Complex detection in protein-protein interaction networks: a compact overview for researchers and practitioners

Clara Pizzuti[1], Simona E. Rombo[1,2], Elena Marchiori[3]

[1] Institute for High Performance Computing and Networking,
National Research Council of Italy, CNR-ICAR,
Via P. Bucci 41C, 87036 Rende (CS), Italy
pizzuti@icar.cnr.it
[2] DEIS, Università della Calabria
Via P. Bucci 41C, 87036 Rende (CS), Italy
simona.rombo@deis.unical.it
[3] Radboud University, Department of Computer Science
Nijmegen, The Netherlands
elenam@cs.ru.nl

Abstract. The availability of large volumes of protein-protein interaction data has allowed the study of biological networks to unveil the complex structure and organization in the cell. It has been recognized by biologists that proteins interacting with each other often participate in the same biological processes, and that protein modules may be often associated with specific biological functions. Thus the detection of protein complexes is an important research problem in systems biology. In this review, recent graph-based approaches to clustering protein interaction networks are described and classified with respect to common peculiarities. The goal is that of providing a useful guide and reference for both computer scientists and biologists.

1 Introduction

In the last few years the development of advanced high-throughput technologies [48] to determine protein interactions has made available large volumes of experimental data that reflect the interplay among proteins in complex cellular networks. *Protein-protein interaction (PPI) networks* can be used for discovering (putative) functional modules, or complexes, consisting of proteins sharing a common function. This is motivated by the observation that proteins are organized into different putative protein complexes each performing specific tasks in the cell [18, 36] and that proteins interacting with each other often participate in the same biological processes. Furthermore, protein modules can be often associated with specific biological functions and proteins belonging to a specific module are more related each other than to the members of other modules [47]. Therefore the detection of putative protein complexes using PPI networks can

help in understanding the mechanisms regulating cell life, in describing the evolutionary orthology signal (e.g., [22]), in predicting the biological functions of uncharacterized proteins, and, more importantly, for therapeutic purposes. It is worth to point out that protein complexes and functional modules have different biological meanings. A protein complex is a molecular machine that consists of several proteins that bind each other at the same place and time. On the contrary, a functional module consists of a few proteins that control or perform a particular cellular function through interactions between themselves (these proteins do not necessarily interact at the same time and place). However, it is hard to distinguish them in many cases because analyzed pair-wise protein interactions do not have temporal and spatial information, thus in the following we will use the two terms as synonyms.

The problem of detecting protein complexes using PPI networks can be computationally addressed by using clustering techniques. Clustering consists in grouping data objects into groups (clusters) such that the objects in the same cluster are more similar each other than with objects in the other clusters [20]. In PPI networks, clustering means grouping together proteins which share a larger number of interactions, which are considered to represent functional modules. Possible uncharacterized proteins in a cluster may be assigned to the biological function recognized for that module. PPI networks have various characteristics which have to be taken into account when developing clustering algorithms for detecting functional complexes. Therefore a number of clustering approaches have been proposed to extract relevant modules from PPI networks.

In this work, we present a short overview of state-of-the-art clustering methods for complex detection in PPI networks, by introducing a classification criterion that is different from those proposed previously. We mainly focus on methods that use only the topology of the graph for detecting clusters, and do not employ similarity measures between proteins as described by vectors of features (for instance, features derived by the protein aminoacid sequences or by functional domain composition of proteins). Our goal is twofold: (a) to guide researchers in the development of new methods for clustering PPI networks by providing a description of the main algorithmic approaches of state-of-the-art methods; and (b) to guide practitioners in the application of methods by providing information about their availability.

In this respect our contribution differs from that contained in other surveys, whose main goal is either to describe and compare experimentally methods presented in the literature, such as [2, 8, 40, 28, 41, 49, 27], or to highlight the computational aspects of graph-based analysis of networks [34].

2 Methods

Clustering approaches for detecting protein complexes in PPI networks can be broadly categorized as distance-based and graph-based ones [28]. Distance-based clustering approaches employ the concept of distance between two proteins as described by vectors of features (for instance, derived by their aminoacid se-

quence) [7, 43, 4, 35]. Graph-based clustering techniques (mainly) consider the topology of the network. These latter techniques are deeply studied in other research fields, such as physics and data mining, and are known as community detection methods [17].

We distinguish the following five main types of algorithmic approaches employed in methods for complex detection in PPI networks:

1. Local neighbourhood Density search (LD);
2. Cost-based Local search (CL);
3. Flow Simulation (FS);
4. Statistical-based Measures (SM);
5. Population-based Stochastic search (PS).

For each of the categories listed above, we describe a selection of methods by focusing on those that can be directly used by practitioners, that is, whose software is publicly available.

2.1 Local Neighborhood Density Search (LD)

Many methods, including the most popular ones, are based on local neighbourhood density search. Their objective is to find dense subgraphs (that is, where each of its nodes is connected to many other nodes in the same subgraph) within the input network. We summarize in the following six representative methods of this approach, and include a pointer to the software when publicly available.

One of the most popular methods for finding modules in PPI networks based on the LD approach is **MCODE** [6]. This method employs a node weighting procedure by local neighbourhood density and outward traversal from a locally dense seed protein, in order to isolate the dense regions according to given input parameters. The algorithm allows fine-tuning of clusters of interest without considering the rest of the network and allows examination of cluster interconnectivity, which is relevant for protein networks. It is implemented as Cytoscape plug-in. With a user-friendly interface, it is suited for both computationally and biologically oriented researchers.
`http://baderlab.org/Software/MCODE`.

In [3] the **DPClus** method for discovering protein complexes in large interaction graphs was introduced. It is based on the concepts of *node weight* and *cluster property* which are used for selecting a seed node to be expanded by iteratively adding neighbours, and to terminate the expansion process, respectively. Once a cluster is generated, its nodes are removed from the graph and the next cluster is generated using only the remaining nodes until all the nodes have been assigned to a cluster. The algorithm allows also to generate overlapping clusters by keeping the nodes already assigned to clusters.
`http://kanaya.naist.jp/DPClus/`.

SWEMODE was introduced in [30]. It identifies dense sub-graphs by introducing two network measures that combine functional information with topological properties of the networks. These measures, weighted cluster coefficient and weighted nearest-neighbours degree, compute the strengths of interactions between the proteins by using their semantic similarity based on the Gene Ontology terms of the proteins.
No publicly available implementation.

DECAFF [26], is an algorithm to mine protein complexes in PPI networks that tries to address two major limitations plaguing protein interaction data, namely incompleteness and noise. The method consists of three main steps: detection of local dense neighbourhoods of each protein, merging of the local sub-graphs on the base of the similarity degree between neighbourhoods, filtering away possible false complexes detected.
No publicly available implementation.

CFinder is a program for detecting and analyzing overlapping dense groups of nodes in networks; it is based on the clique percolation concept (see [12, 33, 1]). The idea behind this method is that a cluster can be interpreted as the union of small fully connected sub-graphs that share nodes, where a parameter is used to specify the minimum number of shared nodes.
`http://hal.elte.hu/cfinder/wiki/?n=Main.Manual`.

The greedy local expansion method **PINCoC** was introduced in [38]. It expands a single protein randomly selected by adding/removing connected proteins that best contribute to improve a given quality function based on the concept of co-clustering [32] that favors the detection of maximal dense groups. In order to escape poor local maxima, with a given probability, the protein causing the minimal decrease of the quality function is removed. An extension of PINCoC for detecting multi-functional protein complexes, called MF-PINCoC, was introduced in [39].
`http://wwwinfo.deis.unical.it/~rombo/pincoc/download.html`.

PCP is a method proposed in [11] that exploits the shared interaction partners of proteins, i.e., the level-2 neighbours. The method transforms the input graph by adding edges between level-2 neighbours and by removing edges, using a criterion that quantifies the likelihood that the two proteins of an edge share functions. Any clustering method can then be applied to the resulting graph. The authors proposed a clustering method that iteratively merges dense subgraphs.
`http://www.comp.nus.edu.sg/~wongls/projects/complexprediction/PCP-3aug07/`.

In [16] the **DME** method for extracting dense modules from a weighted interaction network was introduced. The method detects all the node subsets that satisfy a user-defined minimum density threshold. The method returns only lo-

cally maximal solutions, i.e. modules where all the direct supermodules (containing one additional node) do not satisfy the minimum density threshold. The obtained modules are ranked according to the probability that a random selection of the same number of nodes produces a module with at least the same density. An interesting property of this method is that it allows to incorporate constraints with respect to additional data sources.
`http://people.kyb.tuebingen.mpg.de/georgii/dme.html`.

The methods based on the LD approach here briefly described have as common objective that of finding dense subgraphs within the network and to maximize the density of each subgraph. MCODE and DPClus adopt a rather similar search strategy. They define the weight of each node, the node with highest weight is chosen as seed cluster, and neighbouring nodes are added to the current cluster if threshold parameters are satisfied. The main difference between the methods lies in the definition of weight. The originality of PCP mainly relies in the procedure for transforming an interaction graph by adding and removing edges. Both CFinder and the extended version of PINCoC, generate overlapping clusters, and use the concepts of k-clique and co-cluster to find dense subgraphs, respectively. DME is somewhat different from all other methods since it enumerates *all* node subsets that satisfy a user-defined minimum density threshold. Each of the above mentioned methods require setting the values of some parameters; this influences the number and resolution of the discovered clusters. Other recent algorithms based on this approach include SPICi [23] and DEEN [21], two seed-based fast algorithms for complex detection in PPI networks.

2.2 Cost-based Local Search (CL)

Methods based on cost-based local search extract modules from the interaction graph by partitioning the graph into connected subgraphs using a cost function for guiding the search towards a best partition. We describe here in short three methods based on this approach with different characteristics.

A typical instance of this approach is **RNSC** [24], which explores the solution space of all the possible clusterings in order to minimize a cost function that reflects the number of inter-cluster and intra-cluster edges. The algorithm begins with a random clustering, and attempts to find a clustering with best cost by repeatedly moving one node from a cluster to another one. A list of tabular moves is used to forbid cycling back to previously examined solutions. In order to output clusters likely to correspond to true protein complexes, thresholds for minimum cluster size, minimum density, and functional homogeneity must be set. Only clusters satisfying these criteria are given as the final result. This obviously implies that many proteins are not assigned to any cluster.
`http://www.cs.toronto.edu/~juris/data/rnsc/`.

Several community discovery algorithms have been proposed based on the optimization of a modularity-based function (see e.g. [15]). Modularity measures

the fraction of edges falling within communities, subtracted by what would be expected if the edges were randomly placed. In particular, **Qcut** [44] is an efficient heuristic algorithm applied to detect protein complexes. Qcut optimizes modularity by combining spectral graph partitioning and local search. By optimizing modularity, communities that are smaller than a certain scale or have relatively high inter-community density may be merged into a single cluster. In order to overcome this drawback, the authors introduce an algorithm that recursively applies Qcut to divide a community into sub-communities. In order to avoid over-partitioning, a statistical test is applied to determine whether a community indeed contains intrinsic sub-community.
`http://cs.utsa.edu/~jruan/Software.html`

Recently, the notion of **ModuLand** [25], has been introduced. **ModuLand** is an integrative method family for determining overlapping network modules as hills of an influence function-based, centrality-type community landscape, and including several widely used modularization methods as special cases. Several algorithms obtained from ModuLand provide an efficient analysis of weighted and directed networks, determine overlapping modules with high resolution, uncover a detailed hierarchical network structure allowing an efficient, zoom-in analysis of large networks, and allow the determination of key network nodes. It is implemented as Cytoscape plug-in.
`http://www.linkgroup.hu/modules.php`

2.3 Flow Simulation (FS)

Methods based on the flow simulation approach mimic the spread of information on a network. We report four methods based on this approach. The first two are based on the concept of random walk and are popular methods with available software. The other two methods exploit biological knowledge for passing information between proteins in the network in order to cluster proteins. Unfortunately, we could not find publicly available software for these two methods.

One of the first flow simulation method for detecting protein complexes in a PPI network is the *Markov Clustering algorithm MCL* [13]. **MCL** simulates the behaviour of many walkers starting from the same point, that move within the graph in a random way.
`http://micans.org/mcl/`

A more recent method based on flow simulation is **RRW** [31]. RRW is an efficient and biologically sensitive algorithm based on repeated random walks for discovering functional modules, which implicitly makes use of network topology, edge weights, and long range interactions between proteins.
`http://www.cs.ucsb.edu/~kpm/software.html`

IFB [10] proposed an Information Flow-Based approach to identify overlapping functional modules. The algorithm integrates topological and biological

knowledge to select a number of informative proteins and simulates the information flow through the network from each informative protein. The weighted degree of a node is defined as the sum of the weights of the edges containing that node, and the weight of an edge is computed using the correlation between the expression profiles of the two genes encoding the proteins linked by that edge. This weighted degree provides the semantic information of a node.
No publicly available implementation.

An interesting method based on flow simulation is **STM** [19], which finds clusters of arbitrary shape by modelling the dynamic relationships between proteins of a PPI network as a signal transduction system. The overall signal transduction behaviour between two proteins of the network is defined in order to evaluate the perturbation of one protein on the other one both biologically and topologically. The signal transduction behaviour is modelled using the Erlag distribution.
No publicly available implementation.

2.4 Statistical Measures (SM)

The two following approaches rely on the use of statistical concepts to cluster proteins. They are based on the number of shared neighbours between two proteins, and on the notion of preferential attachment of the members of a module to other elements of the same module, respectively.

SL [45] is a clustering method based on the idea that if two proteins share a number of common interaction partners larger than what would be expected in a random network, then they should be clustered together. The method assesses the statistical significance of forming shared partnership between a pair of proteins using the concept of p-value of a pair of proteins. The p-values of all proteins pairs are computed and stored in a similarity matrix. The protein pair with the lowest p-value is chosen to form the first group and the corresponding rows and columns of the matrix are merged in a new row and column. The new p-value of the merged row/column is the geometric mean of the separate p-values of the corresponding elements. This process is repeated by adding new proteins to the actual cluster until a threshold is reached. The process is repeated on the remaining proteins until all the proteins have been clustered.
No publicly available implementation.

In [14] a statistical approach for the identification of protein clusters is presented, here called **Farutin** (the name of the first author). This method is based on the concept of preferential interaction among the members of a module. The authors use a novel metric to measure the community strength. The community strength is gauged by the preferential attachment of each member of a module to the other elements of the same module. This concept of preferential attachment is quantified by how unlikely it is observed in a random graph. Since it is necessary to count the number of edges in the graph, the authors assume a

random graph as the null model where an edge is the random variable. This measure of community strength is local, since it is a function of the sub-graph induced by a set of proteins and their degrees. To identify the clusters a greedy approach that searches for a set of nodes in the network with small values of community strength is adopted. At the beginning a list of two adjacent nodes is considered. The list is then grown by adding the node that leads to the largest decrease of the community score until no such node exists. This is repeated for each connected node pair, thus the obtained clusters can partially overlap.
No publicly available implementation.

2.5 Population-based Stochastic search (PS)

Population-based stochastic search has been used for developing algorithms for community detection in networks (see, e.g., [46, 37]). However, we are aware of only two works that apply this approach to detect protein complexes in PPI networks.

Specifically, in [29] the authors proposed an algorithm based on evolutionary computation, here called **CGA**, for enumerating maximal cliques and apply it to the Yeast genomic data. The advantage of this method is that it can find as many potential protein complexes as possible.
No publicly available implementation.

Recently, in [42] an immune genetic algorithm, here called **IGA**, is described to find dense subgraphs based on efficient vaccination method, variable-length antibody schema definition and new local and global mutations. The algorithm is applied to clustering protein-protein interaction networks.
No publicly available implementation.

3 Discussion

We summarize the characteristics of each method in Table 1, with respect to few features: the structure of the clusters found by a method, the kind of approach it uses, whether the clusters are found simultaneously or one at a time, the capability of the method to detect overlapping clusters, if the method assigns each protein to a cluster, and if software for that method is publicly available.

All the considered methods have some input parameters that influence the number of clusters produced, the size, the density, and the structure. The LN methods, except CFinder, obtain the modules one at a time because they select a seed node and expand it until a condition, generally related to cluster density, is satisfied. Thus they can be considered bottom-up approaches: individual nodes are grouped together until all the graph has been examined. Methods that simultaneously find the clusters can be considered top-down. They consider the whole graph and try to partition it in connected components. Because of the threshold constraints incorporated in many methods in order to decide when

a group of connected nodes is a cluster, nodes with few interactions are often discarded.

The elimination of sparsely connected nodes could result in the elimination of important information on the network structure and possibly prevent the detection of clusters of different topological shapes. Nevertheless, it is not clear whether the assumption that each protein has to belong to a cluster (representing a putative protein complex) is realistic, given the actual incompleteness of the PPI network data available, and forcing every node into a community could distort results [51].

Several challenges for the topic discussed in this work are still open. Notably among them, the necessity of diminishing the clustering methods dependence on many input parameters. A first step in this direction has already been done by all the methods discussed in this review, avoiding the number of output clusters to be required a priori. Further improvements could be achieved by making a method able to set automatically some of its parameters, for example according to the density and/or characterization of the input PPI network.

Another interesting issue is that of finding a suitable compromise between the accuracy of the proposed method, and the portion of input graph that is involved in the final clustering. Indeed, the most accurate clustering methods are often able to assemble only a small percentage of the PPI network they analyze (e.g., **MCODE** [6]).

Furthermore, biological graphs are affected by inaccuracy also due to the methods exploited in order to discover protein-protein interactions (e.g., high throughput and computational methods). Although several techniques are able to exploit the specific reliability indices provided by the available interaction datasets (e.g., MINT [9]) as suitable filters during the clustering process, many efforts are still needed to make the clustering techniques more robust to such a kind of noise.

Finally, all the considered methods, with the exception of **SWEMODE** [30], cluster the input biological graph only on the basis of topological connections. An interesting challenge would be that of combining the main advantages of the considered approaches with taking into account also possible properties of the nodes, such as protein sequence similarity, Gene Ontology annotations [5] or functional domain composition of proteins [50].

4 Conclusion

In this paper, we presented a compact survey of graph-based clustering methods for detecting protein complexes in a PPI network. We proposed a classification based on five main categories, that are, local neighbourhood density search, cost-based local search, flow simulation, statistical measures and population-based stochastic search. We summarized the main algorithmic features and software

Table 1. Summary of some characteristics of the methods. The first column report the method acronym and reference, in chronological order. The second column reports the topological structure a method searches (a = arbitrary, d = dense sub-graphs). The approach each method is based on is reported in the third one. The fourth column (Simult.) 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).

Method	Structure	Approach	Simult.	Overlap	Un. Prot.	Software
MCL [13]	a	FS	yes	no	no	yes
SL [45]	a	SM	no	no	no	no
MCODE [6]	d	LN	yes	no	yes	yes
RNSC [24]	d	CL	yes	no	yes	yes
STM [19]	a	FS	yes	yes	yes	no
SWECODE [30]	d	LN	no	no	yes	no
DPClus [3]	d	LN	yes	no	yes	yes
IFB [10]	a	FS	no	yes	yes	no
Farutin [14]	a	SM	no	yes	no	no
CFinder [1]	d	LN	yes	yes	yes	yes
CGA [29]	d	PS	yes	yes	yes	no
PCP[11]	d	LN	no	yes	yes	yes
DECAFF [26]	d	LN	no	yes	yes	no
MF-PINCoC [38]	a	LN	no	yes	no	yes
Qcut [44]	d	CL	yes	no	no	yes
DME [16]	d	LN	no	yes	yes	yes
RRW [31]	a	FS	yes	no	no	yes
ModuLand [25]	d	CL	yes	yes	no	yes
IGA [42]	d	PS	yes	yes	no	no

availability of the considered methods, by also discussing their possible limitations. Finally, we pointed out some open issues related to the problem of clustering PPI networks.

We hope that the overview presented in this paper will be used by both computer scientists and practitioners as a quick reference for guiding the selection, use and development of algorithms for discoverying protein complexes and functions through the analysis of PPI networks.

References

1. B. Adamcsek, G. Palla, I. J. Farkas, I. Dernyi, and T. Vicsek. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.
2. B. Aittokallio and B. Schwikowski. Graph-based methods for analyzing networks in cell biology. *Briefing in Bioinformatics*, 7(3):243–255, 2006.

3. M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 7(207), 2006.
4. V. Arnau, S. Mars, and I. Marìn. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 21(3):364–378, 2005.
5. S. Asburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, and et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25:25–29, 2000.
6. G. Bader and H. Hogue. An automated method for finding molecular complexes in large protein-protein interaction networks. *BMC Bioinformatics*, 4(2), 2003.
7. M. Blatt, S. Wiseman, and E. Domany. Superparamagnetic clustering of data. *Phisical Review Letters*, 76(18):3251–3254, 1996.
8. S. Brohèe and J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, 2006.
9. A. Ceol et al. Mint, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(Database issue):D532–D539, 2010.
10. Y.-R. Cho, W. Hwang, and A. Zhang. Identification of overlapping functional modules in protein interaction networks: Information flow-based approach. In *Proc. of the Sixth Int. Conf. on Data Mining-Workshops (ICDMW'06)*, 2006.
11. H.N. Chua, K. Ning, W.K. Sung, H.W. Leong, and L. Wong. Using indirect protein-protein interactions for protein complex prediction. In *Proceedings of Computational Systems Bioinformatics Conference (CSB07)*, pages 97–109, 2007.
12. I. Derenyi, Gergely Palla, and Tamas Vicsek. Clique percolation in random networks. *Physical Review Letters*, 94(16):160–202, 2005.
13. A.J. Enright, S.V. Dongen, and C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–84, 2002.
14. V. Farutin, K. Robinson, E. Lightcap, V. Dancik, A. Ruttenberg, S. Letovsky, and J. Pradines. Edge-count probabilities for the identification of local protein communities and their organization. *Proteins: Structure, Function, and Bioinformatics*, 62:800–818, 2006.
15. Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.
16. E. Georgii, S. Dietmann, T. Uno, P. Pagel, and K. Tsuda. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics*, 25(7):933–940, 2009.
17. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. In *Proc. National. Academy of Science. USA 99*, pages 7821–7826, 2002.
18. L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. Clustering algorithm based graph connectivity. *Nature*, 402:C47–C52, 1999.
19. W. Hwang, Y.-R. Cho, A. Zhang, and M. Ramanathan. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms for Molecular Biology*, 1(24), 2006.
20. R. D. A. Jain. *Algorithms for Clustering Data*. Prentice Hall, 1988.
21. P. Jancura and E. Marchiori. Detecting high quality complexes in a PPI network by edge deletion and node expansion. In *CIBB*, 2011.
22. P. Jancura, E. Mavridou, E. Carrillo-De Santa Pau, and E. Marchiori. A methodology for detecting the orthology signal in a ppi network at a functional complex level. *BMC Bioinformatics*, 2011. accepted for publication.
23. Peng Jiang and Mona Singh. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics (Oxford, England)*, 26(8):1105–1111, 2010.

24. A. D. King, Natasa Przulj, , and Igor Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004.
25. Istvan A. Kovacs, Robin Palotai, Mate S. Szalay, and Peter Csermely. Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS ONE*, 5(9), 2010.
26. XL Li, CS Foo, and SK Ng. Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. In *Proceedings of Computational Systems Bioinformatics Conference (CSB07)*, pages 157–168, 2007.
27. XL Li, Min Wu, CK Kwoh, and SK Ng. Computational approaches for detecting protein complexes from protein interaction network: a survey. *BMC Bioinformatics*, 9, 2010.
28. C. Lin, Y. Cho, W. Hwang, P. Pei, and A. Zhang. Clustering methods in protein-protein interaction network. *in Knowledge Discovery in Bioinformatics: Techniques, Methods and Application, John Wiley & Sons,Inc*, 2006.
29. H. Liu and J. Liu. Clustering protein interaction data through chaotic genetic algorithm. In *Simulated Evolution and Learning*, volume 4247 of *Lecture Notes in Computer Science*, pages 858–864. Springer Berlin / Heidelberg, 2006.
30. Z. Lubovac, J. Gamalielsson, and B. Olsson. Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, 64:948–959, 2006.
31. K. Macropol, T. Can, and A. Singh. Rrw: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, 10(1):283, 2009.
32. S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Trans. on Comp. Biol. and Bioinf.*, 1(1):24–45, 2004.
33. G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.
34. G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, Jan Aert, R. Schneider, and P. G. Bagos. Using graph theory to analyze biological networks. *BioData Mining*, 4(10), 2011.
35. P. Pei and A. Zhang. A two-step approach for clustering proteins based on protein interaction profiles. In *IEEE Int. Symposium on Bioinformatics and Bioengeneering (BIBE'2005)*, pages 201–209, 2005.
36. J. B. Pereira, A.J. Enright, and C.A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins: Structure, Fuctions, and Bioinformatics*, (20):49–57, 2004.
37. C. Pizzuti. GA-NET: a genetic algorithm for community detection in social networks. In *Proc. of the 10th Intenational Conference on Parallel Problem Solving from Nature (PPSN 2008)*, pages 1081–1090, 2008.
38. C. Pizzuti and S. E. Rombo. Pincoc: a co-clustering based approach to analyze protein-protein interaction networks. In *Proc. of the 8th Intern. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL'07)*, pages 821–830, 2007.
39. C. Pizzuti and S. E. Rombo. Multi-functional protein clustering in ppi networks. In *Proc. of the 2nd Int. Conf. on Bioinf. Res. and Dev. (BIRD'08)*, pages 318–330, 2008.
40. C. Pizzuti and S. E. Rombo. *Discovering Protein Complexes in Protein Interaction Networks in Biological Data Mining in Protein Interaction Networks, Xiao-Li Li & See-Kiong NG Eds.* IGI Global- Medical Inf. Science Ref., 2009.
41. N. Przulj. Functional topology in a network of protein interactions. *in Knowledge Discovery in Proteomics, edited by I. Jurisica and D. Wigle, CRC Press*, 2005.

42. H. Ravaee, A. Masoudi-Nejad, S. Omidi, and A. Moeini. Improved immune genetic algorithm for clustering protein-protein interaction network. In *Proceedings of the 2010 IEEE International Conference on Bioinformatics and Bioengineering*, BIBE '10, pages 174–179. IEEE Computer Society, 2010.
43. A. W. Rives and T. Galitski. Modular organization of cellular networks. *Proc. of the National Academy of Science*, 100(3):1128–1133, 2003.
44. Jianhua Ruan and Weixiong Zhang. Identifying network communities with a high resolution. *Physical Review E*, 77(1), January 2008.
45. M.P. Samantha and S. Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. of the National Academy of Science*, 100(22):12579–12583, 2003.
46. M. Tasgin and H. Bingol. Community detection in complex networks using genetic algorithm. In *arXiv:0711.0491, 2007*, 2007.
47. S. Tornw and H.W. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Research*, 31(21):6283–6289, 2003.
48. D. von Mering, C. Krause, and *et al.* Comparative assessment of a large-scale data sets of protein-protein interactions. *Nature*, 31:399–403, 2002.
49. J. Wang, M. li, Y. Deng, and Yi Pan. Recent advances in clustering methods for protein interaction networks. *BMC Genomics*, 11(S10), 2010.
50. S. Zhang, H. Chen, K. Liu, and Z. Sun. Inferring protein function by domain context similarities in protein-protein interaction networks. *BMC Bioinformatics*, 10:395, 2009.
51. Y. Zhao, E. Levina, and J. Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.