The centrality of cancer proteins in human protein-protein interaction network: a revisit

Wei Xiong, Luyu Xie and Shuigeng Zhou*

School of Computer Science
and Shanghai Key Lab of Intelligent Information Processing,
Fudan University,
Shanghai 200433, China
E-mail: wxiong@fudan.edu.cn
E-mail: taoistly@gmail.com
E-mail: sgzhou@fudan.edu.cn
*Corresponding author

Hui Liu

The Research Lab of Information Management,
Changzhou University,
Changzhou, Jiangsu 213164, China
E-mail: liuhui@fudan.edu.cn

Jihong Guan

Department of Computer Science and Technology,
Tongji University,
Shanghai 201804, China
E-mail: jhguan@tongji.edu.cn

Abstract: Topological analysis of protein-protein interaction (PPI) networks has been widely applied to the investigation on cancer mechanisms. However, there is still a debate on whether cancer proteins exhibit more topological centrality compared to the other proteins in the human PPI network. To resolve this debate, we first identified four sets of human proteins, and then mapped these proteins into the yeast PPI network by homologous genes. Finally, we compared these proteins' properties in human and yeast PPI networks. Experiments over two real datasets demonstrated that cancer proteins tend to have higher degree and smaller clustering coefficient than non-cancer proteins. Experimental results also validated that cancer proteins have larger betweenness centrality compared to the other proteins on the STRING dataset. However, on the BioGRID dataset, the average betweenness centrality of cancer proteins is larger than that of disease and control proteins, but smaller than that of essential proteins.

Keywords: protein-protein interaction network; cancer proteins; network topological centrality.

Reference to this paper should be made as follows: Xiong, W., Xie, L., Zhou, S., Liu, H. and Guan, J. (2014) 'The centrality of cancer proteins in human protein-protein interaction network: a revisit', *Int. J. Computational Biology and Drug Design*, Vol. 7, Nos. 2/3, pp.146–156.

Biographical notes: Wei Xiong received his Bachelor and Master degrees in computer science from Henan University, Kaifeng, Henan, China, in 2004 and 2008 respectively. He is currently a PhD candidate majored in Bioinformatics in Fudan University, Shanghai, China. His research interests include data mining, machine learning and Bioinformatics.

Luyu Xie is currently an undergraduate student at the School of Computer Science, Fudan University, Shanghai, China. His research interests include data mining and Bioinformatics.

Shuigeng Zhou received PhD in Computer science from Fudan University, Shanghai, China, in 2000. He is currently a Professor at the School of Computer Science, Fudan University. His research interests include data management, data mining and machine learning, and Bioinformatics. He has published more than 200 papers in domestic and international journals and conferences. He is a member of the IEEE, ACM and IEICE.

Hui Liu received PhD in Computer Science from Fudan University, Shanghai, China, in 2009. He is currently an Assistant Professor at the Research Laboratory of Information Management, Changzhou University, Jiangsu, China. He has published several papers in domestic and international journals. His research interests include machine learning, data mining, and computational biology.

Jihong Guan received PhD from Wuhan University, Wuhan, in 2002. She is currently a Professor in the Department of Computer Science and Technology, Tongji University, Shanghai, China. She has published more than 100 papers in domestic and international journals and conferences. Her research interests include databases, data mining, distributed computing, bioinformatics, and geographic information systems.

This paper is a revised and expanded version of a paper entitled 'The centrality of cancer proteins in human protein-protein interaction network: a revisit' presented at the *International Conference on Intelligent Biology and Medicine (ICIBM 2013)*, Nashville, TN, USA, 11–13 August, 2013.

1 Introduction

Proteins are essential macromolecules of life and protein-protein interactions play important roles in many biological processes in the cell such as DNA replication, transcription and translation, signal transduction and recognition of foreign molecules etc. In recent years, the rapid development of high-throughput experimental techniques, especially yeast two-hybrid (Y2H) (Suter et al., 1998) and affinity purification coupled to mass spectrometry (AP-MS) (Ho et al., 2003), has led to the production of vast amounts of PPI

data. Meanwhile, a large number of computational methods have been developed to predict PPIs, such as gene neighbourhood (Ferrer et al., 2010), gene fusion (Reid et al., 2010), gene co-expression (Lukk et al., 2010) and text-mining techniques (Tikk et al., 2010) etc. PPI data are always represented as networks, where each node corresponds to a protein and each edge corresponds to an interaction between a pair of proteins. To date, PPI data have been widely applied to predicting disease-causing genes (Oti et al., 2006; GÖstlund and Sonnhammer, 2010), searching dynamic modularities (Taylor et al., 2009), identifying disease-related subnetworks (Pujana et al., 2007), and discovering drug targets (Hormozdiari et al., 2010) etc.

Cancer is one of the most severe human diseases. With the large amounts of PPI data available, a systematic examination on the proteins encoded by cancer genes (cancer proteins) in the human PPI network may help us to gain a deeper understanding of the underlying biological mechanisms of cancer. Thus, topological analysis of the human PPI network to explore the essentiality of cancer proteins has received more and more attention in the post-genomic era. Nevertheless, there is still a debate on whether network properties of cancer proteins exhibit some topological centrality compared to the other proteins in the human PPI network (Ideker and Sharan, 2008). For example, several studies reported that cancer proteins are central to the human PPI network and have stronger connectivity than non-cancer proteins (Jonsson and Bates, 2006; Sun and Zhao, 2010; Xia et al., 2011). However, this observation has not been universally accepted. Goh et al. (2007) constructed the human disease network and the disease gene network, and found that cancer proteins are more likely to be located at the functional periphery of the human PPI network and are less connected than the non-cancer proteins.

Two major reasons maybe account for the incongruence of cancer protein properties in the human PPI network (Cheng et al., 2012). On the one hand, there may exist a bias favourable to the proteins relevant to cancer because proteins associated with cancer may have been more extensively studied due to their perceived medical importance and investigators' special interest (Platzer et al., 2007), which leads to relatively more proteins and interactions associated with cancer in the human PPI network. On the other hand, PPI data often contain many false positives (noise) and false negatives (incompleteness) (Hart et al., 2006; Stumpf et al., 2008), and the presence of noise and incompleteness in PPI data may lead to inaccurate or even incorrect computation results.

In this paper, to help resolve the debate above, we revisited the centrality problem of cancer proteins by carrying out a study on the network characteristics of cancer proteins with the PPI data available so far. We also proved the existence of protein bias and checked its effect on network properties of proteins in the human PPI network, and validated the effect of PPI data quality on computation results by comparing topological properties of proteins in PPI networks of different quality. That is, we constructed four PPI networks with two PPI datasets: BioGRID (Stark et al., 2011) and STRING (Szklarczyk et al., 2011). For the BioGRID dataset, we constructed two PPI networks (one for human and another for yeast) using only physical interactions. For the STRING dataset, we constructed two PPI networks using both physical and predicted interactions.

Concretely, to evaluate and compare the network characteristics of cancer proteins and other proteins in human and yeast PPI networks of different quality, we first identified four sets of human proteins: cancer proteins, non-cancer disease proteins, non-cancer essential proteins and control proteins, and then mapped these proteins into the yeast PPI network by homologous genes. Finally, we compared these proteins' properties in both the human PPI network and the yeast PPI network, in terms of three common network topological

centrality measures: *degree centrality*, *clustering coefficient* and *betweenness centrality*. Experimental results over the two real datasets showed that

- Cancer proteins tend to have a higher degree and a smaller clustering coefficient than non-cancer proteins. This result indicates that cancer proteins are central to the human PPI network and have strong connectivity, but their neighbours have less likelihood to connect each other.
- Cancer proteins tend to have a larger betweenness centrality compared to other proteins on the STRING dataset.
- The average betweenness centrality of cancer proteins is larger than that of disease and control proteins, but smaller than that of essential proteins on the BioGRID dataset.
- Bias of protein and interaction does exist in human PPI data, but it does not substantially impact the topological centrality of cancer proteins.
- The quality of PPI data has considerable impact on the topological centrality of cancer proteins in PPI networks.

2 Methods

2.1 Constructing protein-protein interaction networks

We examined the network topology properties of four sets of proteins with two PPI datasets. Protein interactions of the first PPI dataset were downloaded from the biological general repository for interaction datasets (BioGRID) (Stark et al., 2011). BioGRID is a public database that archives and disseminates genetic and protein interaction data from model organisms and humans, it currently holds 347,966 interactions (170,162 genetic, 177,804 physical) obtained from both high-throughput datasets and individual focused studies, which were derived from over 23,000 publications in the literature.

With the BioGRID dataset, we constructed two PPI networks (one for human and another for yeast) using only physical interactions. In the two networks, a node corresponds to a protein and a un-weighted edge corresponds to an interaction between two proteins. The details of the two PPI networks are shown in Table 1.

Protein interactions of the second PPI dataset were downloaded from the STRING database (Szklarczyk et al., 2011), which stores not only experimentally verified interactions but also predicted interactions that have been annotated by computational methods. These interactions were mainly derived from four data sources: genomic context, high-throughput experiments, conserved co-expression and previous knowledge. The most recent version of STRING covers about 5.2 million proteins from 1133 organisms.

For the STRING dataset, we also constructed two PPI networks (one for human and another for yeast) where a node corresponds to a protein and a weighted edge corresponds to an interaction between two proteins, and the weight of each edge is based on the interaction confidence. The details for these two PPI networks are also shown in Table 1. Compared to the BioGRID dataset, the STRING dataset contains both physical and predicted interactions, it is quite larger and more complete, but possibly noisier.

Table 1 The numbers of proteins and interactions in the four PPI networks

Organism (*dataset*)	*Number of proteins*	*Number of interactions*
Human (BioGRID)	12463	110734
Yeast (BioGRID)	5421	291827
Human (STRING)	14556	1605917
Yeast (STRING)	5638	432075

2.2 Classification of human protein-coding genes

For comparison, human genes were classified into four sets: cancer genes, non-cancer disease genes, non-cancer essential genes and control genes.

First, we obtained 488 human cancer genes from the Cancer Gene Census database. The CGC database is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer. The original census and analysis was published in Nature Reviews Cancer (Futreal et al., 2004) and supplemental analysis information related to the paper is also available. Among the 488 cancer genes, there are 398 genes in the BioGRID dataset and 405 genes in the STRING dataset whose proteins can be found in the human PPI network. We considered these genes as cancer genes.

Second, we obtained human disease genes from the Online Mendelian Inheritance in Man (Hamosh et al., 2005). OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. We mapped these disease genes to the human PPI network with official gene symbols in NCBI, then we excluded the cancer genes retrieved from the CGC database. Finally, we got 2119 genes in the BioGRID dataset and 2724 genes in the STRING dataset respectively, we considered them as non-cancer disease genes.

Third, we obtained human essential genes based on mouse lethality phenotype. As in Goh et al. (2007), we considered embryonic lethality, prenatal lethality and postnatal lethality as lethal phenotypes and the rest of phenotypes as non-lethal ones. We downloaded mouse phenotype data and human-mouse orthologs from the Mouse Genome Informatics (Bult et al., 2008). After excluding the cancer genes, there are 1990 genes in the BioGRID dataset and 2198 genes in the STRING dataset whose proteins can be found in the human PPI network. We considered them as non-cancer essential genes.

Finally, we excluded the cancer genes, the disease genes and the essential genes from all protein-coding genes mapped in human PPI network, this results in 7479 control genes in the BioGRID dataset and 8330 control genes in the STRING dataset, respectively. Additionally, the homologous of human cancer genes, disease genes, essential genes and control genes in yeast organism were downloaded from NCBI HomoloGene database. Table 2 summarises the four gene sets and their corresponding proteins in the human and yeast PPI networks.

2.3 Network topological centrality measures

In this study, three common network topological centrality measures, i.e., *degree centrality*, *clustering coefficient* and *betweenness centrality*, were used to examine network topology properties of the four sets of proteins in PPI networks.

Table 2 The four sets of genes (proteins)

Organism (*dataset*)	*Number of genes* (*proteins*)			
	Cancer	*Disease*	*Essential*	*Control*
Human (BioGRID)	398(436)	2119(2371)	1990(2308)	7479(8156)
Yeast (BioGRID)	39(39)	287(297)	242(249)	691(710)
Human (STRING)	405(447)	2724(3215)	2198(2559)	8330(9134)
Yeast (STRING)	40(40)	293(304)	247(254)	708(729)

Degree centrality. Graph degree centrality is perhaps the simplest measure of centrality, it is defined as the number of links incident upon a vertex (i.e., the degree of a vertex). So the graph degree centrality of a vertex v_i is defined as follows:

$$D(v_i) = \deg(v_i). \tag{1}$$

Clustering coefficient (Watts and Strogatz, 1998). Graph clustering coefficient is a measure of degree to which vertices in a graph tend to cluster together. Given a graph $G = \{V, E\}$ where V and E are the sets of vertices and edges in the graph. The local clustering coefficient C_i for a vertex v_i is given by the proportion of links between the vertices within its neighbourhood divided by the number of links that could possibly exist between them. For a undirected graph, if a vertex v_i has k_i neighbours, $\frac{k_i(k_i-1)}{2}$ edges could exist among the vertices within the neighbourhood. Thus, the local clustering coefficient of a vertex v_i is given as:

$$C(v_i) = \frac{2\left|\left\{\mathrm{e}_{jk} : v_j, v_k \in N_i, \mathrm{e}_{jk} \in E\right\}\right|}{k_i(k_i - 1)} \tag{2}$$

where $N_i = \{v_j | v_j \in V, \mathrm{e}_{ij} \in E, \mathrm{e}_{ji} \in E\}$ is the immediately connected neighbours of the vertex v_i. Thus, the average of the local clustering coefficients of all the vertices n can be defined as:

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C(v_i). \tag{3}$$

Betweenness centrality (Freeman, 1977). Graph betweenness centrality is perhaps one of the most prominent measures of centrality, it quantifies the number of times a vertex acts as a bridge along the shortest path between two other vertices. That is, vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness. Graph betweenness centrality of a vertex v_i is evaluated as follows:

$$B(v_i) = \sum_{s \neq v_i \neq t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}. \tag{4}$$

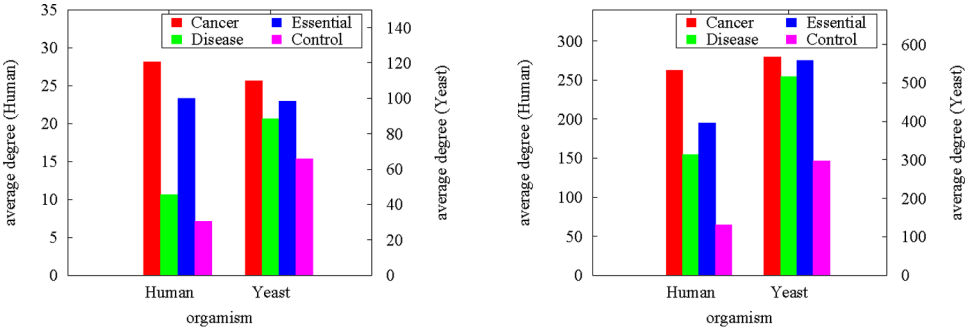
Above, σ_{st} is the total number of shortest paths from vertex s to vertex t and $\sigma_{st}(v_i)$ is the number of those paths that pass through v_i.

3 Results and Discussion

3.1 Cancer proteins tend to have a higher degree centrality

First, we examined degree centrality of the four sets of proteins in human and yeast PPI networks. Degree centrality is perhaps the simplest measure of centrality, Figure 1(a) and (b) present the average degrees of the cancer, disease, essential and control proteins over the BioGRID and the STRING dataset respectively. It can be seen from Figure 1 that the average degree of cancer proteins is higher than that of disease, essential and control proteins in the human PPI network, the results clearly indicate that proteins encoded by cancer genes tend to interact strongly with other proteins in the human PPI network. We then examined this property in the yeast organism whose PPI data should not be biased towards cancer or other disease. As shown in Figure 1, cancer proteins tend to have much stronger interactions than other kinds of proteins in the yeast PPI networks. The observation reveals that data bias is not the primary factor for the strong interaction of cancer proteins. In addition, it is worth noting that the gap between the degree centrality values of cancer proteins and other proteins is narrowed in the yeast PPI network, implying that the bias in protein interaction data caused by prevalent cancer research in human indeed exists. For example, on the BioGRID dataset, the average degree of cancer proteins is 28.13, which is approximately 4 times that of the control proteins (7.15) in the human PPI network. However, in yeast PPI network, the ratio is only 1.66.

Figure 1 The average degree centrality values of the cancer, disease, essential and control proteins: (a) BioGRID and (b) STRING (see online version for colours)

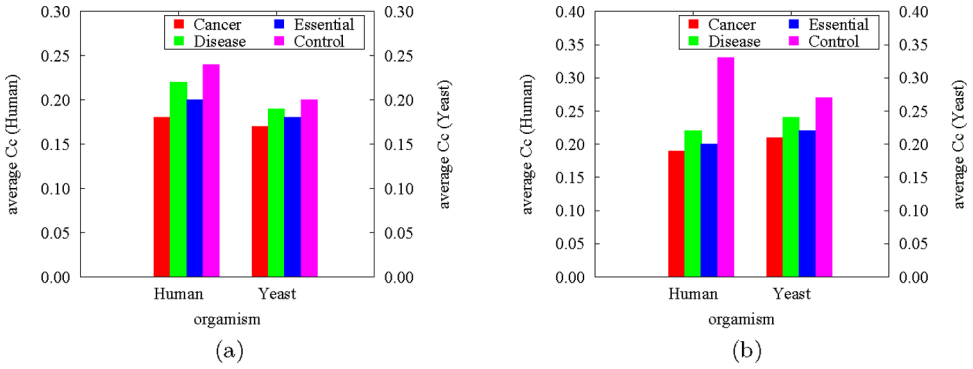


3.2 Cancer proteins tend to have a smaller clustering coefficient

Next, we examined the clustering centrality of the four sets of proteins in human and yeast PPI networks. Clustering coefficient of a node in the PPI network measures how well

connected among its direct interactors, and a larger clustering coefficient of a node indicates a higher density of its network connections. Figure 2(a) and (b) present the average clustering coefficients of the cancer, disease, essential and control proteins over the BioGRID and the STRING dataset respectively. We can see that the cancer proteins tend to have a smaller clustering coefficient compared to the disease, essential and control proteins in both human and yeast PPI networks, and the average clustering coefficient of cancer proteins is almost equal to that of the essential proteins in the four PPI networks. This means the neighbours of the cancer proteins have less likelihood to connect each other in both human and yeast PPI networks.

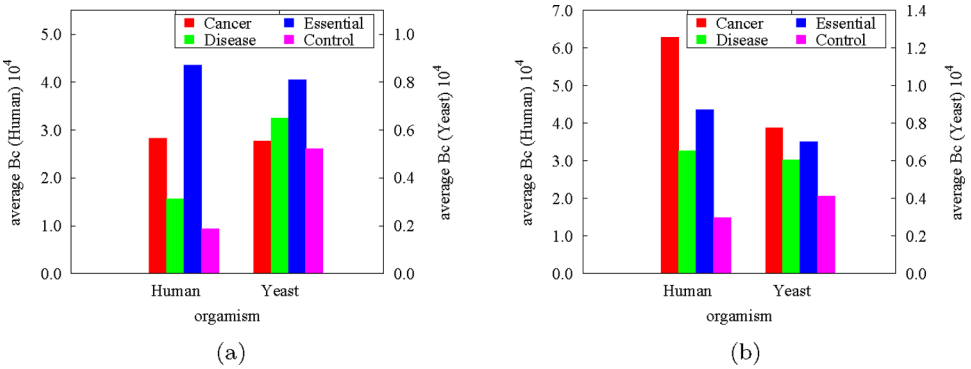
Figure 2 The average clustering coefficients of the cancer, disease, essential and control proteins: (a) BioGRID and (b) STRING (see online version for colours)



3.3 Cancer and essential proteins tend to have a larger betweenness centrality

Finally, we examined the betweenness centrality of the four sets of proteins in human and yeast PPI networks. Betweenness centrality measures how many shortest paths go through a particular node in a specific network. It can be used to reflect how much information might pass through the node in a network. Figure 3(a) and (b) present the average betweenness centrality of the cancer, disease, essential and control proteins over the BioGRID and the STRING dataset respectively. We found that the cancer and essential proteins tend to have a larger betweenness centrality compared to the disease and control proteins in both human and yeast PPI networks. This result suggests that the cancer and essential proteins tend to occupy the network positions of global importance in communication between protein pairs in the PPI network. However, very interestingly we also noticed that on the BioGRID dataset, the average betweenness centrality of cancer proteins is larger than that of disease and control proteins, but smaller than that of essential proteins. This is because the STRING database is an integrated protein interaction database containing both known (experimentally validated) and predicted protein interactions. However, for BioGRID dataset, we constructed two PPI networks using only physical interactions. This result indicates that the quality of PPI data has considerable impact on network properties of the four sets of proteins.

Figure 3 The average betweenness centrality values of the cancer, disease, essential and control proteins: (a) BioGRID and (b) STRING (see online version for colours)



4 Conclusion

This study revisited the centrality of cancer proteins in the human PPI network by conducting a comprehensive investigation on the network characteristics of cancer proteins with the PPI data available so far. For this purpose, we first identified four sets of human proteins, and then mapped these proteins into the yeast PPI network by homologous genes, finally we evaluated and compared their network properties in both human and yeast PPI networks by using three common network topological centrality measures. Experiments over two real datasets demonstrated that the cancer proteins have significantly different topological properties compared to the disease, essential and control proteins. Specifically, cancer proteins tend to have a higher degree and a smaller clustering coefficient than the other proteins. This result indicates that cancer proteins are central to the human PPI network and have strong connectivity, but the neighbours of the cancer proteins have less likelihood to connect each other. Furthermore, it is worth noting that the cancer proteins tend to have a larger betweenness centrality on the STRING dataset that contains both known and predicted interactions. However, the average betweenness centrality of cancer proteins is smaller than that of essential proteins on the BioGRID dataset that contains only physical interactions. This means the quality of PPI data has considerable impact on the network properties of proteins. In summary, our study suggests that topological characteristics of cancer proteins in PPI network is important for our understanding the underlying biological mechanisms of cancer. It can also help biologists accurately predict disease-causing genes and effectively discover drug targets.

Acknowledgements

This study was supported partially by China 863 Program (grant No. 2012AA020403) and NSFC (grant No. 61272380). Hui Liu was supported by NSFC (grant No. 31100954), and Jihong Guan was supported by NSFC (grant No. 61173118).

References

Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. and Blake, J.A. (2008) 'The Mouse Genome Database (MGD): mouse biology and model systems', *Nucleic Acids Research*, Vol. 36, Suppl. 1, pp.D724–D728.

Cheng, T.M.K., Gulati, S., Agius, R. and Bates, P.A. (2012) 'Understanding cancer mechanisms through network dynamics', *Briefings in Functional Genomics*, Vol. 11, No. 6, pp.543–560.

Ferrer, L., Dale, J. and Karp, P.(2010) 'A systematic study of genome context methods: calibration, normalization and combination', *BMC Bioinformatics*, Vol. 11, No. 1, p.493.

Freeman, L.C.(1977) 'A set of measures of centrality based upon betweenness', *Sociometry*, Vol. 40, p.35041.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) 'A census of human cancer genes', *Nature Reviews Cancer*, Vol. 4, No. 3, pp.177–183.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A.L. (2007) 'The human disease network', *Proceedings of the National Academy of Sciences*, Vol. 104, No. 21, pp.8685–8690.

GÖstlund, M.L. and Sonnhammer, E.L.L. (1997) 'Network-based identification of novel cancer genes', *Molecular and Cellular Proteomics*, Vol. 9, No. 4, pp.648–655.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Research*, Vol. 33, Suppl. 1, pp.D514–D517.

Hart, G.T., Ramani, A.K. and Marcotte, E.M. (2006) 'How complete are current yeast and human protein-interaction networks', *Genome Biol.*, Vol. 7, No. 11, p.120.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., SOrensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W.V., Figeys, D. and Tyers, M. (2002) 'Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry', *Nature*, Vol. 415, No. 6868, pp.180–183.

Hormozdiari, F., Salari, R., Bafna, V. and Sahinalp, S.C. (2010) 'Protein-protein interaction network evaluation for identifying potential drug targets', *Journal of Computational Biology*, Vol. 17, No. 5, pp.669–684.

Ideker, T. and Sharan, R.(2008) 'Protein networks in disease', *Genome Research*, Vol. 18, No. 4, pp.644–652.

Jonsson, P.F. and Bates, P.A.(2006) 'Global topological features of cancer proteins in the human interactome', *Bioinformatics*, Vol. 22, pp.2291–2297.

Lukk, M., Kapushesky, M. and Nikkilä (2010) 'A global map of human gene expression', *Nature Biotechnology*, Vol. 28, No. 4, pp.322–324.

Oti, M., Snel, B., Huynen, M.A. and Brunner, H.J. (2006) 'Predicting disease genes using protein-protein interactions', *Journal of Medical Genetics*, Vol. 43, No. 8, pp.691–698.

Platzer, A., Perco, P., Lukas, A. and Mayer, B. (2007) 'Characterization of protein-interaction networks in tumors', *BMC bioinformatics*, Vol. 8, No. 1, p.224.

Pujana, M.A., Han, J.D.J., Starita, L.M., Stevens, K.N., Tewari, M., Ahn, J.S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., ElShamy, W.M., Rual, J.F., Levine, D., Rozek, L.S., Gelman, R.S., Gunsalus, K.C., Greenberg, R.A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sol, X., Herndez, P., Lzaro, C., Nathanson, K.L., Weber, B.L., Cusick, M.E., Hill, D.E., Offit, K., Livingston, D.M., Gruber, S.B., Parvin, J.D. and Vidal, M. (2007) 'Network modeling links breast cancer susceptibility and centrosome dysfunction', *Nature Genetics*, Vol. 39, No. 11, pp.1338–1349.

Reid, A.J., Ranea, J.A.G., Clegg, A.B. and Orengo, C.A. (1994) 'CODA: accurate detection of functional associations between proteins in eukaryotic genomes using domain fusion', *PloS One*, Vol. 5, No. 6, p.e10908.

Stumpf, M.P.H., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M. and Wiuf, C. (2008) 'Estimating the size of the human interactome', *Proceedings of the National Academy of Sciences*, Vol. 105, No. 19, pp.6959–6964.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., *et al.* (2011) 'The BioGRID interaction database: 2011 update', *Nucleic Acids Research*, Vol. 39, Suppl. 1, pp.D698–D704.

Sun, J. and Zhao, Z.(2010) 'A comparative study of cancer proteins in the human protein-protein interaction network', *BMC Genomics*, Vol. 11, Suppl. 3, p.S5.

Suter, B., Kittanakom, S., Stagljar, I. (2008) 'Two-hybrid technologies in proteomics research', *Current Opinion in Biotechnology*, Vol. 19, pp.316–323.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J. and Mering, C. (2011) 'The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored', *Nucleic Acids Research*, Vol. 39, Suppl. 1, pp.D561–D568.

Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, G., Faria, D., Bull, S., Pawson, T., Morris, Q. and Wrana, J. (2009) 'Dynamic modularity in protein interaction networks predicts breast cancer outcome', *Nature Biotechnology*, Vol. 27, No. 2, pp.199–204.

Tikk, D., Thomas, P., Palaga, p., Hakenberg, J. and Leser, U. (2010) 'A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature', *PLoS Computational Biology*, Vol. 6, No. 7, p.e1000837.

Watts, D. and Strogatz, S.(1998) 'Collective dynamics of small-world networks', *Nature*, Vol. 393, No. 6684, pp.440–442.

Xia, J., Sun, J., Jia, P. and Zhao, Z. (2011) 'Do cancer proteins really interact strongly in the human protein-protein interaction network?', *Computational Biology and Chemistry*, Vol. 35, No. 3, pp.121–125.