RESEARCH ARTICLE

Localized network centrality and essentiality in the yeast–protein interaction network

Keunwan Park[1] *and Dongsup Kim*[1,2]

[1] Department of Bio and Brain Engineering, KAIST, Daejeon, South Korea
[2] KAIST Institute for BioCentury, KAIST, Daejeon, South Korea

Received: May 27, 2009
Revised: July 22, 2009
Accepted: September 1, 2009

It has been suggested that a close relationship exists between gene essentiality and network centrality in protein–protein interaction networks. However, recent studies have reported somewhat conflicting results on this relationship. In this study, we investigated whether essential proteins could be inferred from network centrality alone. In addition, we determined which centrality measures describe the essentiality well. For this analysis, we devised new local centrality measures based on several well-known centrality measures to more precisely describe the connection between network topology and essentiality. We examined two recent yeast protein–protein interaction networks using 40 different centrality measures. We discovered a close relationship between the path-based localized information centrality and gene essentiality, which suggested underlying topological features that represent essentiality. We propose that two important features of the localized information centrality (proper representation of environmental complexity and the consideration of local subnetworks) are the key factors that reveal essentiality. In addition, a random forest classifier showed reasonable performance at classifying essential proteins. Finally, the results of clustering analysis using centrality measures indicate that some network clusters are closely related with both particular biological processes and essentiality, suggesting that functionally related proteins tend to share similar network properties.

Keywords:
Essentiality / Localization / Network centrality / Protein interaction network / Systems biology

1 Introduction

That highly connected nodes (hubs) in a protein interaction network tend to be more essential is a well-known relationship between network topology and gene essentiality [1–5]. This centrality–lethality rule states that essential proteins are likely to have a high degree of centrality. Network centrality measures the importance of a node for signal flow in various ways and has been used to reveal more precise relationships between hubs and their role in the network [6]. Importantly, however, because most network centrality measures depend on global network topology, changes in network contents or parts of network structure can have substantially altered the centrality of a node.

Many studies have shown differences between the network relationships identified, depending on the specific interaction networks used [1–3, 7, 8]. In particular, Yu *et al.* [7] reported an uncorrelated relationship between degree centrality and essentiality in yeast–protein interaction networks. As stated in their study, interactions generated with the yeast two hybrid (Y2H) technique tend to be transient binary interactions, while interactions identified with affinity purification followed by MS (AP/MS) typically represent multi-protein complexes (co-complexes). In their study, neither of the two networks generated from these two types of data showed a close relationship between degree

Correspondence: Professor Dongsup Kim, Department of Bio and Brain Engineering, KAIST, Daejeon 305-701, South Korea
E-mail: kds@kaist.ac.kr
Fax: +82-42-869-4310

Abbreviations: AP/MS, affinity purification followed by mass spectrometry; **GO,** Gene Ontology; **OOB,** out-of-bag; **Y2H,** yeast-two hybrid

and essentiality. Furthermore, in the case of the AP/MS network, a negative relationship was observed. However, the network composed of literature-curated interactions displayed a high linear correlation between centrality and essentiality ($R^2 = 0.84$), leading to the speculation that social preference or interest could have an artificial impact on network topology.

The sampling problem is another serious concern for proper representation of network structure [9]. Because protein–protein interaction data sets only contain a portion of the total number of interactions, modeling networks on these data sets can provide incorrect information about network properties unless the experimental sampling procedure is appropriate and the number of nodes sampled is large enough to reveal the architecture of the entire network. In addition, as reported by Zotenko *et al.* [8], essential proteins tend to form highly connected clusters (modular structures) that share similar functions, implying that network centrality measures that can represent environmental complexity have a better capacity to reveal essentiality. Considering these various difficulties, we propose that a better approach to linking network centrality to essentiality would involve a method that considers both global and local structures and uses different centrality measures for different kinds of networks.

In this study, various network centrality measures were calculated and applied to two recent yeast–protein interaction networks, the Y2H network and the AP/MS network, both from the study by Yu *et al.* [7]. Intrinsically, each centrality measure has unique attributes. We categorized the centrality measures by their properties, an approach that helped to reveal hidden factors about network topology and signal communication in the analyzed networks. Moreover, localized versions of these centrality measures were devised to examine sub-network effects. Using all of these measures, we more precisely explored the relationship between network topology and essentiality. Our results indicate that some centrality measures, particularly localized information centralities, show high correlation with essentiality, which naturally suggests the presence of underlying topological features representing essentiality.

From the results of our study, we concluded that essentiality in the AP/MS network is closely related to the local and dense clusters. The global positions of high-rank nodes (of localized information centrality) and the decision tree model for essential nodes support this conclusion about essentiality. In addition, a random forest classifier model showed reasonably good sensitivity and coverage for classifying essential nodes in the AP/MS network. Finally, we uncovered a connection between specific network clusters and biological processes using clustering analysis.

2 Materials and methods

2.1 Network centrality measures used in this study and their properties

In this study, 40 centrality measures, including the new localized centrality measures that we developed, were applied to yeast–protein interaction networks. The measures used were: shortest path betweenness (s_bet, lx_s_bet), shortest path closeness (s_clo, lx_s_clo), eigenvector centrality (eig, lx_eig), Harary graph centrality (graph), information centrality (info, lx_info), stress centrality (stress, lx_stress), random walk betweenness (rw_bet, lx_rw_bet),

Table 1. Summary of the centrality measures used in this study. The centrality measures that we developed are represented in boldface

Centrality	Global	Local	Signal flow	Complexity	Neighbor effects	Hub-related	Reference
Shortest path betweenness	s_bet	**l2_s_bet, l3_s_bet, l4_s_bet**	Shortest path				[10]
Shortest path closeness	s_clo	**l2_s_clo, l3_s_clo, l4_s_clo**	Shortest path				[10]
Eigenvector	eig	**l2_eig, l3_eig, l4_eig**	Random walk	Yes			[22]
Harary graph	graph		Shortest path				[23]
Information	info	**l2_info, l3_info, l4_info**	Path	Yes			[20]
Stress	stress	**l2_stress, l3_stress, l4_stress**	Shortest path				[24]
Random walk betweenness		**l2_rw_bet, l3_rw_bet, l4_rw_bet**	Random walk	Yes			
Random walk closeness		**l2_rw_clo, l3_rw_clo, l4_rw_clo**	Random walk	Yes			
Degree		degree			Yes	Yes	
Clustering coefficient		CC			Yes		[25]
Subgraph	SC	**l2_SC, l3_SC, l4_SC**	Random walk	Yes			[14]
Complexity		BI		Yes			[26]
Hub-relatedness		**maxdeg**				Yes	
Assortative mixing		ASS		Yes	Yes	Yes	[25, 27]

random walk closeness (rw_clo, lx_rw_clo), degree centrality (degree), clustering coefficient (CC), subgraph centrality (SC), complexity measure (lx_BI), sub-network maximum degree (lx_maxdeg), and assortative mixing (ASS) centralities. In our notation for the localized centrality measures, using lx_info for localized information centrality as an example, x can be 2, 3, or 4, specifying the size of the localized sub-networks centered at a specific node. We categorized the centrality measures by their properties, such as assumption of signal flow and effective distance. Simple descriptions of each centrality measure are shown in Table 1 and in the next section.

2.2 Development of new localized centrality measures

Localized versions of well-known global centrality measures were generated to test whether the localized version of a certain centrality measure was more meaningful in representing essentiality than the global metric. In this way, we were able to compare localized centrality measures with their global counterparts and examine what kind of information each provided. The sub-networks around a particular node were extracted according to the given path length. For example, a prefix of l2 indicates that the corresponding centrality measure was calculated from a sub-network with diameter 2 (for example, l2_eig is the eigenvector centrality calculated from a sub-network defined by limiting the length between the node of interest and the outermost node to 2). Nearly all centrality measures used in this study were defined both globally and locally.

2.3 The network centrality measures not based on shortest path

We also considered signal transmission as an underlying feature of the centrality measures. In graph-theoretic terms, a walk is any sequence of nodes and edges from any node to any node, a trail is a walk with distinct edges, and a path is a trail with distinct nodes.

Freeman closeness [10], a well-known network centrality, is related to the sum of the shortest path lengths from a given node to the others. This measures how fast a signal from one node travels to the other nodes. Freeman betweenness, another famous network centrality, is related to the number of shortest paths passing through a given node, a quantity that measures the volume of signal traffic for each node. Thus, both measures assume that signals in the network flow along the ideal path (shortest path), but this may be an inadequate assumption in some cases. In fact, Borgatti [11] has discussed the importance of correct assumptions regarding signal flows and how they should depend on specific types of networks. Specifically, what the node entities represent and how the nodes communicate

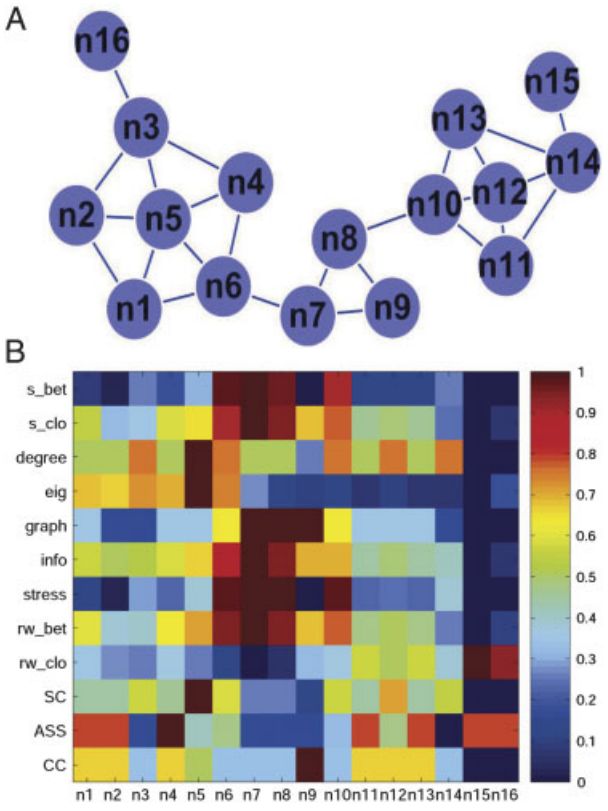

Figure 1. (A) Toy network example for understanding different centrality measures. (B) Heat map of centrality measures in the example network (red color means high centrality).

should be considered precisely when measuring information flow in the network.

Another important consideration is whether centrality can represent the complexity of a network structure. The shortest-path-based measures often miss this network property. For example, in our sample network (Fig. 1), node 9 has a zero Freeman betweenness value, while the other centrality measures give relatively high scores to the node. Moreover, the Freeman closeness centrality for node 9 may not change when the complexity of the two modules on either side increase or decrease. Therefore, we considered centrality measures based on random walks or paths (Table 1), approaches that can more properly represent network complexity, as shown in the example in Fig. 1.

2.4 Development of new centrality measures based on random walks

Although a betweenness centrality measure based on random walks has been developed by Newman [12], we developed the more intuitive centrality measures, rw_bet and rw_clo, calculated in the same manner as shortest path betweenness and closeness, respectively. In essence, our method is similar to the "absorbing model" presented by

Stojmircivic and YU [13], specifically an absorbing random walk of Newman [12], in that we assumed that a sink node ends signal transition.

We calculated expected visiting time for every node when starting signal flows by random walking until an absorbing node (or sink node) was reached. Using the expected visiting time for every node, rw_bet measures how many times the node was visited as an intermediate node for all pairwise signal transmissions. On the other hand, rw_clo measures how many walk-steps were needed to arrive at every other node from a given node, a value that is closely related to closeness centrality (for more details, see the Supporting Information material, Supp1). In the example network (Fig. 1), the rw_bet centrality of node 9 has a high value, while the rw_clo centrality of the same node has a low value. In addition, the subgraph centrality shows localized scores, which may be influenced by the scaling effect based on distance (by the factorial of the order of the spectral moment) [14].

Importantly, the walk- (or path-) based measures that consider multiple signal trajectories compute somewhat different scores than the shortest-path-based measures. Because the walk-based signal choose each step randomly, the nodes having high connectivity (or the nodes near the highly connected nodes) usually display high scores. In other words, the walk-based measures, rw_bet and rw_clo, consider the complexity of network topology and conceptually correspond to well-known betweenness and closeness centrality. The only difference between these methods is whether the signal propagates along the shortest path or a random walk. Thus, rw_bet and rw_clo can be applied for other uses, and interpretation of their calculations is dependent on the specific problem.

2.5 Network centrality measures for hubs or hub-relatedness

In addition to typical ASS and degree (ASS is a centrality measure about degrees of neighbor nodes) centrality, we developed a hub-related measure designated 'maxdeg'. It finds a node having the maximum degree in an extracted sub-network around a specified node. The maximum degree is divided by the shortest path length to that node from a specified node. Therefore, 'maxdeg' represents how closely a specified node is located to a hub node. As the maximum degree becomes larger and distance becomes shorter, the measure has a higher value. This measure was used to test whether essential nodes are located near hubs.

2.6 Yeast–protein interaction networks used in this study

We used two yeast protein interaction networks that were constructed with two different techniques (Supporting Information, Supp2). As previously described, these two networks exhibit different aspects of the total protein interaction map. Briefly, the Y2H network (union of Uetz-screen [15], Ito-core [16], and CCSB-YI1 [7]) was constructed from binary physical interactions and contained a large number of transient interactions. The AP/MS network (combined-AP/MS data set [17, 18]) was largely composed of co-complex machineries. In this study, we used the largest connected components of the two networks, resulting in 356 essential nodes from 1647 total nodes in the Y2H network and 401 essential nodes from 1004 total nodes in AP/MS network. In addition, the mean clustering coefficient, which measures connectivity of neighboring nodes, is very different for the two networks, 0.17 for the Y2H network and 0.72 for the AP/MS network.

2.7 Classification for essential protein prediction

The random forest approach, developed by Leo Breiman [19] and Adele Cutler, was to test classification capability for essential proteins. A random forest is a collection of tree-based classifiers where each tree construction depends on the independent feature-sampling procedure. The voting results from the ensemble of decision trees are used to determine the most popular objective class. For more details about the Random forest, see Breiman's study [19]. The random forest classifier has been shown to be relatively free from the over-fitting problem, as compared with other machine learning methods, making this approach the most appropriate for our performance test. In addition, the random forest performs a type of cross-validation in parallel with the training step by using the out-of-bag (OOB) error estimate. Specifically, the left-out samples (about one-third of samples) after bootstrapping in the training step constitute OOB samples. Because OOB samples have not been used in the tree construction, they are used to estimate test set errors (OOB error). We investigated the maximum performance in classifying essential proteins using only network centrality. The classification of 'Common' (Table 2) was performed by combining all network centrality measures for the AP/MS and Y2H networks.

The variable importance measure of the random forest classifier can be useful in finding influential network centrality as related to essentiality. In this study, measurement of variable importance was based on the Gini index (mean decrease in node impurities from splitting on the variable) obtained from all classification trees (forest).

3 Results and discussion

3.1 Enrichment of essential proteins in high-rank nodes by different centrality measures

We calculated 40 network centrality measures for the two yeast–protein interaction networks, Y2H and AP/MS. Our

Table 2. Classification performances by the random forest classifiers

Network (♯Essential/♯Non)	♯ feature	OOB error	Confusion matrix			
Y2H (356/1647)	40	25.72%		Essential	Non	Error
			Essential	14	336	0.96
			Non	77	1179	0.06
AP/MS (410/1004)	40	21.91%		Essential	Non	Error
			Essential	270	131	0.32
			Non	89	514	0.14
Common (139/333)	80	21.62%		Essential	Non	Error
			Essential	95	44	0.31
			Non	28	166	0.14
AP/MS (410/1004)	6	23.31%		Essential	Non	Error
			Essential	268	133	0.32
			Non	101	502	0.14

results indicate that the statistics of the two networks are substantially different. In the case of the Y2H network, some network centrality measures, including l2_maxdeg, l2_eig, l4_rw1, l4_eig, l2_s_bet, ASS, and l3_info (see Table 1 for centrality naming), showed better correlation with essentiality than the other measures (*i.e.* the nodes with higher scores for those centrality measures were more likely to be essential) (Fig. 2A and B). For example, 50% of the nodes with the top 1% of l2_eig scores were essential nodes, a much higher percentage than in a random set (21.6%). However, most centrality measures were not significantly correlated with essentiality. That is, the percentage of essential proteins among high-rank nodes (for example, up to 50th rank in Fig. 2B) for many centrality measures was about 40%, but those percentages decreased rapidly to a level similar to randomly chosen sets of nodes (21.6%). We also found that those centrality measures showing better correlation with essentiality were not correlated with each other, indicating that high-rank nodes for each centrality measure were not overlapping and that essentiality cannot be explained by one dominant factor. Together, these data indicate that there is no single best centrality measure, among the 40 measures tested, for predicting essentiality in the Y2H network.

In the case of the AP/MS network, the high-ranked nodes identified by the l2_info were the most highly correlated with essentiality and the l4_info performed second best (Fig. 2C and D). In the case of the l2_info, 90% of the top 1% nodes were essential, a much higher percentage than in a random set (39.9%). Similar to the Y2H network, most centrality measures showed low performance, with essentiality classification performance close to that of random sets. Among all information centrality measures, l2_info performed better than l3_info or l4_info. However, when considering more than the top 100 ranked nodes, the accuracy of l2_info decreased drastically (Fig. 2D). When the top 300 nodes identified in the AP/MS network by l2_info were considered, the accuracy decreased to 50%. These results suggest that the different centrality measures cover different regions for essentiality, although all measures performed better than random selection.

3.2 Classification using centrality measures

A random forest classifier was used to test whether essential proteins could be predicted from network centrality measures alone. Each node was represented by a feature vector composed of 40 network centrality measures, which was then used as an input vector for a random forest classifier (Table 2). In the Y2H network, the classification results showed very low performance in predicting essential proteins. The error rate was 96% for essential nodes, indicating that network centrality has no ability to predict essentiality. On the other hand, the classification result in the AP/MS network showed better performance. Of the 401 essential proteins, 270 nodes were correctly predicted (67.4%), and the OOB estimate of the error rate (see Section 2) was 21.91%. It should be noted that the primary reason for performing this classification procedure was to validate the relationship between network topology and essentiality, rather than to construct a precise classifier model for other uses. In this study, the OOB error rate, which is conceptually comparable to the cross validation procedure, is used to access the capability of our classifier model to explain the data.

Because classification performance was reasonable in the AP/MS network, influential variables in classification were listed by importance measure of the random forest classifier. Variable importance was based on the Gini index, representing the mean decrease in node impurities from splitting on the variable (see Section 2). The most important variable was l4_info, followed by ASS, info, l3_info, l2_info, and degree in decreasing order. The network centrality measures showing different distributions of essential and non-essential nodes could be useful for classification. The *p*-values calculated by two-sampled *t*-test estimation of the different mean values of the two distributions indicated statistical significance for those influential centrality measures: *p*-value of ASS: 0.08, l2_info: 4.839e-14, l3_info: 2.2e-16, l4_info: 2.2e-16. This result suggests that these information centrality measures are superior to the others for predicting essentiality.

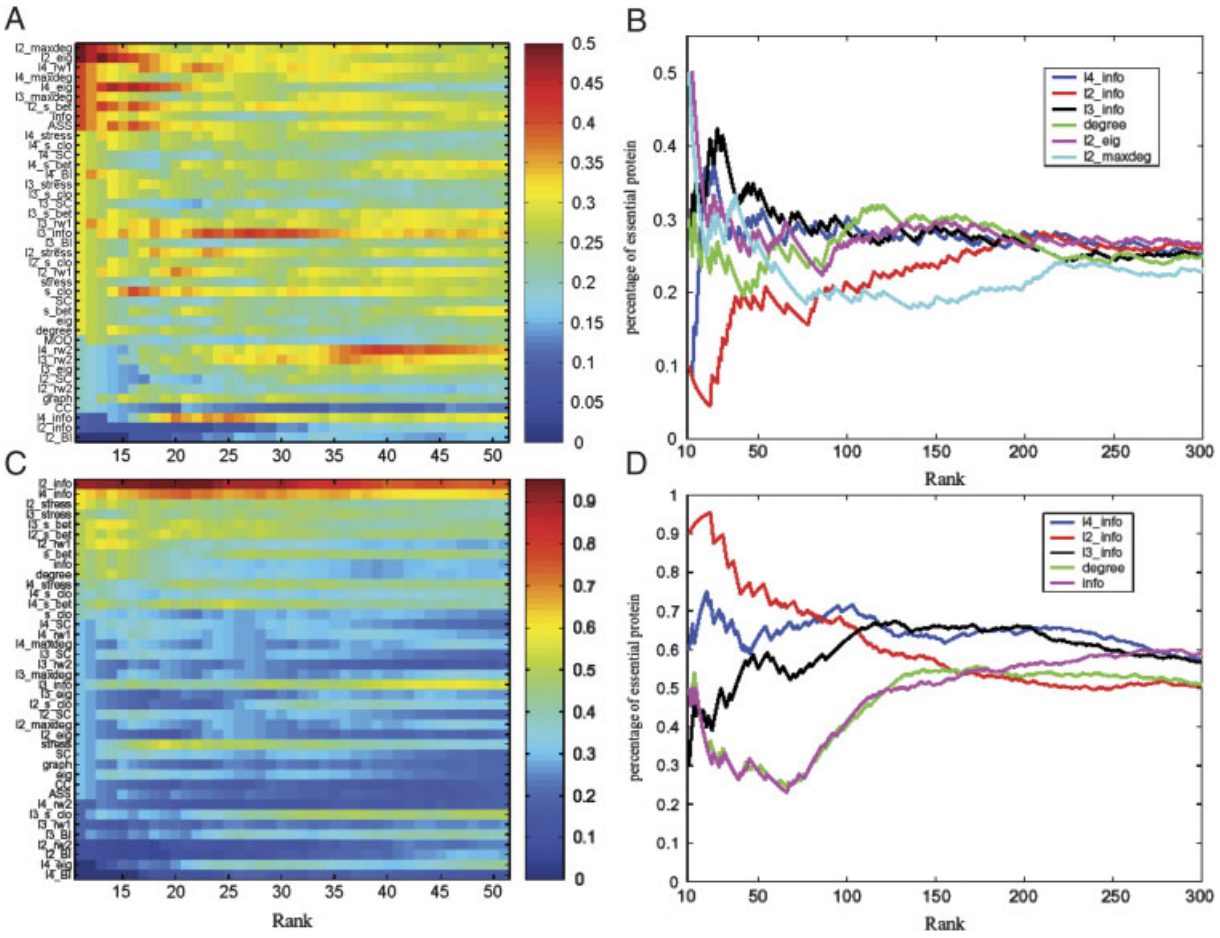

Figure 2. Heat map showing the fraction of essential nodes among high-rank (up to 50th) nodes by different centrality measures in (A) the Y2H network and (C) the AP/MS network. Line plots of a few measurements in (B) the Y2H network and (D) the AP/MS network.

We also tested classification accuracy considering only those top six centrality measures in the AP/MS network (see the fourth row in Table 2). The classification accuracy by the OOB estimate was 76.69% and the sensitivity was 67.3% (268/401), similar values to those of the classifier using all centrality measures. This result suggests that sub-networks of different lengths provide different information and that considering them together is helpful for inferring essentiality.

3.3 The size of sub-network affects the prediction accuracy

Comparing the patterns obtained from application of conceptually similar measures to sub-networks of differing size can be helpful in examining the effect of sub-network size and understanding meaningful internal network structure. In the Y2H network analysis, the l2_info measure correlated more poorly with essentiality than either l3_info or l4_info, implying that the nearest neighbors provide little information for predicting centrality. This result also implies that the appropriate network size for revealing essentiality in Y2H network should be at least 3. In other words, the essentiality information of the Y2H network is encoded in longer-ranged sub-networks.

In contrast, the results observed in the AP/MS were somewhat different. All localized information centrality measures showed good discrimination. While l2_info fit best to more high-scoring nodes (up to 100th rank), l3_info and l4_info fit well to mid-range nodes (after 100th rank). Surprisingly, global information centrality measures showed similar patterns to degree centrality, with no discriminatory patterns regarding essential nodes. In fact, global measures performed worse than random selection in some ranges. Thus, we can infer that local sub-networks provide sufficient information to estimate essentiality, and that the global centrality is often not ideal for predicting essentiality. One plausible explanation for these results is that the network nodes are not sampled uniformly or sufficiently. While global centrality measures would be severely affected by the sampling problem, localized centrality measures are relatively free from this issue. Our results suggest that global centrality measures should only

be used when network nodes are sampled uniformly and sufficiently from the entire population. In the case of insufficient sampling, global measures may give incorrect information and a well-devised local centrality measure may be a better solution in these cases.

3.4 Localized path-based information centralities have the best correlation with essentiality

For both networks, information centrality measures were the best predictors of essential proteins. Although many centrality measures that involved different network properties were tested for their relationships with essentiality, localized information centrality measures were the only metric that showed reasonable prediction accuracy. Information centrality is based on multiple paths, not random walks or the shortest path. It considers all paths from a source node to a target node, and the paths are weighted by their distance [20]. Similarly, the random walk-based centrality measure (rw_bet) traces a random walk from the source node to the target node. The critical difference between path-based and walk-based measures is that the walk-based measures give much higher scores to the highly connected nodes than path-based measures do because a path does not visit the same edge or the same node multiple times. In very extreme cases, random walks can go back and forth infinitely, giving much higher scores to high degree nodes that form large dense clusters. However, path-based measures detect local modular centers and simultaneously consider environmental complexity, so that they are not affected by dense clusters in the same way that other centrality measures are.

Global positions in the AP/MS network may provide some intuition about why path-based information centralities correlate well with essentiality. In Fig. 3, those nodes ranked highly by four different centrality measures are represented by a red color. The network displays dense and large hub nodes in the global center (Fig. 3B), and the degree centralities of these nodes are very high (Fig. 3A). Therefore, most centrality measures give high scores to the nodes belonging to this cluster. For example, the centrality measures assuming ideal path (shortest path) cannot capture the complexity, and as a result, do not represent essentiality well (Fig. 3C). In addition, the centrality measures assuming random walks focus mainly on the central cluster, due to its large number of complex connections. However, the localized information centrality gives high scores to the local highly connected nodes, as well as to the central hub nodes (Fig. 3D). These differences between centralities may arise from the different assumptions about signal transmission. As described by Zotenko *et al.*, [8] the critical feature revealing essentiality is the node's local neighborhood, rather than its global position. The main difference between our work and theirs is that we considered a longer range when defining local modules and have estimated essentiality by using only centrality within the modules.

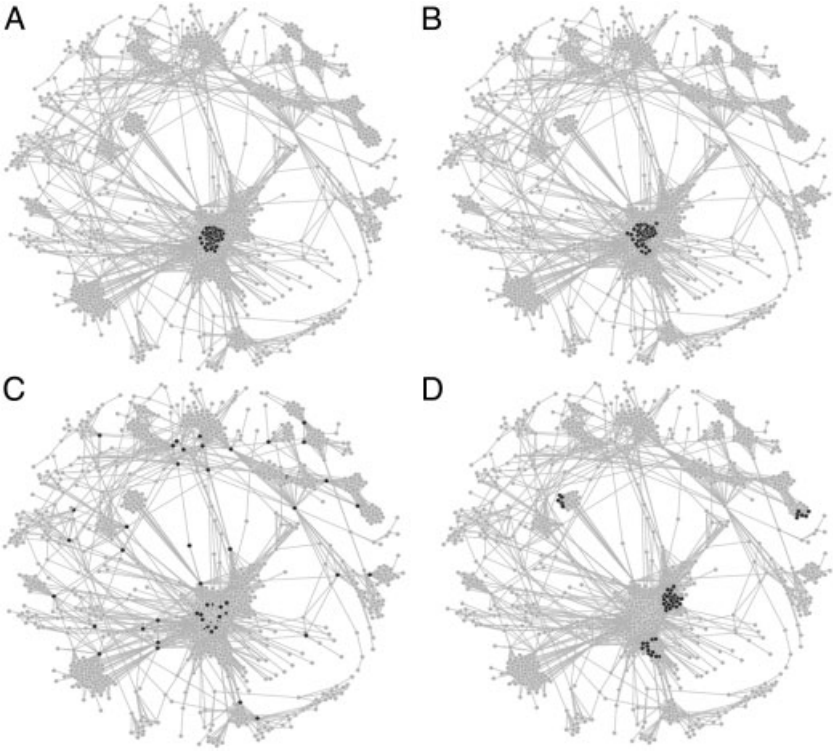

Figure 3. Topological positions of the top 50 nodes (red) for (A) degree, (B) info, (C) s_bet and (D) l2_info centralities in the AP/MS network.

Topologically, attacking the central dense cluster would effectively perturb the global network and should be sufficient to damage important network functions. However, these perturbation effects on local dense clusters cannot be explained if we assume that all signals in the network flow along the shortest paths. If we assume that signal transmissions travel by random walks, or along multiple paths, the importance of local dense nodes increases as shown in our toy example network (see n5 in Fig. 1). The center nodes in highly dense clusters have high impact on global structure in the centrality measures that assume random walks or paths (*e.g.* rw_bet, info in Fig. 1B). In addition, the shortest path is not a realistic measurement for describing communications in protein interaction networks because signals would be transmitted along all possible paths, through all possible neighbors, not along the single ideal route. For those signals traveling only through the ideal route, information about the ideal path should be known in advance. Furthermore, because of the static co-complex nature of the AP/MS network, the nodes in each cluster tend to perform similar functions together. Therefore, local dense clusters may act as important messenger modules for transmitting functional signals.

3.5 Network clustering analysis using centrality measures

To further investigate the relationship between network topology and gene essentiality, we performed *k*-means clustering analysis for the Y2H and AP/MS networks. As in the classification procedure, different centrality measures were used for representation of the network nodes. Because groups of nodes in the resulting clusters share common topological features within the same clusters, we used these clusters to test whether the specified groups could represent essentiality.

Although the classification performance for the Y2H network was poor, we presumed that a more specified network cluster might reveal network topology related to essentiality. However, *k*-means clustering (using $k = 5, 7$) results indicated that the essentiality proportion in the clustered groups was very close to random groups for most clusters (Table 3), indicating that essentiality (or non-essentiality) is unlikely to be encoded in the Y2H network topology. The global map of the Y2H network can be rewired in time and space, as a response to dynamic environmental forces. Therefore, to identify connections between essentiality and network topology, we may need to study a series of specific snapshots of the network, not the static interaction map that considers all possible interactions together.

In contrast, essentiality in the AP/MS network seemed to show some network characteristics arising from global positions of high rank nodes identified with the lx_info measure (compare the high-rank nodes in Fig. 3D with essential nodes in Fig. 4A). The clustering results for the AP/MS network show that the proportion of essential nodes in some clusters is much higher (or lower) than in randomly selected groups (Fig. 4B, Table 3). Surprisingly, the proportion of essential genes in the g3 cluster (using $k = 7$) is low (27.9%), even though this cluster contains many hub nodes. The mean degree of nodes of the g3 cluster is 125.86 and the node with the largest degree, 254, is a member of this cluster. In addition, 85.9% of the nodes with degree greater than 87 belong to the g3 cluster. Our findings and analysis of the g3 cluster suggests that high degree nodes or globally centered nodes are not necessarily related to essential genes. This feature may be the primary reason that prevents most centrality measures from accurately predicting essential nodes.

3.6 Functional implications of network clusters

We performed Gene Ontology (GO) analysis on the seven clusters of the AP/MS network to identify functional relationships of each network cluster that might connect specific network topology to particular biological functions

Table 3. Proportion of essential genes in each cluster constructed through the *k*-means clustering method

♯ cluster	Arbitrary cluster name	Essentiality (%) in Y2H(356/1647)	Essentiality (%) in AP/MS(401/1004)
$K = 7$	g1	22(43/192)	36.5(76/208)
	g2	19.7(70/356)	68.8(66/96[a)])
	g3	29.8(20/67)	27.9(24/86)
	g4	29.7(33/111)	66.3(59/89)[a)]
	g5	17.9(70/391)	38.3(67/175)
	g6	22.4(98/438)	24.7(67/271)
	g7	23.9(22/92)	53.2(42/79)[a)]
$K = 5$	g1	21.5(67/311)	59(85/143)[a)]
	g2	19.4(103/530)	24.9(67/269)
	g3	21(117/555)	39.3(162/412)
	g4	28.8(36/125)	67.7(63/93)[a)]
	g5	26.2(33/126)	27.6(24/87)

a) Represents the cluster of high essentiality.

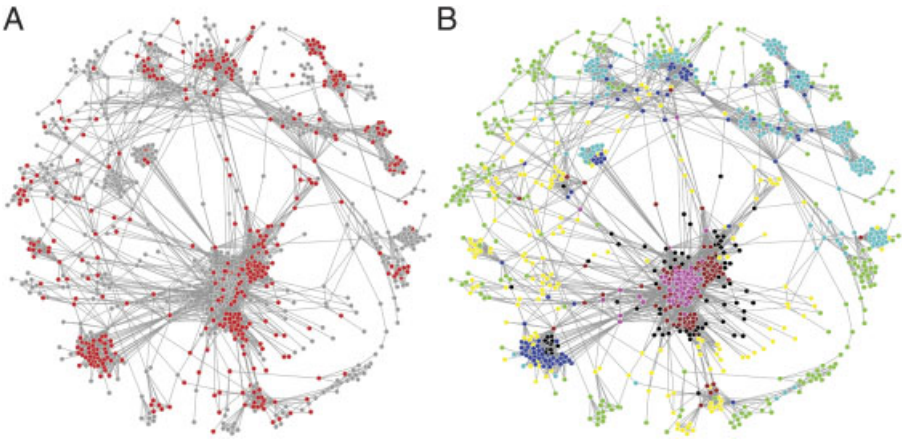

Figure 4. (A) Essential nodes (red) in the AP/MS network. (B) The seven clusters identified through *k*-means clustering methods are represented by different colors (cyan: g1, blue: g2, pink: g3, red: g4, yellow: g5, green: g6, black: g7).

as well as essentiality. The term enrichment tool of AmiGO [21] (the official tool for searching and browsing the GO database) was used to find significantly shared GO terms in a given gene list. The gene lists for each cluster were used as query gene products and all genes in the AP/MS network were used as a background set. The *p*-value, a significance measure of overlapping GO terms, was determined by considering the sample frequency and the background frequency. Among the top five GO terms (by non-decreasingly ordered *p*-values), only the terms with more than 40% sample frequency (*i.e.* covering more than 40% nodes within the corresponding cluster) were considered in this study. The cases in which no terms satisfying the above criteria were identified are denoted in the tables by 'No term'. Because GO terms are categorized into biological process, cellular component, and molecular function, we calculated significant GO terms for each category (Tables 4 and 5).

Among the seven clusters from the AP/MS network, the g3 and g6 clusters showed low proportions of essential genes. The g3 cluster, which included many nodes with very high degrees located at the global center, was enriched for the following common GO terms: 'translation', 'cellular protein metabolic process', and 'cellular biopolymer biosynthetic process' in the biological process category, 'cytosolic ribosome' and 'cytosolic part' in the cellular component category, and 'structural constituent of ribosome' and 'structural molecule activity' in molecular function category (Tables 4 and 5). These enrichments suggest that the proteins in the g3 cluster are mainly located in the cytoplasm and generally involved in the chemical reactions and pathways involving a specific protein. The nodes involved in translation (g3 cluster) formed a large highly connected cluster, but they were unlikely to be essential genes. In addition, the nodes of the g6 cluster were enriched for the 'biological regulation' term in the biological process category and no significant terms in other GO categories.

In contrast to the g3 and g6 cluster, genes in the g2, g4, and g7 clusters, all of which show a high proportion of essential genes, were enriched for GO functions suggesting nuclear components and generally performed functions at the transcription level (g2 cluster: ''RNA splicing'', g4 cluster: ''ncRNA processing'', ''ribosome biogenesis'', g7 cluster: ''ribonucleoprotein complex'', ''ribosome biogenesis'').

In summary, essential nodes tend to perform functions at the transcriptional level in the nucleus, while non-essential nodes are more likely to be involved in biological regulation or translation in the cytosol. Moreover, our results suggested that some network groups (*e.g.* the g2 and g4 clusters) are closely related to specific biological functions, especially biological processes in GO categories.

3.7 Putative classification model for essential nodes in the AP/MS network

Our classification results implied that some topological features are related to essentiality even though they may not be revealed through a single high centrality value. Based on this point of view, the random forest classifier has a shortcoming, in that it is difficult to understand the topology of essential nodes clearly due to the intrinsic complexity of the machine learning technique. Therefore, we constructed a tree-based model using centrality measures that would provide clearer interpretation in terms of network topology. We utilized the essentiality classification model using a decision tree method and excluding the g3 cluster. Results of this model suggests that essential nodes tend to be located in dense local clusters (l3_SC > -0.64 and ASS < 1.69) and not at the edge of the network (info > 0.56) (Fig. 5). The criterion values for each centrality measure were scaled by subtracting sample means and dividing by standard deviations. The inequality about the global information centrality, info, suggests that the majority of essential nodes tend to be located at or near the global center of the network. In addition, high l3_SC and low ASS values represent highly connected environments and low degrees of neighboring nodes, attributes that together form a dense local cluster.

Table 4. Significant GO terms (biological process) for the seven clusters

Cluster	GO	*p*-Value	Sample frequency	Background frequency
g1	GO:0006996 organelle organization	1.29e−17	107/204 (52.5%)	272/984 (27.6%)
	GO:0006351 transcription, DNA-dependent	2.34e−15	92/204 (45.1%)	228/984 (23.2%)
	GO:0032774 RNA biosynthetic process	2.34e−15	92/204 (45.1%)	228/984 (23.2%)
	GO:0006350 transcription	4.76e−15	92/204 (45.1%)	230/984 (23.4%)
g2	GO:0000377 RNA splicing, *via* transesterification reactions with bulged adenosine as nucleophile	2.17e−24	39/96 (40.6%)	71/984 (7.2%)
	GO:0000398 nuclear mRNA splicing, *via* spliceosome	2.17e−24	39/96 (40.6%)	71/984 (7.2%)
	GO:0000375 RNA splicing, *via* transesterification reactions	2.17e−24	39/96 (40.6%)	71/984 (7.2%)
	GO:0008380 RNA splicing	8.98e−24	39/96 (40.6%)	73/984 (7.4%)
	GO:0006397 mRNA processing	1.10e−20	41/96 (42.7%)	94/984 (9.6%)
g3	GO:0006412 translation	3.23e−27	71/83 (85.5%)	304/984 (30.9%)
	GO:0044267 cellular protein metabolic process	9.12e−13	71/83 (85.5%)	487/984 (49.5%)
	GO:0019538 protein metabolic process	1.36e−12	71/83 (85.5%)	490/984 (49.8%)
	GO:0034961 cellular biopolymer biosynthetic process	3.03e−11	72/83 (86.7%)	529/984 (53.8%)
	GO:0034645 cellular macromolecule biosynthetic process	5.56e−11	72/83 (86.7%)	534/984 (54.3%)
g4	GO:0034470 ncRNA processing	3.74e−22	50/87 (57.5%)	150/984 (15.2%)
	GO:0042254 ribosome biogenesis	1.45e−21	65/87 (74.7%)	274/984 (27.8%)
	GO:0022613 ribonucleoprotein complex biogenesis	3.69e−21	69/87 (79.3%)	318/984 (32.3%)
	GO:0006364 rRNA processing	1.67e−20	47/87 (54.0%)	141/984 (14.3%)
	GO:0044085 cellular component biogenesis	7.91e−17	71/87 (81.6%)	392/984 (39.8%)
g5	No term			
g6	GO:0065007 biological regulation	3.45e−14	141/266 (53.0%)	334/984 (33.9%)
	GO:0050789 regulation of biological process	2.98e−10	126/266 (47.4%)	313/984 (31.8%)
	GO:0050794 regulation of cellular process	3.60e−10	124/266 (46.6%)	307/984 (31.2%)
g7	GO:0022613 ribonucleoprotein complex biogenesis	2.98e−16	59/78 (75.6%)	318/984 (32.3%)
	GO:0042254 ribosome biogenesis	1.67e−11	49/78 (62.8%)	274/984 (27.8%)
	GO:0044085 cellular component biogenesis	2.37e−11	59/78 (75.6%)	392/984 (39.8%)
	GO:0006396 RNA processing	3.52e−09	42/78 (53.8%)	238/984 (24.2%)

The proportion of essential genes in the g6 cluster (using $k = 7$) is also low (24.7%) and the cluster is topologically located at the border of the AP/MS network. The inequality conditions, l3_SC < -0.8 and l2_rw_bet < -0.486, represent 246 nodes of the 271 total nodes of the g6 cluster. Because both l3_SC and l2_rw_bet are local centrality measures, these negative values indicate that the g6 cluster is located at the edge of the network. Taken together, our data show that in the AP/MS network, essential nodes tend to be found in dense local clusters not located at the edge of the network, but at the local centers that connect neighbor clusters.

4 Concluding remarks

In this study, we investigated whether essential proteins could be inferred from network centrality alone. Further, we studied which centrality measures described gene essentiality well. The underlying assumption of our investigation was that the various centrality measures make their own assumptions about signal transmission and that particular centrality measurements would reveal characteristic topological features. A more thorough analysis of those centrality measures that properly reveal node essentiality would be helpful for understanding essentiality in terms of network topology.

As shown by a previous study [7], the two networks, Y2H and AP/MS, which were constructed using different experimental techniques, showed distinct topological features. Accordingly, we expected that the two networks would represent different aspects of the protein interaction map and might display different capacities to reveal essentiality from topology. In our results, various network centrality measures of the Y2H network seemed to have little utility at predicting essential nodes (Fig. 2). Our finding is consistent with recent experimental results indicating that degree centrality is closely related to the phenotypic variance, but not to essentiality [7]. The inability of the Y2H network to predict essentiality may be the result of insufficient network data or the fact that Y2H techniques may not effectively predict essentiality. On the other hand, the topological centrality of the AP/MS network explains essentiality to a greater extent. Because of the co-complex nature of the AP/MS network, nodes tend to form modular dense clusters. These topological features of this type of network may be more apt at revealing essential nodes.

Our results suggest that path-based and localized information centrality measurements predict essentiality in both networks. Conversely, our findings imply that global centrality measures and hub-related measures might be not appropriate for revealing essentiality. Moreover, localized

Table 5. Significant GO terms (cellular component and molecular function) for the seven clusters

Cluster	GO (cellular component)	*p*-value	Sample frequency	Background frequency
g1	GO:0044451 nucleoplasm part	6.26e−23	97/204 (47.5%)	205/984 (20.8%)
	GO:0005654 nucleoplasm	1.35e−21	98/204 (48.0%)	215/984 (21.8%)
	GO:0043233 organelle lumen	4.56e−20	142/204 (69.6%)	407/984 (41.4%)
	GO:0070013 intracellular organelle lumen	4.56e−20	142/204 (69.6%)	407/984 (41.4%)
	GO:0031974 membrane-enclosed lumen	1.60e−19	142/204 (69.6%)	411/984 (41.8%)
g2	GO:0044428 nuclear part	1.01e−09	73/96 (76.0%)	462/984 (47.0%)
g3	GO:0022626 cytosolic ribosome	6.55e−53	65/83 (78.3%)	115/984 (11.7%)
	GO:0044445 cytosolic part	1.29e−44	65/83 (78.3%)	143/984 (14.5%)
	GO:0033279 ribosomal subunit	1.97e−40	65/83 (78.3%)	161/984 (16.4%)
	GO:0005829 cytosol	1.37e−39	65/83 (78.3%)	165/984 (16.8%)
	GO:0005840 ribosome	5.46e−38	68/83 (81.9%)	196/984 (19.9%)
g4	GO:0005730 nucleolus	3.09e−26	53/87 (60.9%)	145/984 (14.7%)
	GO:0030684 preribosome	2.96e−25	46/87 (52.9%)	109/984 (11.1%)
	GO:0043228 non-membrane-bounded organelle	9.00e−15	73/87 (83.9%)	446/984 (45.3%)
	GO:0043232 intracellular non-membrane-bounded organelle	9.00e−15	73/87 (83.9%)	446/984 (45.3%)
g5	No term			
g6	No term			
g7	GO:0030529 ribonucleoprotein complex	1.07e−12	60/78 (76.9%)	382/984 (38.8%)
Cluster	GO (molecular function)	*p*-value	Sample frequency	Background frequency
g1	No term			
g2	GO:0003824 catalytic activity	2.08e−08	60/96 (62.5%)	353/984 (35.9%)
g3	GO:0003735 structural constituent ofribosome	9.45e−42	65/83 (78.3%)	155/984 (15.8%)
	GO:0005198 structural molecule activity	1.42e−38	65/83 (78.3%)	170/984 (17.3%)
g4	No term			
g5	No term			
g6	No term			
g7	No term			

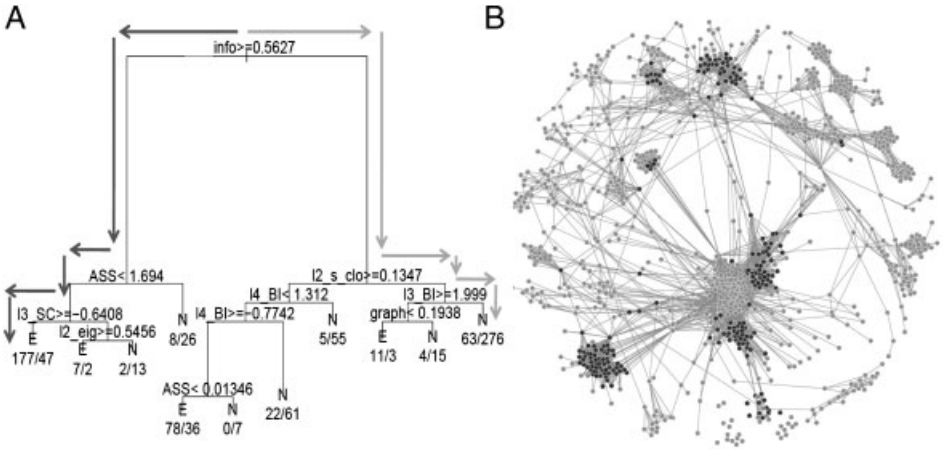

Figure 5. (A) Essentiality model using the decision tree classification method. The majority of essential nodes are represented by red arrow paths and the majority of non-essential nodes by blue arrow-paths. '*E*' denotes essential nodes and '*N*' denotes non-essential nodes. The numbers in the leaf nodes represent '(number of essential nodes)/(number of non-essential nodes)'. (B) The network view of the nodes that belong to the left red arrow path (represented by violet colors).

information centrality measures covering different ranges provide relevant information about essential nodes. The localized centrality measures that assume ideal paths or random walks show weaker correlations with essentiality than the information centrality measures. That is, those centrality measures that represent environmental complexity and consider the local sub-network around a particular node are a better measurement for predicting essential nodes in the protein interaction network, especially in the AP/MS network. Based on our finding that localized information centrality measures contain the most relevant information for predicting essentiality, we infer that local dense clusters tend to contain essential nodes, as the effects of perturbation on the clusters could be substantial under the plausible assumption that signal flows through multiple paths utilizing its neighboring environments, and not by a single shortest path. Furthermore, the results from our clustering analysis indicate that specific biological processes

are consistent with specific network clusters, suggesting a close relationship between specific network topology and biological function. In conclusion, despite recent controversy regarding the relationship between centrality and essentiality, our study demonstrates that cellular functions, including essentiality, are closely related to network topology.

This work is supported by CHUNG Moon Soul Center for BioInformation and Bioelectronics (CMSC), and by Korea Science and Engineering Foundation.

The authors have declared no conflict of interest.

5 References

[1] Jeong, H., Mason, S. P., Barabasi, A. L., Oltvai, Z. N., Lethality and centrality in protein networks. *Nature* 2001, *411*, 41–42.

[2] Batada, N. N., Hurst, L. D., Tyers, M., Evolutionary and physiological importance of hub proteins. *Plos. Comput. Biol.* 2006, *2*, e88.

[3] Hahn, M. W., Kern, A. D., Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 2005, *22*, 803–806.

[4] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., Gerstein, M., The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *Plos. Comput. Biol.* 2007, *3*, e59.

[5] Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., Gerstein, M., Genomic analysis of essentiality within protein networks. *Trends Genet.* 2004, *20*, 227–231.

[6] Estrada, E., Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 2006, *6*, 35–40.

[7] Yu, H., Braun, P., Yildirim, M. A., Lemmens, I. *et al.*, High-quality binary protein interaction map of the yeast interactome network. *Science* 2008, *322*, 104–110.

[8] Zotenko, E., Mestre, J., O'Leary, D. P., Przytycka, T. M., Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *Plos. Comput. Biol.* 2008, *4*, e1000140.

[9] Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E., Vidal, M., Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.* 2005, *23*, 839–844.

[10] Freeman, L. C., Centrality in social networks I: conceptual clarification. *Soc. Networks* 1979, *1*, 215–239.

[11] Borgatti, S. P., Centrality and network flow. *Soc. Networks* 2005, *27*, 55–71.

[12] Newman, M. E. J., A measure of betweenness centrality based on random walks. *Soc. Networks* 2005, *27*, 39–54.

[13] Stojmircivic, A., Yu, Y. K., Information flow in interaction networks. *J. Comput. Biol.* 2007, *14*, 1115–1143.

[14] Estrada, E., Rodriguez-Velazquez, J. A., Subgraph centrality in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 2005, *71*, 056103.

[15] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, *403*, 623–627.

[16] Ito, T., Chiba, T., Ozawa, R., Yoshida, M. *et al.*, A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Nat. Acad. Sci. USA* 2001, *98*, 4569–4574.

[17] Gavin, A. C., Aloy, P., Grandi, P., Krause, R. *et al.*, Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006, *440*, 631–636.

[18] Krogan, N. J., Cagney, G., Yu, H. Y., Zhong, G. Q. *et al.*, Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006, *440*, 637–643.

[19] Breiman, L., Random forests. *Mach. Learning* 2001, *45*, 5–32.

[20] Stephenson, K., Zelen, M., Rethinking centrality: methods and examples. *Soc. Networks* 2989, *11*, 1–37.

[21] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, *25*, 25–29.

[22] Bonacich, P., Some unique properties of eigenvector centrality. *Soc. Networks* 2007, *29*, 555–564.

[23] Hage, P., Harary, F., Eccentricity and centrality in networks. *Soc. Networks* 1995, *17*, 57–63.

[24] Shimbel, A., Structural parameters of communication networks. *Bull. Math. Biophys.* 1953, *15*, 501–507.

[25] Watts, D. J., Strogatz, S. H., Collective dynamics of 'small-world' networks. *Nature* 1998, *393*, 440–442.

[26] Danail Bonchev, G. A. B., *Quantitative Measures of Network Complexity*, Springer US 2005.

[27] Newman, M. E. J., Assortative mixing in networks. *Phys. Rev. Lett.* 2002, *89*, 208701.

