International Journal of Multidisciplinary Engineering in Current Research
Volume 6, Issue 5, May 2021, http://ijmec.com/

A Survey on Protein-Protein Interaction Network Methods and Challenges

RamalingamShanmugam
*Chief Consultant, ArutJothi Siddha Clinic, Padiyanallur,
Redhills, Chennai, Tamil Nadu, India

ABSTRACT

Biological processes may be studied using a bioinformatics system. When it comes to illness detection and analysis, bioinformatics is the most important field. An analysis of the structure of a protein sequence and its visualisation in 3D structure is done using PPI (Protein Protein Interactions). The identification of cancer-causing proteins via the PPI network has been studied using a variety of methods. For the development of drugs for a specific human ailment, protein interactions in Human Interaction Networks are employed. When analysing enormous datasets, there are several benefits and downsides to using various methodologies. As a consequence, we obtain varied analyses and outcomes. To incorporate multiple machine learning and deep learning approaches as well as 3D visualisation in data modelling and analysis, it is employed for feature directions. Different techniques of analysis have been examined here for potential future approaches.

Keywords

1.In bioinformatics and protein-protein interaction, machine learning, deep learning, 3D visualisation, proteins, network design, and protein-protein interactions.

2.INTRODUCTION

One of the most important areas of information technology study is the identification of biological data that can be properly analysed. Genomic information that can be utilised to find the disease's genes via the use of information relevant to the disease's biology

Knowing what the medications are can be done. Biological and computer science are intertwined in this field of study. In order to comprehend the relationship between molecules on a huge data set, several biological terminologies and information methods are used. This is also known as a management information system for biological sciences. It is possible to preprocess enormous amounts of biological data using a variety of computer-based and mathematically-based methods in molecular biology. These technologies preprocess enormous databases of complicated data at a rapid pace to conduct complex biological sequences.

In bioinformatics research, the primary purpose is to retain the data in such a manner that it can be accessed in a simple format as needed, and to retrieve new information as soon as it is created. That's why we need to create a programme that can analyse the information, then provide it to the user in an understandable fashion. The use of PPI networks in bioinformatics research has received much attention and discussion. Protein-protein interaction (PPI) networks may be defined by the biochemical or electrostatic forces that link two or more proteins. To that end, the primary goal of this work is to present the many methodologies[22] used in the identification of breast cancer by purification of gene expression. Tools for breast cancer PPI network analysis are discussed in this research.

The majority of cellular functions are performed by several proteins, not just one. So many people these days

Data on protein-protein interactions[1] may be found online and analysed by anybody. Analysis of complex networks produced by the combination of two or more proteins requires modern computational and mathematical approaches. Analyzing the protein sequence with more precision while saving time and money is the goal of this approach.

The proteome's interaction and functional structure can only be improved with the aid of an effective framework. It is the primary purpose of PPI's in proteomics to uncover the molecular functioning underpinning biological proteins in order to comprehend human illnesses at the cellular level. We need to study the PPI network in order to determine the function of a protein. These analyses are used to prepare the target medication. There are a lot of unbound proteins in huge numbers of cells in the PPI network[5].

The Importance of Protein Interactions

Among the more than 50000 different proteins that genes generally make are those found in the body's molecules and cells, which are also known as molecules or cells. There are 22,000 different protein-coding genes in the human body. One of the most important aspects of the cell's activities is its ability to transmit messages from the extracellular environment, which is facilitated by PPI signal molecules.

This occurs when a protein is transporting another protein across membranes.

3) Cell metabolism is the result of enzyme interactions, which creates tiny macromolecules.

4) Several gene disease-related interactions take place during muscle contraction.

Methods for analysing PPI networks:

Use a biological database to gather information

Secondly, use the tools to forecast.

Analyze and visualise the data.

This may be done with the use of a bioinformatics technique.

The network graph should be designed.

Figure out how protein modules differ from one another.

Find the link between the protein modules.

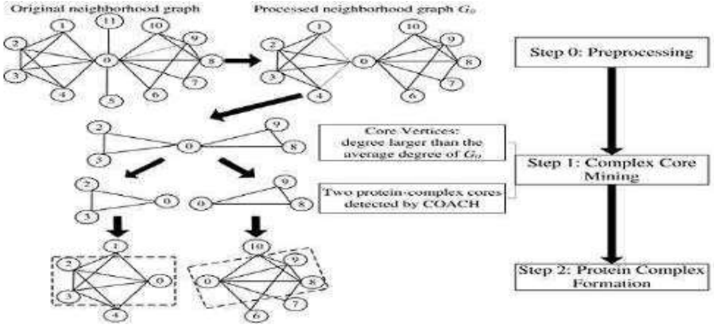
Original neighborhood graph
Processed neighborhood graph G_0
Step 0: Preprocessing
Core Vertices: degree larger than the average degree of G_0
Step 1: Complex Core Mining
Two protein-complex cores detected by COACH
Step 2: Protein Complex Formation

Fig. 1 shows the results of COACH's protein complex core detection.

PROTEIN PROTEIN INTERACTION METHODS FOR ANALYZING PROTEIN SEQUENCES

3.EXPERIMENTAL METHODS

Procedures like CO-IP (co-immune precipitation) are used to discover protein interactions in experimental methods.

YEAST-TWO-HYBRID SCREENING

This technique may be used to discover the interactions between two proteins. This method divides yeast cells into prey and bait, with no interaction between the two if no transcription of a reporter protein is detected. In order to determine whether a relationship exists between two proteins, a consequent reporter protein expression must be taken into account. There are a lot of false positives and false negatives with this procedure. Physical interactions may differ between Yeast and humans because protein folding and expression are not same in both species.

AFFINITY PURIFICATION OR MASS SPECTROMETRY

An affinity purification or mass spectrometry is used to find the constant interactions between proteins and detect the functional relationship between the protein sequences. Purification of the protein of interest is the first step in this process. which has already been taken. Typically, the protein of interest is expressed as a cell in this technique. Both the amount and the quality of PPI may be assessed by using this technique [24].

4.COMPUTATIONAL METHODS

The PPI networks may be analysed and evaluated in a variety of ways. Data mining, soft computing, and biological structures all have some influence on these techniques. The following are a few of the most popular approaches:

PHYLOGENETIC PROFILING

Phylogenetic profiling is a technique that identifies patterns in protein families that are shared by a large number of species. This technique posits that if two proteins interact, they must have co-evolved through time. Each protein was given its own phylogenetic profile[5]. We can use this strategy to figure out how previously unidentified enzymes work.

PREDICTION OF PROTEIN PAIRS USING SIMILAR PHYLOGENETIC TREES

An interaction between two or more proteins is detected using phylogenetic trees constructed from a protein dataset[5]. Proteins may be traced back in time by using pylogenetic trees. The phylogenetic technique for predicting protein interactions follows the mirror tree method. Co-evolution between proteins is identified and the degree of similarity is calculated[8].

ROSETTA STONE METHOD

An organism's single-domain protein may fuse with another's multi-domain protein to create a new protein in the other organism. Gene fusion is another term for this procedure. This technique identifies the most common fusion activity. It is based on this phenomenon that proteins are mixed and tested to see whether they interact with one other or not. Homologous qualities are detected when they come into contact.proteins[22]. There is a constraint to this method: it can only be used on proteins that have a domain organisation.

HOMOLOGOUS STRUCTURE INDICATIONS OF THE INTERACTION

This model takes into account both the sequence structure of a protein that is already known as well as proteins that are close neighbours to that protein. Complete linkage clustering is used to compare protein clusters based on their similarity.Volume VIII, Issue I, January/ 2019 Page No:976

5.In order to create a matrix, our programme measures the distance between the clusters.

Then look at the clusters that are closest to each other. Method of classifying
6. PPI networks often use the classification approach. Here is a data mining methodology, and this is how it's done. Classifying protein pairings into interacting and non-interacting pairs is done in this manner. Most
8 used classifiers are Support Vector
10.Machine(SVM) and Random Forest
12.Decision(RFD).
Structural patterns may be identified.
As a result of the Protein Data Bank, a predefined set of protein interfaces is used in this method: (PDB
Interfaces are defined as polypeptide pairs that are smaller than the Van der Waals radius of their respective atoms in this database.

INVESTIGATIONS INTO DOMAIN-PAIR EXCLUSIONS

Non-specific, indiscriminate interaction may be identified using Bayesian approaches; however, identifying connections across specified domains is more challenging. Domain interactions are discovered using E-Scores.
25.–Analysis of exclusion by pair. Interactions between proteins within a domain are very rare if E-score is low; conversely, a high E-score suggests that two domains may interact. False positives and false negatives are not taken into account when analysing experimental data in this manner[23]. THE PROBLEM OF SUPERVISED LEARNING
26.An important data mining approach is supervised learning. Using known protein interactions as input, it is possible to predict the PPI network by using the function of supervised learning to determine whether or not there is an interaction between two proteins. [9][10].

28 **ASSESSMENT OF THE PPIS EXAMPLE DATABASE**

Depending on the protein interaction network, a single interaction or hundreds of interactions might occur. All of this information was stored in databases that were created expressly for storing biological data. In order to offer users with up-to-date and accurate interaction data, these databases are regularly updated. The amount of biological databases grows exponentially every day. Primary databases, meta-databases and prediction databases[12] are the three basic types of databases. Databases that make up the majority of the system
All experimentally confirmed protein interactions are stored in primary databases. Molecular Interaction Database (MINT), MIPS Mammalian Protein Protein Interaction Database (MIPS – MPPI), the Biological General Repository for Interaction Datasets (BioGRID), the IntAct Molecular Interaction Database, the MIPS Protein Interaction Resource on Yeast (MIPS Mpact), and the Human Protein Reference Database (HPRD) are some examples of these types of databases. META-DATABASES
Additional data is stored in these databases in addition to the information found in the core databases. Protein Interaction Network Analysis (PINA) and Agile Protein Interaction Data Analyzer are examples of meta databases (APID). APID is a handy tool for displaying protein sequences in a graphical form [26]. APID employs apinBrowser, a multi-stage tool, to provide graphical analysis. TABLE OF AGREEED-UPON **STRATEGIES**
To do a search and aggregate the results, this database requires just a minimal amount of time to calculate and combine. CPDB employed

a mapping standard that supports the PPI analysis tool to filter the dataset. It has a large number of alternative outputs and also accepts a large number of input datasets. Scalable and dynamic graphs are supported by the CPDB.

generating the outcome as the goal. TRANSACTION DATABASE FOR CHEMICALS (MINT)

The Java browser must be installed in order to run any query against this database. Short summaries of results and database use are provided by the MINT databse UI. As a result of this, the MINT database does not have graphical representation. The HDOCK SERVER

Third party applications, docking algorithms, and scoring systems are all included in the HDOCH[18] collection. It accepts input from both the protein's sequence and structure. The homologous protein sequence is discovered by running a sequence similarity test against the Protein Data Bank (PDB). Then, look for sequences that appear in both datasets and remove them from the analysis.

The AUTO ENCODER IS STACKED

Unlabeled input in artificial neural networks is processed using an unsupervised learning approach called Stacked Auto Encoder. Unlabeled data is transformed into hidden structures using an auto encoder. The encoder accepts 'X' as an input and generates 'X' as output. Each layer is trained independently, but the preceding layer's output is taken into account as a parameter[25] before creating output. EPSILON-CP[17]

Using this approach, you may figure out the structure of a given object in advance. Sequence-based information combined with evolutionary data from numerous sequence alignments and physicochemical data from structure prediction methods is used in this way of contacting the atoms and molecules in a biological system. When compared to CASP11, we were able to improve our prediction accuracy by incorporating data from numerous sources. Stacking and neural network training are used in EPSILON – CP approach to discover the association between the data. INTERACTOME OF PROTEOME SCALE

Network maps are described to find out how the resulting interactions affect the biological system's functioning. Some researchers believe that disease proteins are clustered together in topological modules, where they interact with each other more often than with other proteins outside of this area. This concept is known as the disease module hypothesis. We will get a better knowledge of the interplay between proteins and isoforms by integrating quantitative, geographical and temporal information into our mapping efforts.

TOOLS APPLIED IN PPI NETWORK ANALYSIS

The protein-protein interaction network may be analysed using a variety of techniques. Analyzing tools include:

Cytoscape is a free and open-source programme for creating and analysing PPI networks of any size. Multiple protein networks may be combined into one. It creates a three-dimensional model of the PPI network. It may be used to analyse the PPI network using specified tools as well as tools that are already in use.

Protein pathways may be discovered using PathBLAST, a PPI network analyzer and search tool for comparing PPI networks across various species.

PPI network analysis may also be done using APID (Agile Protein Interaction Data Analyses), another tool.

Importing data from the BioGrid Database was done using the BiogridPlugin2 plugin.

To import a particular PPI network, we may use filters.

33. CONCLUSION

Data mining, soft computing, and experimental approaches such as Y2H and co-IP may be used to analyse protein-protein interactions. Protein-protein interaction datasets may be found in a variety of databases. Some of these databases are based on protein structure, while others are based on sequence. When compared to the machine learning approaches mentioned in this work, the accuracy of experimental methods is lower. Analyzing on a huge scale is possible.

protein datasets by using deep learning techniques. This survey paper gives overview of methods available in PPI network analysis.

32. REFERENCES

[1] Antonio Mora, Katerina Michalickova and Ian M Donaldson, “A survey of protein interaction data and multigenic inherited disorders”, BMC Bioinformatics, vol. 14, pp. 1-7, 2013.

[2] Fiona Browne, 1 Huiru Zheng,1 HaiyingWang,1 and Francisco Azuaje, “From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions”, Advances in Artificial Intelligence, Hindawi, vol. 2010, pp. 1-15, 2010.

[3] Tord Berggård1, Sara Linse1 and Peter James2, “Methods for the detection and analysis ofprotein–protein interactions”,Proteomics, vol.7 2007, pp. 1-10, 2007.

[4] Ulrich Stelzl, Uwe Worm, MaciejLalowski, Christian Haening,”A human protein-protein interaction network: A resource for annotating the Proteome”, Cell Press, Vol.122, issue 6, pp. 957-968, 2005.

[5] Joan Planas-Iglesias,JaumeBonet,Javier Garcia- Garcia,Manuel A. Marin-Lopez,ElisendaFelliu and BaldoOliva, ”Understanding Protein–Protein Interactions Using Local Structural Features”, JMB Article, pp.1-3, 2013.

[6] Marco Wiltgen, “Structural Bioinformatics: From the Sequence to Structure and Function”,CurrentBioinformatics,pp. 1-2, 2009.

[7] Hung Xuan Ta and Liisa Holm, “Computational approaches for predicting protein interaction networks: The wiring of protein”, The Biochemical Society,pp.1- 3,2011., S. Mullender

[8] Sun, Tanlin, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm." BMC bioinformatics, vol. 18.1, pp.277, 2017.

[9] Akkoyun, Emrah, and Tolga Can, "Parallelization of the functional flow algorithm for prediction of protein function using protein-protein interaction networks." High Performance Computing and Simulation (HPCS), International Conference on. IEEE, 2011.

[10] Hu, Lun, "Efficiently predicting large- scale protein-protein interactions using MapReduce." Computational Biology and Chemistry 2017.

[11] Sun, Peng, et al. "Towards Distributed Machine Learning in Shared Clusters: A Dynamically-Partitioned Approach." Smart Computing (SMARTCOMP), IEEE International Conference on. IEEE, 2017

[12] Tovchigrechko, Andrey, and Ilya A. Vakser. "GRAMM-X public web server for protein–

protein docking." Nucleic acids research, vol. 34.suppl 2,pp.W310-W314,2017.

[13] May, Andreas, and Martin Zacharias. "Protein–protein docking in CAPRI using ATTRACT to account for global and local flexibility." Proteins: Structure, Function, and Bioinformatics , vol. 69.4,pp.774-780, 2017.

[14] Zacharias, Martin. "Protein–protein docking with a reduced protein model accounting for side-chain flexibility." Protein Science , vol.12.6,pp. 1271- 1282,2003

[15] Carter, Phil, "Protein–protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI." Proteins: Structure, Function, and Bioinformatics ,vol.60.2 ,pp. 281- 288,2005.

[16] Smith, Graham R., and Michael JE Sternberg. "Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial." Proteins: Structure, Function, and Bioinformatics,vol.52.1,pp.74- 79,2003.

[17] Stahl, Kolja, Michael Schneider, and Oliver Brock. "EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction."