Effect of community structures in protein–protein interaction network in cancer protein identification

Sminu Izudheen*, Eljose S. Sajan, Ivan George, Jeevan John and Chris Shaju Attipetty

Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology, Kochi 682 039, India

Protein interactions determine molecular and cellular mechanisms which control healthy and diseased states in organisms. Hence, a protein interaction network can be used to make scientific abstractions to understand mechanisms that trigger the onset and progress of diseases like cancer. Tumour-promoting function of several aberrantly expressed proteins in the cancerous state depends on their ability to interact with their protein-binding partners. Therefore, exploring more about these abnormal protein–protein interactions (PPIs) can help in identifying the disease pathway. This study examines the effect of community structures in the PPI network in cancer protein identification. It also provides a detailed analysis of topological properties of cancer, cancer chance and non-cancer proteins in the PPI network.

Keywords: Biological networks, cancer, protein–protein interaction, topological characteristics.

PROTEINS and genes act as the basic building blocks of any living organism, influencing its phenotype and genotype respectively. Any dysfunction or mutation in them can lead to genetic diseases and disorders. Therefore, identifying genes whose expression is associated with a specific phenotype is a key step in understanding disease mechanisms and developing targeted diagnostic and therapeutic interventions[1]. Traditional methods such as positional cloning via linkage analysis were applied for disease–gene mapping; but they encountered challenges, the most significant being the large number of genes among large family datasets that need to be analysed[2]. This is a labour-intensive task, which costs a staggering amount of manpower and resources to complete[3]. So, a large number of alternative methods have been devised for mapping or predicting the disease–gene relation; such as gene-functional annotations[4], sequence-based[5] and network-based analysis[6]. Due to the advent of high-throughput computational methods, computational approaches towards human genome sequencing have improved over the years. This has resulted in the creation of a number of protein–protein interaction (PPI) networks, like the human protein reference database (HPRD), molecular interaction NeTwork (MINT), UniProt database, etc. and has greatly helped in improving the computational approach to disease–gene mapping[7,8]. Many biological processes are involved in the formation of protein–protein complexes, and each of these functions consists of a specific PPI[9]. Hence a number of candidate gene discovery methods have been proposed based on PPI network analysis[4,5,7,10,11]. These techniques are based on the principle that genes associated with the same or similar disease phenotype are not randomly distributed in the interaction network, but rather they cluster together and have common topological features[12,13]. Based on these topological features, several gene scoring criteria and methods have been developed. For example, Izudheen and Mathew[14] developed a cancer protein identifier based on five graph centrality measures. A long-held and partially proved theory by biologists is that genes associated with some or similar disease phenotypes are likely to be functionally related and hence reside close to each other in a molecular network[3]. Hence module structures are an important property of these PPI networks[15], i.e. proteins or nodes with high interactivity tend to cluster together. In network theory, this clustering of highly interactive nodes can be characterized by the concept of community structures[16,17]. Communities (also called clusters or modules) are groups of vertices which most likely share certain common features or properties, and/or play similar roles within the network[16]. Every community detection algorithm makes different assumptions on the definition of a community. Some algorithms base their definition upon removal of high-betweenness edges[18], while some others aim at mining dense subgraphs[17], modularity measure[19], etc. Though the concept of community structures is a well-known feature in network theory, serious research focus from a computer science perspective came after the work done by Girvan and Newman[18]. The main limitation of the Newman–Girvan algorithm was that it could not detect overlapping community structures; a

*For correspondence. (e-mail: sminu_i@rajagiritech.edu.in)

standard and well-observed feature in most real-world datasets[16]. Hence, research in finding overlapping community structures had gained wide attention in the past few years. The most popular and widely used technique for detecting overlapping community structures, according to Fortunato[16], is the clique percolation technique by Palla *et al.*[17]. A clique is a group of nodes in a network such that every node is connected to every other node. Palla *et al.*[17] define a community or more specifically a percolated k-clique community as a group of k-cliques which are connected to each other by adjacent k-cliques; where adjacent implies that they share at least $k-1$ nodes. This definition of a community ensures that its member nodes are reachable through well-connected subsets of nodes. Therefore, it is quite possible that certain nodes can be part of subsets which may belong to another community. Hence a single node can belong to several communities; resulting in a number of overlapping communities. Though clique percolation is the most widely used overlapping community structure detection method, it is computationally challenging as it requires testing cliques against other cliques with which they share some nodes, but do not percolate[20]. The algorithms proposed by Reid *et al.*[20] consistently outperformed other clique percolation-based algorithms, namely CFinder[21] and SCP[22]. Hence, the method proposed by Reid *et al.*[20] to extract community structure from the PPI network is used in this study. As proteins perform their functions in a modular fashion, mutations of proteins in the same module may lead to similar disease phenotype[23]. These modular structures can be characterized by community structures and therefore community structures should be a more direct and robust property to capture the functional modularity in PPI networks. Genes associated with the same or similar disease phenotypes commonly reside in the same community and hence community structures may greatly help in the disease gene mapping.

Methodology

A protein or gene can be tested for its competence or similarity with other proteins (or genes) based on its attributes; if two proteins exhibit similar functionalities and attributes, then it is assumed that they are functionally similar. Presently, however, the available functional attributes of proteins and genes are limited. On the other hand, due to improved high performance computational methods of the last few years, and expertly curated and verified PPI network, the research focus for similarity testing has turned towards network topological attributes for PPI. In graph theory, centrality measures are key components in answering the following question – which are the most important nodes in the network? The centrality measures that have been used here are listed below.

Degree centrality

Degree centrality (DC) of a node is defined as the number of edges incident upon that node, which indicates the number of direct neighbours of that node.

$$\mathrm{DC}(i) = \sum_{j=1}^{n} A_{ij}, \tag{1}$$

where A is the adjacency matrix and n is the total number of vertices in graph $G = (V, E)$. Here, DC values are normalized by dividing them by the maximum possible degree (i.e. $n-1$), where n is the number of nodes in the graph.

Eigenvector centrality

Eigenvector centrality (EC) of a node is the measure of influence of a node in the network. It computes the centrality of a node based on the centrality of its neighbours. EC can be calculated as

$$\mathrm{EC}(i) = \frac{1}{\lambda} \sum_{t \in N(i)} x_t, \tag{2}$$

where λ is the largest eigenvalue of A produced by the algorithm and v is a non-zero vector which is the corresponding eigenvector of λ.

Closeness centrality

Closeness centrality (CC) of a node is defined as the reciprocal of average length of the shortest path between a particular node and all the remaining nodes in the graph. Hence if CC of a node is large, the closer it is to all other nodes[24].

$$\mathrm{CC}(u) = \frac{(n-1)}{\sum_{v=1}^{n-1} d(v, u)}, \tag{3}$$

where $d(v, u)$ is the shortest path between v and u, and n is the number of nodes in the connected part of the graph containing the node. If the graph is not completely connected, CC for each connected part is computed separately, scaled by the part size.

Betweenness centrality

The betweenness centrality (BC) measure of a vertex quantifies the number of shortest paths between two other nodes that pass through this node. Hence it gives a sense about how important this node is in terms of its function as a bridge between two nodes. Therefore, BC of a node v is the sum of all pairs of shortest paths that pass through v.

$$\mathrm{BC}(v) = \sum_{s,t \in v} \frac{\sigma(s, t/v)}{\sigma(s, t)}, \quad (4)$$

where v is the set of nodes, $\sigma(s, t)$ the number of shortest paths between (s, t) and $\sigma(s, t/v)$ is the number of those paths passing through some node v other than s, t (ref. 25).

Community structures

Graphs representing real-world data are objects where order exists with disorder. Real-world networks display larger inhomogeneities, revealing a high level of order and organization[17]. The edge distribution is locally inhomogeneous, indicating that there is a higher frequency of edges within a certain boundary of nodes and low frequency between nodes on either side of these boundaries. This nature of real-world networks can be perfectly explained using the concept of community structures. The *k*-clique percolation method of detecting overlapping community structures as described by Reid *et al.*[20] is used here, which, according to its authors, performs consistently better than other clique percolation algorithms like CFinder[22] and SCP[23], especially real-world data like PPI networks. According to *k*-clique percolation method, two cliques of size k percolate with each other, if they share $k - 1$ nodes. Communities generated by the method are the maximal set of cliques satisfying the property that every clique in the set is reachable from every other clique in the set through a path connecting percolating pairs.

Scoring criteria employed

This study focuses on ranking all proteins from the PPI network based on path-based centrality measures. For this, all the four centrality measure values for the 9608 proteins in the network were found and four separate ranked lists created, in which each entry corresponds to a protein and the list is ordered in decreasing values of centrality measures. A correlation matrix was then generated to find the dependence between these centrality measures. Using the inference developed from the correlation matrix, a final rank reflecting various centrality measures was generated. This ranking scheme was developed on the assumption that highly active nodes would possess a more central position within the PPI network and hence would have higher centrality scores. Hence it is presumed that proteins occupying the top positions of the ranking scheme are highly active and hence it would be reasonable to speculate that these top-ranked entries are cancerous in nature. The next methodology used for protein–disease mapping is based on the concept of community structures in network theory, as these can be used to connect topological structures and real-world functional protein modules. After the communities were mined using *k*-clique percolation method[20], the total number of proteins and the number of disease proteins within each community were aggregated. As mutated genes which have been proven to cause cancer and tumour growth are highly active in their interactions with other genes, it was hypothesized that a majority of genes residing inside a community are cancerous in nature. This assumption was based on the fact that proteins within each community are densely interconnected, since definition of a community is based on the percolated clique theory. It can be speculated that any protein lying outside but connected to at least one of the proteins within the community is also highly active and has high affinity with the complexes formed by proteins comprising the community it has a direct link to. Hence, this study specifies a novel approach to predict cancer proteins based on their relationship with protein complexes.

Testing and observations

From the PPI data, a protein network was created. Using the various centrality values calculated, the following observations were made. As discussed before, the main limitation of the clique percolation method for overlapping community detection is striking a balance between good coverage of the input graph and preventing the formation of a super-cluster. For this, the community detection algorithm was run with k value ranging from 5 to 10. Figure 1 provides details about the communities detected.

The aforementioned number of communities for the different k values was observed after rerunning the algorithm 200 times. The readers can observe that, for $k = 5$, 70 communities were detected which, as a whole contained a total of 657 protein entries from the graph. For $k = 6$, the number of communities detected reduced to half, with only 30 communities being detected and comprising only 277 protein entries within them. This inverse relationship of decrease in the number of communities being detected with the increase in k value is observed for the rest of the tests. From these observations, choosing a k value of 5 gives reasonable coverage of the PPI network while at the same time preventing the formation of a super-community containing all the nodes.

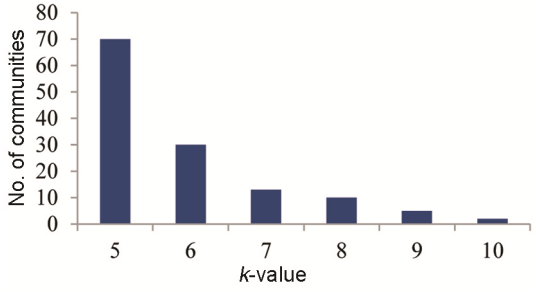

Figure 1. Communities detected based on k value.

Hence the reader should note that further discussions will be based on observations of $k = 5$. After the various communities of proteins were obtained, we had to test whether these communities or rather, the proteins within these communities, play any significant part in cancer formation or propagation. We had collected a list of oncogenes and TSP genes from various on-line resources like OMIM and GeneSignDB. After cross-referencing between the databases, 4630 true cancer proteins and about 33,673 cancer chance proteins were discovered. Among the 4630 true disease proteins, we eliminated multiple copies and aliases to obtain 1234 cancer proteins which are represented in the HPRD PPI network. Using the lists of true and chance disease proteins, we compared proteins within the communities and those outside but directly interacting with each of the communities (Figure 2).

Using five-clique percolation, over 657 protein entries across the 70-odd communities were detected. And 218 of these were true cancer proteins, while almost 417 matched with cancer chance proteins. This showed that 33.18% of proteins detected within the communities were known cancer-causing oncoproteins or TSPs, and 63.47% were cancer chance proteins having a high probability of being an oncoprotein or TSP. Hence almost 96.6% of the proteins within the communities that were detected when $k = 5$, were cancerous in nature.

This observation is in line with the initial assumption that cancer proteins are highly active and cluster together to form protein modules which are cancerous in nature. These protein modules can be represented mathematically using the community structures observed which, as stated previously, are also cancerous in nature. The clique percolation method permits the discovery of overlapping community structures. This feature allows us to capture proteins that are highly active, and can have special properties and importance. Hence, it was speculated that these highly overlapping proteins play an important role in various cancers.

From Figure 3, it may be noted that overlapping proteins are either known cancer proteins or have matched with an entry in the cancer chance protein lists. Since overlapping feature of a node implies that it is highly active in nature, it is reasonable to expect that almost all of the overlapping proteins are cancerous in nature and the predicted chance proteins are actual oncoproteins or TSPs. One particular protein that is to be considered is TP53. For $k = 5$, TSP was present in almost nine communities; the highest for any proteins. *TP53* is a known cancer gene that has been experimentally verified to be present in almost 63% of all known cancer types. In many cases, this acts as a propagator or starting point to the particular cancer mutation pathway. This nature can be explained perfectly by the high overlapping nature of TP53, being a single point of connection between nine different communities. Hence, TP53 (and other overlapping nodes for that matter) acts as a bridge between these protein complexes and plays an important role in the different cancer mutation pathways.

From the result obtained, it can be observed that almost 1000 cancer proteins and roughly 5000 cancer chance proteins do not lie inside any of the communities that we had detected. As stated before, proteins never function alone to cause any biological processes. Hence these cancer proteins lying outside any community can either form functional modules on their own, or they have an interactive relationship with some of the detected protein modules. Then the first case, though not entirely false, goes against the definition of communities acting as functional modules. This is because if a functional protein module is to be formed, then this would require a group of proteins showing high interaction with each other. This is the basic definition of a community. So the more logical method is to check the interaction between the cancer proteins outside any community and the communities themselves. For this, proteins that lie outside any community based on their hop distance from any one of the communities were detected with $k = 5$.

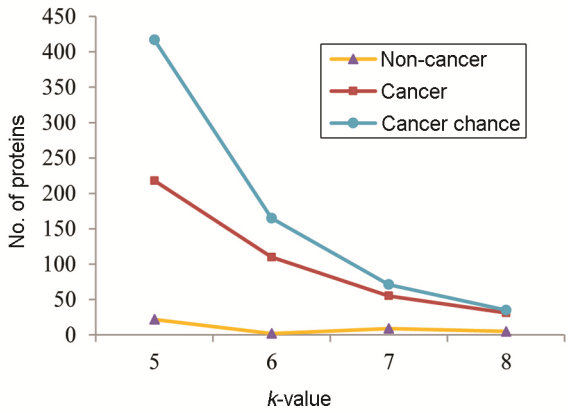

Figure 2. Protein distribution within the communities.

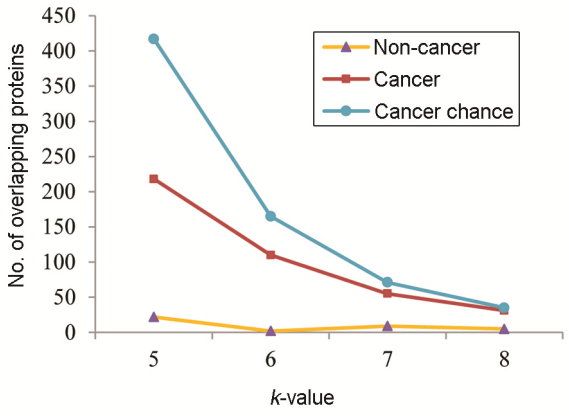

Figure 3. Overlapping proteins within the communities.

Figure 4 shows the statistics regarding cancer proteins lying outside any community. Of the 1016 known cancer proteins lying outside of any community, nearly 70% are directly interacting with at least one of the detected communities. The next 25% of true cancer proteins can be reached in two hops from some community. This implies that cancer proteins have a high interaction potential with the other proteins contained within the communities. One noteworthy observation is that, of all known cancer proteins lying outside any community, only 2.08% did not have any interaction with the communities that were detected. This indicates that almost 98% of all known cancer proteins belonging outside any of the communities will be a direct or an indirect neighbour to these communities. Based on the above observation, it can concluded that, of the remaining proteins lying outside of the communities which are not an entry in the known cancer protein list, if they have a hop distance of one, then they have a high probability of being cancerous in nature. Hence, using this metric of hop distance, it can predicted as a novel cancer protein.

Almost 7944 proteins lying outside of any community did not belong to the true cancer protein list. Running the same test, it was found that 46% of these non-cancer proteins are direct neighbours of the communities. Since cancerous proteins have high interactivity capability, it can be speculated that a majority of these non-cancerous proteins with one hop distance are novel cancer proteins or cancer chance proteins. From Figure 4 it may be noted that, of the 3691 non-cancer proteins within one hop distance, 66% are in the cancer chance list. This confirms our assumption that oncoproteins and TSPs will have a high interaction capability and are more likely to have a closer interaction with protein modules. It is also reasonable to conclude that some of the remaining non-cancerous proteins with one hop distance are novel cancer proteins.

As already observed, highly active nodes would possess a more central position within the PPI network and hence would have higher centrality scores. Therefore, ranking based on centrality measure must bring cancerous proteins to the top of the list. Figure 5 shows the distribution of top n proteins for various centrality measures.

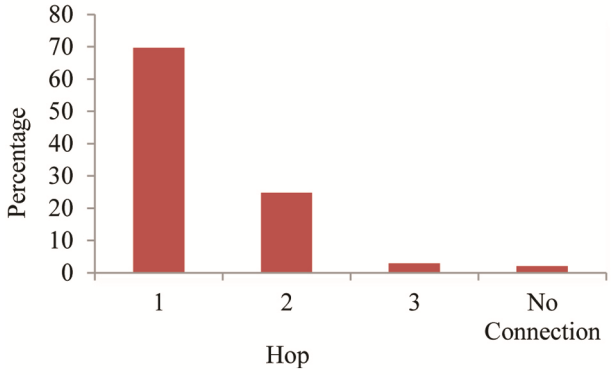

Figure 4. Cancer proteins outside the communities.

Among the four centrality measures considered, EC provided the highest percentage of disease protein identification, both in terms of true cancer proteins and cancer chance proteins. The second highest prediction accuracy was observed with CC. Both DC and BC have almost similar predictive capabilities. From the correlation matrix given in Table 1, it may be noted that there exists strong correlation between betweenness and degree. It may be noted that the prediction accuracy for EC, CC, DC and BC was 93.3, 80, 66.7 and 66.1 respectively. By considering these centrality measures, a weighted rank was generated as given in eq. (5)

$$r_i = w_1 e_i + w_2(1 - c_i) + w_3 d_i + w_4 d_i, \quad (5)$$

where w_1, w_2, w_3 and w_4 were 0.93, 0.8, 0.67 and 0.66 respectively, and represent the weights assigned based on prediction accuracy of EC, CC, DC and BC respectively. Here, e_i, c_i, d_i and b_i represent EC, CC, DC and BC for a protein i, normalized to the range [0, 1]. Using the rank generated from the weighted score, one would be able to predict whether a protein is cancerous or not; higher the rank, more likely it is to be a disease protein. The algorithm was tested on a dataset consisting of 657 proteins, of which 218 were true cancer proteins. From the confusion matrix given in Table 2, one can find that the precision and accuracy of the algorithm are 89.4% and 92.5% respectively.

To find the effect of community structures in cancer protein identification, community count of the proteins, i.e. the number of communities in which a protein is present was calculated. As cancer proteins are highly active, it would be reasonable to speculate that they may be present in more communities and hence have higher rank. A modified rank with this community count as the fifth parameter in addition to the four centrality measures was generated as given in eq. (6).

$$r_i = w_1 e_i + w_2(1 - c_i) + w_3 d_i + w_4 d_i + w_5 u_i, \quad (6)$$

where u_i is the community count of protein i and $w_5 = 0.69$, is the weight assigned based on prediction accuracy of community count. From the confusion matrix given in Table 3, precision and accuracy of the modified algorithm are 90.4% and 93% respectively, which is an improvement over previous ranking schemes.

As already mentioned, when $k = 5$, maximum 70 communities among 657 proteins were detected. To confirm the statistical significance of the result obtained, a comparison on centrality measures of proteins within a community with equal number of randomly selected proteins from the network was done. Table 4 presents the average value for normalized centrality measures for the five largest communities. It may be noted that communities

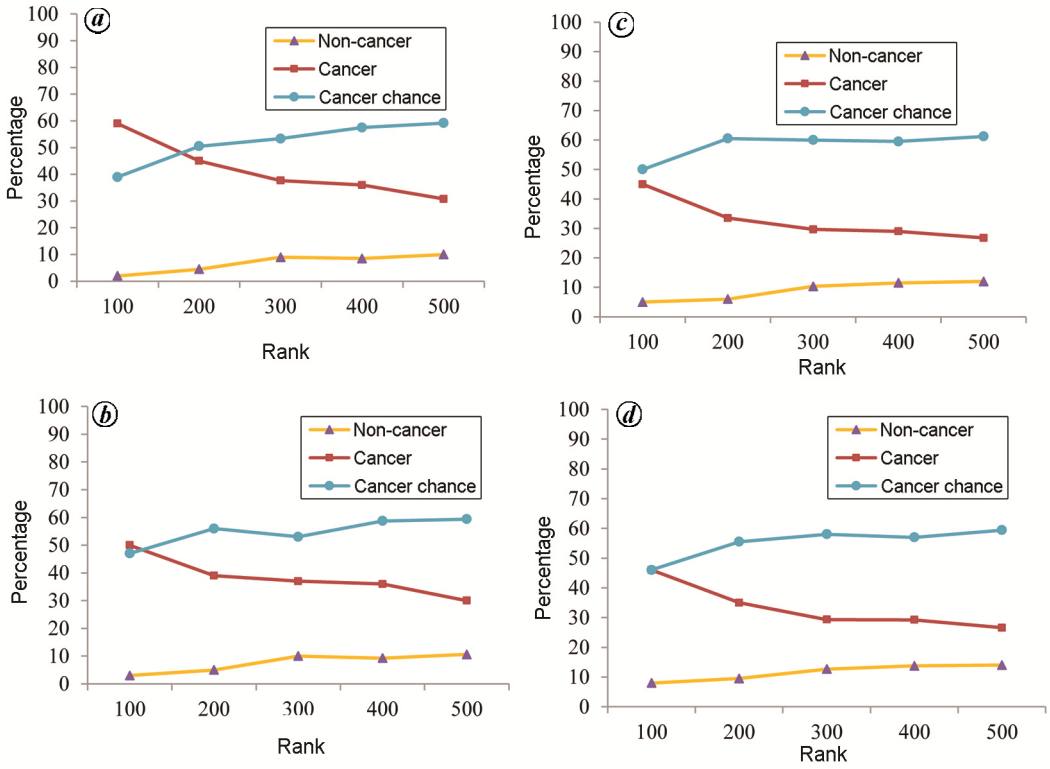

Figure 5. Protein distribution for: (***a***) eigenvector centrality, (***b***) closeness centrality, (***c***) betweenness centrality and (***d***) degree centrality.

Table 1. Correlation matrix for centrality measures

	Eigen vector	Betweenness	Closeness	Degree
Eigen vector	1	–0.01794	–0.0081	0.0039
Betweenness	–0.01794	1	–0.0499	0.8745
Closeness	–0.0081	–0.0499	1	–0.0218
Degree	0.0039	0.8745	–0.0218	1

Table 2. Confusion matrix without community count

		Predicted	
		Cancer	Non-cancer
Actual	Cancer	195	26
	Non-cancer	23	413

Table 3. Confusion matrix with community count

		Predicted	
		Cancer	Non-cancer
Actual	Cancer	198	25
	Non-cancer	21	411

are enriched with cancer and cancer chance proteins. Higher DC, higher BC, higher EC and lower CC of proteins within the community, which are the salient features of a cancer protein, assert the impact of community structures in cancer protein identification.

The functional relevance of the result obtained was evaluated by verifying the annotation of the protein using gene ontology tool. Table 5 gives the top 10 proteins listed by the algorithm and their details obtained from GeneCards[26]. It may be noted that all the proteins listed in the table are associated with some disorder leading to cancer. Observations that TP53 is an important protein in many of the mutation pathways occupying highest position in the consolidated centrality score and member of nine communities, reaffirm the correctness of our ranking criteria.

Conclusion

Cancer remains as a high-risk disease and the number of cancer cases reported in the past few years is alarming. Hence, a fast and efficient way of predicting the proteins involved in cancer formation and propagation is required. A systematic analysis on the topological properties, with a stress on community structures in the PPI network towards cancer protein identification is presented here. One of the most popular and scientifically verified overlapping community structure detection algorithms, viz. *k*-clique percolation method was used for identifying the communities. From the results it can be noted that most of the proteins lying inside these communities are cancerous in nature; either being known cancer proteins or having a high probability of being a novel cancer protein. Another

Table 4. Comparison of centrality values within community with random proteins from the network

Community size		Protein percentage		Total (cancer + cancer chance)	Average centrality			
		Cancer	Cancer chance		Eigen vector	Closeness	Degree	Betweenness
103	Within community	52.4	29.1	81.5	0.1419	0.3038	0.228	0.4455
	Random sampling	28.2	31.1	59.3	0.1213	0.3176	0.2175	0.0123
84	Within community	47.6	25	72.6	0.1683	0.3126	0.2258	0.346
	Random sampling	21.4	25	46.4	0.0945	0.3284	0.2132	0.134
67	Within community	44.8	23.9	68.7	0.1492	0.3128	0.217	0.234
	Random sampling	25.4	25.4	50.8	0.0913	0.3176	0.1912	0.114
52	Within community	34.6	34.6	69.2	0.1419	0.3054	0.208	0.152
	Random sampling	32.7	25	57.7	0.1113	0.3193	0.1846	0.122
45	Within community	24.4	33.3	57.7	0.1265	0.3009	0.1977	0.216
	Random sampling	28.9	24.4	53.3	0.0882	0.3184	0.1769	0.148

Table 5. Functional relevance of top centrality proteins

Symbol	Description	Disorder	No. of communities
TP53	Tumor protein P53	Li-Fraumeni syndrome	9
BRCA1	BRCA1, DNA repair associated	Breast and ovarian cancer	7
EP300	E1A-associated cellular p300 transcriptional co-activator protein	Colorectal cancer	5
SRC	SRC Proto-oncogene, non-receptor tyrosine kinase	Colorectal cancer	5
CREBBP	CREB binding protein	Acute myeloid leukaemia and neonatal leukaemia	4
ESR1	Estrogen receptor 1	Breast cancer and endometrial cancer	4
SMAD3	SMAD family member 3	Loeys–Dietz syndrome 3	4
EGFR	Epidermal growth factor receptor	Lung cancer	3
PRKCA	Protein kinase C alpha	Asbestos-related lung carcinoma	3
ATXN1	Ataxin 1	Cervical cancer	3

noteworthy observation is that almost all overlapping proteins are cancerous in nature and those with the highest overlap play a significant role in many mutation pathways. High overlapping nature of TP53, being a single point of connection between nine different communities, asserts that these overlapping proteins play an important role, either acting as the source of disease, or as a bridge protein in the many mutation pathways. Using the novel approach of hop distance, it has been shown that proteins having a smaller hop have higher interactivity capabilities and hence have high probability of being cancerous in nature. This new metric can also be used to predict novel cancer proteins, as majority of non-cancer proteins within a smaller hop pose cancerous capabilities and belong to cancer chance lists. This study also presents an analysis of the role of centrality measures in cancer protein identification.

1. Shah, S. D. and Braun, R., GeneSurrounder: Network-based identification of disease genes in expression data.
2. Drumm, M. L. *et al.*, Correction of the cystic fibrosis defect in vitro by retrovirus-mediated gene transfer. *Cell*, 1990, **62**, 1227–1233.
3. Hu1, Ke, Hu1, Jing-Bo, Xiang, Ju, Li, Hui-Jia, Zhang, Yan, Chen, Shi and Yi, Chen-He, Predicting disease-related genes by path-based similarity community structure in protein-protein interaction network; doi:10.1088/1742-5468/aae02b.
4. Liu, B., Jin, M. and Zeng, P., Prioritization of candidate disease genes by combining topological similarity and semantic similarity. *J. Biomed. Informat.*, 2015, **57**, 1–5.
5. Turner, F. S. and Clutterbuck, D. R., POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, 2003, **4**(11), R75.
6. Luo, J. and Liang, S., Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data. *J. Biomed. Informat.*, 2015, **53**, 229–236.
7. Navlakha, S. and Kingsford, C., The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 2010, **26**, 1057–1063.
8. Tiffin, N., Andrade-Navarro, M. A. and Perez-Iratxeta, C., Linking genes to diseases: its all in the data. *Genome Med.*, 2009, **1**(8), 77.
9. Idekar, T. and Sharan, R., Protein networks in diseases. *Genome Med.*, 2008, **18**(4), 644–652.
10. Wu, S.-Y., Shao, F.-J., Sun, R.-C., Sui, Y., Wang, Y. and Wang, J.-L., Analysis of human genes with protein–protein interaction network for detecting disease genes. *Physica A*, 2014, **398**, 217–228.
11. Li, M., Zhang, J., Liu, Q., Wang, J. and Wu, F.-X., Prediction of disease related genes based on weighted tissue-specific networks by using DNA methylation. *BMC Med. Genomics*, 2014, **7**, S4.

12. Oti, M. and Brunner, H. G., The modular nature of genetic diseases. *Clin. Genet.*, 2007, **71**(1), 1–11.
13. Liu, W., Sun, Z. and Xie, H., The analyses of human inherited disease and tissue-specific proteins in the interaction network. *J. Biomed. Informat.*, 2016, **61**, 10–18.
14. Izudheen, S. and Mathew, S., Cancer gene identification using graph centrality. *Curr. Sci.*, 2013, **105**(8), 1143–1148.
15. Gagneur, J., Krause, R., Bouwmeester, T. and Casari, G., Modular decomposition of protein–protein interaction networks. *Genome Biol.*, 2004, **5**, R57–R57.
16. Fortunato, S., Community detection in graphs. *Phys. Rep.*, 2010, **486**, 75–174.
17. Palla, G., Derenyi, I., Farkas, I. and Vicsek, T., Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, **435**, 814–818.
18. Girvan, M. and Newman, M. E. J., Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 2002, **99**(12), 7821–7826.
19. Newman, M. E. J., Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 2004, **69**, 066133.
20. Reid, F., McDaid, A. and Hurley, N., Percolation computation in complex networks. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM'12), 2012, pp. 274–281.
21. Adamcsek, B., Palla, G., Farkas, I., Derenyi, I. and Vicsek, T., C Finder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 2006, **22**(8), 1021–1023.
22. Kumpula, J., Kivela, M., Kaski, K. and Saramaki, J., Sequential algorithm for fast clique percolation. *Phys. Rev. E*, 2008, **78**(2), 026109.
23. Goh, K.-I. and Choi, I.-G., Exploring the human diseasome: the human disease network. *Brief. Funct. Genomics*, 2012, **11**, 533542.
24. Freeman, L., Centrality in social network: conceptual clarification. *Soc. Networks*, 1979, **1**, 215–239.
25. Brandis, U., On variants of shortest path betweenness centrality and their generic computations. *Soc. Networks*, 2008, **30**(2), 136–145.
26. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D., GeneCards: integrating information about genes, proteins and diseases. *Trends Genet*., 1997, **13**(4), 163.

ACKNOWLEDGEMENT. This work was funded by CERD Research Scheme Money of A. P. J. Abdul Kalam Technological University, India.

Received 25 January 2019; revised accepted 10 September 2019

doi: 10.18520/cs/v118/i1/62-69