A Survey on Prediction of Protein-Protein Interactions

Nivedha S¹, Bhavani S²

¹ Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore-641062, India
² Department of Electronics and Communication Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore-641062, India

Email: nivedha.nivi6@gmail.com, bhavanisridharan7@gmail.com

Abstract: Proteins interacts with one another to form complexes for performing the biological process in a cell. Understanding the interactions of proteins helps in recognizing the disease mechanisms. It plays a major part in the catalytic reaction, drug ability of the molecules. The wet-lab methods are tedious and expensive. However, the data that is identified in experimental methods can be processed and used. Advancement in the field led the prediction using computational approaches for signaling pathways and the recognizable proof for specific diseases. Identification of protein-protein interaction offers a way for advancements of therapeutic methods and drug design. In this paper, an analysis of prediction methodologies that are used for protein-protein interactions was described. Experimental methods and six different computational methods were discussed.

1. Introduction

Proteins are made up of large units of amino-acids, attached to each other to form longer chains. There exist 20 different types of amino acids (AA), joined to form a protein. Structure of the cell and capacity plays a major role in Protein-Protein Interaction (PPI). Proteins involved in vital processes in the cell such as Immune response, cell signaling and metabolism. Even though Advancements in the field of bioinformatics and genome biology, function of larger sequenced proteins remain uncharacterized. The study of the intercommunication of a known target protein with unknown proteins helps us to identify the functionality of unknown proteins. Interaction pattern among proteins suggests designing new drugs by providing the biological pathways that exist in neighbourhoods of the target. Identifying the interaction is not only important for drug design, but it also helps in understanding the working of proteins in a cell to perform cellular functions. Recognition of protein interaction leads to Artificial design of Interactions [2]. Interaction, metabolic and development control of cell are the procedures where PPIs have an extensive scope. PPI refers to the physical interaction between at least two protein molecules resulting in a biochemical event [1]. Experimental methods that aided in predicting PPIs with all possible PPIs in the cell. Still, the generated data often contains false positives and negative values along with the missing values with few overlaps. This recommends that the data is incomplete, erroneous or both. It has been predicted that only 50% of yeast PPI and 10% of human PPI network have been identified. Limitations in wet-lab methods makes the necessity to determine the PPI using computational methods. Computational methods can be Machine Learning (ML) or statistical
approach that has been enforced to the PPI network to determine the interaction. This paper described the methodologies that involved in Prediction in PPIs.

2. Protein – Protein Interaction

In PPI networks, most of the proteins interact with others to perform their functions, whereas some of the proteins may work in isolation. Interaction of protein implies the formation of active protein complexes through physical contact between proteins. It involves signal transduction, replication of DNA, and transport. In signal transduction, PPI is involved in transmitting signals between the cell exterior to the interior. Protein modification can be done through interaction. Example, phosphorylation process [3]. A kinase interacts with protein target to make it add to the group of protein complexes. PPI is responsible for nearly all the functionalities within the cell. Still, larger proportion of interactions among proteins remain unknown. Hence, understanding protein-protein interaction is important. PPI patterns illustrate pathways that surround the drug targets. PPI prediction can be observed as binary classification, identifying the pairs as either interacting or non-interacting [4,5,6]. Various PPI prediction tasks include direct PPI prediction, direct PPI and indirect functional association prediction, pathway membership prediction. First refers to direct interaction between proteins. Second, involves interacting proteins may not have direct interaction, but interact with the target indirectly. The term “interacting” is considered as proteins within the subcellular complex, though they are not interacting directly. But connected through other proteins within complex [7]. Pathway membership prediction involves interaction that occurs in a logical way like signaling pathways [8]. Network of PPIs can be designed by applying module-based interaction and pair-wise interaction methods. The objective of pair-wise interaction is to identify whether two proteins are located in the same complex. That is, prediction of the direct interaction of proteins. It occurs in proteins that appear in the same cellular complex. Module-based interaction deals with the interaction of a group of proteins. Here, direct contact with proteins is not required. These prediction approaches aim to classify the proteins either interacting or non-interacting and used in the construction of PPI Networks.

3. Methods

3.1 Experimental methods

Conventional biological methods were used in PPI prediction. These methods focus on biochemical or biophysical properties of protein complexes. These experiments are performed as vitro or in vivo. In vitro experiments are done outside the living organisms in a controlled environment. In vivo experiments are done inside the organisms. Some of the experiments use to screen across the whole proteome. The wet-lab methods are used to characterize, determine and validate PPI. Experimental methods like Y2H methods are used to measure the protein interactions. Some of the experimental methods for the prediction of PPI are listed in Table 1. However, experimental methods are time-consuming and expensive. Thus, it creates a requirement of a computational approach in the PPI prediction. Estimation error-rates of experimental methods range 41% to 90%. These biological data are integrated along with the computational approaches to improve interactome coverage and prediction of PPIs.
Table 1. Experimental methods for Prediction of PPI

| Method            | Description                                                                 |
|-------------------|-----------------------------------------------------------------------------|
| Co-immunoprecipitations | To Identify the interaction among proteins, precipitation procedures were used |
| Surface Plasmon Resonance | To measure the change in protein binding in identification of interaction, a laser light is reflected where the protein is attached to the surface. |
| X-ray Crystallography | It helps in constructing the interaction structure through crystallization of PPIs |
| Label Transfer | known protein is “tagged”. Proteins that interact with this tagged protein can be obtained by identifying the tag. |
| Y2H | It employs in “two-plasmid” approach in yeast. This technique is an example of direct interaction between proteins |
| MS TAP | Target proteins are attached with affinity tags. Proteins are separated based on the mass to discover purified protein complexes. MS technique is used to analysed [9]. |
| Gene Co-expression | Protein -pairs that are co-expressed are mostly interacting proteins. This method is an example of indirect interaction between proteins. |
| Synthetic lethality | PPIs can be identified in Synthetic interactions between gene products, their occurrence in a pathway [9]. |

3.2 Computational methods

In past decades, various computational approaches have been identified to predict PPI. Many proteins have no information on structure, except the information about protein sequences. With the advancement of sequencing techniques, an increasing number of protein sequences are obtained. In this paper, six computational models for predicting of PPI were discussed.

3.2.1 EnsDNN

Protein-protein Interaction prediction is done using Ensemble Deep Neural Networks (EnsDNN) [11]. EnsDNN uses local descriptor, auto covariance descriptor, MCD to explore the interaction patterns. It trains the DNN with different configurations of each descriptor. It integrates Deep Neural network into an ensemble predictor to influence complementary data. Performance of EnsDNN is 95.29% of accuracy, 95.12% of sensitivity, and 95.45% precision on predicting PPIs. The proposed model is based on AA sequences. It includes three steps. First, it uses AC descriptor, LD descriptor, and MCD descriptor to convert the protein-pairs interaction information into numeric vectors. Second, train individual deep neural networks on each of the three types of vectors. Finally, its ensembles 27 individual deep neural networks using a two-hidden layer neural network.
3.2.2 DeepPPISP

DEEPPISP, a framework to predict PPI site through the features of local and global using DNN [10]. Deep Learning (DL) techniques were used to select and combine global sequence features. The main objective is to extract both global and local features from sequence and combine it to a DL framework. For local features of target AA, a sliding window is used to get the neighboring AA features. For global, different deep learning structures are used. DeepPPISP framework includes feature extraction and classification parts. Input to the framework is local and global sequence features. First part involves preprocessing and extracting the patterns and features to predict the sites. It uses Text Convolutional Neural Network to capture sequences allows to easily model the target AA and protein sequence. The limitation of DeepPPISP is slow speed, not good in lengthy proteins prediction. For DeepPPISP, sequence-based site prediction is still a challenging problem.

3.2.3 SMOTE and EL-SMURF

SMOTE and EL-SMURF is used for predicting sites [13]. Existing models have good prediction performance. Still, imbalance issue can reduce the performance of conventional algorithms. Ensemble learning is applied to the ML field due to its faulttolerant rate for better classification prediction results. EL-SMURF model is designed using Ensemble learning for better prediction accuracy. This model includes acquisition of data and extraction of features, sampling and feature selection, classifier modelling and classification. To combine RF classifier and learning model ELSMURF, majority voting method is used as learning strategy. In feature extraction of neighboring residues, the combination of sequence profile feature and evolution rate of residue was applied. ELSMURF has prediction accuracies of 77.7%, 79.1% and 77.1%, for three different datasets. Still, more feature selection methods can be applied to improve the performance.

3.2.4 SCRIBER

SCRIBER is sequence-based predictor for protein binding residues which minimize the cross predictions [14]. Dataset used in the model takes an advantage of covering multiple binding residue types. Inputs used are similar to the model and construction that is personalized to decrease the cross-predictions. In this model, two-layered architecture is used. First, produces a prediction of DNA-binding, RNA-binding, protein-binding along with small-ligand binding residues. It converts the protein sequence into a comprehensive profile. This profile is processed by five different algorithms to produce the propensity of each binding. Second, again predicts protein binding residues by decreasing overlap between other binding residues and PBRs. It utilizes propensities produce in the first and a ML model to predict the protein-binding propensities with the objective to decrease cross-predictions.
3.2.5 SLSTM Network with DLPred

SLSTM network with the DL model, DLPred for refining the imbalanced prediction of PPI sites is proposed. In this model, three different ideas were proposed. First, to hold the entire sequential completeness, the training dataset is used to build the group of protein sequences, instead of single residues. Second, to the activation function a new factor is attached, which decreases the non-interaction site loss. Third, multi-task learning of interaction sites and residues is used for the prediction model on non-interaction sites. This model explores new designs to address the imbalance issue. Performance achieved is 38.9% F-measure, 69.1% accuracy and 80.1% in AUC for the first dataset. For imbalance classification, it has an improved performance, when compared with other models.

3.2.6 Convolutional Neural Networks

PPI site prediction can be done using Convolutional Neural Networks (CNN) with improved datasets [14]. Protein-Protein Interaction site prediction can be divided as sequence-based methods, structure-based methods and methods based on integrated information. First method uses feature extraction to predict protein interaction sites. Second method requires information on the 3D appearance of protein complexes which gives data about prediction sites. The third category will be a combination of both structure and sequence-based methods. In this model, CNN and propensity of residue binding is used to increase the positive samples in the result. The model has 0.912 AUC on improved datasets. Thus, it provides fewer false-positive PPI sites. Number of negative samples acquired is higher than positive which leads to imbalanced data. Hence Easy Ensemble algorithm is used to build the equal negative and positive samples training set. Finally, the model is compared with the result obtained from propensity to that is used to randomly sampled interacting pairs.

Table 2. Comparison of Computational methods

| Method                        | Merits                                                                 | Performance metrics                        |
|-------------------------------|------------------------------------------------------------------------|--------------------------------------------|
| EnsDNN                        | Uses LC, AC, MCD to explore the interaction patterns                   | Accuracy: 95.29%                           |
|                               |                                                                        | Precision: 95.45%                          |
| DeepPPISp                    | Combines the features of local and global from the sequence for prediction | Accuracy: 0.655                            |
|                               |                                                                        | Precision: 0.303                           |
| SMOTE & EL SMURF             | Handles imbalanced data                                               | Accuracy: 77.7%                            |
| SCRIBER                      | Minimizes the cross-predictions                                       | Accuracy: 0.821 +/- 0.007                  |
|                               |                                                                        | Precision: 0.332 +/- 0.013                 |
|                               |                                                                        | AUC: 0.715 +/- 0.013                       |
| SLSTM Network with DLPred   | A New penalization factor identified and joined with the loss function which decreases the non-interaction site loss | Accuracy: 73.68                            |
|                               |                                                                        | Precision: 0.418                           |
|                               |                                                                        | F-Measure: 47.61                           |
|                               |                                                                        | AUC (%): 81.81                             |
The above comparison stated in Table 2 identifies that the six computational techniques or methods have their own merits and demerits. When the merits of the individual methods are integrated into one, then the resulting algorithm has higher efficiency.

4. Conclusion

Analyzing and identifying the interaction that occur among proteins plays an important part in biological functions and recognizing the drug targets. Data obtained from the wet-lab methods has functional information and supports the prediction of PPI networks. PPI networks can be designed from number of principles which include module-based and pairwise interaction prediction. However, Experimental methods are time-consuming, labor-intensive and expensive. ML based computational methods are widely used to predict PPI. In this paper, a survey on Prediction of PPI for six different methods along with their performance metrics is analyzed. Each technique that was discussed under computational methods has its own advantage. Data obtained from Gene expression and protein interaction can also be used to improve the prediction of existing models.

References

[1]. Umbrin, Hina, and Saba Latif. "A survey on Protein Interactions (PPI) methods, databases, challenges and future directions." 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). IEEE, 2018.

[2]. Ding, Ziyun, and Daisuke Kihara. "Computational Methods for Predicting Protein Protein Interactions Using Various Protein Features." Current protocols in protein science 93.1 (2018): e62.

[3]. Browne, Fiona, et al. "From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions." Advances in Artificial Intelligence (16877470) (2010).

[4]. N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao, “Information assessment on predicting protein-protein interactions,” BMC Bioinformatics, vol. 5, article 154, 2004 [5]. C. L. Myers and O. G. Troyanskaya, “Context-sensitive data integration and prediction of biological networks,” Bioinformatics, vol. 23, no. 17, pp. 2322–2330, 2007.

[6]. O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, “A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae),” Proceedings of the National Academy of Sciences of the United States of America, vol. 100, no. 14, pp. 8348–8353, 2003

[7]. R. Jansen, H. Yu, D. Greenbaum, et al., “A Bayesian networks approach for predicting protein-protein interactions from genomic data,” Science, vol. 302, no. 5644, pp. 449–453, 2003 [8]. C. von Mering, R. Krause, B. Snel, et al., “Comparative assessment of large-scale data sets of protein-protein interactions,” Nature, vol. 417, no. 6887, pp. 399–403, 2002.

[9]. B. A. Shoemaker and A. R. Panchenko, “Deciphering protein protein interactions— part I: experimental techniques and databases,” PLoS Computational Biology, vol. 3, no. 3, article e42, 2007

[10]. Zeng, Min, et al. "Protein–protein interaction site prediction through combining local
and global features with deep neural networks." Bioinformatics 36.4 (2020): 1114-1120.
[11]. Zhang, Long, et al. "Protein–protein interactions prediction based on ensemble deep neural networks." Neurocomputing 324 (2019): 10-19.
[12]. Wang, Xiaoying, et al. "Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique." Bioinformatics 35.14 (2019): 2395-2402. [13]. Zhang, Jian, and Lukasz Kurgan. "SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences." Bioinformatics 35.14 (2019): i343-i353. [14]. Xie, Zengyan, Xiaoya Deng, and Kunxian Shu. "Prediction of protein–protein interaction sites using convolutional neural network and improved data sets." International journal of molecular sciences 21.2 (2020): 467.