Protein Complex Detection: A Study

D. Ramyachitra
Assistant Professor
Department of Computer Science
Bharathiar University
Coimbatore, India
jaichitra1@yahoo.co.in

D. Banupriya
M.Phil Research Scholar
Department of Computer Science
Bharathiar University
Coimbatore, India
d.banupriyamphil@gmail.com

***Abstract*-Protein complexes play an important role in cell biology. Identification of protein complexes from protein-protein interaction networks is the first step in understanding the cell functions. Computational advances for noticing protein complexes from protein interaction facts and figures are helpful complements to the limited experimental methods. The increasing amount of accessible protein–protein interaction (PPI) data endows us to evolve accurate and scalable computational procedures for protein complex detections. Although, experimentally determined protein complex data, particularly of those involving more than two protein-protein interactions, are relatively restricted in the current experimental methods. This paper gives a study on several new computational methods, tools, and protein-protein interaction databases for protein complex detection and the performance metrics are used to evaluate these algorithms.**

Keywords-Protein complex, Cell biology, Protein-protein interaction, Clustering techniques, Tools, Databases

I. INTRODUCTION

Biologically protein complex is a group of two or more associated polypeptide chains that densely interact with one another. These complexes are used in many biological processes and they perform the vast amount of functions such as, cell cycle control, differentiation, signaling, protein folding, translation, transcription, post-translational modification, control of gene expression, inhibition of enzymes, antigen–antibody interaction and transportation [1, 2]. For example, the complex RNA polymerase II transcribes genetic information into messages for ribosomes to produce proteins. Another example is complex Proteasome core particle involved in the degradation of proteins, which is an essential process within the cell [3]. Observations show many genetic diseases caused by the same or similar diseases tend to lie close to one another in a network of protein–protein or functional interactions. Such interaction relatedness could be exploited by measuring the evolutionary relationships between genes to help in the finding of novel protein complexes which ultimately leads to the finding of disease genes.

There exist many different topologies of interaction among proteins considering the biochemical nature of the interactions. The common interaction involves the direct contact of molecules, but proteins may also interact through a medium or even through the exchange of ions [4].

Commonly protein complexes are identified from the protein-protein interaction networks. High-throughput detection methods include yeast two-hybrid screening and affinity capture mass spectrometry that produce large amount of protein-protein interaction data. It is desirable to use this data to predict the protein complexes. Development of a generic computational algorithm for protein complex assembly is challenging mainly due to the variety of topological connotations of protein complexes.

There are several methods and algorithms to predict the protein complexes from PPI data. Li et al [5] reviews the state-of-the-art techniques to mine protein complexes from protein interaction networks and then describes classical graph clustering for complex mining methods and some new emerging techniques. Moschopoulos et al. [6] evaluated four different clustering algorithms using six different interaction datasets. Recently Sriganesh Srihari [7] made a survey on computational methods developed till date for the identification of protein complexes from PPI networks.

In this paper, we study new emerging techniques and tools to predict the protein complexes from protein-protein interaction databases. The growing PPI database helps both biological and computational scientists to predict gene functions, functional pathways, protein complexes and improve the diagnosis and treatment of diseases [8-16].

The remaining sections of the paper are organized as follows. Section 2 deals with the challenges in protein complex detection. Section 3 describes various types of protein complex detecting techniques. Section 4 provides some tools. Section 5 discusses the protein-protein interaction databases. Performance measures are discussed in section 7.

II. CHALLENGES IN PROTEIN COMPLEX DETECTION

General observations show that the following are the fundamental problems to detect the protein complexes in PPI network.

A. Noise

Protein interaction data are very noisy. Instead of traditional unweighted graph the weighted and filtered graph to represent a PPI network is proven to be an effective way. Then the next problem is how to obtain the reliable interactions in PPI data.

B. Multiple Interactions

Proteins may participate in multiple protein complexes. As a result, protein complexes may overlap and sometimes that may not consider the properties and features of protein complexes.

C. Representation

Most existing clustering methods assume protein complexes as dense subgraphs, which is not always true for the protein complexes in the PPI networks [17]. In addition, all kinds of topologies present in protein complexes, and tremendous variation of the sizes of protein complexes pose a further problem for identifying the specific topologies.

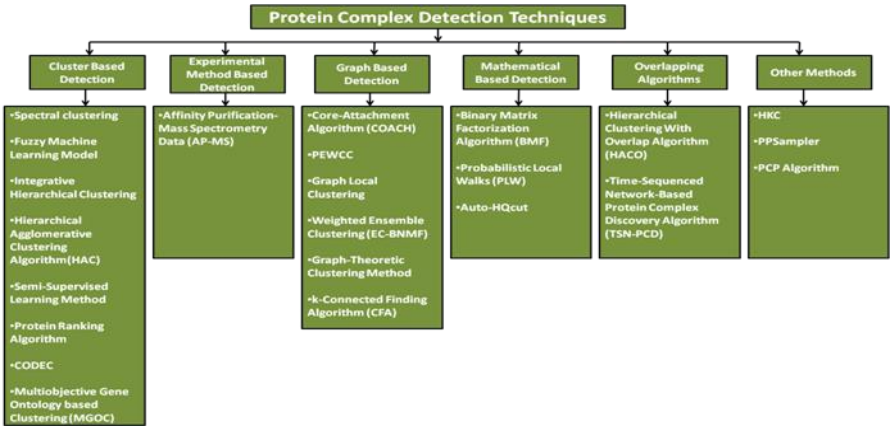

Figure 1. Classifications of protein complex detection techniques

III. PROTEIN COMPLEX DETECTION TECHNIQUES

A. Cluster Based Detection

a) Spectral clustering

G. Qin, L. Gao [18] proposed this method to detect the protein complexes from PPI network. It mainly focuses on two issues (i) constructing similarity graphs and (ii) determining number of clusters. G. Qin, L. Gao used four similarity graphs based on adjacency matrix, namely adjacency similarity, common neighbor similarity, transmission similarity and commute similarity to construct the similarity graphs. Maslov et al. [19] further found that most interactions occur between highly connected (i.e. hub) and lowly connected proteins. Based on this, the number of complexes is determined by the number of hub nodes. Using biological knowledge and scale-free networks the number of clusters is determined. Figure 2 shows the scale-free network with hub nodes. This method requires pre-process step before clustering because of noise in PPI data.

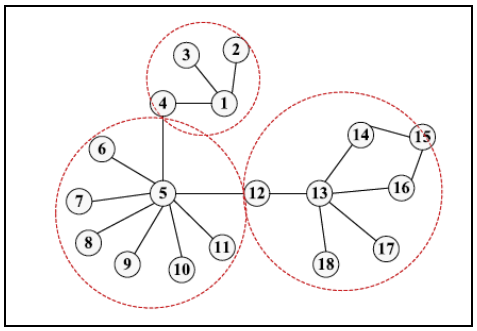

Figure 2. Scale-free network [18].

b) Fuzzy Machine Learning Model

Xu et al. [20] applied this algorithm to classify the protein complex from candidate subgraph. To improve the performance of identifying protein complexes it incorporates Genetic Algorithm Fuzzy Naïve Bayes (GAFNB) as a filter in protein complex identification. This method first detects the candidate proteins from the existing protein complex detection methods. Secondly, the candidate protein complexes are filtered by GBFNB method.

c) Integrative Hierarchical Clustering

M. Wu et al. [21] proposed this approach to diagnose the problem of noise (e.g, false negative and false positive interactions) in PPI data and to detect the protein complexes from various data sources such as PPI data, gene expression profiles, GO terms, and TAP-MS data.

This method consists of following steps:

i. Identify the protein complexes from variety of sources.
ii. Calculate affinities between the proteins to show how they interact.
iii. Using support vector machine to calculate weights for each source of data.
iv. Calculate sum of weighed scores and results as final score matrix.
v. Use the hierarchical clustering algorithm on final score matrix to generate the predicted protein complex.

d) Hierarchical Agglomerative Clustering Algorithm (HAC)

L. Yu et al. [22] presents PPI networks based on within-module and between-module edges of subgraphs and degree distribution. Then they develop a hierarchical agglomerative clustering algorithm (ADHAC) to identify the protein complex in genome scale protein-interaction network. Using this method PPI networks are characterized by hierarchical modularity. This framework can identify complexes with high biological significance and more valuable information regarding cellular function. ADHAC can discover both dense and sparser biologically significant complexes.

Z.Xie et al. [23] used co-complex scoring matrix method to predict the protein complex from AP-MS data. It is an unsupervised method. In the process of deriving a PPI network from the AP-MS data, primary information that two proteins are unlikely to be co-complexed is eliminated. Then a hierarchical clustering algorithm is applied directly on this merged score matrix.

Gavin et al., [24] created a socio-affinity scoring system to weight logical interactions between pairs of proteins in AP-MS data. There are several clustering methods have been used to cluster the PPI networks. Collins et al. [25] developed a scoring system and applied hierarchical clustering methods to weighted networks to derive complexes.

e) Semi-Supervised Learning Method

Shi et al. [26] proposed semi-supervised learning method to detect the protein complex from noisy protein-protein interaction data. This method use a multi-layer neural network based semi-supervised method to detect the hidden protein complexes. This algorithm first gains the weights for different features from the limited known protein complexes. Then it will assign a score to subgraph in the graph. With a setup threshold, it could label some of the subgraphs as complexes. Recursively, it will find all protein complexes in the PPI network.

f) Multiobjective Gene Ontology based Clustering (MGOC)

Sumanta Ray et al. [27] used NSGA-II [28] as underlying multi-objective algorithm for finding protein complexes and functional modules. The searching is performed on a number of objectives. Here they used three different classifications such as biological process, cellular component, molecular function of the Lin [29] measure and some graphical properties of protein-protein interaction network are used as objective function.

Sumanta Ray et al. [30] proposed Protein Complex Detection using Multi-objective Evolutionary Approach based on Semantic Similarity (PROCOMOSS) to optimize both GO-semantic similarities based metric and graph based density metric concurrently to find dense protein complexes containing functionally similar proteins and then used three semantic similarity measures such as Lin, Jiang and Conrath and Kappa's measure [31-33] to compute the similarity measures. It can able to group similar proteins as clusters.

g) Protein Ranking Algorithm

N. Zaki et al. [34] proposed ProRank to detect the protein complexes based on their importance in the network and relationships between them. The essential protein is expected to interact and to have high similarity to most proteins within a complex. Sequence similarity often suggests evolutionary relationships between protein sequences which are important for inferring similarity of structure or function [35]. This method is similar to Google's page rank algorithm which is used to identify the important proteins in the network.

The ProRank algorithm consists of five steps: (i) Pruning (ii) Filtering (iii) Protein Similarity Calculating (iv) Protein Ranking (v) Complex Detection.

This method is based on PageRank algorithm. Second, incorporates evolutionary relationships between proteins. Third, it uses strong methods to analyze the topology of network which helps to remove noise and unreliable interactions in the network.

h) CODEC

G.Geva and R.Sharan [36] used CODEC method to cluster the AP-MS data. It displays the AP-MS data as bipartite graph where one set of vertices corresponds to prey protein and other corresponds to bait proteins. Edges show the interaction between these two kinds of proteins. Then it detects the protein complex from that graph.

B. Graph Methods

a) Core-Attachment Algorithm (COACH)

Gavin et al. [24] have revealed that a complex consists of a core component and attachments to predict protein complexes by discovering the core and attachments. A core component is the 'heart' of a protein complex and has relatively more interactions among proteins, while each attachment protein binds to a core to form a biologically meaningful complex.

Leung et al. [37] proposed this method to directly predict protein complexes from the PPI network. The key idea behind his approach consists of three main steps: (1) predict core components; (2) identify attachments for the cores and eliminate insignificant cores; and (3) compute and rank the significance of predicted complexes.

Wu et al. [38] introduced a novel method called CACHET to detect protein complexes with Core-AttaCHment structures directly from bipartitE TAP data. CACHET selects protein-complex cores from bicliques and then simultaneously assembles all the attachments into cores to form the complexes. If it regard cores as seed graphs, this approach for adding proteins into seed graphs to form protein complex. CACHET exploits three state-of-the-art reliability measurements, such as Socio-Affinity (SA), Purification Enrichment (PE), and Dice Coefficient (DC), which are used to assess the reliability of bait-prey relationships more accurately.

X. Ma, L. Gao [39] introduced this algorithm to find the cores and attachments from the network. It first characterized the core component of a protein complex by the graph communicability. Then constructed a virtual network whose size is equal to that of the original PPI network. In that case, it transformed the detection of core in the original network into a well-known all-clique problem in the virtual network. Finally, the attachments were included to form protein complexes.

Srihari et al. [40] coupled core-attachment method to MCL for finding complexes from weighted PPI networks. MCL-CAw algorithm consists of two phases. In the first phase, divide the PPI network into multiple dense clusters using MCL. In the second phase, it refines these clusters to obtain meaningful complexes.

The MCL-CAw algorithm consists of the following steps:

i. Cluster the PPI network using MCL
ii. Categorize the core proteins within clusters
iii. Filtering noisy clusters
iv. Recruiting proteins as attachments into clusters
v. Extracting out complexes from clusters
vi. Ranking the predicted complexes

b) PEWCC

In addition to improving graph mining techniques, it is necessary to obtain high quality benchmarks by assessing protein interaction reliability. Zaki et al [41] proposed a novel method for assessing the reliability of interaction data and the concept of weighted clustering coefficients as a measure to define which subgraph is the closest to the maximal clique. The clustering coefficient of a vertex in this case is the density of its neighbourhood [42]. Here they used the PE-measure, a new measure for protein pair's interaction reliability which reduces the noise in the PPI.

c) Graph Local Clustering

Y.Qi et al. [43] used graph topology to model each complex subgraph by a probabilistic Bayesian network (BN). The main objective is to recover the protein complexes from the undirected PPI graph. Rather than the clique assumption, it obtains several properties from known complexes, and uses these properties to search for new complexes and learns the importance of each of the features. This method relies on real complexes; it does not assume any past model for complexes. The main strength of this method is that it considers the possibility of multiple factors defining complexes in protein interaction graphs.

d) Weighted Ensemble Clustering (EC-BNMF)

Ou-Yang et al. [44] proposed Bayesian Nonnegative Matrix Factorization (NMF)-based weighted Ensemble Clustering algorithm (EC-BNMF) to detect protein complexes from PPI networks. EC-BNMF can integrate multiple clustering results features of a PPI network and produce a more accurate and informative clustering. Also, EC-BNMF allowed proteins to be shared among complexes, which is much closer to the reality.

EC-BNMF consists of two phases firstly, extracts useful information from several base clustering results and generates an ensemble PPI network. Secondly, Bayesian NMF-based ensemble clustering is used to detect protein complexes from the ensemble PPI network.

e) Graph-Theoretic Clustering Method

S.H.Jung et al. [45] developed a network model that incorporates interaction information drawn from protein domain data. This method use graph-theoretic method to find the protein complexes. The network model, simultaneous protein interaction network (SPIN) captures different sets of non-competitive mutual exclusion interactions (MEIs) which are extracted from the original PPIN. After creating SPINs, naive clustering algorithm is applied to the SPINs for protein complex predictions. Redundant proteins in predicting protein complexes are excluded from the network.

f) k-Connected Finding Algorithm (CFA)

Habibi et al. [46] proposed k-Connected Finding Algorithm (CFA) to find the protein complexes based on k-connected subgraphs. The union of all connected subgraphs forms the candidate clusters of proteins. The clusters which contain less than four proteins and clusters having large in diameter are filtered and removed.

C. Mathematical Methods

a) Binary Matrix Factorization Algorithm (BMF)

Tu et al [47] proposed a binary matrix factorization (BMF) algorithm based on Bayesian Ying-Yang (BYY) learning [48, 49] to predict protein complexes from PPI networks. The BMF represents the binary adjacent matrix of the PPI interaction graph as a result of two low rank matrices with binary entries. The clusters consist of proteins that share similar interaction patterns. The algorithm has the following merits: the input of the known cluster number required by most of the existing BMF algorithms is not necessary and BYY-BMF has no dependence on any parameters or thresholds.

b) Probabilistic Local Walks (PLW)

Wong et al. [50] designed a novel method called Probabilistic Local Walks (PLW) which clusters regions in a PPI network with high functional similarity to find protein complex cores with high precision and efficiency in $O(|V| \log |V| + |E|)$ time. This approach is able to detect the less dense protein complexes and it used a seed selection strategy and devises a topological measure called common neighbour similarity to estimate the functional similarity between two proteins. Based on these PLW performs probabilistic local walks efficiently to excavate protein complex cores by identifying areas of high common neighbour similarity.

c) Auto-HQcut

Lei et al. [51] developed a random walk based algorithm that converts the PPI network into similarity matrix which is then used to construct weighted networks. Two proteins sharing some high-order topological similarities interacts with each other and be involved in the same biological processes. Using the reconstructed weighted network, it measures the interactions of all protein pairs using real value which differs from connected/non connected measure in the PPI network. This method can identify the uncovered significant biological interactions and helps to reduce noise in the network. Then they used a parameter free modularity based community finding algorithm (Auto- HQcut) to identify protein complexes from PPI network by optimizing the modularity function.

D. Overlapping Methods

a) Hierarchical Clustering With Overlap Algorithm (HACO)

Wang et al. [52] proposed HACO, a hierarchical clustering with overlap algorithm, to reconstruct complexes. It used to build the Complex-Net, an interaction network of proteins and complexes, in order to study the higher-level organization of complexes.

b) Time-Sequenced Network-Based Protein Complex Discovery Algorithm (TSN-PCD)

Li et al. [53] developed TSN-PCD algorithm to identify protein complexes from the TSNs. As protein complexes are appreciably related to functional modules, a new algorithm DFM-CIN is proposed to discover functional modules based on the identified complexes. This algorithm not only explores the protein complexes and functional

modules but also study their relationships. Figure 6 shows the general frame work of this algorithm.

E. Other Methods

a) HKC

XiaominWang et al. [54] presents a new topology-based algorithm, HKC it mainly uses two important concepts, highest k-core and cohesion. Using these concepts this method detects protein complexes by identifying overlapping clusters in large-scale PPI networks. Based on k-core it selects the node with highest score and degrees as seeds. Expands this seed core to include nodes which are highly possible to form a cluster based on the criteria of node score and cohesion; finally by filtering the clusters, the protein complex is detected.

b) PPSampler

Tatsuke and Maruyama [55] proposed PPSampler based on the Metropolis-Hastings algorithm, in which all participating proteins are generated as a sample based on the probability distribution which is specified by a scoring function. There are three scoring functions f1 is a sum of PPI weights within predicted clusters, f2 is a frequency size of predicted clusters that follow power-law distribution and f3 is the gap between the number of proteins within predicted clusters.

Then Widita and Maruyama [56] have improved the scoring functions, f1, f2, and f3, of PPSampler in order to predict protein complexes more accurately. Firstly scoring function of f1 is improved by replacing the sum of the weights of PPIs within a cluster. The remaining scoring functions, f2 and f3, are also newly modeled, using Gaussian distributions. Secondly, the new entire scoring function is devised as the negative of the sum of the resulting scoring functions g1, g2 and g3. Lastly performs a random walk over the states.

c) PCP Algorithm

H. N. Chua et al. [57] used indirect interactions with FS-Weight to modify the existing PPI network as a preprocessing step to complex prediction. The original PPI network is expanded by including indirect interactions i.e. relationships between protein pairs that do not interact, but commonly shares interaction partners. FS-Weight (functional similarity weight), is then computed for both direct and indirect interactions. Interactions with weights below a threshold are removed. It uses the FS-Weight information during the merging of cliques (clusters). It is more feasible based on reliable PPI networks.

IV. TOOLS

A. IMproving PREdiction of Complexes (IMPRECO)

M. Cannataro et al. [4] used a new complexes meta-predictor which is capable of predicting protein complexes by integrating the results of different predictors. Cannataro presents a distributed architecture that implements the IMPRECO prediction algorithm and demonstrates its ability to predict protein complexes. The proposed metapredictor first invokes different available predictors wrapped as services in a parallel way. In second it integrates the results using graph analysis, and finally evaluates the predicted results by comparing them against external databases storing experimentally determined protein complexes.

B. GIBA

Moschopoulos et al. [58] presented GIBA (named by the first characters of its developers' nicknames). GIBA first applies the MCL or the RNSC clustering algorithm on interaction data and then applies individual or combination of four filtering methods such as a) density, b) haircut operation, c) best neighbour and d) cutting edges to generate the final candidate list of predicted complexes. It provides a user-friendly environment and any user could perform clustering without any difficulties.

C. ProCope

J.Krumsiek et al. [59] presented ProCope java based extensible software package for predicting protein complexes from purification datasets which integrates efficient implementations of the major prediction methods. This package provides a graphical user interface, command line tools for job processing and java application programming interface. ProCope developed an unsupervised bootstrap approach. Apart from the purification experiments it does not require additional training data

D. Cytoscape

Cytoscape [60] is to building open-source network visualization and analysis software. It is used for integrating biomolecular interaction networks with high-throughput expression data and other molecular state information. It supports many use cases in molecular and systems biology, genomics, and proteomics. Cytoscape is most powerful when used in conjunction with large databases of protein-protein, protein-DNA, and genetic interactions. Some of the features are: (i). Loads interaction datasets in many standard formats. (ii). Integrate global datasets and functional annotations. (iii). Performs advanced analysis and modeling using Cytoscape Apps. (iv). It can able to analyze human curated pathway datasets.

V. DATABASES

The set of all binary interactions is spread across different repositories, such as BIND [61], MIPS [62] and DIP [63]. These databases usually contain interaction information determined in wet labs via one or more experimental technologies.

A. Yeast Proteome Database (YPD)

YPD [64] is the first database to describe the complete proteome of any organism. Now the complete genome sequence of yeast is available in YPD and provides description of Saccharomyces cerevisiae. Each yeast protein, characterized either by experimentally or known only as an ORF (open reading frame).

B. Molecular INTeraction database (MINT)

MINT [65] the Molecular INTeraction database, spotlight on experimentally verified protein-protein interactions. Understanding the physical and functional interactions between the cell molecules is one of the main objective in cell biology. MINT does not specialize in

selected model organisms and in the present version contains interactions between proteins from more than 30 different species.

C. IntAct

InAct [66] is a open source database and software used to model, store and analyze the molecular interaction data populated by data either curate from the literature or from direct data depositions.

D. Database of Interacting Proteins (DIP)

The DIP [67] database experimentally determined interactions between proteins. It combines variety of information's from various sources to create a single, reliable set of protein-protein interactions. The data stored in the DIP database with both, manually by expert curators and also automatically using computational approaches that utilize the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data.

E. Biomolecular Interaction Network Database (BIND)

BIND [68] is a database that contains interaction, molecular complex and pathway records. Interaction between two objects is stored in interaction record. Molecular complexes are defined as collections of more than two interactions that form a complex with descriptions. Pathways are defined as collections of two or more interactions that form a pathway.

F. Biological General Repository for Interaction Datasets (BioGRID)

The Biological General Repository for Interaction Datasets (BioGRID) [69] is a public database that archives and disseminates genetic and protein interaction data from model organisms and humans. Complete coverage of the entire literature is maintained for budding yeast (S. cerevisiae), fission yeast (S. pombe) and thale cress (A. thaliana), and efforts to expand curation across multiple metazoan species are underway.

G. Mammalian Protein-Protein Interaction Database(MIPS)

The MIPS [70] Mammalian Protein-Protein Interaction Database is a collection of manually curated high-quality PPI data collected from the scientific literature by expert curators. It includes only data from individually performed experiments since they usually provide the most reliable evidence for physical interactions.

H. Human Protein Interaction Database (HPID)

The Human Protein Interaction Database [71] was designed to provide protein interaction information of humans. It integrates with BIND, DIP and HPRD to provide the human protein interactions and to find proteins from the databases. HPID allows the user to use the protein IDs in EMBL, Ensembl, IM,RefSeq, HPRD and NCBI to search protein interactions of interest.

I. Drosophila Interactions Database (DroID)

DroID [72] assembles variety of sources of protein or gene interactions data into one location. Drosophila interactome data in DroID can be downloaded at the DroID home page. The data also can be searched, graphed, integrated, and downloaded using IM Browser or the DroID Cystoscape plugin. DroID is updated periodically.

J. STRING

STRING [73] is a database of known and predicted protein interactions. The interactions include direct and indirect associations; they are derived from four sources: High-throughput Experiments, Genomic Context, Previous Knowledge and Co expression. STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable.

K. HIV Interaction Database

The goal of this database is to provide yet detailed, summary of all known interactions of HIV-1[74] proteins with host cell proteins, other HIV-1 proteins, or proteins from disease organisms associated with HIV/AIDS. To this end, the database has been designed to track the following information for each protein-protein interaction identified in the literature: (i). NCBI Reference Sequence (RefSeq) protein accession numbers. (ii). NCBI Entrez Gene ID numbers. (iii). Amino acids from each protein that are known to be involved in the interaction. (iv). Brief description of the protein-protein interaction. (v). Keywords to support searching for interactions. (vi). National Library of Medicine (NLM) PubMed identification numbers (PMIDs) for all journal articles describing the interaction.

VI. PERFORMANCE MEASURES

Figure 3. Common Performance metrics.

Precision (p): measures the fraction of the predicted clusters that match the positive complexes among all predicted clusters.

Recall (r): measures the fraction of known complexes matched by predicted clusters, divided by the total number of known complexes.

Ncp in Equation (1) is defined as the number of predicted complexes that match at least one benchmark complex and Ncb in Equation (2) to be the number of benchmark complexes that match at least one predicted complex.

$$Ncp = |\{p|p \in P, \exists b \in B, NA(p,b) \geq \omega\}| \quad (1)$$

$$Ncb = |\{b|b \in B, \exists p \in P, NA(p,b) \geq \omega\}| \quad (2)$$

$$Precision = \frac{N_{cp}}{|P|} \quad (3)$$

$$Recall = \frac{N_{cb}}{|B|} \quad (4)$$

F measure: The F-measure is the harmonic mean of Recall and Precision. It combines the precision and recall scores. It is defined as

$$f = 2\ X\ \frac{Precision\ X\ Recall}{Precision + Recall} \quad (5)$$

Sensitivity: Sensitivity is the fraction of proteins of complex i found in predicted cluster j defined in Equation 6.

$$Sn = \frac{\sum_i {}^{max}_{j}\ T_{i,j}}{\sum_i |b_i|} \quad (6)$$

where b_i is the number of proteins belonging to complex i. Ti;j is the number of proteins shared by bi and cj. A complex wise sensitivity Sn_{co_i} may be defined as

$$Sn_{co_i} = max_{j=1}^{m} Sn_{i,j} \quad (7)$$

Positive predictive value: The positive predictive value is the proportion of members of predicted cluster j which belong to complex i, relative to the total number of members of this cluster assigned to all complexes. It is shown in Equation 8.

$$PPV = \frac{\sum_j {}^{max}_{i}\ T_{i,j}}{\sum_j |\cup_i (b_i \cap c_j)|} \quad (8)$$

Accuracy: The geometric accuracy (Acc) represents a trade off between sensitivity and the positive predictive value and is defined in Equation 9.

$$Accuracy = \sqrt{Sn \times PPV} \quad (9)$$

VII. CONCLUSIONS

Identifying protein complexes is important for biological processes since all biological processes in the cells are carried out through the formation of protein complexes. Protein complex detection still remains a challenging problem and it is important to develop accurate approaches for predicting protein complexes from PPI data. This paper studied new techniques to detect the protein complexes and it provides brief details on complex detecting tools and databases. Also this paper studied some common performance metrics for evaluating the algorithms.

REFERENCES

[1]. Poyatos J, Hurst L " How biologically relevant are interaction-based modules in protein networks?" Genome Biology 2004, 5(11):R93.

[2]. Maggio ET, Ramnarayan K. "Recent developments in computational Proteomics". Trends Biotechnol 2001;19:266–272.

[3]. Gavin A, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M, Hoffman V, Hoefert C, Klein K: Proteome survey reveals modularity of the yeast cell machinery. Nature 2006, 440(7084):631-6.

[4]. Mario Cannataro, Pietro H. Guzzi , Pierangelo Veltri "IMPRECO: Distributed prediction of protein complexes" Future Generation Computer Systems 2010, 26 434_440

[5]. Xiaoli Li, Min Wu, Chee-Keong Kwoh and See-Kiong Ng "Computational approaches for detecting protein complexes from protein interaction networks: a survey" BMC Genomics 2010, 11 S3.

[6]. Charalampos N Moschopoulos, Georgios A Pavlopoulos, Ernesto Iacucci, Jan Aerts, Spiridon Likothanassis, Reinhard Schneider and Sophia Kossida, "Which clustering algorithm is better for predicting protein complexes?" BMC Research 4 (2011) 549.

[7]. Sriganesh srihari and Hon wai leong "A survey of computational methods for protein complex prediction from Protein interaction networks" 2013,2- 1230002.

[8]. Bader G, Hogue C "Analyzing yeast protein-protein interaction data obtained from different sources." Nat Biotechnol 2002, 20:991-7.

[9]. Wang C, Ding C, Yang Q, Holbrook S "Consistent dissection of the protein interaction network by combining global and local metrics." Genome Biol 2007, 8:R271.

[10]. King A, Przulj N, Jurisica I "Protein complex prediction via cost-based clustering." Bioinformatics 2004, 20:3013-20.

[11]. Asthana S, King O, Gibbons F, Roth F "Predicting protein complex membership using probabilistic network reliability." Genome Res 2004, 14:1170-1175.

[12]. Wang J, Li M, Deng Y, Pan Y " Recent advances in clustering methods for protein interaction networks." BMC Genomics 2010, 11(Suppl 3):S10.

[13]. Ulitsky I, Shamir R "Identifying functional modules using expression profiles and confidence-scored protein interactions." Bioinformatics 2009, 25:1158-64.

[14]. Chua HN, Sung WK, Wong L: "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions." Bioinformatics 2006, 22(13):1623-1630.

[15]. Sharan R, Ulitsky I, Shamir R "Network-based prediction of protein function." Molecular Systems Biology 2007, 3:88.

[16]. Friedel C, Krumsiek J, Zimmer R: "Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast." Journal of Computational Biology 2009, 16(8):1-17.

[17]. Watts DJ, Strogatz SH: "Collective dynamics of 'small-world' networks." Nature 1998, 393(6684):409–410.

[18]. Guimin Qin, Lin Gao, "Spectral clustering for detecting protein complexes in protein_protein interaction (PPI) networks" Mathematical and Computer Modelling 2010(52) 2066_2074.

[19]. S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks, Science 296 (2002) 910_913.

[20]. Bo Xu, Hongfei Lin, Kavishwar B Wagholikar, Zhihao Yang, Hongfang Liu "Identifying protein complexes with fuzzy machine learning model" Proteome Science 2013, 11(Suppl 1):S21

[21]. Min Wu, Zhipeng Xie, Xiaoli Li, Chee-Keong Kwoh, and Jie Zheng "Identifying protein complexes from heterogeneous biological data", Proteins 2013; 00:000–000.

[22]. Liang Yua, Lin Gaoa, Kui Li, Yi Zhao, David K.Y. Chiu "A degree-distribution based hierarchical agglomerative clustering algorithm for protein complexes identification" Computational Biology and Chemistry 2011,35 298–307.

[23]. Zhipeng Xie, Chee Keong Kwoh, Xiao-Li and Min Wu "Construction of co-complex score matrix for protein complex prediction from AP-MS data" Vol. 27 ISMB 2011, pages i159–i166.

[24]. Gavin A, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M, Hoffman V, Hoefert C, Klein K: "Proteome survey reveals modularity of the yeast cell machinery." Nature 2006, 440(7084):631-6.

[25]. Fields, S. and Song, O. "A novel genetic system to detect protein–protein interactions.", Nature 1989340:245–246.

[26]. Lei Shi1, Xiujuan Lei, Aidong Zhang "Protein complex detection with semi-supervised learning in protein interaction networks" Proteome Science 2011, 9(Suppl 1):S5

[27]. Sumanta Raya, Moumita Deb, Anirban Mukhopadhyayc "A Multiobjective GO based Approach to Protein Complex Detection" Procedia Technology 2012 4 555 – 560.

[28]. K. Deb, A. Pratap, S. Agrawal, T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II", IEEE Transactions on Evolutionary Computation 6 (2002) 182–197.

[29]. D. Lin, "An information-theoretic definition of similarity, in: Proc.", 15th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

[30]. Anirban Mukhopadhyay, Sumanta Ray* and Moumita De "Detecting protein complexes in a PPI network: a gene ontology based multi-objective evolutionary approachw" Mol. BioSyst., 2012, 8, 3036–3048.

[31]. D. Lin, Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 296–304.

[32]. J. J. Jiang and D. W. Conrath, Proceedings of the International Conference Research on Computational Linguistics, 1997.

[33]. D. Huang, B. Sherman, Q. Tan, J. Collins, W. Alvord, J. Roayaei, R. Stephens, M. Baseler, H. Lane and R. Lempicki, "The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists,", Genome Biol., 2007, 8, R183.

[34]. Nazar Zaki, Jose Berengueres, and Dmitry Efimov "Detection of protein complexes using a protein ranking algorithm" Proteins 2012 80:2459–2468.

[35]. Kuang R, Weston J, Noble WS, Leslie CS. "Motif-based protein ranking by network propagation" Bioinformatics,2005, 21- 3711–3718

[36]. Geva G, Sharan R: "Identification of protein complexes from coimmuno precipitation data." Bioinformatics 2011, 27(1):111-117.

[37]. H.C. Leung, Q. Xiang, S.M. Yiu, F.Y. Chin, "Predicting protein complexes from PPI data: a core-attachment approach" ,J. Comput. Biol. 2009 16 (2) 133–144.

[38]. M. Wu, X. Li, C.K. Kwoh, S. Ng, "A core-attachment based method to detect protein complexes in PPI networks", BMC Bioinform. 10 (2009) 169.

[39]. Xiaoke Ma⇑, Lin Gao "Predicting protein complexes in protein interaction networks using a core-attachment algorithm based on graph communicability" Information Sciences 2012, 189 233–254

[40]. Sriganesh Srihari, Kang Ning, Hon Wai Leong "MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure" BMC Bioinformatics 2010, 11:504.

[41]. Watts DJ, Strogatz SH: "Collective dynamics of 'small-world' networks." Nature 1998, 393(6684):409–410.

[42]. Nazar Zaki, Dmitry Efimov and Jose Berengueres "Protein complex detection using interaction reliability assessment and weighted clustering coefficient, BMC Bioinformatics 2013, 14:163.

[43]. Yanjun Qi, Fernanda Balem, Christos Faloutsos, Judith Klein-Seetharaman and Ziv Bar-Joseph, "Protein complex identification by supervised graph local clustering" Vol. 24 ISMB 2008, pages i250–i258.

[44]. Le Ou-Yang, Dao-Qing Dai, Xiao-Fei Zhang, "Protein Complex Detection via Weighted Ensemble Clustering Based on Bayesian Nonnegative Matrix Factorization", PLoS ONE 2013 8(5): e62158.

[45]. Suk Hoon Jung, Bora Hyun, Woo-Hyuk Jang, Hee-Young Hur and Dong-Soo Han "Protein complex prediction based on simultaneous protein interaction network" Vol. 26 no. 3 2010, pages 385–391.

[46]. Mahnaz Habibi, Changiz Eslahchi, Limsoon Wong: "Protein complex prediction based on k-connected subgraphs in protein interaction network", BMC Systems Biology 2010, 4:129.

[47]. Shikui Tu, Runsheng Chen, Lei Xu "A binary matrix factorization algorithm for protein complex prediction" Proteome Science 2011, 9(Suppl 1):S18.

[48]. Xu L: "Bayesian Ying-Yang System, Best Harmony Learning, and FiveAction Circling. A special issue on Emerging Themes on Information Theory and Bayesian Approach", Journal of Frontiers of Electrical and Electronic Engineering in China 2010, 5(3):281-328.

[49]. Xu L: "Bayesian-Kullback coupled YING-YANG machines: unified learning and new results on vector quantization.", Proceedings of International Conference on Neural Information Processing Beijing, China; 1995, 977-988.

[50]. Daniel Lin-Kit Wong, Xiao-Li Li, Min Wu, Jie Zheng, See-Kiong Ng "PLW: Probabilistic Local Walks for

detecting protein complexes from protein interaction networks" BMC Genomics 2013, 14(Suppl 5):S15
[51]. Chengwei Lei1, Saleh Tamim, Alexander JR Bishop, Jianhua Ruan "Fully automated protein complex prediction based on topological similarity and community structure" Proteome Science 2013, 11(Suppl 1):S9
[52]. Wang H, kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walter T, Krogan NJ, Koller D "A complex-based reconstruction of the Saccharomyces cerevisiae interactome." Mol Cell Proteomics 2009, 8:1361-1377.

[53]. Min Li, Xuehong Wu, Jianxin Wang and Yi Panl "Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data" BMC Bioinformatics 2012, 13:109.
[54]. XiaominWang, ZhengzhiWang, and Jun Ye "HKC: An Algorithm to Predict Protein Complexes in Protein-Protein Interaction Networks" Journal of Biomedicine and Biotechnology Volume 2011, Article ID 480294,
[55]. Tatsuke D, Maruyama O: "Sampling strategy for protein complex prediction using cluster size frequency." Gene 2013, 518:152-158.
[56]. Chasanah Kusumastuti Widita, Osamu Maruyama "PPSampler2: Predicting protein complexes more accurately and efficiently by sampling" BMC Systems Biology 2013, 7(Suppl 6):S14.
[57]. Hon Nian Chua "Using Indirect Protein–Protein Interactions for Protein Complex Prediction" Journal of Bioinformatics and Computational Biology 2008) Vol. 6, No. 3 435–466
[58]. Charalampos N Moschopoulos, Georgios A Pavlopoulos,Reinhard Schneider, Spiridon D Likothanassis and Sophia Kossida "GIBA: a clustering tool for detecting protein complexes" BMC Bioinformatics 2009, 10(Suppl 6):S11
[59]. Jan Krumsiek, Caroline C. Friedel and Ralf Zimmer "ProCope—protein complex prediction and evaluation" Vol. 24 no. 18 2008, pages 2115–2116

[60]. https://www.cytoscape.org
[61]. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW: "BIND–the biomolecular interaction network database." Nucleic Acids Res 2001, 29(1):242-245.
[62]. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, et al: "MIPS: analysis and annotation of proteins from whole genomes." Nucleic Acids Res 2004, 32(Database issue):D41-D44.
[63]. Lukasz Salwinski, Christopher S. Miller, Adam J. Smith, Frank K. Pettit, James U. Bowie, David Eisenberg, The database of interacting proteins: 2004 update, Nucleic Acids Res. 2004 32 (suppl 1) D449_451.
[64]. www.proteome.com/databases/
[65]. mint.bio.uniroma2.it/mint/
[66]. www.ebi.ac.uk/intact/
[67]. dip.doe-mbi.ucla.edu
[68]. bind.ca/
[69]. thebiogrid.org/
[70]. http://mips.helmholtz-muenchen.de/proj/ppi/
[71]. wilab.inha.ac.kr/hpid/
[72]. www.droidb.org/
[73]. string-db.org/
[74]. http://www.ncbi.nlm.nih.gov/projects/RefSeq/HIVInteractions/