ARTICLE

A STUDY ON IDENTIFICATION OF STATIC AND DYNAMIC PROTEIN COMPLEX AND FUNCTIONAL MODULE IN PPI NETWORK

Subham Datta\textsuperscript{1*}, Dinesh Karunanidhy\textsuperscript{1}, J. Amudhavel\textsuperscript{2}, Thanizh Selvam Datchinamurthy\textsuperscript{1}, Subramanian Ramalingam\textsuperscript{1}

\textsuperscript{1}Department of Computer Science and Engineering, Pondicherry University, Puducherry, INDIA
\textsuperscript{2}Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA

ABSTRACT

Background: The availability of large-scale high-throughput experimental data has provided the opportunity of exploring biological networks to reveal complex structure and organization in the cell. Proteins are the main actors to perform cellular functions and majority of them do not carry out their functions in isolation but by interacting together in a complex manner. Protein complexes and functional modules are two important constructs formed by physical interactions among proteins. Identifying protein complexes and functional modules are crucial to understand the principles of cellular organization with important applications in disease diagnosis and therapy. However, experimental detection of protein complexes and functional modules is highly limited in the current state-of-the-art high-throughput experimental techniques. Thus computational approaches for detecting protein complexes and functional modules from protein-protein interaction data are valuable complements to the experimental techniques. From computational viewpoint, mining objects like protein complexes and functional modules from protein-protein interaction data (equivalently from protein-protein interaction networks) is computationally challenging task, since many problems related to determining structural properties of graphs are often NP-hard in nature. In this survey, we review the current state-of-the-art as computational techniques as well as recent emerging techniques for detecting protein complexes and functional modules, and discuss some promising research directions.

INTRODUCTION

The components of living systems can be organized into hierarchical and modular fashion. Although functionality and organization of higher levels are quite easily understandable, complexity increases drastically as we move towards lower levels. From several decades, researchers have been trying to understand living systems at molecular level of detail. If we see the molecular level of living systems, we find mainly four major families of small organic molecules (sugars, fatty acids, amino acids, and nucleotides) and their corresponding macromolecules (polysaccharides/carbohydrates, lipids, proteins, and nucleic acids, respectively, formed by linking them into long chains) [1]. Majority of the cell mass are accounted by these molecules only. The biological cells are basic structural and functional unit of living systems. They are basic building blocks and smallest units of life. A collection of similar type cells (performing similar function) form a tissue and multiple tissues constitute an organ. Similarly, multiple organs make an organism.

The components at each of these levels (of living systems) interact among them-selves to maintain the dynamic of the system; hence system-level understanding (systems biology) is becoming increasingly important to understand the essence of life. Networks (graphs) are useful mathematical models to describe complex systems, where the elements or components of the system are represented by nodes (vertices) and their relationships or interactions are represented by edges (arcs or links). Typically biological networks at the molecular level are metabolic networks, gene regulatory net-works, signal transduction networks, and protein-protein interaction networks. Different properties of these networks can be analyzed using graph theoretic methods. Further this inferred information can be used to generate important hypotheses about the possible behaviors of the system.

Proteins are the main actors to perform cellular functions and to determine the structure of the cell. Most of the biological functions are carried out by complex interactions among multiple proteins than that by any single protein. A little modification of a single protein may lead dramatic effects on the cell's overall functionality due to cascading effects. In fact, many diseases are the consequence of small changes to a single protein. Thus protein-protein interactions are fundamental to almost all biological functions.

Protein-protein interaction (PPI) networks play a central role towards system-level understanding of cellular processes with important applications in disease diagnosis and therapy. Recently, a large amount of protein interaction data is becoming increasingly available due to high-throughput technologies. Currently, two experimental strategies, yeast two-hybrid (Y2H) system [2], and Tandem affinity purification followed by mass spectrometry (TAP-MS) [3] are widely used to generate protein interaction data at high-throughput. Each of these two methods provides fundamentally different sources of protein interaction data. However, all these interaction data involve a considerable fraction of false positive and false negative interactions [4]. Hence, a major challenge is to distinguish between useful and noisy data.

Generally, protein-protein interactions are temporal and spatial depending on certain conditions of the particular cell. In literature, three important constructs are defined in PPI networks: protein complexes, functional modules, and signaling path-ways. These three constructs are highly interrelated and it is very difficult to clearly define and distinguishes among them. Many researchers believe the following distinction between protein complexes and functional modules given by Spirin and Mirny [5]. A protein complex or...
functional module is a group of proteins that physically interact (bind) among themselves, but in a protein complex all interactions among proteins occur simultaneously at one location, whereas in a functional module interactions may not necessarily take place at same place and time. Complexes can be either permanent or transient. Functional modules may involve more than one complex along with other individual proteins. In many cases it is hard to distinguish between protein complexes and functional modules, because most of the underlying protein interaction data do not have temporal and spatial information. To distinguish between protein complexes and functional modules additional information (e.g., gene expression data) need to be integrated. A signaling pathway is a sequence of proteins that interact in ordered succession and involve in signal transduction process. Similar to functional modules, proteins involved in a signaling pathway also do not necessarily interact at the same location and time. Some authors treat signaling pathways as a special type of functional modules.

Proteins are organized into different complexes, each of which performs specific tasks in the cell [6] and the interacting proteins within a complex often participate in the same biological processes. Furthermore, functional modules can often be associated with specific biological functions and proteins belonging to a specific functional module are more related to each other than to the members of other functional modules [7]. Therefore, identifying protein complexes and functional modules are important for understanding the principles of cellular organization and annotating functions to uncharacterized proteins with important applications in disease diagnosis and therapy.

While there are few experimental techniques (with several limitations) for detecting complexes, no high-throughput experimental technique is available (in our knowledge) for detecting functional modules. Tandem Affinity Purification with Mass Spectrometry (TAP-MS) [3] is the most frequently used technique for detecting protein complexes experimentally. However, there are several limitations to this technique, such as its multiple washing and purification steps tend to eliminate transient low affinity complexes; it may miss complexes that are not present under the given in vitro (in an artificial environment outside the living organism or outside the cell) experimental conditions [8]. Therefore, accurately detecting protein complexes remains a challenge. Thus, specialized computational approaches are treated as valuable complements to the experimental techniques for identifying complexes.

From computational viewpoint, identification of protein complexes and functional modules in PPI networks are quite related to graph-based problems, such as sub-graph isomorphism, graph-based clustering, and problems related to graph partitioning. These problems are computationally challenging due to their NP-hard property [2]. In fact, the development of efficient algorithms to determine structural properties of graphs is a classic challenge in theoretical computer science. Therefore, efficient computational methods for detecting complexes and functional modules are strongly demanded. The problem of detecting protein complexes in PPI networks is quite similar to the problem of detecting communities in complex networks. In fact, several community detection algorithms have been directly applied for detecting protein complexes in PPI networks.

A wide range of computational approaches have been developed for detecting complexes and functional modules from PPI networks [9,10]. But the accuracy of these methods is limited due to several reasons, in particular, the presence of false positive and false negative interactions, and the complex topological structure of the networks. The complexity of PPI networks caused by cross-talk between modules also makes functional module detection more challenging.

The scope of our current research is to design and develop effective computational techniques for detecting protein complexes and functional modules mainly from PPI networks.

DIFFICULTIES IN DETECTION OF PROTEIN COMPLEX

Following are some of the challenges that occur in general observation during detection of protein complexes in protein-protein interaction network.

Representation of protein complexes
The main problem with the existing methods basically clustering methods is that it assumes that every protein complex are dense sub-graph which may not be the case always. Actually, further analysis shows that all type of topologies is present in protein complexes in PPI network. Another problem is to identify the specific topologies.

Multiple interactions
Many existing methods shows poor performance when one protein interacts with multiple protein complexes. Because of this protein complexes may overlap and sometimes it may not show the properties and features of protein complexes.

Noise effect
Generally protein interaction data are very noisy. Instead of traditional un-weighted 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.
PROTEIN COMPLEX DETECTION TECHNIQUES

In this section we present a short overview of state-of-the-art computational techniques as well as recent work for detecting protein complexes and functional modules. First we will discuss state of the art computational techniques. It can be classified into following categories as in [9,10]:

State of art computational techniques

- Local neighborhood Density Search (LD)
- Cost-based Local Search (CL)
- Flow Simulation (FS)
- Statistical-based Measures (SM)
- Network Alignment (NA)
- Other Methods

Local neighborhood Density Search (LD)
The objective of Local neighborhood Density search (LD) is to find dense sub-graphs in PPI networks and to maximize the densities of each sub-graph using some kind of local density measures. This approach is mainly used to detect complexes. Several clustering algorithms have been developed based on LD approach. Algorithms based on LD include MCODE [11], DPClus [12], SWEMODE [13], DECAFF [14], CFinder [15], PINCoC [16], MFPINCoC [5], and PCP [17]. All of these algorithms need setting the values of some parameters which impact the number and resolution of discovered clusters. LD is based on the assumption that complexes form dense sub-graphs in the interaction graph. While this is true for many complexes, there are many others which are not dense structure and thus they will not be detected by this approach [18]. We explain two representative algorithms MCODE and DECAFF in brief.

MCODE (Molecular Complex Detection)
One of the first widespread used computational methods for identifying protein complexes based on LD approach is MCODE. It has three main steps:

Step1: Generate weight of each vertex based on their local neighborhood densities.
Step2: Select seed vertices having high weights as preliminary clusters.
Step3: Finally, augment neighboring vertices to clusters using outward traversal from the seeds.

DPClus is quite similar algorithm to MCODE. The main difference between the two algorithms lies in defining weight of vertices. In MCODE each vertex is assigned a weight based on its local neighborhood density, whereas in DPClus each edge is assigned a weight according to their common neighbors between the two vertices and finally each vertex is assigned a weight using their weighted degree.

DECAFF (Dense-neighborhood Extraction using Connectivity and confidence Features)
DECAFF incorporates functional information to detect dense and reliable sub-graphs as protein complexes. The steps involved are:

Step1: Find the local dense neighborhood graph of each vertex. A local clique mining method is used to locate the local cliques, and then a novel hub-removal technique is used to identify local dense sub-graphs in local neighborhood graph of each vertex heuristically.
Step2: Repeatedly merge two local sub-graphs if they have a large overlap.
Step3: Eliminate probable false complexes detected with low reliability estimated.

DECAFF tries to address two key constraints in PPI network viz. incompleteness and noise. The reliability of a sub-graph is the average reliability of the edges within the sub-graph. A probabilistic model with functional information of interacting proteins is used to estimate the reliabilities of edges.

Cost-based Local search (CL)
The objective of Cost-based Local search (CL) is to partition the network into connected sub-graphs using some kind of cost functions to guide towards the best partition. This approach can be used to extract more general functional modules, since cost functions are not only limited to density measures. The common CL based algorithms are RNSC [19], Qcut [20], and ModuLand [21]. Here we explain RNSC and Qcut briefly.

RNSC (Restricted Neighborhood Search Clustering)
RNSC is a classic example of CL approach. It searches the solution space of all possible clustering, each of which is assigned a cost according to a cost function, in order to find out a clustering with low cost. It has the following basic steps:
Step 1: Select an initial clustering randomly.

Step 2: Search a clustering with optimal or near optimal cost by repeatedly moving one node from a cluster to another one randomly to improve the cost.

Step 3: Filter the results according to cluster size, cluster density and functional homogeneity.

The common problem of CL approach is its pre-convergence to poor local minima. This problem can largely be avoided by using diversification and multiple experiments. RNSC makes the diversification moves by shuffling clustering occasionally. A list of forbidden moves is maintained to prevent cyclic back to the already explored partitioning. Proper thresholds for minimum cluster size, minimum cluster density, and functional homogeneity need to be set to come across a good clustering comparable to the true complexes.

QCut
Several community detection algorithms have been developed based on the optimization of a modularity function \( Q \) [2]. Modularity is a measure that calculates the fraction of edges present within communities, subtracted by the fraction of edges we would expect if edges were positioned randomly. However, it has been shown that optimizing \( Q \) is NP-hard problem. So, heuristics are proposed. Qcut is a good heuristic that optimizes modularity by combining spectral graph partitioning and local search. It involves the following two steps:

Step 1: (Partitioning): Divide the network recursively using a spectral graph algorithm until no improvement to \( Q \) can be achieved.

Step 2: (Refinement): Move vertices among communities or merge communities to improve \( Q \). Go to step 1 to check any community affected by refinement stage (step 2) can further be partitioned.

These two steps are carried out alternatively until neither of them can improve \( Q \).

Flow simulation (FL)
The objective of Flow Simulation (FS) is to simulate the spreading of some information on a network. This is a useful approach to extract functional modules. The common FS based algorithms are MCL [22], RRW [23], IFB [24], and STM [25]. MCL and RRW use the notion of random walk, while IFB and STM exploit biological knowledge for spreading information among proteins in the network. The FS approach is also biased towards dense sub-graphs, since the flow tends to gather towards dense sub-graphs.

Markov Clustering Algorithm (MCL)
The fundamental notion of MCL is that if we start from a particular node and travel randomly to a connected node, we are more likely to stay within a cluster than travel between clusters. MCL exploits the concept of random walks upon the graph to discover where the flow tends to gather. It assumes flows tend most likely towards clusters. MCL works by simulating many random walkers starting from the same node that travel in the graph randomly. It involves the following steps:

Step 1: Represent the weighted (or un-weighted) PPI network as weighted (or un-weighted) adjacency matrix. The matrix could be all-vs.-all protein-protein similarity matrix using BLAST E-values as weights (log \( E \)).

Step 2: Transform weights into column-wise transition probabilities, i.e., the matrix elements (on each column) correspond to probability values (Markov Matrix).

Step 3: (Expansion): Expand the matrix by taking square of the matrix.

Step 4: (Inflation): Take the Hadamard power of the matrix, i.e., calculating power entry-wise and then transform the matrix into stochastic again.

Step 5: Repeat steps 3 and 4 alternatively until a steady state is reached (convergence) (i.e., until there is very small or no variation in the matrix).

Step 6: Interpret the resulting matrix to determine clusters.

New probabilities are assigned to all pairs of nodes during the expansion operator, whereas the probabilities to all walks in the graph are changed by the in action operator, consequently the probabilities of intra-cluster walks and inter-cluster walks are being increased and decreased, respectively. Iterative expansion and in action will disconnect the network into many connected components as predicted complexes.
**IFB (Information Flow-Based)**

IFB first find some informative proteins by exploiting a sort of topological and biological information and then simulates the information flow in the network starting from each of those informative proteins. It can be used for identifying overlapping functional modules in PPI networks. It has three major steps:

Step1: Select informative nodes.

Step2: Identify preliminary modules by using information flow simulation for each informative node.

Step3: Merge the initial modules.

The amount of information of a node in a PPI network is defined as a measure of topological and biological essentiality of the corresponding protein. The weight of each edge is calculated using the correlation between the expression profiles of the two genes that encode the corresponding proteins connected by the edge. The weighted degree of a node is calculated as the sum of weights of adjacent edges of the node. The proteins with high weighted degrees are selected as informative nodes. The information moves through the network starting from these informative nodes and tends to gather towards dense sub-graphs, which are identified as preliminary modules. Lastly similar preliminary modules are merged to generate final modules.

**Statistical-based Measures (SM)**

The SM approach clusters the network using some kind of statistical measures. Algorithms based on SM include SL [26] and Farutin (by the name of the authors) [27]. SL is devised based on the number of shared neighbors between a pair of proteins, while Farutin uses the notion of preferential attachment among the members of a module.

**SL (Samanta and Liang)**

SL (here called SL by the name of the authors) is a network-based statistical method that tries to overcome the difficulty of false positive present in PPI network. The hypothesis of SL is that if two proteins share significantly larger number of common interaction partners than what is expected from a random network, they have close functional association; hence they should be clustered together. SL involves the following four steps:

Step1: Compute p-values of all protein pairs and store in a similarity matrix.

Step2: Pick the protein pair with the smallest p-value and assign them in the same group (cluster).

Step3: Merge the corresponding rows and columns of the selected pair (in step 2) into a new row and column. The p-values for this newly merged row/column (group) are geometric means of the individual p-values (or arithmetic mean of the log p values).

Step4: Step 2 and 3 are repeated by adding more and more clusters or making the existing ones bigger until a chosen threshold is reached.

**Network Alignment (NA)**

Protein sequence analysis for identifying conserved patterns (sequences) across species leads biological meaningful results. Similar idea is extended to PPI networks with similar hypothesis conserved sub-graphs across species may have biological significance and thus are expected to correspond to real protein complexes or functional modules [28]. Given two or more networks, the aim of network alignment algorithms is to identify modules that are conserved across the networks. Most Network Alignment algorithms first define an alignment graph (orthology graph) where each node represents a set of orthologous proteins (related genes/proteins present in different species and are derived from a common ancestral gene/protein after a speciation event) one from each network and each edge represent conserved interactions between the corresponding protein sets in each species. A search is then carried out over this alignment graph to detect high scoring sub-graphs.

Several algorithms have been devised for aligning networks, both locally and glob-ally. Most of these algorithms exploit either sequence similarity or network topology, or sometimes both. PathBLAST [28], NetworkBLAST [29], MaWiSH [30], Graemlin 1.0 [31], the Bayesian method [32], the match and split algorithm [33], and Phunkee [34] are local network alignment algorithms; whereas global network alignment algorithms include: Graemlin 2.0 [35], Markov random field-based method [9], IsoRank [16], and GRAAL [36]. Recently, Ali et al. [37] have exploited functional similarity along with sequence similarity of proteins in the local alignment match and split algorithm [38]. MI-GRAAL [39] is another recent global alignment algorithm in which different kind of information, such as topological features, sequence similarity and functional similarity, can be incorporated. However, how the addition of functional similarity effect the resulting alignment yet to be assessed. Recently, Phan and Sternberg have been developed a global alignment algorithm, PINALOG [40], which integrates both protein sequence similarity and functional similarity along with network topology.
Although Network Alignment is a good choice for identifying conserved modules in multiple species, clustering methods outperform in identification of protein complexes and functional modules. It also highly depends on the graph topology for correct results. This is a critical issue due to the presence of significant amount of false positive and false negative interactions in current available PPI networks. Although PINALOG may not be a proper representative of Network Alignment approach, here we describe it briefly due to its novelty of exploiting different kind of information in quite different ways along with network alignment.

**PINALOG**

PINALOG is a recent global network alignment algorithm, which integrates information from protein sequence, function and network topology. It uses communities in the networks to identify seed protein pairs and a sort of scoring schemes for neighborhood similarity of mapped protein pairs. Instead of constructing traditional alignment graph, PINALOG first discovers highly similar protein pairs (seed protein pairs) from densely connected sub-graphs (communities) in the network and then expand the alignment to other proteins in the neighborhood of seed protein pairs.

**Step1:** Community detection: Identify overlapping dense sub-graphs of input networks using CFinder [41].

**Step2** Community mapping to obtain seed protein pairs: Map similar communities that have high similarity score, i.e., containing many inter-species proteins with high similarity scores and extract similar protein pairs from mapped communities to form a list of core pairs (seed pairs).

**Step3**: Extension mapping: Map proteins in the neighborhood of the core protein pairs which are then added to the core.

**Step4**: Repeat step 3 until no more pair is added.

In step 2, the similarity between two proteins is computed as a combination of sequence and functional similarity. The Hungarian algorithm, a polynomial time combinatorial optimization algorithm, is used to map communities between two species. In step 3, a kind of topological similarity is used in the form neighborhood similarity.

**OTHER METHODS**

There are few other methods which are important to mention. Though some of them might be categorized into above classification it is important to describe briefly according to their key features.

**Supervised graph Clustering**

Most of the graph clustering methods are unsupervised in nature. The main hypothesis of these methods is that dense sub-graphs (e.g., cliques or defective cliques - i.e., nearly complete cliques) are likely to be complexes in PPI networks. The flow-based clustering methods [42] are also biased towards dense sub-graphs, since the flow tends to gather towards dense sub-graphs. In reality, many protein complexes are not dense structure [38] and thus they will not be detected by those methods. Recently, Qi et al. [18] developed a supervised graph clustering method without assuming such topological properties of complexes to detect complexes. They used many important topological and biological properties (e.g., sub-graph size, density, degree statistics, clustering coefficient statistics, first eigenvalues, protein weight/size statistics etc.) of known complexes as feature vector. A set of known complexes and a set of random sub-graphs (as non-complexes) are used as training data. All these features are integrated by using a probabilistic Bayesian network (BN) and the parameters of the BN model are estimated from the training data set. Given a candidate sub-graph, a log-likelihood score is calculated by the BN model to evaluate whether it is qualified as a complex. If the log-likelihood score exceeds a certain threshold the sub-graph is predicted to be complexes. A simulated annealing search is further used to modify the candidate sub-graph if possible.

**Step1:** Extract property features from positive (a set of known complexes) or negative (a set of random sub-graphs) training examples.

**Step2:** Discretize the continuous features.

**Step3:** Calculate the BN MLE parameters for different features properties on the multinomial distribution.

**Step4:** Apply simulated annealing search (starting from seed sub-graphs) for expanding and detecting candidate complexes.

**Step5:** Output those sub-graphs (as clusters) that exceed their ratio scores a certain threshold.

However, the knowledge learned from the limited set of training data could be highly incomplete and thus may affect the complex detection. Again, it is not easy to identify a concise set of features which characterize the true complexes.
**Complex detection using Protein core attachments**

Recently, Gavin et al. [43] recommended that a complex consists of two parts: a core and an attachment. Core proteins form the center of a protein complex and have comparatively more interactions among themselves. Each attachment protein binds to a subgroup of core proteins and forms the final complex. They identified that core proteins have three major properties. They have comparatively more interactions among themselves. The attachment proteins bind to the core proteins to form protein complex. Each protein complex has a unique set of core proteins. Few methods such as CORE [44] and COACH [45] have been proposed based on the concept of core attachments.

**CORE Algorithm**

The CORE algorithm is a statistical framework for detecting cores of protein complexes. The algorithm uses the p-value defined as the probability for each pair of proteins to present in the same core and is calculated according to the factors that whether the pair interact or not and the number of shared neighbors between them.

Step1: Calculate p-values for all pairs of proteins.

Step2: If p-value of (p1, p2) is the lowest among p-value of (p1, pk) and p-value of (p2, pk) for all possible protein pk then form p1 and p2 as a core. This step creates some cores of size two only.

Step3: Assign protein pj (which is not yet assigned to any core) to the core C if the p-value of pj and any protein in C is smaller than the p-value of pj and any other protein not in C.

Step4: Repeat step 3 until no further improvement of the size of any core is possible.

Step5: Given a core C, select those proteins that are common neighbors of more than half of the core proteins of C as the attachment proteins of the core C. If there is only one core protein, select all neighbors of the protein as the attachment.

Step6: Remove the cores that highly overlap with the attachment of another core or the cores whose attachment proteins overlap with other cores.

Step 3 merges some proteins to a set of core proteins of size 2, 3, etc. gradually, until no more improvement of the size of any core is being achieved. The sets of core proteins are disjoint, since each protein can associate with a unique set of proteins with the smallest p-values. The degree of protein is not considered for determining if it is an attachment of a complex or not, since an attachment protein may appear in multiple protein complexes. Step 6 removes some of the noisy (insignificant) cores according to the assumption that normally core proteins should not appear as attachment proteins to other cores [43]. However, the CORE algorithm does not consider the issues of false positive or false negative interactions.

**Complex detection by considering exclusive or co-operative interactions**

Most of the traditional methods assume that all interactions in PPI networks can occur simultaneously. Two adjacent interactions (i.e., interactions with a common partner) are said to be mutually exclusive [46, 47] if they cannot occur simultaneously and the situation (of mutually exclusiveness) may arise when the binding interfaces on the shared protein may be overlapped. Recently, Jung et al. [48] introduced a method to extract Simultaneous Protein Interaction Clusters (SPIC) by using similar strategy and they exploited structural interface data of protein domains for mutually exclusive interaction information. However, there are still many problems that need to be solved, for instance, how to get large amount of conflicting interactions depend on 3D structure data and how to devise more efficient methods for SPIC. Jin et al. [41] incorporated time series information of gene expression pro les to distinguish between exclusive or cooperative interactions. They assumed that two adjacent interactions are cooperative (i.e., they can occur simultaneously) when the shared protein has overlapping time-ranges for both the interactions.

**Complex detection by incorporating gene expression**

It is expected that proteins which interact with each other are activated or repressed under same conditions, i.e., interacting proteins are likely to exhibit similar gene-expression pro les. In fact, gene expression pro le has been widely used to annotate protein functions and to predict novel protein-protein interactions. Recently, few methods, such as GFA [49], and DMSP [50], have been developed based on the hypothesis - genes (equivalently proteins) that exhibit similar gene-expression profile are likely to have a functional relationship.

**Complex detection from TAP data with or without constructing the PPI network**

Almost all the methods discussed above use pairwise physical interactions detected by high-throughput experiments such as yeast two-hybrid (Y2H) system as PPI data for detecting complexes. Recently, few researchers have tried to detect protein complexes from interaction data taken merely from TAP (Tandem Affinity Purification) experiments. There are two ways to use TAP data for detecting complexes: TAP data with constructing the PPI network, TAP data without constructing PPI network. To construct PPI networks from TAP data, the detected links need to be carefully evaluated, since besides detecting direct
interactions TAP also detects indirect interactions in protein complexes. Korga et al. [51] used machine learning techniques to learn the edge weights for constructing the PPI network. Caroline et al. [52] exploited boot-strap sampling along with socio-affinity indices and Pu et al. [53] utilized a scoring function to infer the reliability of interactions. However, all the three methods applied MCL finally to detect protein complexes. Geva et al. [54] introduced another new method, CODEC, for detecting complexes from TAP data by modeling TAP data as a bipartite graph instead of transforming TAP data into PPI networks. The PPI networks built from TAP data are not flawless, since TAP data does not detect direct pairwise binary interactions. Recently, few methods have been developed for detecting complexes directly from the high-throughput TAP data without constructing PPI networks. Rungsaratorn et al. [55] and Chu et al. [56] used Markov Random Fields (MRF) and Bayesian approaches, respectively, for detecting complexes directly from TAP data.

RECENT EMERGING TECHNIQUES

a. Network centrality measure techniques.
   b. Essential minimum dominating set (e-MDSet) [57].
   c. BiCAMWI: a genetic based Biclustering algorithm. [33]
   d. Bottleneck based graph partitioning method.
   e. United complex centrality for essential protein detection.
   f. Multi-stage kernel extension for mining of protein complex.
   g. Supervised graph local clustering.
   h. Mutually exclusive protein-protein interaction identification for protein complex detection.
   i. Core-Attachment Approach

Network centrality measure techniques
Complex biological system may be represented and analyzed as computed network. A biological network is any network that applies to biological system. Several biological networks are available in which PPI network, metabolic network, gene regulatory networks are popular. Jugal K Kalita et al. [58] in his paper deals with PPI network in which centrality techniques like degree, eigenvector, pageRank, closeness, betweenness, p-value, radiality are applied associated with graph. Among all of these measures PageRank and radiality measure shows dominating result.

Essential minimum dominating set (e-MDSet)
Sawsan Khuri et al. [57] proposed Essential minimum dominating sets (e-MDSet) to detection of the essential protein complex in PPI network. Determination of minimum dominating sets (MDSet) of protein interaction network of E. Coli, S. Cerevisiae and H. Sapiens that have been entirely determined by Yeast two hybrid approaches which is basically defined as sub-graph, can control the underlying protein interaction network. The proteins that are present in the sub-graph can reach the other proteins of the network in a single interaction. Several parameters and functional parameters of essential, MDSet and e-MDSet were compared in which it is found that e-MDSet contains all essential proteins that enhances the enrichment signals of biological functions of essential proteins.

BiCAMWI: a genetic based Biclustering algorithm
Amir Lakizadeh et al. [33] proposed a new method called BiCAMWI which is a genetic based biclustering algorithm to employ dynamicity in detecting protein complexes. Experimentally BiCAMWI effectively models the dynamicity inherent in static PPI network and achieves significantly better result than state of art methods. Emad Ramadan et al. [59] proposed a genetic based algorithm in which the objective function was mainly for exclusive clustering and overlapping clustering. It also shows better result than MCL, ClusterOne, and Mcode in terms of quality of predicted complexes.

Bottleneck based graph partitioning method
For more detail and precise detection of protein complexes Jaegyoon Ahn et al. [60] proposed a novel data structure which employs bottleneck proteins as partitioning points. In networks like Saccharomyces, Cerevisiae, and Homo Sapiens, this method results in improved F1 score which is the indication of higher precision.

United complex centrality for essential protein detection
Min Li et al. [51] proposed a new method called United Complex (UC) that identifies essential proteins by integrating the protein complexes with the topological features. In this paper a modified version of United Complex method is also proposed. The proposed algorithm performed better then degree centrality, eigenvalue centrality, PageRank, betweenness centrality etc.

Multi-stage kernel extension for mining of protein complex
Mining algorithms like CPM, MCL algorithms concentrate on mining dense protein sub-graphs as protein complexes falling to take account the inherent organizational structure within protein complexes. Xianjun Shen et al. [62] proposed a new mining algorithm called Multistage Kernel Extension (MKE) which is based on the formation process of cliques in complex social networks and centrality-lethality rule. It shows better performance than classical clique percolation method both on ontology semantic similarity and co-localization.
**Supervised graph local clustering**

Yanjun Qi et al. [18] proposed an algorithm for inferring protein complexes from weighted interaction graphs. This method is applied to protein interaction data in yeast. This algorithm achieved improvement over clique based algorithms in terms of ability to recover known complexes.

**Mutually exclusive protein-protein interaction identification for protein complex detection**

Protein-protein interaction network includes static as well as dynamic interactions. Most of the conventional complex detection approach uses graph theoretic clustering methods on static PPI data. Sk Md Mosaddek Hossain et al. [63] proposed a multi-objective framework which is motivated from PROCOMOSS [64] which considers dynamic interaction for detecting protein complexes.

**Core-Attachment Approach**

Most of the computational approach relies on the assumption that proteins within the same complex would have relatively more interactive. According to Galvin [45], protein complexes consist of two parts: a core and an attachment. Qian Xiang et al. [45] proposed a novel method based on the concept of core-attachment which identifies cores and attachments of a complex separately. The result shows this method has 30% more accuracy than the other computational techniques. Using this method cores from the protein complexes are predicted more accurately.

**DISCUSSION AND RESEARCH DIRECTIONS**

Accuracy of computational techniques for detecting protein complexes and functional modules is highly dependent on the quality of available interaction data. Despite the abundance, PPI data is noisy with high false positive and false negative rates [4] due to limitations of experiments involved for detecting protein interactions. Although several methods (described previously) tried to address the reliability issues of interaction data, still computational approaches need to be developed for validating false positive interactions and predicting novel false negative interactions.

Fundamentally, mining PPI networks is computationally challenging, since many graph-based problems, such as sub-graph isomorphism, graph-based clustering and problems related to graph partitioning are often NP-hard [2]. Furthermore PPI net-works are very large graphs with thousands of vertices and edges, even for a simple organism like yeast. Although identifying objects like complexes or functional modules from PPI networks are computationally challenging problems, it may be possible to reduce the search space and time complexity of such problems and also better mining results may be produced by exploiting additional biological knowledge along with the novel graph mining techniques specialized on such networks [65]. Additional biological information may include protein sequence similarity, Gene Ontology annotations, protein domain context similarity, protein structural similarity etc. However, all proteins do not have all types of biological information available. Thus an interesting challenge would be to model the availability issue of different biological data for individual proteins.

It may be possible to integrate various biological evidences into mining process to obtain better results. Biological evidences may include - metrics like reproducibility of the interactions from multiple experimental methods, support from non-interaction data such as co-expression, co-localization, shared functions, conservation of the protein interactions across other genomes etc. Few computational methods, such as Bayesian network models [35] and kernel methods [67] already have been proposed to integrate different biological resources.

It may be helpful to explore the notion of exclusive or cooperative interactions, since very few methods exploit this notion. Moreover, in our knowledge, currently there is no available method that exploits both binary interactions (e.g., data from Y2H system) as well as higher order interactions (e.g., data from TAP-MS). Further-more, an interesting challenge would be that of combining the main advantages of the different approaches described previously.

Living systems are highly dynamic and responsive to external stimuli from the environment. Cellular functions and responses to external signals are regulated by a complex web of various molecular interactions like protein-protein interactions, protein-DNA interactions, and metabolic processes. Despite the availability of large-scale high-throughput experimental data, understanding the dynamic nature of cellular activity remains a challenging task due to static nature of most available biological data that only correspond to snapshots of cellular activities. Thus an important and challenging problem in post-genomic era is to investigate the dynamic organization of PPI networks and explore biologically significant clusters. The time-series gene expression pro ling data may help to detangle the temporal complexity of biological networks as in [41, 68].

Another interesting challenge is to distinguish between protein complexes and functional modules, as most existing methods cannot make distinction between the two. One reason for this could be less focus on interaction dynamics. So far, little progress has been made on this perspective.
Table 2: Summary of some characteristics of computational methods. The first column reports the method acronym and reference, in chronological order. The second column reports the topological structure a method searches (A = arbitrary, D = dense subgraphs). The approach each method is based on is reported in the third one. The fourth column (Simulation) specifies if the method finds all clusters simultaneously and the fifth column (Overlap) reports if the method generates overlapping clusters. Finally, the last two columns specify if the method returns some unassigned proteins (Un. Prot), and if software implementing that method is (publicly) available (Software).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Approach</th>
<th>Formula</th>
<th>Application</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>It is the number of other nodes with which a given node is connected.</td>
<td>$C_{deg}(v_i) =</td>
<td>v_j</td>
<td>.$</td>
</tr>
<tr>
<td>Betweenness Centrality</td>
<td>Quantifies a node’s ability to monitor information flow between other Vertices.</td>
<td>$C_{between}(v_i) = \sum_{v_j \neq v_i} \frac{\alpha_{v_j}}{\alpha_{v_j}}.$ where $\alpha_{v_j}$ is the number of shortest paths from node $v_i$ to node $v_j$ and $\alpha_{v_j}(v_i)$ is the number of those paths that pass through $v_i.$</td>
<td>Estimate biological significance when applied to regulatory networks of mammals [20]. This measure has also been used to suggest modular property of the yeast interactome.</td>
<td>A great proportion of nodes do not lie on the route of any shortest path, i such cases, they get a score of 0.</td>
</tr>
<tr>
<td>Eigenvector Centrality</td>
<td>Assigns a higher rank to nodes that are connected to more important neighbors. It ensures that a node affects all its neighbors in a similar way.</td>
<td>$C_{eig}(v_i) = \frac{1}{\sum_{v_j} d(v_i, v_j)E}$ where $d(v_i, v_j)$ is the shortest distance from node $v_i$ to node $v_j.$</td>
<td>Used in biological networks to identify gene-disease associations and to discover unknown gene-disease associations for further analysis [20].</td>
<td>Repeated reflection of centrality from central nodes to its neighbours result in accumulation of large centrality near hub nodes in the network.</td>
</tr>
<tr>
<td>Closeness Centrality</td>
<td>Measures the significance of a node based on its degree of closeness to other nodes in the network. This measure assigns higher values to nodes which can communicate quickly with other nodes in the network.</td>
<td>$C_{closeness}(v_i) = \frac{1}{\sum_{v_j} d(v_i, v_j)}.$</td>
<td>Used for ranking pathways and for identifying core metabolite molecules in metabolic networks [20].</td>
<td>Cannot be applied to networks with disconnected components.</td>
</tr>
</tbody>
</table>

Table 1: Comparison between various centrality measure techniques
Radiality
Based on the reachability of a node to all other nodes in the network.

It uses the reverse distance matrix for its calculation which is defined as

\[ R_d(x,y) = \text{diam}(G) + 1 - \text{Dist}(x,y) \]

where \( \text{Dist}(x,y) \) is the distance between nodes \( x \) and \( y \) in the network and \( \text{diam}(G) \) is the length of the shortest path between the most distant nodes, \( v_x \) and \( v_y \) in the whole network. Radiality of a node \( v_i \) is then given as

\[ C_{R_d}(v_i) = \frac{\sum_{v_n \neq v_i} R_d(v_n, v_i)}{(n-1)} \]

where \( R_d(v_n, v_i) \) is the reverse distance between node \( v_n \) and \( v_i \) and \( n \) is the total number of nodes.

PageRank Centrality
It is a variation of the eigenvector centrality measure. It considers both the number and quality of links to decide the score of a node.

It was used to interpret the possibility of certain proteins to be highly relevant to a subset of other proteins while being completely dissimilar to another subset of proteins in the network [23].

PageRank centrality of node \( v_i \) is given as

\[ C_{PR}(v_i) = \sum_{v_n \in B_n} \frac{P_g R(v_n)}{L(v_n)} \]

where \( B_n \) is the set of all nodes linked to node \( v_i \) and \( L(v_n) \) is the number of links from node \( v_n \).

**Table 2: Summary of some characteristics of computational methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Structure</th>
<th>Approach</th>
<th>Simulation</th>
<th>Overlap</th>
<th>Un. Prot</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCL[22]</td>
<td>A</td>
<td>FS</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>SL</td>
<td>A</td>
<td>SM</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>RNDC[19]</td>
<td>D</td>
<td>CL</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>STM[28]</td>
<td>A</td>
<td>FS</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>SWECODE[13]</td>
<td>D</td>
<td>LN</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>DPCLus[12]</td>
<td>D</td>
<td>LN</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>IFB</td>
<td>A</td>
<td>FS</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>FARUTIN[27]</td>
<td>A</td>
<td>SM</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>CFINDER[15]</td>
<td>D</td>
<td>LN</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>CGA</td>
<td>D</td>
<td>PS</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>PCP</td>
<td>D</td>
<td>LN</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>DECAFF[14]</td>
<td>D</td>
<td>LN</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>MF-PINCoC[69]</td>
<td>A</td>
<td>LN</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>QCut[20]</td>
<td>D</td>
<td>CL</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>DME</td>
<td>D</td>
<td>LN</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>RRW</td>
<td>A</td>
<td>FS</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>MODULAND[21]</td>
<td>D</td>
<td>CL</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>IGA</td>
<td>D</td>
<td>PS</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>

**CONCLUSION**

Identification of protein complexes and functional modules is utmost important to understand the principles of cellular organization and predicting protein functions. Several computational methods have...
been developed for this task due to limitation of experimental techniques. From this survey, it is clear that no computational method is ubiquitous with all respects and thus researchers are tireless in developing new computational methods for detecting protein complexes and functional modules. The scope of our current research is not only to design and develop effective computational techniques for detecting protein complexes and functional modules but also distinguish between them. In other words, we shall not only be exploring complexes and functional modules but also studying the relationship between them. At the same time, we shall be investigating the dynamic organization of PPI networks too.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
This work is a part of the Research Projects sponsored by Visvesvaraya Ph.D. Scheme for Electronics & IT, Ministry of Electronics & Information Technology, India, and Reference No: PHD-MLA(4/44)/2015-16, dated August 2015. The authors would like to express their thanks for the financial supports offered by the Sponsored Agency.

FINANCIAL DISCLOSURE
None

REFERENCES


[58] Sharma P. [2016]. Centrality Analysis in PPI Networks.


