Report

Akshat Sharma
sharma06@ads.uni-passau.de
Universität Passau

Amit Manbansh
manban01@ads.uni-passau.de
Universität Passau

Lovesh Bishnoi
bishno01@ads.uni-passau.de
Universität Passau

Mihir Shah
shah02@ads.uni-passau.de
Universität Passau

1 INTRODUCTION[AKSHAT SHARMA]

In the body of all living creatures, proteins are building blocks of all cells. Some specialized cells in the human body bind with each other to identify what is new for the body. Such an identification process is possible via protein-protein interaction on the surface of the immune system cells. Thus, the affinity of binding cells can determine whether or not the immune process will be initiated. Protein and its interaction with other proteins are remarkably very complex phenomena. Knowledge of protein interaction can help pharmacists and microbiologists understand the function and behaviour of protein, to assign a new function. Adding to this, a cluster of proteins with the same function can be generated. Biologists and pharmacists can study protein interaction so that they can characterize protein complexes [2]. Study of protein-protein interaction can give answers to many questions of life on this planet, such as how one gets cancer, how one grows old, why one gets affected to disease, and how an antibody to a particular disease can be developed. These are the very few examples of protein functions, and that is why it is very important to study protein, its composition, its structure, its interaction with other proteins.

1.1 Graph Convolution Network[Akshat Sharma, Amit Manbash]

Graph Convolution Networks (GCNs) [6] are the neural network models that share the filter parameters over all locations in the graph. The goal of these neural network models is to learn a function of features on a graph. Let $G = (V, E)$ be the graph where V is the set of the Nodes and E is the set of the Edges. The inputs to the Graph Convolution Network are a feature description x_i for every node i, where $x_i \in X^{N\times D}$ and X is the feature matrix and N is the number of the nodes and D is the number of the input features, and A is the Adjacency Matrix which represents the description of the graph structure in matrix form. The Graph Convolution Network, i.e., G produces a node level output $Z^{N\times F}$, where F is the output features per node.

Every Graphical Neural Network layer can be represented mathematically as a non linear function [6]

$$H^{(l+1)} = f(H^{(l)}, A). \tag{1}$$

where $H^{(0)} = X$ and $H^{(l)} = Z$ and $'l'$ is the number of layers and $'f'$ is the function chosen for a specific task.
A simple layer wise Propagation rule [6] is given by

$$f(H^{(l)}, A) = \sigma(AH^{(l)}W^{(l)}) \tag{2}$$

where $W^{(l)}$ is a weight matrix for the l-th neural network and $\sigma(.)$ is a non linear activation function which could be ReLU, sigmoid, tanh, softmax, et cetera depending on the choice. Mathematically, these functions are defined as respectively:

$$Relu(x) = \begin{cases} 0 & \text{if x} < 0 \\ \text{x} & \text{if x} \geq 0 \end{cases}, \tag{3}$$

$$\sigma(x) = \frac{1}{1+e^{x}}, \tag{4}$$

$$tanh(x) = \frac{2}{1+e^{2x}} - 1, \tag{5}$$

$$softmax = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}. \tag{6}$$

There are two main limitations of this model and their possible adjustments are:

(1) Directly multiplication with A means that for every node all the feature vectors of all the neighbouring nodes are summed but not of that node under consideration. This limitation is overcome by adding the Identity Matrix, i.e., I [7] to the Adjacency matrix, i.e., A, mathematically:
$$\tilde{A} = A + I \tag{7}$$
where $\tilde{A}$ is the Adjacency Matrix with self loop and it is done so that each node also includes its own features at its next representation and it also helps with the numerical stability.
(2) A is not normalised which can lead to the change of scale of the feature vectors. This limitation is adjusted by normalising A such that all rows sum to one, i.e., $D^{-1}A$, where D is the diagonal node degree matrix of A and $D^{-1}A$ denotes the averaging of the neighbouring node features. Practically averaging of the neighbouring node features is not sufficient therefore symmetric normalisation or spectral approximation of A is done, i.e., $D^{\frac{-1}{2}}AD^{\frac{-1}{2}}$ [7].

Combining these two adjustments to the above limitations we get a propagation rule [7]:

$$f(H^{(l)}, A) = \sigma(\tilde{D}^{\frac{-1}{2}}\tilde{A}\tilde{D}^{\frac{-1}{2}}H^{(l)}W^{(l)}). \tag{8}$$

Where $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$ and it is used to normalise the nodes with large degrees.
In the research paper "Semi-Supervised Classification with Graph

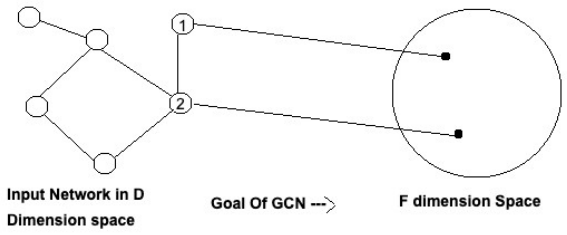

Figure 1: Mapping of Graph nodes from higher dimension to lower dimension

Convolution Networks" by Thomas Kipf et al [7], the authors tried to solve the problem of "Citation Networks" by using the the above mentioned techniques of the GCN propagation in a two layer GCN model. In the citation network the documents were considered nodes and edges were the documents cited in the published documents and on this semi-supervised classification the loss was computed as:

$$L = -\sum_{l \in Y_L} \sum_{f=1}^{F} Y_{lf} \ln Z_{lf} \tag{9}$$

Where Y_L is the set of node indices with labels.

1.2 Protein-Protein Interaction[Akshat Sharma, Mihir Shah]

A protein is a large organic bio-molecule [1] which is formed from the combination of Amino Acids which combine with each other using a peptide bond [1] to create a protein molecule. The length of the protein depends on its genetic code which specifies the number and types of amino acids responsible in the formation of a particular protein. The genetic expression in any organism is directly related to the proteins associated with that genetic code. When these proteins interact with each other then this interaction is known as Protein-Protein Interaction (PPIs) [5]. Cellular functions are defined by the interaction between the proteins, therefore if we want to understand the basic cell functions, we need to map the protein-protein interactions. Protein interactions have a huge potential within an organism as a whole, e.g., total interactions of human PPIs [3] are estimated to be around 650,000.

There are mainly two methods to detect the protein-protein interactions [3], namely Experimental and Computational. Experimental methods include Y2H, TAP-MS, protein microarrays, mbSUS, pull-down arrays, DPI, etc. Computational methods include PTMs, gene fusion, co-expression, GO annotation, gene neighbourhood, Phylogenetic profile, topological features, sequence features, Domain interactions, protein fold, etc.

Computational techniques consider "protein–protein interactions [5]" or PPIs as the associations or links between proteins. These techniques overcome the limitations of experimental techniques, e.g., Experimental findings are often incomplete even for well-studied organisms, therefore computational methods are used to complete the missing or incomplete part of the experimental PPI data and thus help in finding the clues to map out PPI mechanisms. These methods mainly focus on individual evidence for prediction and have certain specificities and biases. The various evidence sources are integrated in a statistical learning framework, such methods are called "prediction of protein-protein interactions by evidence-combining methods". The machine learning techniques can be applied on such a statistical framework.
The machine learning algorithm on the prediction of protein-protein interactions by evidence-combining methods mainly consists of three steps:

(1) Defining Gold Standard Datasets [3] (training datasets of interacting and non-interacting protein pairs).
(2) Characterising the interactions between proteins by annotating the Gold Standard Datasets [3] with carefully chosen and diverse evidence.
(3) Determining the probability of a particular interaction or interactions by individual evidence and then combining the probabilities of all the evidences.

1.2.1 Gold Standard Datasets. They are created for training or testing the PPI predictions and the datasets for training and testing are separate. These datasets could be either GSP datasets (Gold Standard Positive) or GSN datasets (Gold Standard Negative).

GSPs are PPIs with high experimental confidence or reliability or reference evidence, e.g., BioGRID, IntAct, etc, are some examples which are available in public databases. They are mainly the repositories of protein complexes and interactions are varied in terms of size and species-specificity and they contain information from both the experimental and the computational sources.

GSNs are usually not obtained by direct experimental methods or techniques. Negatome Database (2.0) provides a collection of proteins and domain pairs which are unlikely to engage in direct interaction but it wasn't able to satisfy the diverse GSP datasets of different users. GSNs can be obtained using certain methods, e.g., Negative examples can be chosen from the categories of their particular functions like annotations and subcellular localisation, etc.

1.2.2 Annotations of protein pairs with diverse evidence. Protein pairs are annotated based on their interactions with each other which could be based on cell physiology, biochemical environment, structures of the protein complexes, etc. To detect PPIs experimentally, certain conditions and criteria are met depending on the nature of protein interaction. For prediction of the PPIs by machine learning algorithms, we need to extract the protein interaction-based features. As there are several conditions for different PPIs therefore features are categorised into different categories and each category can provide a different view of protein interactions. Some of these categories and the features contained in them are:

(1) Evolutionary relationship (EVO) [3]: It uses the genomic context of organisms to categorise the PPIs. Features in this category are:
    - Gene Neighbourhood (GN)
    - Gene Fusion Event (FE)
    - Gene Cluster (GCL)
    - Polygenetic Profile (PP)
(2) Functional Features (FF) [3]: This method is based on the assumption that two proteins functioning in the same or a

similar biological process are more likely to interact with each other. Features in this category are:
- GO Cellular Component (COM)
- Co-essentiality (ESS)
- Gene/Protein Expression (Exp)
- Colocalization (Loc)

(3) Sequence based code signatures (SEQ) [3]: This method is based on the sequences of the amino acids that form the protein and if a set of proteins which interact with a set of proteins based on their sequences, then they are categorised in this category. Features in this category are:
- Conjoint Triads (COT)
- N-gram (NGR)

(4) Structure based signatures [3]: It is based on the structural interactions of the proteins which could be in the form of a covalent bond, ionic bond, etc. some of the features in this category are:
- Domain Domain Interaction (DDI)
- Van Der Waals Forces (VDW)
- Proteinfold (Fold)
- Electrostatics (ELE)

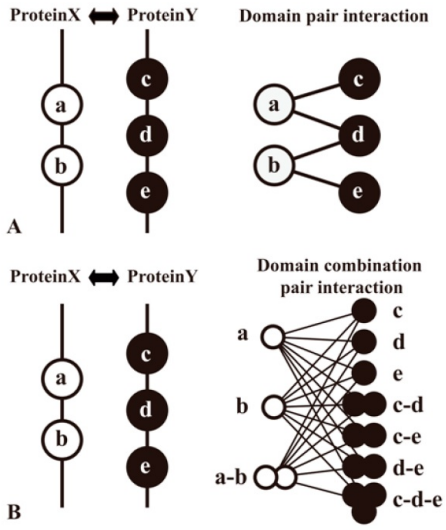

Figure 2: Types of Domain Domain Interactions [3]

Figure 2 shows two methods to predict Domain Domain Interactions from PPIs. Consider two proteins with domains {a, b} and {c, d, e} respectively, here PPIs are interpreted as the interaction amongst the domains of the two proteins.
In method A one domain of a protein will interact with one domain of the other protein.
In method B one or more domains of a protein will interact with one or more domains of the other protein.

1.2.3 Strategy for Integrative Analysis. Classification Algorithm is used to integrate the protein interaction related features, with these available features, classifiers are trained to differentiate between positive and negative examples. The usual process of PPI prediction by evidence-combining techniques [3] includes mainly three steps which are:

(1) Step I Choose appropriate evidence.
(2) Step II Encode protein pairs with evidence.
(3) Step III Find strategy to merge the classifiers into the integrative datasets.

The strategies could be the use of Artificial Neural Networks, Naïve Bayes, Decision Tree, K Nearest Neighbours, Support Vector Machine, et cetera.

1.2.4 Performance evaluation of PPI Prediction . The following techniques are considered to perform evaluation: Precision, Recall, Specificity, Overall Prediction Accuracy, Matthews's Correlation Coefficient (MCC), et cetera. They are defined as:

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{FP+TN}$

1.2.5 PPI Prediction through Domain Domain Interactions. A domain is mainly one or more submolecular parts of protein, described as a structural and functional module and usually an evolutionarily conserved unit. Recent studies about domains have concluded that abnormalities of domains can cause various diseases. Therefore, studies on the protein domain can help in developing disease models and tools to diagnose them. However, experimental techniques of the domain-domain interaction for predicting PPI have often shown more false-positive and false-negative results. The researchers have started studies on PPI prediction using computational techniques, that are mainly based on different features of protein such as protein sequence, domain information, three-dimensional structure, and protein evolution. The current state of the art approach does not fully consider domain information, instead only works on important domain and domain co-occurrences. However, an overall view of the PPI network can be better understood by domain-domain interaction [8].

1.3 Related work[Lovesh Bishnoi]

Our model takes inspiration from the conventional computation approach for PPI prediction and recent development on the graph convolution network for PPI. In what follows, we provide a brief description of related work in both the fields.

1.3.1 Conventional Approach for PPI. In the paper "Prediction of Protein-Protein interaction based on Domain" by Xue Li. et.al [8] proposed a novel approach based on protein domain for Protein-Protein Interaction. In this approach, the authors have used a state-of-the-art SVM model, in which physochemical properties of domain and domain-domain interaction score are used as the features for the prediction model. The outcome of the SVM model and the domain-domain score were used to design a Protein interaction prediction model. Accuracy, sensitivity, specificity, precision, Matthews correlation coefficient, and the F1 score are used as the evaluation matrix of the prediction model. The drawback of this approach was that it was implemented on a very small scale of the dataset. Adding to this, for the unavailability of negative domain-domain pair data, the noise was added instead.

1.3.2 Graph Convolution Network for PPI prediction . The spatial Graph convolution approach was used for protein interface prediction by Alex Fout et.al. [4] in the paper "Protein Interface Prediction using Graph Convolutional Networks". The prediction of the protein interface was based on the graph structure of the protein where amino acid residues are nodes. A set of k residues determined by mean distance between the atoms is used as a neighborhood-based convolution, which is able to detect the node features accurately. The edge features were also derived by their network, but in limited amount and were static compared to the node feature. Thus, the model learned the latent pattern for node features only. For evaluation, the authors compared their approach with State-of-the-art SVM based protein prediction approach. The AUC score for novel approach is 0.89 whereas, AUC score for the SVM model was 0.81. The authors suggest that their approach can be improved by using a large set of protein dataset for better representation for CNN and along with node feature, the model can be improved to learn edge feature representation.

1.4 Problem Definition[Akshat Sharma]

In this project, we are trying to solve the problem of the time-consuming experimental process to find out the possibility of PPI for given proteins with a faster computational method. We will be using GCN to design a model which can predict the possibility of a link between two given proteins.

Mathematically, the problem can be summarized as: For a graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, given two nodes $x, y \subseteq V$, where V represents proteins, we will implement a Graphical Neural Network model to predict if Edge {x, y} $\subseteq E$ | Edge {x, y} $\not\subset E$.

2 ACKNOWLEDGEMENT

This article was written during the Data Science Lab 2020 at the University of Passau. Our team comprises:

Table 1: DSL2020 team members

Name	Ownership of Phase
Akshat Sharma	Phase I Introduction
Lovesh Bishnoi	Phase II Data Acquisition and data Prepossessing
Mihir Shah	Phase III Model Implementation
Amit Manbansh	Phase IV Evaluation

REFERENCES

[1] Stryer L. Berg JM, Tymoczko JL. 2002. *Biochemistry.* Section 3.2: Primary Structure: Amino Acids Are Linked by Peptide Bonds to Form Polypeptide Chains, Vol. 5th edition. W H Freeman, New York. https://www.ncbi.nlm.nih.gov/books/NBK22364/

[2] Norberto de Souza O et al Breda A, Valadares NF. 2006 May 1 [Updated 2007 Sep 14]. Protein Structure, Modelling and Applications. *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach [Internet].* 1 (2006 May 1 [Updated 2007 Sep 14]), Chapter A06. https://www.ncbi.nlm.nih.gov/books/NBK6824/

[3] Ul Qamar MT Chen LL Ding YD. Chang JW, Zhou YQ. Nov 22, 2016. Prediction of Protein-Protein Interactions by Evidence Combining Methods. *Int J Mol Sci. 2016;17(11):1946.* 6 (Nov 22, 2016). https://doi.org/10.3390/ijms17111946

[4] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. 2017. Protein interface prediction using graph convolutional networks. In *Advances in neural information processing systems.* 6530–6539.

[5] S Jones and J M Thornton. 1996. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences* 93, 5 (1996), 13–20. https://doi.org/10.1073/pnas.93.1.13 arXiv:https://www.pnas.org/content/93/1/13.full.pdf

[6] Thomas Kipf. September 13, 2016. Graph Convolution Networks. https://tkipf.github.io/graph-convolutional-networks/.

[7] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:cs.LG/1609.02907

[8] Xue Li, Lifeng Yang, Xiaopan Zhang, and Xiong Jiao. 2019. Prediction of Protein-Protein Interactions Based on Domain. *Computational and mathematical methods in medicine* 2019 (2019).