cates the conformity of the claimed uncertainty of force realized by 50 kN dead weight force machine. The conformity of *En* value also confirms the equivalence of force realization by both the force machines. The conformity also states the equivalence of the force realization by both force realizing machines.

The 1 MN force standard machine with uncertainty of force realized 0.002% ($k = 2$) for forces between 1 and 100 kN using dead weights and 0.009% ($k = 2$) for forces between 10 and 1000 kN using lever multiplication system, has been recently established by GTM, Germany and metrologically characterized by PTB, Germany, by means of intercomparison. Hence, an intercomparison has been carried out between 50 kN dead weight force machine and 1 MN force standard machine serving as primary standard force machines in the range 1–50 kN to affirm the metrological capabilities of the former. The study confirms that the normalized error (*En* value) computed is within the permissible limits and hence, the claimed uncertainty of 50 kN dead weight force machine has been justified.

1. NPLI, Technical Bulletin, National Physical Laboratory, New Delhi, 3–4 April 1987.
2. Jain, K. K., Jain, S. K., Dhawan, J. K. and Anil Kumar, Realization of force scale up to 50 kN through dead weight force. *Mapan – J. Metrol. Soc. India*, 2005, **20**, 249–257.
3. Jain, S. K., Harish Kumar and Titus, S. S. K., Tegtmeier, F., Prenzlow, N. and Schwind, D., Metrological characterization of the new 1 MN force standard machine of NPL India. *Measurement*, 2012, **45**, 590–596.
4. Woger, W., Remarks on the *En*-criterion used in measurement comparisons. *Internationale Zusammenarbeit, PTB-Mitteilungen*, 1999, **109**, 24–27.
5. Sawla, A., Uncertainty scope of the force calibration machines. In Proceedings of XVIth IMEKO World Congress, Vienna, 25–28 September 2000.
6. Metallic materials – calibration of force proving instruments used for verification of uniaxial testing machines. ISO 376:2011.
7. International Standards Organization, Guide for expression of uncertainty in measurement. ISO GUM Document, 1995.
8. Kumme, R. and Brito, C., Investigations of the measurement uncertainty of the force standard machines of the IPQ by intercomparison measurements with PTB. In Proceedings of the International Conference on Force, Mass, Torque and Pressure Measurement, IMEKO TC 3, Istanbul, Turkey, September 2001.
9. Heamawatanachai, S., Chaemthet, K., Sumyong, N. and Amornsakun, C., Analytical study on the uncertainty of load cells calibrated with dead weight force comparator machine. In Proceedings of the International Conference on Mechanical Engineering, Krabi, October 2011.
10. Rajesh Kumar, Harish Kumar, Anil Kumar and Vikram, Long term uncertainty investigations of 1 MN force calibration machine at NPL, India (NPLI). *Meas. Sci. Rev.*, 2012, **12**, 149–152.

ACKNOWLEDGEMENTS. We thank the Director, the Head, Apex Level Standards & Industrial Metrology Group, and the members of Force and Hardness Standard Group, CSIR-National Physical Laboratory, New Delhi for support.

Received 4 March 2013; revised accepted 8 August 2013

Cancer gene identification using graph centrality

Sminu Izudheen[1,]* and Sheena Mathew[2]

[1]Department of Computer Science,
Rajagiri School of Engineering and Technology, Kochi 682 039, India
[2]Division of Computer Engineering, School of Engineering,
Cochin University of Science and Technology, Cochin 682 022, India

One of the most significant challenges of modern bioinformatics is in the development of computational tools to understand and treat diseases like cancer. So far, a variety of methods have been explored for identifying candidate cancer genes. Since protein interactions carry out most biological processes, we propose an algorithm for identifying cancer genes from graph centrality values of the human protein–protein interaction network. The precision and accuracy of the results obtained while applying the method on actual protein–protein interaction data assert that it can be used as an effective model to identify novel cancer proteins.

Keywords: Biological networks, cancer gene identification, graph centrality, network characteristics, protein–protein interaction.

PROTEIN–PROTEIN interactions (PPIs) are fundamental to virtually every cellular process[1]. They can inactive a protein, alter the kinetic properties of proteins, result in the formation of a new binding site or change the specificity of a protein for its substrate. The past few decades have marked many major milestones in understanding PPIs and thereby exploring more about these complex biological systems[2]. Protein complexes performing a specific biological function often contain highly connected protein modules[3]. Study about these protein modules plays a crucial role in understanding the pathophysiological properties of complex diseases like cancer.

Cancer is a disease caused by uncontrolled growth of abnormal cells in the body. There are over 200 types of cancers and it is estimated that about 9 million new cancer cases are diagnosed every year and over 4.5 million people die from the disease each year in the world. Early detection of cancer can greatly improve the odds of successful treatment and survival. It is an extremely complex genetic disease and almost 5–10% of human genes contribute to the genesis of cancer, but only 1% has been identified so far. As cancer is caused by uncontrolled growth of cells, a systematic examination of the proteins encoding cancer genes in the protein–protein network may help us to identify new candidate genes.

In this communication, we made an approach to identify cancer genes from PPIs. The algorithm focuses on

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

five different graph centrality values of a protein network, viz. degree, shortest path distance between two proteins, betweenness centrality, eigen centrality and clustering coefficient. When we applied the method to real PPI data, we were able to obtain an accuracy of 83%.

Here we propose a method to identify novel cancer proteins using different graph centrality values of protein interaction network. An overview of the proposed method is given in Figure 1. We collect PPI data from public databases. Since the size of the PPI data is large, we sample the data by randomly selecting *k* cancer proteins and *k* non-cancer proteins from the PPI data. A subset of interactions is then generated by selecting all interactions of these selected proteins from the PPI data. The data are then represented as an adjacency matrix and various graph centrality values are calculated. Rank of the protein for each centrality parameter is then generated. From these individual ranks the final score for each protein is calculated. The process is repeated on samples generated from another *k* cancer proteins and *k* non-cancer proteins. After each iteration the final rank is modified. The process is repeated until there is no change in the final score. Figure 1 provides a schematic overview of the method showing steps involved in a single iteration. Details of the steps are discussed in the following and then validation methods are presented.

PPI network is constructed by extracting data based on experimental evidence from Human Protein Reference Database (HPRD)[4]. The dataset used for the study is the latest dataset of HPRD created on April 2010 consisting of 39,240 interactions. After removing redundancy and self-interactions, we had 37,080 protein interactions among 7932 proteins.

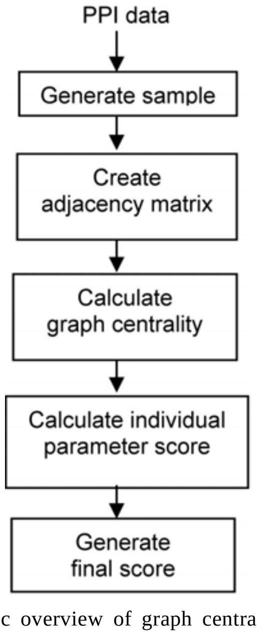

Figure 1. Schematic overview of graph centrality based method to identify cancer genes.

To identify the importance of a node in a graph, different centrality values are suggested. In the proposed method, we prioritize the nodes by calculating five centrality measures. A protein network $G(V, E)$, is represented as an *nxn* adjacency matrix *A*, where *n* is $|V|$, and various graph centrality values calculated are as follows.

Degree centrality (DC) of a node is the number of edges connected to that node, which represents the number of immediate neighbours of the node[5].

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

Shortest path (SP) between a pair of nodes is the shortest distance between the nodes. Average shortest path is the average of the shortest paths from this node to all other nodes in the network.

$$\mathrm{SP}(i) = \frac{1}{n}\sum_{j=1}^{n} p_{ij}, \tag{2}$$

where $\sum_{j=1}^{n} p_{ij}$ denotes the shortest path between *i* and *j*.

Betweenness centrality (BC) of the node is the number of shortest paths passing through that node[6]. It shows how important the node is in the network.

$$\mathrm{BC}(i) = \sum_{s\neq i\neq t} \frac{\sigma_{s,t}(i)}{\sigma_{s,t}}, \tag{3}$$

where $\sigma_{s,t}(i)$ is the shortest path between *s* and *t* passing through *i*.

Clustering coefficient (CC) of a node is the ratio between the number of neighbouring edges present to the number of possible neighbours. It shows how close the neighbours of the given node are among themselves. If l_i, represents the number of possible edges to a node *i*, then the clustering coefficient of the node is given by

$$\mathrm{CC}(i) = \sum_{j,k\in N} \frac{2\cdot A_{ij}\cdot A_{jk}\cdot A_{ki}}{l_i\cdot(l_i-1)}, \tag{4}$$

Eigenvector centrality (EC) of a node is the measure of the influence of a node in the network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. It is the *i*th component of the principal eigenvector of adjacency matrix *A* (ref. 7). Eigen centrality of a node *i* is given by

$$\mathrm{EC}(i) = \frac{1}{\lambda}\sum_{j\in G} A_{ij} EC(j), \tag{5}$$

where λ represents the largest eigenvalue of *A*.

When considering the above parameters, cancer genes score high for degree, betweenness and eigen centrality. At the same time, the scores for shortest path distance and clustering coefficient are low.

Proteins are ranked based on five centrality values in such a way that we expect the cancer protein to precede the list. For validation, we extracted a collection of 4630 cancer proteins and 33,673 proteins which are classified as cancer chance proteins from public databases like GeneSignDB[8] and compared those with the results.

From the PPI data, a protein network is created and we represent it as an adjacency matrix. From the various centrality values calculated, the following observations are noted.

The average degree of the cancer genes reported by the method is 15.52, which is 2.66 times higher than that of non-cancer genes, with an average degree of 5.83. This ratio is slightly higher than that reported by Jonsson and Bates[9], which was based on the predicted PPIs. The ratio of cancer and non-cancer proteins for each degree is given in Figure 2. From the figure it is clear that cancer proteins interact strongly with other proteins and show higher connectivity in the whole network.

Another parameter considered was the shortest path distance from a node to all other nodes in the network. The average value obtained for cancer and non-cancer proteins was 3.32 and 3.67 respectively. This indicates that the path from cancer proteins to other proteins is shorter than that from non-cancer proteins to other proteins. Figure 3 shows the ratio of cancer and non-cancer proteins for various shortest path distances. From this figure it is clear that the shortest path distance of cancer proteins is much shorter than that of non-cancer proteins.

We have also calculated the shortest path distance from cancer chance proteins to all other proteins in the network. Cancer chance proteins also show similar characteristics as cancer proteins. The average shortest path distance of cancer chance proteins reported by the method was 3.36. Comparison for the shortest path distances for cancer protein versus cancer chance protein and cancer chance protein versus non-cancer protein is given in Figure 3 *b* and *c* respectively. It may be noted that even though cancer chance proteins have higher average shortest path distance than cancer proteins, when compared with the non-cancer proteins, they have lower shortest path distance.

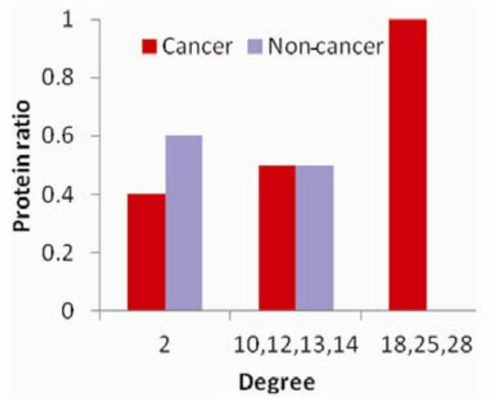

Figure 2. Degree distribution: cancer versus non-cancer proteins.

The next parameter we calculated was betweenness centrality. Betweenness centrality of a node is the number of shortest paths passing through the node, or it represents how important is the node in the network. The average betweenness value reported for cancer and non-cancer proteins was 470.3 and 41.36 respectively. Figure 4 *a* shows the ratio of cancer and non-cancer proteins for various betweenness values. From this figure it is clear that betweenness value for cancer proteins is higher than that of non-cancer proteins. Betweenness value for non-cancer proteins was 179.84, which is again greater than non-cancer proteins. Figure 4 *b* shows the comparison between cancer chance proteins and non-cancer proteins.

Clustering coefficient of a node is the ratio between the number of neighbouring edges present to the number of possible neighbours. It shows how well a node is connected among its direct neighbours. Results from our method shows that average clustering coefficient of cancer proteins is lesser than that of non-cancer proteins. Similar finding was observed between cancer chance proteins and non-cancer proteins. Average clustering coefficient for cancer proteins, cancer chance proteins and non-cancer proteins was 0.11, 0.09 and 0.23 respectively. Figure 5 *a* and *b* show the ratio of cancer to non-cancer proteins and cancer chance protein to non-cancer protein respectively.

The final parameter considered was eigenvector centrality, which measures the influence of a node in the network. We have calculated the eigenvector centrality of each node. Average centrality value for cancer proteins was 780.36, which is 3.9 times more than the non-cancer genes with an average value of 201.42. Cancer chance proteins also have larger eigen centrality (568.88) when compared with non-cancer proteins. Ratio of cancer to non-cancer proteins and cancer chance proteins to non-cancer proteins for various centrality values is given in Figure 6 *a* and *b* respectively.

The significance of the results obtained was statistically verified by performing the *t*-test[10]. For testing the hypothesis, we assume that the random variable *x* follows normal distribution with mean zero and standard deviation one. A significance level of 10% (i.e. 90% confidence) was used in the analysis. The null hypothesis (H_0) was selected so that there is no difference between centrality values of various groups. For example, the null hypothesis for testing the degree centrality for cancer and non-cancer proteins is given below.

$$H_0: \mu_{\text{cancer}} - \mu_{\text{non-cancer}} = 0.$$

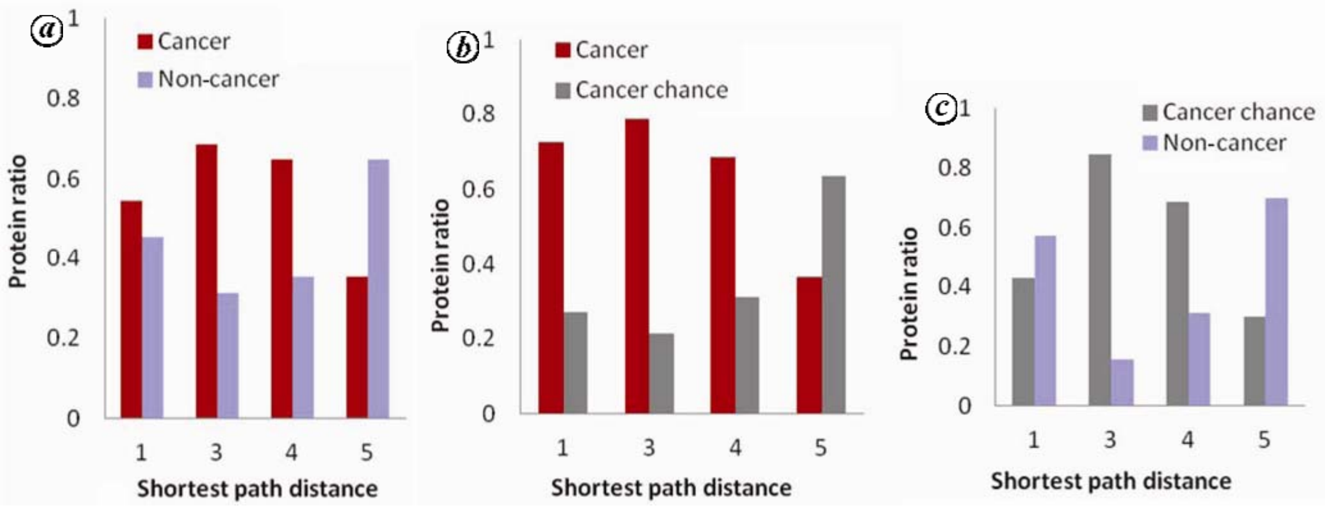

Figure 3. Shortest path distance. ***a***, Cancer versus non-cancer proteins; ***b***, Cancer versus cancer chance proteins; ***c***, Cancer chance versus non-cancer proteins.

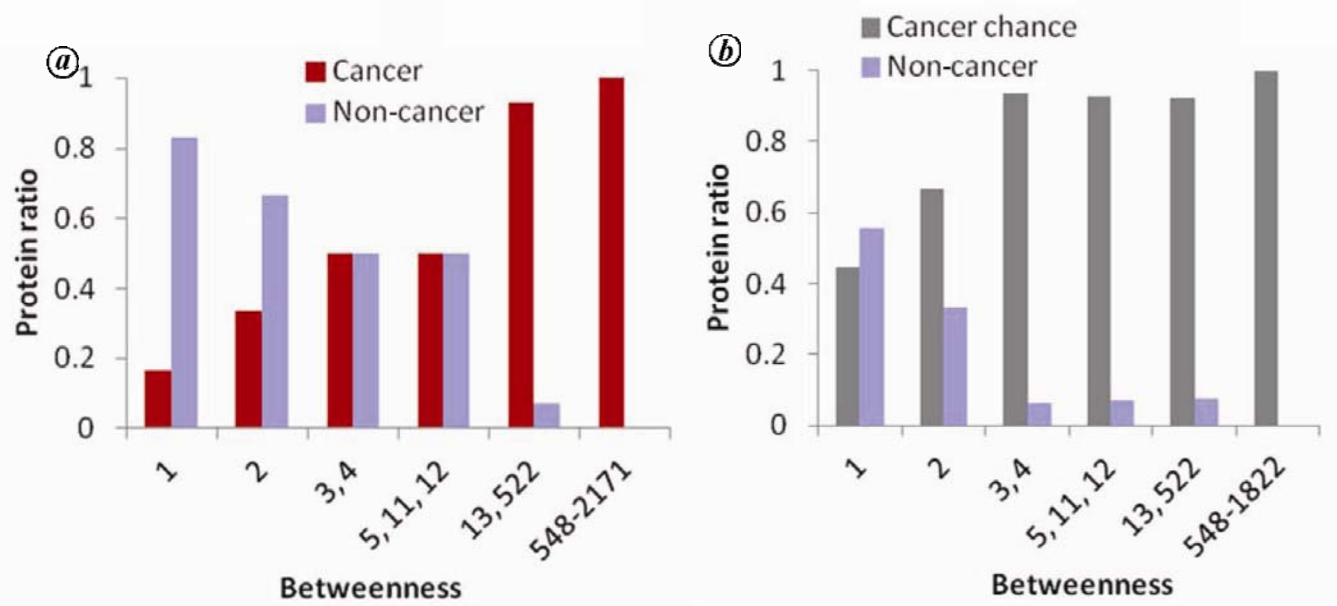

Figure 4. Betweenness centrality. ***a***, Cancer versus non-cancer proteins; ***b***, Cancer chance versus non-cancer proteins.

For testing the degree and betweenness centrality we performed right-tailed test. Here the critical point at 10% level of significance is 1.28. So we reject the null hypothesis, if the computed test statistics value is greater than 1.28; otherwise we accept the null hypothesis. The results of *t*-test for various centrality parameters is given in Table 1.

The computed value for degree centrality using the test statistics is 1.725. Since the value is greater than the observed table value, we reject the null hypothesis with 95% confidence. Hence we conclude that the degree of cancer proteins is greater than the degree of non-cancer proteins.

The test statistic for betweenness centrality between cancer and non-cancer proteins is 5.360 and for cancer chance proteins and non-cancer proteins is 2.880. Both values are greater than the observed value and we reject the null hypothesis in both cases with 99% confidence. Hence we conclude that betweenness centrality of both cancer and cancer chance proteins is greater than non-cancer proteins.

For testing the clustering coefficient, eigen centrality and shortest path distance, we performed left-tailed test. Here the critical point at 10% level of significance is –1.28. So we reject the null hypothesis, if the computed

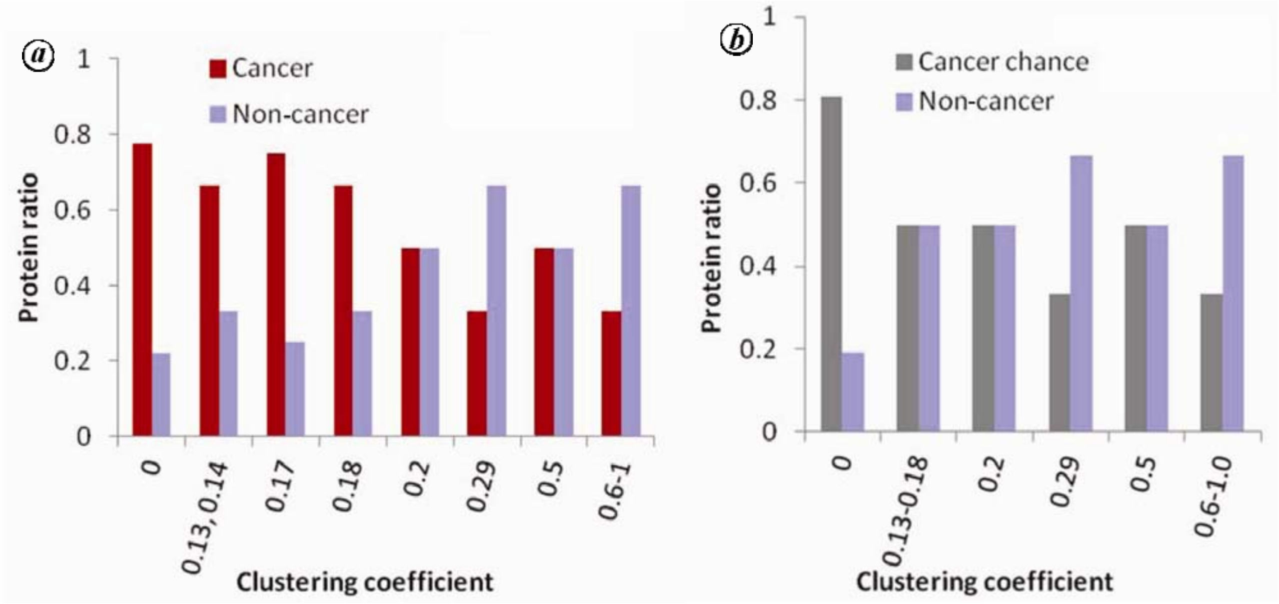

Figure 5. Clustering coefficient. ***a***, Cancer versus non-cancer proteins; ***b***, Cancer chance versus non-cancer proteins.

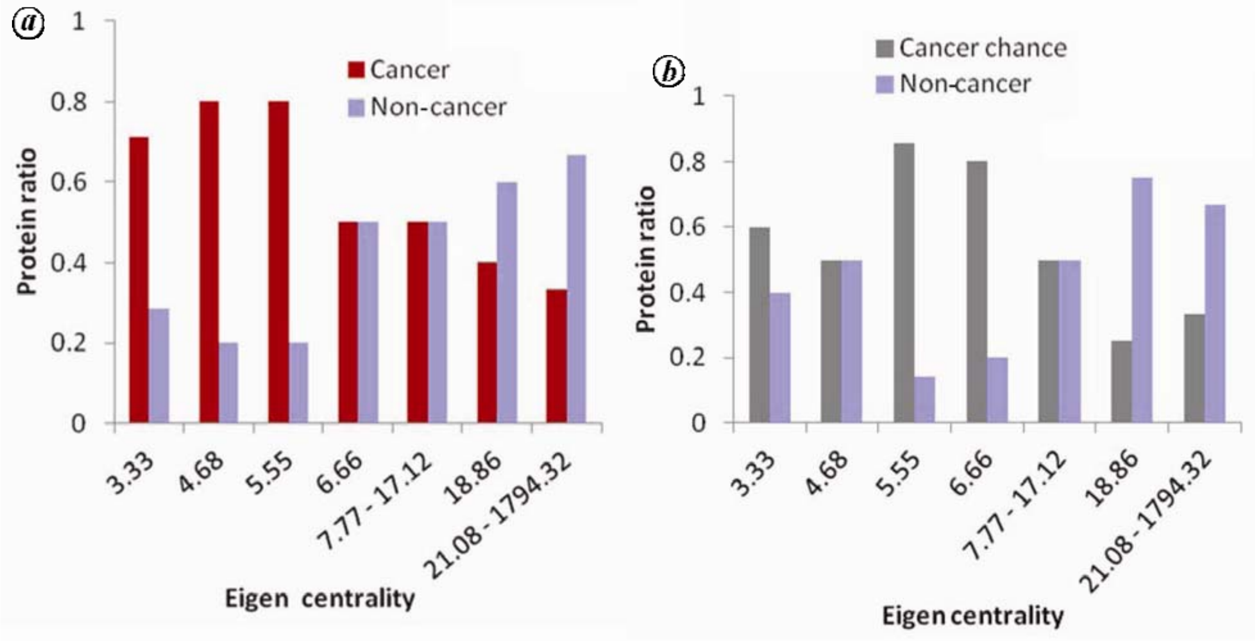

Figure 6. Eigen centrality. ***a***, Cancer versus non-cancer proteins; ***b***, Cancer chance versus non-cancer proteins.

test statistics value is less than –1.28; otherwise we accept the null hypothesis.

The test statistic for clustering coefficient between cancer and non-cancer proteins is –1.941 and for cancer chance proteins and non-cancer proteins is –2.109. Both values are less than the observed value and we reject the null hypothesis with 97% and 98% confidence respectively. Hence we conclude that clustering coefficient of both cancer and cancer chance proteins are less than non-cancer proteins.

The value obtained for eigen centrality between cancer and non-cancer proteins is –1.843 and for cancer chance proteins and non-cancer proteins is –1.909. Here also the values are less than the observed value and the null hypothesis is rejected with 96% and 97% confidence respectively. Hence we conclude that clustering coefficient

Table 1. Results of *t*-test for various centrality parameters

Parameter considered	*t*-statistic	Confidence (%)
degree$_{\text{cancer}}$ versus degree$_{\text{non-cancer}}$	1.725	95
BC$_{\text{cancer}}$ versus BC$_{\text{non-cancer}}$	5.360	99
BC$_{\text{cancer chance}}$ versus BC$_{\text{non-cancer}}$	2.880	99
CC$_{\text{cancer}}$ versus CC$_{\text{non-cancer}}$	–1.941	97
CC$_{\text{cancer chance}}$ versus CC$_{\text{non-cancer}}$	–2.109	98
EC$_{\text{cancer}}$ versus EC$_{\text{non-cancer}}$	–1.843	96
EC$_{\text{cancer chance}}$ versus EC$_{\text{non-cancer}}$	–1.909	97
SP$_{\text{cancer}}$ versus SP$_{\text{non-cancer}}$	–1.299	90
SP$_{\text{cancer chance}}$ versus SP$_{\text{non-cancer}}$	–1.313	90
SP$_{\text{cancer}}$ versus SP$_{\text{cancer chance}}$	–1.681	95

Table 2. Confusion matrix

	Predicted	
	Cancer	Non-cancer
Actual		
Cancer	75	10
Non-cancer	9	20

Table 3. Comparison matrices

Metrics	Result (%)
Sensitivity	88.23
Specificity	68.96

of both cancer and cancer chance proteins is less than non-cancer proteins.

The last parameter tested was shortest path distance. The test statistic obtained for cancer versus non-cancer proteins, cancer chance proteins versus non-cancer proteins and cancer versus cancer chance proteins was –1.299, –1.313 and –1.681 respectively. All the values are less than the observed value of –1.28. Hence the null hypothesis for the first two cases is rejected with 90% confidence and for the third case it is rejected with 95% confidence. From the values it can be concluded that shortest path distance for cancer and cancer chance proteins is less than non-cancer proteins. Similarly, the shortest path distance of cancer proteins is lesser than cancer chance proteins.

The present communication presents a method to predict cancer proteins from PPI data. These data were downloaded from HPRD database. Actual dataset extracted from HPRD contains 39,240 interactions. After removing redundancy and self-interactions, we had 37,080 protein interactions pairs among 7932 proteins. Because of computational limitations, we sampled the data and subsets of interactions were extracted. The above specified five centrality parameters were calculated for each protein and based on these scores we ranked the proteins, so that we expect cancer protein to precede the list. The process is repeated for *k* different samples and each time rank of a protein is modified, if necessary. Based on these *k* iterations, the final score is generated. A list of cancer proteins and those marked as cancer chance proteins was collected from various public databases. When we mapped the result with the data extracted from these public databases, the method marked a precision and accuracy of 88.23% and 83.3% respectively. The confusion matrix and comparison metrics of the method are presented in Tables 2 and Table 3 respectively.

Another quality measure calculated to assess the prediction was Matthews correlation coefficient (MCC)[11]. It takes into account true and false positives and negatives, and represents correlation coefficient between the observed and predicted binary classifications. It is a linear correlation coefficient which scales between –1 and +1, where +1, –1 and 0 represents perfect correlation, anti correlation and randomness respectively. MCC for our method was 0.566, which shows that it can be used as an effective method to identify novel cancer genes from protein interaction network.

The results of the method on actual PPI data showed that it can be used as an effective model to identify novel cancer proteins.

1. Hakes, L. *et al.*, Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 7999–8004.
2. Alm, E. and Arkin, A. P., Biological networks. *Curr. Opin. Struct. Biol.*, 2003, **13**, 193–202.
3. Harwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W., From molecular to modular cell biology. *Nature*, 1999, **402**, c47–c52.
4. Peri, S. *et al.*, Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, 2003, **13**, 2363–2371.
5. Jeong, H., Mason, S. P., Barabási, A. L. and Oltvai, Z. N., Lethality and centrality in protein networks. *Nature*, 2001, **411**, 41–42.
6. Joy, M. P., Brock, A., Ingber, D. E. and Huang, S., High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.*, 2005, **2005**, 96–103.
7. Bonacich, P., Power and centrality: a family of measures. *Am. J. Soc.*, 1987, **92**, 1170–1182.
8. Culhane, A. C. *et al.*, GeneSigDB – a curated database of gene expression signatures. *Nucleic Acids Res.*, 2010, **38**, D716–D725.
9. Jonsson, P. F. and Bates, P. A., Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 2006, **22**, 2291–2297.
10. Levin, R. I. and Rubin, D. S., *Statistics for Management*, Prentice-Hall, 1997.
11. Powers and David, M. W., Evaluation: from precision, recall and *F*-factor to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.*, 2011, **2**, 37–63.

Received 4 October 2012; revised accepted 10 July 2013