IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Extracting the Interaction Relation between Proteins from Biomedical Literature using NLP Techniques for Drug Repurposing

Kanika K[1], Rohini R[1], Srinithi B[1], Saranya M[1], Arockia Xavier Annie R[2]

[1]*Department of Computer Science and Engineering, CEG, Anna University, Chennai-600025, India*

[2]*Assistant Professor, Department of Computer Science and Engineering, CEG, Anna University, Chennai-600025, India*

Abstract

Protein-protein interaction identification is essential to reveal the functional mechanism in living cells. Hence, the identification and prediction of protein- protein interaction are one of the essential needs in biology. Various experimental and computational methods are developed for predicting the interactions. Most protein-protein interaction detection systems have made predictions only based on the evidence from a single sentence which affects the model's accuracy.In this paper, we approach protein-protein interaction through a different view where sequence-based approaches are used to identify the interactions which improve the performance of the system. This paper also deals with the protein pairs identification from a large corpus dataset using NLP techniques and text mining methods. These predictions play a significant role in drug repurposing.

Keywords: Protein-Protein Interaction, Drug Repurposing, Human Interactome, NLP,Text Mining

1. Introduction

Proteins are macromolecules, or large molecules consisting of one or more long chains of amino alkanoic acid residues. They vary from each other based on their amino acid sequence, which is the gene's nucleotide sequence and it leads to folding into a selected three-dimensional structure that determines its activity[1]. Protein-protein interaction is an association of more protein molecules as a result of biochemical events which are directed by the interactions which include electrostatic forces, bonding of hydrogen, and hydrophobic effect[2, 15]. These are significant to all the processes in a cell, so the knowledge of PPI is important for understanding cell physiology [3]. Proteins act alone because their functions are regulated [4]. Protein-protein interaction (PPI) identification is a significant task in text mining [5]. Protein–protein interactions (PPIs) play a major role in many biological processes, such as cancer, regeneration and the development of many diseases which makes the identification of these interactions a key point for the understanding of these processes [6]. Protein-protein interactions are important for creating biological models for understanding all the biological processes [7].

Extracting PPIs directly from the scientific literature are often very helpful for providing such context, because the sentences describing these interactions may give insights to researchers in helpful ways [8]. Predicting protein-protein interaction computationally has become a more significant prediction method that can overcome all the obstacles of the existing methodology [5]. Most PPI detection systems make predictions completely based on proof within a single sentence and often endure the heavy burden of manual annotation [9]. At present, protein-protein interaction (PPI) is one of the pivotal topics for the development and advancement of modern structure's biology [1]. This project approaches the PPI detection task from a different pattern by investigating the context of protein pairs collected from a large corpus and their relations [10].

2. Related Works

2.1. Clustering Techniques

Previously, clustering techniques were used to identify protein-protein interaction, but those techniques suffer from some drawbacks. Complex syntactic structures of sentences often make the predictions very difficult. The context of interactions is ignored in these approaches. Some approaches for mining the protein-protein interactions from biomedical literature ranges from co-occurrence analysis to natural language processing systems [8].

2.2. Machine Learning Models

Many approaches explore natural language processing techniques on machine learning (ML) methods [5]. Some approaches investigate various strategies of measuring the distance between two data points and explore it in kernel functions. Jiang et al. make PPI predictions using continuous word representations [6]. ML approach does not require manual construction of rules or patterns and often achieves better accuracy. However, these approaches often suffer from small training datasets.

2.3. Single Sentence based approach

In a single-sentence-based approach, in order to build the training data, every protein pair appearing in a sentence has to be manually annotated as positive (interactions) or negative (non-interaction) [11]. This is very intensive labeling work. As a result, the classifiers are generally trained on small datasets. The two approaches, unsupervised and supervised, for developing computational methods for PPI prediction will be discussed in the following. On the one hand, the unsupervised approaches reconstruct PPI networks solely based on a set of protein attributes. In this category, some approaches investigate the use of topology information [12].

2.4. Text Mining Approach

The AImed corpus consists of plain text abstracts, and human experts identified the protein names and the exact locations of statements expressing PPI in the text. In the BioCreative training data, the same protein names used in the text were not known since only protein identifiers from the UniProt database were provided [6, 16]. With the protein name identifier, the system can be used to produce average results for the Bio- Creative Protein-Protein Interaction task. Some Recent studies used 10-fold cross-validation results to compare their results with other studies that use the AImed corpus dataset.

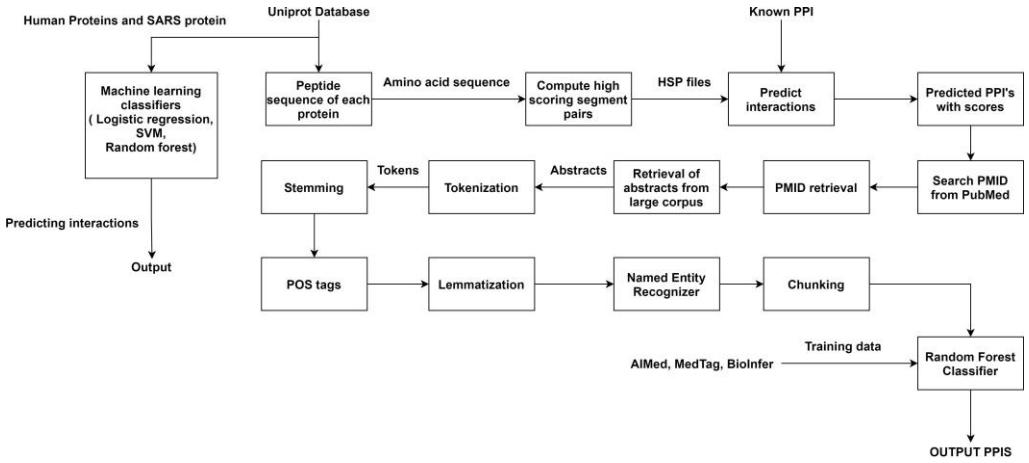

Figure 1: Architecture Diagram

Each study reports different numbers of Protein-Protein Interaction pairs from the AImed corpus, which makes a direct comparison between different systems difficult. Furthermore, recently Yuan et al. developed a generative network model to predict protein-protein interactions in networks. One of the well-established propagation methods, the shortest path propagation, has been recently introduced to predict the PPI in networks [13, 17]. Their approach accomplishes a good performance in PPI prediction with the specificity of 85 percent and sensitivity of 90 percent. However, although it is able to capture the global structure of the network, it should be noticed that the shortest path propagation is known to be sensitive to short-circuit topological noise [14].

2.5. PPI in drug repurposing

Protein-protein interactions are correlated to key activities and are the most promising targets for the discovery of drugs. The depiction of drug-protein interaction networks with biological characteristics has become one of the challenging problems in modern pharmaceutical science toward a better understanding of pharmacology. Polypharmacology is associated with various features of drugs and target proteins (e.g., pharmacophores, functional sites, chemical substructures, and pathways) and complicated associations between heterogeneous features. So, identifying the interaction between proteins plays a major role in DTI prediction and drug-repurposing [18].

3. PROPOSED SYSTEM

The main idea of this project is to identify the protein-protein interactions. By computing the high segment protein pairs and interaction score, we can identify the highly interacted proteins score. By using these proteins, we can retrieve abstracts from biomedical literature and they are processed by using nlp techniques (tokenization, stemming, named entity recognition etc.,). This input is given to the Random Forest classifier which is already trained using AIMed, Bioinfer, MedTag Datasets and these are used to predict the interactions between the proteins. Another approach would be to give the input sequence directly to the machine learning classifier to predict the interactions.

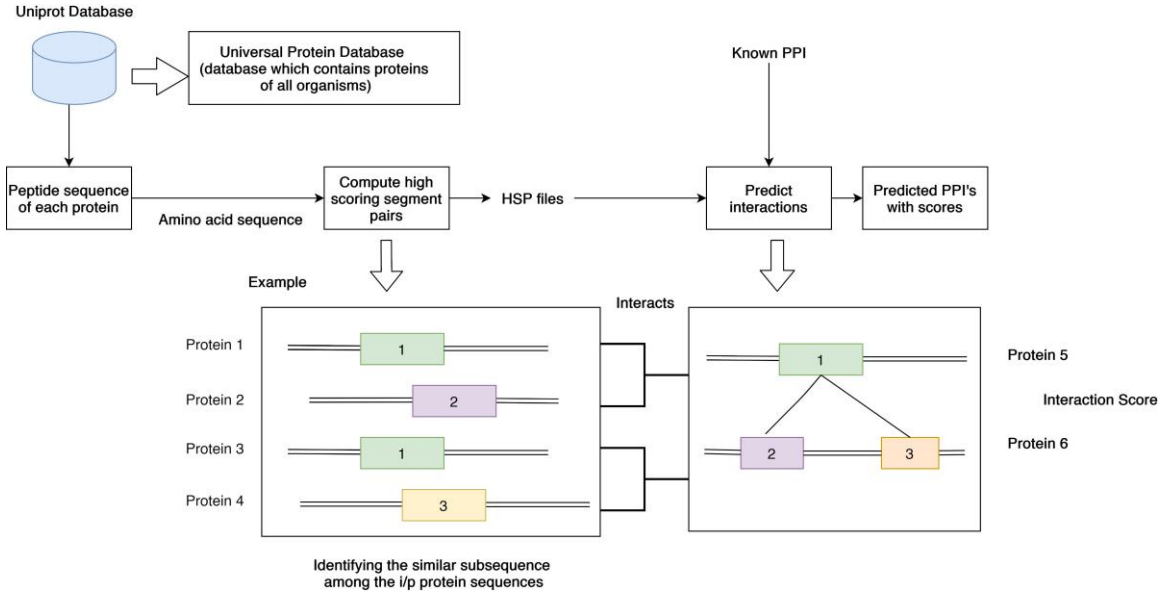

Figure 2: Finding similar subsequences and Interaction Score

3.1. Scoring Protein Interaction

Proteins that are alike to interacting proteins are likely to interact as well. If P1 is known to interact with P2 and the sequences of P1 and P1' are highly similar, and the arrangements of P2 and P2' are highly identical, then P1' and P2' are likely to interact well. The identification of similar subsequences among the input protein sequences is the first step. This is done using spaced seeds. Assume a match of size five is used. In this case, a particular match consists of 5 consecutive matching amino acids between two protein sequences. This is called a hit. Denote the five consecutive matches of a BLAST-like seed by 11111; this is often often called a consecutive seed of weight five. Spaced seeds consist of matches intermixed by don't care positions; this is an example of such a spaced seed: 11****11***1. A spaced match requires only the amino acids in positions like 1's within the seed to match. Note that the quantity of matches (the weight) is that equivalent to the consecutive seed; five in our case. The best value for our problem clothed to be five. Spaced seeds have a better probability of detecting subsequences that are alike, while the amount of hits is that the same as for consecutive seeds; the expected number of hits is specified by the weight of the seed, which is that the same. Several seeds can identify more similar subsequences as they find different similarities.

The distribution of matches and don't care positions is crucial for the quality of the seeds, and we have used SPEED to compute the following seeds; we have experimentally determined that four seeds of weight five are the best choice: SEED 4,5 =11****11***1, 1**1*1***1*1, 11**1***1**1, 1*1******111. To further increase the probability of finding similar subsequences, we consider also hits between similar matches, as against exact ones. We consider also hits consisting of comparable spaced matches. To make this concept precise, we'd like a couple of definitions. Spaced-mers are defined analogously with k-mers but employing a spaced seed.

A k-mer may be a contiguous sequence of k amino acids with spaces, consistent with the seed. For a spaced seed s, we can call the spaced-mers also s-mers. An exact hit, therefore, consists of two occurrences of an equivalent s-mer. An approximate hit, on the opposite hand, requires two similar s-mers. Assume a similarity matrix M is given. Given a seed s and two spaced-mers w and z, the score between the two spaced-mers is given by the sum of the scores of the pairs of amino acids in the two spaced-mers; that is, we calculate the sum over indexes corresponding to 1's in the seed:

$$Ss - mer(w, z) = \sum_{s[i]=1} M(wi, zi) \quad (1)$$

Using (1), we de_ne the set of s-mers that are similar with a given s-mer w:

$$Sim(w) = \{z | zs - mer, Ss - mer(w, z) \geq Thit\} \quad (2)$$

Note that Sim(w) depends on the parameter Thit that controls how similar two spaced-mers have to be to form a hit. It also depends on the seed s and the similarity matrix M, but we do not include them in the notation for clarity. All such hits dues to similar s-mers are found then extended both ways to spot similar regions. That means now we've to gauge the similarity of all the amino acids involved, so we use the regular k-mers. The score between two k-mers A and B is computed because of the sum of all many corresponding amino acids:

$$Sk - mer(A, B) = \sum_{i=1}^{k} M(Ai, Bi) \quad (3)$$

Where Ai is the ith amino acid of A, given a hit that consists of two s-mers w and z, we consider t he two k-mers that contain the occurrences of the two s-mers w and z within the center, denoted k-mer(w) and k-mer(z). If Sk-mer(k-mer(w), k-mer(z)) Tsim, then the two regions are deemed similar. Note the parameter Tsim that controls, together with k-mer size k, how similar two regions should be to be identified as such. What we've computed thus far are similarities, that is, pairs of comparable subsequences of an equivalent length. We now display how to compute the scores. First, we spread the definition of the score from k-mers to arbitrary subsequences of equal length. For two subsequences X and Y of length n, the score is given by the sum of the many all corresponding k-mer pairs; using (3):

$$Se(X, Y) = \sum_{i=1}^{n-k+1} Sk - mer(X[i..i + k1], Y[i..i + k1]) \quad (4)$$

where X[i . . j]= XiXi+1 · · · Xj.

It is important to recall that any two similar sequences we find have the same length; therefore, the above scoring function can be used. Finally, we report how the scores for whole protein sequences are computed. Initially, all scores are set to zero. Each pair of proteins (P1, P2) that are known to interact has its grant to the scores of other pairs. For each calculated similarity (X1, Y1) between P1 and another protein Q1 (X1 is a subsequence of P1 and Y1 is a subsequence of Q1) and for each similarity (X2, Y2) between P2 and another protein Q2, the score between Q1 and Q2, Sp(Q1, Q2), is increased, Using (4), by:

$$Sp(Q1, Q2) = Sp(Q1, Q2) + [Se(X1, Y1)(|X2|k+1) + Se(X2, Y2)(|X1|k+1)] / |Q1||Q2| \quad (5)$$

Where Q indicates the length of the amino acid sequence Q. That means the score of each complement k-mer pair between X1 and Y1 is multiplied by the number of k-mers in X2, that is, the number of times it is used to support the certainly that Q1 is interacting with Q2.Similarly, the score of each corresponding k-mer pair between X2 and Y2 is multiplied by the number of k-mers in X1. The score acquires this way is then normalized by dividing it by the product of the lengths of the proteins involved. Once the score is computed, by considering all given interactions and similar subsequences and computing their impact on the other scores as above, predicting interactions is simply done according to the scores. All protein pairs are sorted decreasingly by the scores; higher scores represent a higher probability of interacting. If a threshold is given, then those pairs with scores above the threshold are reported as interacting.

The next module involves extracting interactions from biomedical literature; it reads a text file with a list of PubMed identifiers and downloads all the necessary articles. It downloads either abstracts from Medline or full-text articles from PubMed Central, depending on the option provided by the user. It uses Stanford CoreNLP for name entity recognition (NER) of proteins and genes. Three datasets were used to train the Stanford Named Entity Recognizer: AImed, MedTag, and BioInfer. 1st each sentence was tokenized by Stanford CoreNLP; then, a Conditional Random Field classifier was trained. Performance of the Named Entity Recognition (NER) tagger was assessed by 2-fold cross-validation. Once the NER is trained, the module extracts all the co-occurring proteins in each sentence and, for each pair of them, computes several features, which will be used to identify pairs as interacting or not. The prediction is based on a Random Forest Classifier trained over the annotated sentences of AImed, LLL-challenge, and BioInfer. Finally, the module produces several possible outputs: an HTML page with all the PPIs, the sentences in which they were found, and a table with all the proteins found in the specified articles.

3.2. FINDING SIMILAR SUBSEQUENCE AND INTERACTION SCORE

Input: Protein sequences Ps, protein interactions Pi .

Output: All protein pairs sorted decreasingly by score .

If proteins A and B are interacting, and if A and A', B and B' are similar then A' and B' are likely to interact as well.

Finding similar subsequence:

Find exact hit and approximate hit.

Exact hit - 2 occurrences of same subsequence in a protein sequence

Approximate hit – 2 similar subsequence in a protein sequence

If it is greater than hit threshold, it is included in similar subsequence and it is extended in both sides to identify the similar regions and again evaluate the similarity of all amino acids

Compute interaction score:

Predict PPIs based on the interaction score (High interaction score indicates high probability for prediction)

3.3. PPI EXTRACTION FROM BIOMEDICAL LITERATURE

Train the random forest classifier using datasets like AIMed, BioInfer etc.

Searching the PubMed ID of the proteins with high interaction score

Retrieve Pubmed ID

Give the input as a text file of PubMed IDs

Retrieving abstracts using e-utilities

Tokenize the abstract

Stemming - removes the subwords

Identify the Parts of the Speech tags

Lemmatize each words in the abstract

Identify the named entity

Rephrase the sentence so that it gives a complete meaning

Provide the sentence as input to Random forest classifier to obtain the prediction.

3.4. INTERACTION BETWEEN HUMAN AND SARS COV 2 AND HUMANPROTEIN

Input: Collect the human proteins that interact with SARS-CoV-2 virus (Positive dataset). Collect non-interacting human proteins from HPRD using the concept of degree distribution (Negative dataset).

Output: Predicted potential human target proteins

Combine both positive and negative dataset and do feature selection to obtain training dataset.

Apply SVM, RF, MLP, AdaBoost, GBoost and Logistic regression classifiers for training the model and do comparative analysis for the prediction of new sets

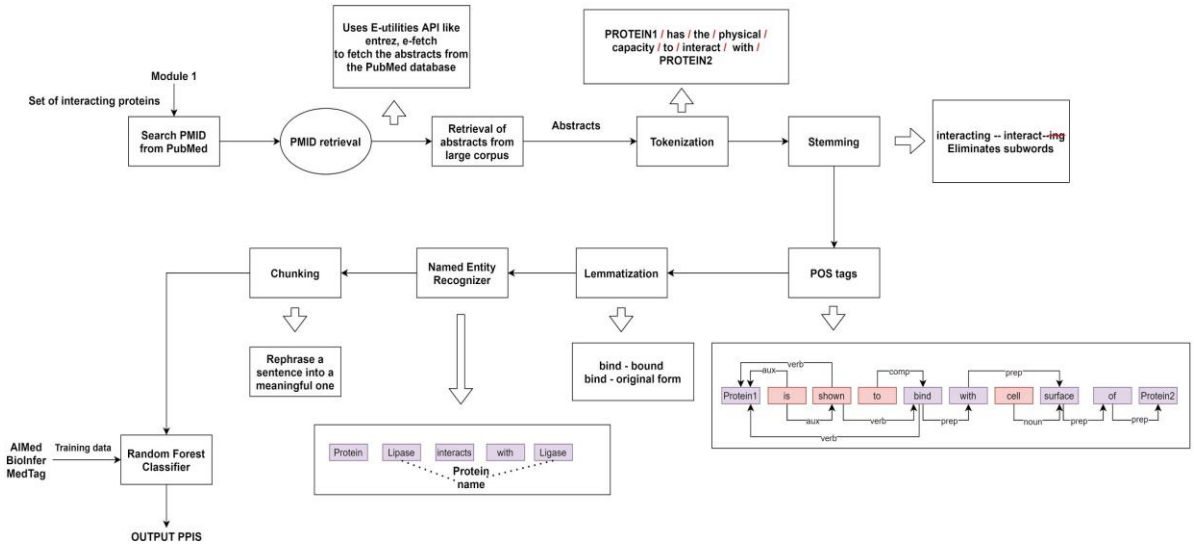

Figure 3: PPI Extraction From Biomedical Literature

```
P41812 P38142
Q3E755 P23900
P37265 P34216
Q99195 P25649
P10961 Q02772
```

Figure 4: A FASTA file example. Each protein consists of the protein name which starts with(greater than symbol) and the amino acid sequence of interactions

Figure 5: An input PPI file example. Each line contains a pair of protein names

3.5. DATASET

To predict the Protein-Protein Interactions (PPIs) of an organism, SPRINT requires two types of data. First, a set of protein names and sequences in FASTA file format.Second, a set of known protein-protein interaction PPIs.The FASTA file uses single letter codes to indicate the peptide sequences of each protein. Each protein has two lines. The first line indicates the protein name, and the second line indicates the amino acid sequence.The line containing protein names starts with any symbol ”¿ ”sign and is followed by the protein name. In the sequence line, each letter scrambles an amino acid. An example is given in Fig. 4, the amino acid sequence of protein P32479 is MKVVKFPWLAHREESRKYEIYTVDVSHDGKRLA.

The raw FASTA file with the protein sequences of most organism can be downloaded from Uniprot at **https://www.uniprot.org/.**

For example, manually annotated all human proteins and their sequences can be downloaded by clicking on “Swiss-Prot” and then “Human” and then “Download.” Uniprot contains additional information such as function and gene ontology information. Each line is a pair of proteins which are known to be interact is shown in fig 5. Note that the protein names in the PPI file should match the ones in the input FASTA file.The PPI dataset is usually obtained through major protein–protein interaction databases such as Biogrid, MINT, InnateDB, IntAct, and HPRD.

3.6. PERFORMANCE METRICS FOR EVALUATION

The performance of the machine learning model is evaluated using Accuracy, precision, recall, and F-score.

(i) Classification accuracy Accuracy = Correct predictions/ Total predictions

(ii) Precision Precision= True Positive/ (True Positive + False Positive)

(iii) F1-Score $F1 = 2 * (Precision * Recall)/(Precision + Recall)$

(iv) Target key-phrase ratio - It is the ratio of occurrences of target and keyphrase in a sentence.

(v) FET (Fisher's Exact Test) p-value - The p value represents the probability of obtaining an effect equal to or more extreme than the one observed considering the null hypothesis is true. Null hypothesis states that there is no association between target key phrase counts and interaction.

Table 1: Result Comparison of Various Models

Model	Accuracy	Precision	F1 score	Recall
SVM	93.74	86.79	84.66	85.71
Logistic Rrgression	78.60	68.31	84.66	75.61
Random Forest	93.94	83.33	88.95	86.05
ADA Boost	61.56	57.03	92.02	70.04
GBoost	82.31	72.39	85.27	78.30

3.7. Results and Discussions

We have presented a new algorithm for predicting PPIs with higher performance than the current programs while using a small amount of memory. The proposed approach is straightforward to use, and we hope it will make PPI predictions for entire interactomes. Since they work directly with the sequence of amino acids, sequence-based methods often have an advantage in finding the actual positions where interaction occurs. Protein and gene name tagging by the Stanford CoreNLP was evaluated by 2-fold cross-validation on the datasets of AImed, MedTag, and Bioinfer; As we don't use any dictionary table in order to identify the proteins, it can tag protein symbols of any species or even newly described proteins and genes. However, it cross-checks the identifiers against the HUGO Gene Nomenclature Committee's aliases as a prior normalization step

The proposed approach can retrieve protein–protein and genetic interactions without defining specific patterns or rules, which makes the application's use broader. We consider only a narrow selection of features, namely POS composition, token distance, and keywords. We include the result as an HTML report output. The output includes a summary table and the retrieved interactions.

4. Conclusion

This paper explores a paradigm in PPI identification that is different from current single-sentence based approaches by investigating the sequence of protein pairs as well as their context. First, the proteins that have high chance of interaction based on sequence are identified. Then, from a large corpus biomedical literature dataset, context where proteins are mentioned are investigated, so that its possible to find whether the proteins with sequence based high interaction score are likely to interact based on biomedical literature as well. Additionally, the proposed approach takes known PPIs in actual PPI databases (e.g., HPRD) as the training data. In the proposed approach, lexical features are extracted based on Natural Language processing. Named Entity Recognizer is used to identify the protein names.

References

[1] D. Zhou, Y. He, Extracting interactions between proteins from the literature, ACM 41(2008). doi:https://doi.org/10.1016/j.jbi.2007.11.008.

[2] B. S. S. Sourav S. Bhowmick, Clustering and summarizing protein-protein nteraction networks: A survey, IEEE Transactions on Knowledge 28 (10) (2016) 638 – 658. doi:10.1109/TKDE.2015.2492559.

[3] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. P. Mäkelä, S. Hauaniemi, Integrated network analysis platform for protein-protein interactions, Nature methods 6 (1) (2009) 75–77.

[4] D. T.-H. Chang, Y.-T. Syu, P.-C. Lin, Predicting the protein-protein interactions using primary structures with predicted protein surface, BMC bioinformatics 11 (1) (2010) 1–10.

[5] H.-J. Zhu, Z.-H. You, W.-L. Shi, S.-K. Xu, T.-H. Jiang, L.-H. Zhuang, Improved prediction of protein-protein interactions using descriptors derived from pssm via gray level co-occurrence matrix, IEEE Access 7 (2019) 49456–49465.

[6] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, D. Eisenberg, The database of interacting proteins: 2004 update, Nucleic acids research 32 (suppl 1) (2004) D449–D451.

[7] V. S. Rao, K. Srinivas, G. Sujini, G. Kumar, Protein-protein interaction detection: methods and analysis, International journal of proteomics (2014).

[8] Y.-N. Zhang, X.-Y. Pan, Y. Huang, H.-B. Shen, Adaptive compressive learning for prediction of protein–protein interactions from primary sequence, Journal of theoretical biology 283 (1) (2011) 44–52.

[9] G. C. Koh, P. Porras, B. Aranda, H. Hermjakob, S. E. Orchard, Analyzing protein–protein interaction networks, Journal of proteome research 11 (4) (2012) 2014–2031.

[10] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, Y. W. Wong, Comparative experiments on learning information extractors for proteins and their interactions, Artificial intelligence in medicine 33 (2) (2005) 139–155.

[11] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, T. Salakoski, Bioinfer: a corpus for information extraction in the biomedical domain, BMC bioinformatics 8 (1) (2007) 1–24. [12] S. P. K. S. W. D. Z. C. Chengbang Huang, Faruck Morcos, J. A. Izaguirre, Predicting protein-protein interactions from protein domains using a set cover approach, IEEE/ACM Transactions on Computational Biology and Bioinformatics 4 (2007) 78 – 87. doi:10.1109/TCBB.2007.1001.

[13] J. Zahiri, A. Emamjomeh, S. Bagheri, A. Ivazeh, G. Mahdevar, H. S. Tehrani, M. Mirzaie, B. A. Fakheri, M. Mohammad-Noori, Protein complex prediction: A survey, Genomics 112 (1) (2020) 174–183.

[14] H. Y. Hang Li, Xiu-Jun Gong, C. Zhou, Deep neural network based predictions of protein interactions using primary sequences, Molecules (2018)20–27, doi:10.3390/molecules23081923.

[15] J. Zahiri, A. Emamjomeh, S. Bagheri, A. Ivazeh, G. Mahdevar, H. S. Tehrani, M. Mirzaie, B. A. Fakheri, M. Mohammad-Noori, Protein complexprediction: A survey, Genomics 112 (1) (2020) 174–183.

[16] S. Abdulkadhar, G. Murugesan, J. Natarajan, Classifying protein-proteinbinteraction articles from biomedical literature using many relevant features and context-free grammar, Journal of King Saud University-Computer and information Sciences (2017)

[17] Z. Zhao, Z. Yang, H. Lin, J. Wang, S. Gao, A protein-protein interaction extraction approach based on deep neural network, International Journal of Data Mining and Bioinformatics 15 (2) (2016) 145–164.

[18] Adhami, Masoumeh, Balal Sadeghi, Ali Rezapour, Ali Akbar Haghdoost, and Habib MotieGhader. "Repurposing novel therapeutic candidate drugs for coronavirus disease-19 based on protein-protein interaction network analysis." BMC biotechnology 21, no. 1 (2021): 1-11.