms of interacting and non-interacting protein pairs to fit ML models. The sequences are presented in String format with 20-unique letters that each specify specific amino-acid with unique chemical properties. The IDs correspond to UniProtKB protein identifiers, so if needed additional information about proteins can be gathered.

Verifying data quality

The data was verified for NAs and the good news it does not contain such values. Not so bad news is that non-interacting protein pair entries in the dataset have been created synthetically instead of using experimental data, which would be a better choice. However, there is little if any experimental data about non-interacting pairs available so there is not much that can be done about it. The high variance of sequence length might throw off embedding NLP models, if that is the case then very long/short proteins will be removed without ruining the dataset.

TASK 4

Task roadmap

- Data gathering, suitability analysis. The PPI data has to be downloaded from the source and source has to be cited. PPI data has to be validated so that it contains rows of protein pairs and indicator of whenever they interact, also data about protein sequences has to be present along. (Frida <5hours)

- Reading in the data. Data might be presented in numerous formats (.csv, .fasta). It has to be read in Python dataframe with standardized formatting. For the dataset/datasets that will be analyzed such a read-in script has to be constructed. (Frida <2hours)

- Analysis of bias. Here numerous ideas might be explored. First one is to search for whenever there are more interacting or non-interacting proteins that result in some bias. Another is to investigate sequence similarity, sequence mass, other hidden structures in the dataset that compromise data in terms of bias. This might take much longer than expected and is a crucial part of success and requires team effort! (Klavs<15h, Frida<15h).

- Fitting baseline Naive model. Naive model has to be constructed that will be fitted on bias. This model has to show that it learns using bias, and results higher than random performance (Accuracy > 50%). (Frida < 5h)

- Formatting data. Data has to be formatted for ML models. Protein sequences have to be converted into fixed size numerical arrays (possible tools: Word2Vec, BERT, other NLP models). (Klavs <5h)

-Fitting ML model. ML has to be fit on the sequence data. The machine learning model has to perform better than random. (Klavs <1h)

- Possible de-biasing of the data and validation that bias is removed. Training the Naive and ML models. (Klavs <3h, Frida <3h)

-Presentation, documentation and report. Presenting the findings in textual, written, and graphical format. The task here is to effectively convey findings in required format to the audience and keep the repository updated. (Frida - graphics, speech <10h, Klavs - text <10h)

Approximate list of tools used in accomplishing tasks

-Python programming language will be used in the research. Pandas library will be used to work with the data along with the scikit-learn library for implementing ML models.

-Word2Vec or any other NLP model available that can embed textual information in learnable hypothesis space

- Sequence similarity analysis in the data might be performed with BioPython library or BLAST.

References

[1] Muhao Chen, Chelsea J -T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, Wei Wang, Multifaceted protein–protein interaction prediction based on Siamese residual RCNN, *Bioinformatics*, Volume 35, Issue 14, July 2019, Pages i305–i314,

[2] Park, Y., & Marcotte, E. M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nature methods*, *9*(12), 1134–1136.

[3] Sequence representations and their utility for predicting protein-protein interactions Dhananjay Kimothi, Pravesh Biyani, James M Hogan. *bioRxiv* 2019.12.31.

[4] Pan XY, Zhang YN, Shen HB. Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J Proteome Res* 2010.10.1.