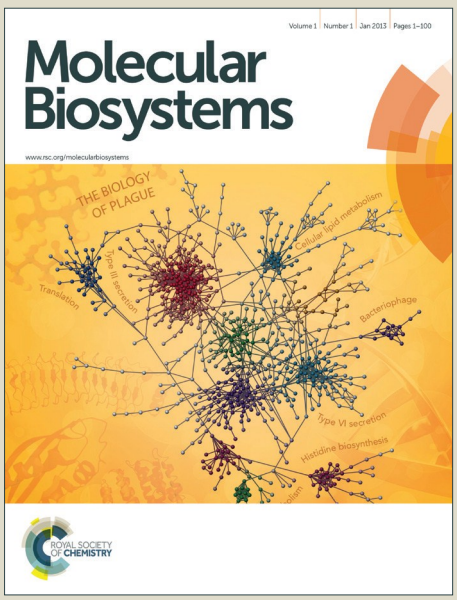
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems

RSCPublishing

ARTICLE

A novel topological centrality measure capturing biologically important proteins

Cite this: DOI: 10.1039/x0xx00000x

Muhammed Erkan Karabekmez[a] and Betul Kirdar[a]

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Topological centrality in protein interaction networks and its biological implications has widely been investigated in the past. In the present study, a novel centrality metric – weighted sum of loads eigenvector centrality (WSL-EC) - based on graph spectra was defined and its performance in identifying topologically and biologically important nodes was comparatively investigated with common centrality metrics in human protein-protein interaction network. The metric can capture nodes from peripherals of the network differently from conventional eigenvector centrality. Different metrics were found to selectively identify hub sets significantly associated with different biological processes. Widely accepted metrics; degree centrality, betweenness centrality, subgraph centrality and eigenvector centrality are subject to a bias towards super-hubs while WSL-EC was not affected by the presence of super-hubs. WSL-EC outperforms other centrality metrics in detecting biologically central nodes such as pathogen interacting, cancer, ageing, HIV-1 or disease related proteins and proteins involved in immune system process and autoimmune diseases in human interactome.

Introduction

The understanding of life at molecular level impressively increased in the last half century due to the technological advances such as microarrays, mass spectroscopy and next-generation sequencing. In addition to these technologies which helped to the identification and quantification of biological molecules at the whole genome level, high through-put technologies were also developed to measure physical protein-protein, protein-DNA or RNA, and enzyme- metabolite interactions All these interactions can be projected as networks or graphs which can provide a good scaffold to model molecular interactions, integrate several omics data and interpret overall physical and functional landscape of cellular function. These developments have provided the scientific basis of a new field known as network biology which combines systems biology, graph theory, statistical and computational analysis[1].

Protein-protein interactions (ppi) networks at the whole genome level (interactome) are considered as an important source to be explored to get further information about the cellular function. These networks are hierarchically organized and consist of tightly clustered groups of proteins working together as part of a biological process or a complex to achieve a specific function in the cell. Protein-protein interactions were identified using yeast two-hybrid (Y2H), tandem affinity purification coupled to mass spectrometry (TAP/MS) and affinity capture mass spectrometry (AC/MS) in several model organisms including humans and are deposited in publically available databases such as Database of Interacting Proteins (DIP)[2], Human Protein Reference Database (HPRD)[3], Biological General Repository for Interaction Datasets (BioGRID)[4] and STRING database[5].

Graph-theoretical analysis of these networks has revealed a strong correlation between topological characteristics of cellular networks and the cellular function. Early studies indicated the scale-free topology of protein-protein interaction networks consisting of small number of hubs with many interactions. Today, molecular interaction networks are considered not scale free, but they are generally heavy tailed consisting of few hubs and many low degree nodes[6,7]. Furthermore the topological analysis of protein-protein interaction networks provided a deeper understanding about the biological systems leading to functional annotation of unknown genes or identification of drug targets or disease related proteins and pathways[8,9]. It has been reported that biologically important proteins in aging, cardio-vascular disorders, metabolic disorders, cancer and infectious diseases have some topological centrality in human interactome[10-14].

A number of different topological centrality metrics were described to define the centrality of the nodes such as degree centrality (DC) which is the number of edges, betweenness centrality (BC) which is the fraction of shortest paths passing through a node and eigenvector centrality (EC). Although DC is the most commonly used centrality measure, it can only give information about local topology of a node. BC was used to determine the bottleneck-nodes with low degree, but detrimental for organism when removed[15]. EC as a global centrality metric can explain latent topology by not only local connectivity but also connectivity of the neighbouring nodes as well[8,16]. Although EC is not a local centrality metric like DC it is limited to the first principal of the graph spectra and therefore it is not a descriptive metric about the peripheral modules in a network[17]. Other centrality measures such as sub-graphs centrality[18] (SC) which accounts for all graph spectra instead of only the first principal was also proposed to represent number of short walks starting and ending at the node of interest. The drawback of the SC is that it converges to EC when the largest eigenvalue breaks away from the second[19].

Other centrality measures such as coreness centrality[20], bipartivity[21], Graphlet Degree Centrality[9], node hierarchy[22] and linear combination of different metrics[6] were also proposed to improve the predictability of the cellular function or biologically central nodes in health or disease. Different centrality measures were extensively used and compared for the topological analysis of biological networks[23,24]. It has been observed that different centrality metrics can be important in different instances. Therefore the development and application of

different metrics is considered to be important in the topological analysis and modelling of networks in systems biology in order to improve the predictability of the cellular function or biologically central nodes in health or disease.

The aim of the present study is to develop a new centrality measure, weighted sum of loads eigenvector centrality (WSL-EC) counting all eigenvectors with a different and more simple weighting strategy in order to capture topologically important nodes not only from the densely populated but from the less densely populated and peripheral parts of the human network. The performance of WSL-EC in the identification of topologically important nodes that contribute to the integrity of network and to capture biologically central nodes were tested in a human global protein-protein interaction network. The performance of this newly introduced centrality measure was compared with the ability of degree centrality (DC), betweenness centrality (BC), eigenvector centrality (EC) and subgraph centrality (SC).

Results and Discussion

We have defined a novel centrality metric called Weighted Sum of Loads Eigenvector Centrality (WSL-EC). WSL-EC is designed as weighted sum of loads of each node to each eigenvector. Absolute values of the loads were used as signs indicates only direction not significance of the load. Instead of exponentials of eigenvalues, e^{λ_i}, eigenvalues themselves, λ_i, were used as weight to prevent dominance of the first principal. We have analyzed then its efficiency or power to identify central proteins and compared with other commonly and extensively used metrics such as Degree Centrality (DC), Betweenness Centrality (BC), Subgraph Centrality (SC) and Eigenvector Centrality (EC). Physical protein-protein interaction network of human (H) consisting of 15,192 nodes and 126,572 binary interactions without self-loops were constructed as described in Materials and Methods. Topological characteristics of the network, the contribution of the nodes scored by this novel and commonly used centrality metrics to network integrity and the ability of the hubs to capture biological function were analyzed and compared.

Topological Analysis of the Network

The power law distribution, hierarchical modularity and degree correlations on connected nodes are commonly used parameters to characterize the biological network organization[25]. Topological characteristics of the biological protein-protein interaction network (H) were calculated to reveal the organization of the network (Table 1). The network was found to have a degree distribution that fit to power law (R^2=0.844) (Fig. S.1).

Table 1. Topological characteristics of the H network

Number of Nodes	15192
Number of Edges	126,572
Density (ρ)	0.001
Diameter (D)	8
Average Clustering Coefficient (CC)	0.256
Characteristic Path Length (*cpl*)	2.670
Network Centralization	0.630
Network Heterogeneity	5.488

Topological characteristics of the network indicated that although nodes are loosely connected (ρ=0.001) and diameter is higher than that of random networks with the same size and density (*p-value*<0.01), average shortest path lengths (*cpl*) is lower and both network centralization and network heterogeneity are higher than characteristics of a randomly wired network with the same number of node and edges (all *p-values* are lower than 0.01). Together with findings that revealed by shortest path length (*spl*) distribution, more than 99% of the *spl*'s are 2, 3 or 4 steps long (Fig. S.1), these numbers point out presence of super-hubs, nodes with off-the-chart high connectivity. In addition to that, observed non-randomly high clustering coefficient (*p-value*<0.01) is a sign of modular organization of the network.

The variation of clustering coefficient is considered to be one of the most commonly used parameter to identify the hierarchical modular structure of biological networks and it has been shown that existence of super-hubs and degree correlation also affect clustering coefficient variation with degree[26]. In order to unveil the hierarchical architectural structure of the network used in the present study, assortativity metrics and variation of the clustering coefficients with degree were investigated (Fig.1). The probability of having a link between two nodes with degrees *k(i)* and *k(j)* was compared with the probability within 100 randomly rewired networks by a z-score. Z-scores greater than zero indicate positive correlation whereas negative z-scores indicate negative correlation between connectivities with respect to random networks. Assortativity coefficient[27] of H network was found to be significantly negative (*r=-0.0717* with a *p-value<0.01*) implying dissortativity of the network. The variation of average clustering coefficient with respect to degree and heat-map of z-scores in H network suggests that nodes with low degree are more likely to be linked with highly connected nodes and nodes with high connections are less likely to be connected to each other, which also reflects scale-free nature of the network. There is a strong repulsion between super-hubs whereas moderately linked nodes with a degree of 100 to 200 have an apparent affinity to each other pointing out the densely connected modules (Fig. 1).

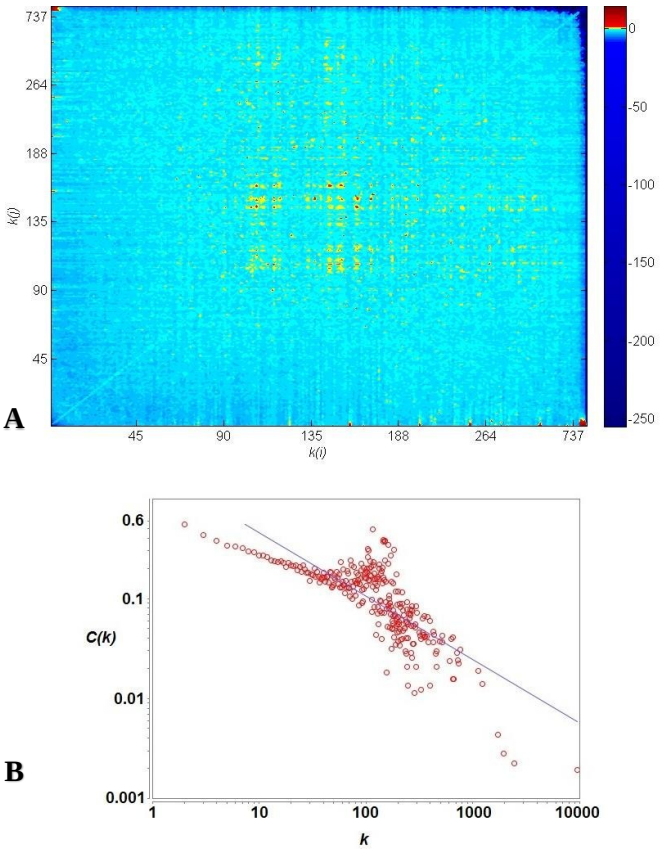


Fig. 1. Correlation profiles of protein-protein interactions (A) and clustering coefficient variations with degree (B) in H network. Color-bars display z-scores for probability of a node with a degree of *k(i)* to be linked with a node with a degree of *k(j)* with respect to 100 randomly rewired networks. Corresponding average clustering coefficient variations were fitted to power law line on a logarithmic scale.

Clustering coefficient variation in H network obeys power-law distribution in parallel with the disassortative nature of the network (Fig. 1) and existence of super-hubs.

Correlation between degree and average nearest neighbor degree is another metric which can identify assortativity of a network[26]. When we plot average nearest neighbor degree as a function of degree the distribution almost perfectly fits to power law with a negative exponent verifying negative assortativity of the Human interactome (Fig. 2).

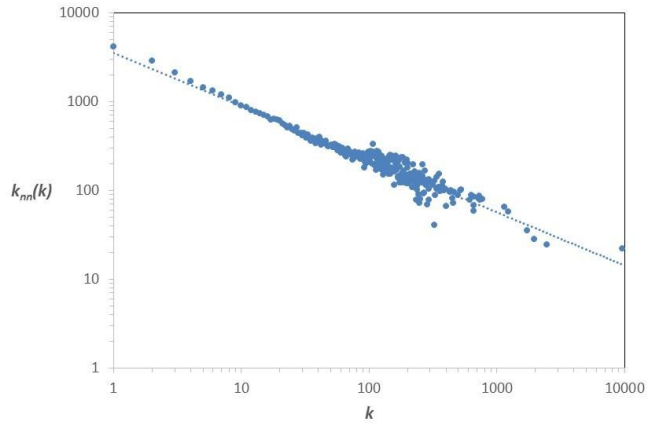


Fig. 2. Average nearest neighbor degree as a function of degree. Distribution fits to power law with R^2=0.934 (slope of the line is -0.597).

Network Integrity

Robustness is an important property of the protein-protein interaction networks which is assumed to emerge from natural selection and refers to the ability of networks to maintain function under perturbation[28]. Efficiency and diameter are the two important topological characteristics to quantify the robustness of the networks[29].

We have scored all the nodes using the newly described topological centrality measure, WSL-EC, in H network and nodes were then removed one by one either by starting with the highest scoring nodes or randomly. Diameter and efficiency recalculated after the removal of each node in each of these three biological networks. The results were compared to the removal of the nodes scored by DC, BC, EC and SC.

The drawback of the SC revealed itself at the beginning that SC based ranking was found to be identical to the ranking found based on EC. The difference between the largest and the second largest eigenvalue is more than 53 which means the weight of the first eigenvector is more than e^{53} times weight of the second eigenvector. Hence the results for SC will not be presented hereafter as it is identical to EC.

Removal of the highest scoring nodes one after the other was found to have more impact to decrease the efficiency when compared to the random removal of the nodes in all networks. Diameter of the networks were observed to become larger followed by a drastic decrease indicating collapse of the network by the targeted removal of the nodes. Whereas diameter of the networks remain unchanged by random removal until the collapse (Fig. S.2).

Percent of nodes that has to be removed in the network to decrease the efficiency or diameter to the half of its original value was calculated and presented in Table 2. A 50% decrease in efficiency occurred after the removal of 12.2 percent of the highest nodes by WSL-EC in H network. This result indicated that the targeted removal of the WSL-EC nodes causes a less severe collapse in this network than the DC, BC or EC targeted attacks (Table 2).

Diameter based criteria on the other hand indicated that WSL-EC targeted attack causes the fastest disintegration in H network, just the opposite of the efficiency based criteria (Table 2).

Drastic difference between these two criteria might possibly be due to the presence of super-hubs in the networks and efficiency might strongly be affected by super-hubs. The most dominant super-hub is *UBC* in H network which is connected to more than 63% of all nodes in the network and it has by far the highest scores with respect to DC, BC, and EC. However WSL-EC rank of *UBC* is only 1861 corresponding to top 12.2% which is exactly the percent of the nodes need to be removed in order to collapse H network in WSL-EC targeted attack (Table 2). Other than *UBC* all top five DC genes and four out of top five BC or EC nodes are not even in the top 10% of the WSL-EC based rank.

WSL-EC seems to be outperforming the other centrality metrics in identifying nodes affecting network diameter without being affected by the presence of super-hubs in H network.

Presence of super-hubs also explains the dominance of the first principal and convergence of SC to EC consequently.

Table 2. Percent of nodes removed to attain defined collapse

Collapse criteria	Random Removal (Avg.)	WSL-EC targeted removal	DC targeted removal	BC targeted removal	EC targeted removal
50% drop in Efficiency	17.1	12.2	0.1	0.1	0.1
50% drop in Diameter	97.9	43.6	45.4	46.2	83.6

Hubs

Four sets consisting of top 10 percent highest scoring nodes with DC, BC, EC, and WSL-EC were identified in each network as hub sets. The networks were visualized by Cytoscape in order to investigate and compare the topological distribution of the hubs. Localization of hub sets identified using different metrics in H network was not clearly distinguishable due to their crowded nature of the network (Fig. S.3). In order to visualize the localization of the hub sets a smaller context specific yeast interaction network Y2 was used. The visualization indicated that WSL-EC seizes the peripheral modules together with the core of the network while EC highlights only the core of the network. Hubs that were identified by DC were also found to be localized at densely connected parts of the network. Hubs identified by BC and WSL-EC were distributed in all over the network (Fig. 3).

Since the visualization of large networks may not give satisfactory results, visually detected differences between hubs dispersions were quantified by the topological analysis of hub-networks constructed using known interactions between the hubs of H network. Four hub-networks constructed using the hubs identified by different centrality metrics were named as degree central hub network (DC-HN), betweenness central hub network (BC-HN), eigenvector central hub network (EC-HN) and WSL eigenvector central hub network (WSL-EC-HN) for H network. The network density of these networks were calculated and compared to quantify dispersion of the hubs (Table 3).

Table 3. Network properties of Hub-Networks in H network

Hub Network Type	Number of Edges	Network Density
DC-HN	43655	0.038
EC-HN	43376	0.038
BC-HN	26106	0.023
WSL-EC-HN	39625	0.034

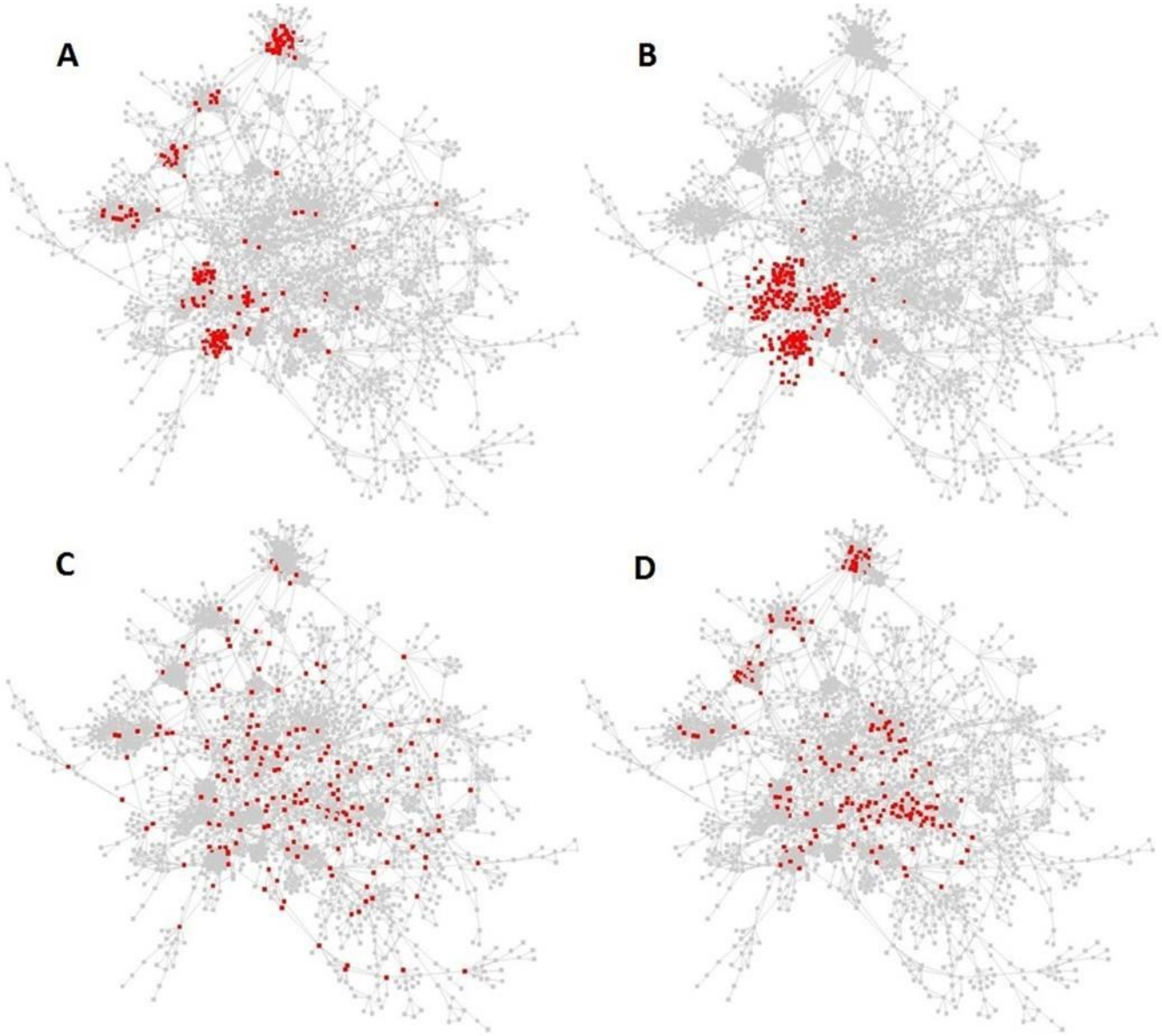


Fig. 3. Distribution of the hubs (red nodes) defined by (A) DC, (B) EC, (C) BC, (D) WSL-EC in Y2 network.

The observation that WSL-EC-HN has lower network density than EC-HN and DC-HN indicated that the WSL-EC hubs are more dispersed than EC and DC hubs confirming the visual observation of the hubs in Y2 network. BC-HN has a lower network density than WSL-EC-HN, again verifying the visual observation that BC hubs in Y2 are more dispersed.

Biological investigation of hubs

Biological role of a node is reported to be strongly related to its topological location within an interaction network; *i.e.* functions of interacting neighbors or processes they involve etc.[30]. We have also investigated GO biological process terms significantly associated with the four hub sets consisting of top ten percent highest scoring nodes with the highest centralities with respect to degree (DC), betweenness (BC), eigenvector centrality (EC) and WSL eigenvector centrality (WSL-EC).

In H network almost half of the any hub set is common to all other sets (Fig. 4). These 747 human genes have significant ontological associations to a wide variety of GO biological process terms related to communication, response, metabolism, catabolism, localization, development, transcription, cell cycle etc. and regulation of all these processes (*p-value* $\leq$ 9.88E-4).

In H network, 53 hubs were uniquely identified by WSL-EC and found to be significantly enriched (*p-value* $\leq$ 8.03E-4) with establishment of organelle localization and regulation of biosynthetic. 285 hubs specifically described by EC were enriched with biological process terms (*p-value* $\leq$ 1.94E-6) such as gene expression, cellular component organization or biogenesis, organelle organization and mRNA transport. 26 hubs which were selectively identified by DC alone were found to be significantly enriched (*p-value* $\leq$ 5.49E-5) with regulation of ligase activity, regulation of protein ubiquitination, protein catabolic process and cell cycle. When BC was used as a centrality measure 570 hubs were

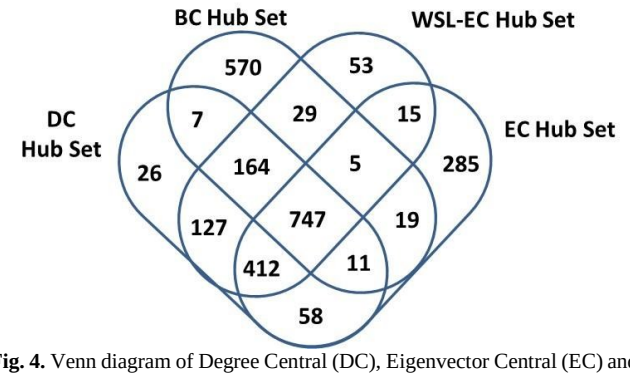


Fig. 4. Venn diagram of Degree Central (DC), Eigenvector Central (EC) and WSL-Eigenvector Central (WSL-EC) Hub Sets and Top 10 Hubs in H Network.

specifically described and these proteins were found to be significantly associated (*p-value* ≤ 6.07E-9) with response to stimulus, biological regulation, multicellular organismal process, signaling, immune system process, developmental process and establishment of localization (Fig. 4).

The top 10 highest scoring hub sets identified by different centrality metrics were compared in order to reveal individual genes favored by specific centrality metrics. *RIOK2*, a kinase related to ribosome biogenesis process is the top central node detected by WSL-EC. Following 4 nodes *VCAM1*, *ITGA4*, *HSPA8* and *HSPA5* are related to stress response and/or immune system (Table S.1). Top 10 hub sets identified by DC, BC and EC were found to be highly overlapping whereas the top 10 hub set of WSL-EC is completely exclusive. Super hubs *UBC* and *NRF1* present at top two for all DC, BC and EC rankings and *APP*, *ELAVL1*, *SUMO2* and *CUL3* are also present in the all three top 10 hub sets. All these nodes are related to ubiquitination processes and/or interact with RNA. Top 10 hub sets and pathways related to the hubs were tabulated in the supplementary tables (Table S.1. to Table S.4). Focus of the novel metric seems to be more on immune system process compare to the other centrality metrics.

Biological and Topological Centrality in Human Interactome

It has been reported that the genes involved in cancer, aging and infectious disorders are also topologically central[9]. Disease related genes, genes involved in immune system process and genes related to autoimmune diseases also have biological importance. The topologically central hub sets identified by the four centrality measures in H network were investigated by the enrichments in HIV-1 interacting proteins, ageing related genes, cancer related genes, pathogen interacting (PI) proteins, disease related genes, genes involved in immune system process and genes related to autoimmune diseases in order to shed light to biological differences of the centrality metrics.

The correlation between biological and topological centralities was investigated by a jackknifing method[31] based on jackknife resampling technique[32] without limitations of thresholds (Fig. 5).

The relative areas under the curve (R-AUC) were calculated for the curves in Figure 5. R-AUC was defined as the ratio of the area under the curve (AUC) to the area under an ideal curve ranking all biologically significant nodes on top (I-AUC) (Fig. 5). As higher R-AUC implies higher correlation with biological significance, R-AUC analyses indicated that WSL-EC based ranks outperform DC, BC, EC or SC based ranks in identifying all biologically central node sets. Statistical significance of the differences was assessed by permutation tests (*p-value*<0.001 for all curves).

GO biological process term enrichments were also used to discriminate subsets of the biologically central nodes with higher ranks with respect to each centrality metric. All subsets of biologically central node sets defined by 7 criteria (pathogen interacting, HIV-1 interacting, cancer related, ageing related, disease related nodes and genes involved in immune system process or related to autoimmune diseases) having the highest ranks in WSL-EC order were found to be enriched in stress response related terms, immune system related terms, transcriptional terms and/or kinase activity related terms. The smallest subsets for all criteria were found to be consisting of nodes favored by DC. The subsets are either not significantly associated with any GO term or enriched in transcriptional, apoptotic and cell cycle processes. The widest range of GO terms were determined for the subsets of high betweenness nodes for all four biological centrality criteria. External processes like exocytosis, endocytosis, cell motility, cell adhesion, cell migration, cell-cell communication, or processes related to differentiation like regulation of neurogenesis and embryonic morphogenesis or immune system related GO terms are some of them. The nodes with the highest order in EC rank have the narrowest range of GO enrichments as they found to be associated with only translational and carbon central metabolism related terms.

Conclusion

In the present study we proposed a novel global centrality metric, weighted sum of loads eigenvector centrality (WSL-EC) counting all eigenvectors. The performance of WSL-EC in the identification of topologically more important nodes that contribute to the integrity of network and to capture essential or biologically central nodes were tested in a biological network and compared with the performance of other four commonly used centrality metrics, DC, BC, EC and SC.

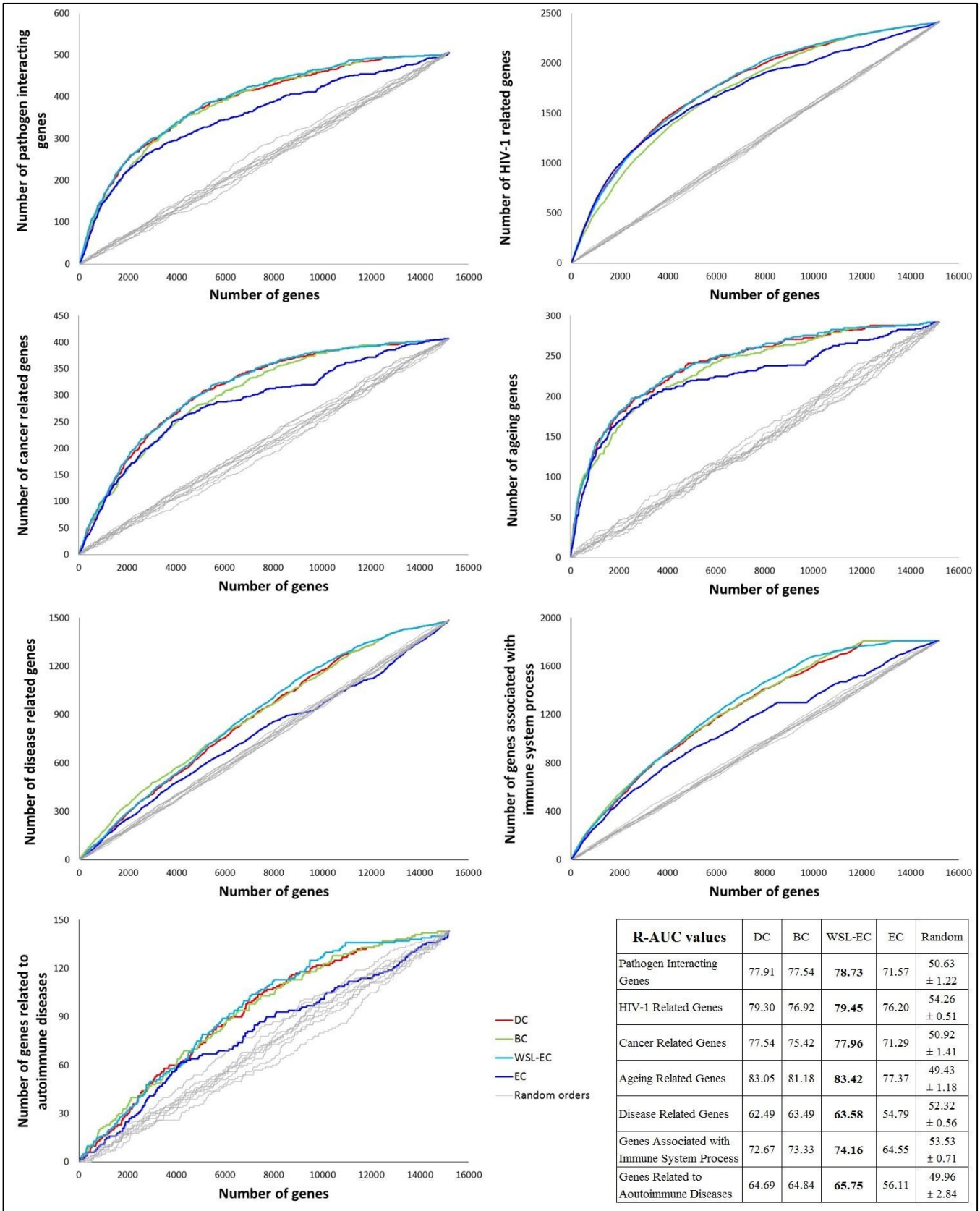
Topological analysis of the network indicated the dissortative and modular architecture for human global protein-protein interaction network (H). WSL-EC outperformed DC, BC, EC and SC in identifying nodes affecting network robustness in human interactome. The topological distributions of hubs in the networks were found to be different for hub-sets identified with different centrality metrics. Hubs identified by BC and WSL-EC were distributed in all over the network whereas EC and DC identified hubs were localized at densely connected parts of the networks.

We have noted that different centrality measures could specifically capture a set of hubs involved in different biological processes.

WSL-EC was found to be outperforming in capturing biologically central nodes like pathogen interacting, HIV-1, cancer, ageing, disease related genes and genes involved in immune system process and related to autoimmune diseases in human interactome compared to DC, BC, EC or SC. The choice of centrality metric is crucial as different metrics focus on different topologies and these topological differences correspond to different biological roles.

Hubs with out-of-the-chart connectivity (super-hubs) create a strong bias in topological centrality for DC, BC, EC or SC while WSL-EC does not seem to be affected from the presence of super-hubs.

WSL-EC is an easy to implement metric which doesn't require a special code or complicated computations. It is promising by this aspect that it can be utilized by diverse researchers.



R-AUC values	DC	BC	WSL-EC	EC	Random
Pathogen Interacting Genes	77.91	77.54	**78.73**	71.57	50.63 ± 1.22
HIV-1 Related Genes	79.30	76.92	**79.45**	76.20	54.26 ± 0.51
Cancer Related Genes	77.54	75.42	**77.96**	71.29	50.92 ± 1.41
Ageing Related Genes	83.05	81.18	**83.42**	77.37	49.43 ± 1.18
Disease Related Genes	62.49	63.49	**63.58**	54.79	52.32 ± 0.56
Genes Associated with Immune System Process	72.67	73.33	**74.16**	64.55	53.53 ± 0.71
Genes Related to Aoutoimmune Diseases	64.69	64.84	**65.75**	56.11	49.96 ± 2.84

Fig. 5. Change in number of biologically significant genes detected by centrality ranks and random ranks in H network and R-AUC values of biologically central node detection performances of DC, BC, WSL-EC and EC

The novel centrality metric, WSL-EC, displays substantial biological relevance and further studies will be required to test the performance of this novel centrality metrics in complex biological networks to reveal the correlation between topology and biological importance. Furthermore integration of other data sets with protein-protein interaction networks should be investigated to improve its performance across different network architectures.

Methods

Human interactome, H, was downloaded from BIOGRID database (version 3.2.108). H is composed of 15,192 nodes and 126,572 physical binary interactions without self-loops[4].

The functional ppi network, Y2, which consists of 1792 proteins related to glucose processes and 6919 bidirectional edges was constructed[33] from STRING database[5] v8.3 with confidence score ≥0.999 using selective permeability algorithm[34] starting with 108 proteins which are associated with glucose metabolic process.

Centrality metrics used for comparison and network parameters were calculated as shown in the Supplementary Methods.

The novel spectral centrality measure (*WCL-EC)* was defined as the weighted sum of loads of the all principles of the graph spectra. Corresponding eigenvalues were used as weight and absolute values of all loads and weights were considered (Eqn. 1).

$$WSL - EC_i = \sum_{j=1}^{N} |\lambda_j| . |v_{ij}| \quad (1)$$

Where, N is the number of nodes in the network, λ_j is the j^{th} eigenvalue of the adjacency matrix and v_{ij} is the load of i^{th} node to the j^{th} principal of the graph spectra.

2417 HIV-1 interacting proteins[35], 292 ageing related genes from Ageing Gene Database (GenAge)[36], 407 cancer related genes collected from The Catalogue of Somatic Mutations in Cancer (COSMIC)[37], 506 pathogen interacting proteins[14], 1485 disease related genes[38], 1811 genes involved in immune system process[39] and 143 genes related to autoimmune diseases[40] were used as biologically important proteins.

Significantly associated Gene Ontology biological process terms (p value<0.001) were determined by GOrilla[41].

Pathways related to the hubs were identified by KEGG release 76.0[42].

Acknowledgements

The authors would like to thank Kazim Yalcin Arga for his valuable comments on the manuscript.

Funding: The research financially supported by Bogazici University Research Fund (BAP) through Project No: 8660 and the Research Council of Turkey (TUBITAK) through Project No: 110M692.

References

1 A.-L. Barabasi and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101-113.

2 I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg, *Nucleic Acids Res.*, 2000, **28**, 289-291.

3 S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. B. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, *et. al.*, *Genome Res.*, 2003, **13**, 2363-2371.

4 C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, *Nucleic Acids Res.*, 2006, **34**, D535-D539.

5 L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork and C. von Mering, *Nucleic Acids Res.*, 2008, **37**, D412-6.

6 S. Roy, *Syst. Synth. Biol.*, 2012, **6**(1-2), 31-34.

7 A. L. G. de Lomana, Q. K. Beg and G. de Fabritiis, *J.Comp. Biol.*, 2010, **17**(7), 869-878.

8 J. Wang, G. Chen, M. Li and Y. Pan, *BMC Syst. Biol.*, 2011, **5**(Suppl 3):S10.

9 T. Milenkovic, V. Memisevic, A. Bonato and N. Przulj, *PLOS one*, 2011, **6**:8, e23016.

10 T. Ideker and R. Sharan, *Genome Res.*, 2008, **18**, 644-652.

11 P. F. Jonsson and P. A. Bates, *Bioinformatics*, 2006, **22**(18), 2291-2297.

12 M. A. Yıldırım, K. Goh, M. E. Cusick, A.-L. Barabasi and M. Vidal, *Nat. Biotechnol.*, 2007, **25**(10), 1119-1126.

13 L. Ferrarini, L. Bertelli, J. Feala, A. D. McCulloch and G. Paternostro, *Bioinformatics*, 2004, **21**(3), 338-348.

14 M. D. Dyer, T. M. Murali and B. W. Sobral, *PLOS Pathog.*, 2008, **4**(2): e32.

15 H. Yu, P. M. Kim, E. Sprecher, V. Trifonov and M. Gerstein, *PLOS Comput. Biol.*, 2007, **3**(4), e59.

16 J. E. Mcdermott, D. L. Diamond, C. Corley, A. L. Rasmussen, M. G. Katze and K. M. Waters, *BMC Syst. Biol.*, 2012, **6**:28.

17 J. Aguirre, D. Papo and J. M. Buldú, 2013, *Nat. Phys.*, 9, 230–234.

18 E. Estrada and J. A. Rodríguez-Velázquez, *Phys. Rev. E*, 2005, **71**, 056103.

19 M. Benzi and Klymko C., 2015, *SIAM J. Matrix Anal. Appl.*, 36:2, 686-706.

20 S. Wuchty and E. Almaas, *Proteomics*, 2005, **5**, 444-449.

21 E. Estrada, *J. Proteome Res.*, 2006, **5**, 2177-2184.

22 N. Bhardwaj, P. M. Kim and M. B. Gerstein, *Sci. Signal.*, 2010, **3**(146), ra79.

23 V. Filkov, Z. M. Saul, S. Roy, R. M. D'Souza and P. T. Devanbu, *EPL*, 2009, **86**, 28003.

24 S. Roy and V. Filkov, *Phys. Rev. E*, 2009, **80**(4), 040902.

25 D. Hao, C. Ren and C. Li, *BMC Syst. Biol.*, 2012, **6**:34.

26 R. Pastor-Satorras, A. Vázquez, and A. Vespignani, 2001, *Phys. Rev. Lett.*, 87: 258701.

27 M. E. J. Newman, 2002, *Phys. Rev. Lett.*, 89: 208701.

28 C. H. Waddington, *Nature*, 1942, **3811**: 563-565.

29 R. Albert, H. Jeong and A.-L. Barabasi, *Nature*, *2000*, **406**, 378-382.

30 X. Zhu, M. Gerstein and M. Snyder, *Genes & Dev.*, 2007, **21**, 1010-1024.

31 A. G. Holman, P. Davis, J. M. Foster, *et al.*, *BMC Microbiol.*, 2009, **9**:43.

32 J. W. Tukey, *Ann. Math. Stat.*, 1958, **29**, 614.

33 O. Yuzuak, *Bogazici University, Chemical Engineering Department*, 2008.

34 K. Y. Arga, Z. I. Onsan, B. Kirdar, K. O. Ulgen and J. Nielsen, *Biotechnol. Bioeng.*, 2007, **97**:5, 1246-1258.

35 W. Fu, B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt, R. G. Ptak, *Nucleic Acids Res.*, 2009, **37**(Database issue), D417-22.

36 R. Tacutu, T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha , J. Costa, V. E. Fraifeld, J. P. de Magalhães, *Nucleic Acids Res.*, 2013, **41**(D1), D1027-D1033.

37 S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, M. R. Stratton, *Curr. Protoc. Hum. Genet.*, 2008, **10,**11.

38 K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabasi, PNAS, 2007, 104:21, 8685-90.

39 M. Ashburner *et al.*, *Nat. Genet.*, 2000, **25(1)**:25-9.

40 M. Cariaso and G. Lennon, *Nucleic Acids Res.*, 2012, 40, D1308-D1312.

41 E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, G. Sherlock, *Bioinformatics*, 2004. **20**(18), 3710-5.

42 M. Kanehisa and S. Goto, *Nucleic Acids Res.*, 2000, 28(1), 27-30.

