Revisiting the General Concept of Network Centralities: A Propose for Centrality Analysis in Network Science

Minoo Ashtiani[1], Mehdi Mirzaie[2*], Zahra Razaghi-Moghadam[3], Holger Hennig[4], Olaf Wolkenhauer[4], Ali Salehzadeh-Yazdi[4*], & Mohieddin Jafari[1*]

[1]Drug Design and Bioinformatics Unit, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, 69 Pasteur St, PO Box 13164, Tehran, Iran

[2]Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Jalal Ale Ahmad Highway, PO Box 14115-134, Tehran, Iran

[3]Faculty of New Sciences and Technologies, University of Tehran, North Kargar Street, Tehran 143995-7131, Iran

[4]Department of Systems Biology and Bioinformatics, University of Rostock, 18051 Rostock, Germany

[*]Corresponding Authors:

Ali Salehzadeh-Yazdi

ali.salehzadeh-yazdi@uni-rostock.de

Mehdi Mirzaie

mirzaie@modares.ac.ir

Mohieddin Jafari

www.jafarilab-pasteur.com

m_jafari@pasteur.ac.ir

Tel: +98 21 6411 2519

Fax: +98 21 66480780

Abstract

Background: In network science, although different types of centrality measures have been introduced to determine important nodes of networks, a consensus pipeline to select and implement the best tailored measure for each complex network is still an open field. In the present study, we examine the node centrality profiles of protein-protein interaction networks (PPINs) in order to detect which measure is succeeding to predict influential proteins. We study and demonstrate the effects of inherent topological features and network reconstruction approaches on the centrality measure values.

Results: PPINs were used to compare a large number of well-known centrality measures. Unsupervised machine learning approaches, including principal component analysis (PCA) and clustering methods, were applied to find out how these measures are similar in terms of characterizing and assorting network influential constituents. We found that the principle components of the network centralities and the contribution level of them demonstrated a network-dependent significancy of these measures. We show that some centralities namely Latora, Decay, Lin, Freeman, Diffusion, Residual and Average had a high level of information in comparison with other measures in all PPINs. Finally, using clustering analysis, we restated that the determination of important nodes within a network depends on its topology.

Conclusions: Using PCA and identifying the contribution proportion of the variables, i.e., centrality measures in principal components, is a prerequisite step of network analysis in order to infer any functional consequences, e.g., essentiality of a node. Our conclusion is based on the signal and noise modeling using PCA and the similarity distance between clusters. Also, an interesting strong correlation between silhouette criterion and contribution value was found which corroborates our results.

Keywords:

Network science, Centrality analysis, Protein-Protein Interaction Network (PPIN), Clustering, Principal Components Analysis (PCA)

Introduction

Network modelling of diverse biological complex processes is a pervasive approach in contemporary biological researches [1]. This application of modelling is divided into two parts; moving from elements and their relationships toward a whole system and getting back from a whole into important and local structures among the system [2, 3]. For instance, researchers experimentally or computationally collect information about a set of proteins and the interactions between them to reconstruct a protein-protein interactions network (PPIN). Then, they define an importance or an influence concept such as degree centrality (hub) in order to identify influential structures among the given PPINs which contains individual nodes, edges and subgraphs such as cliques or pathways. Not only in the biological networks but also in all types of networks such as social or literary, finding center of the network is a chief question which is called centrality analysis [4, 5]. By this analysis, the influential parts of the networks which are useful to predict the behavior of the whole network could be determined. Any change of these parts could be followed through the system and may disturb the whole [6]. Various centrality metrics or measures have been defined during last forty years, mostly in the context of social network analyses [5, 7]. Whilst basic concepts of the graph theory, including number of links, distances, eigenvalues and local structures, are represented in almost all centrality analyses, some of these measures are very popular and mostly used in diverse context using specific tools [8]. There is much research in biological areas which has studied the correlation of the lethality and essentiality with some centrality measures [4, 9-16]. Additionally, these measures have been applied in many social and epidemiological research studies, examples include predicting the details of information controlling or disease spreading within a specific network in order to delineate how to effectively implement target marketing or preventive healthcare [7, 17-21].

However, the selection of the appropriate metric for given networks is still an open question. Which one is better in translating center of the real networks? Do all of them independently highlight the central network elements and encompass independent information or are they correlated? Are computations of all these measures meaningful in all different network models or do they depend on the network architecture and the logic of the network modeling?

In 2004, Koschützki and Schreiber compared five centrality values in two biological networks and showed different patterns of correlations between centralities. They generally concluded that all Degree, Eccentrecity, Closeness, random walk Betweenness and Bonacich's Eigenvector centralities should be considered and could be useful in various applications without explaining any preference among them [22]. Two years later, they showed the independence behavior of centrality measures in a PPIN using 3D parallel coordinates, orbit-based and hierarchy-based comparison [23]. Valente et al. examined the correlation

between the symmetric and directed versions of four measures which are commonly used by the network analysts. By comparing 58 different social networks, they concluded that network data collection methods change the correlation between the measures and these measures show distinct trends [24]. Batool and Niazi also studied three social, ecological and biological neural networks and they concluded the correlation between Closeness-Eccentricity and Degree-Eigenvector and insignificant pattern of Betweenness. They also demonstrated that Eigenvector and Eccentricity measures are better to identify influential nodes [25]. In 2015, Cong Li et al. further investigated the question of correlation between centrality measures and introduced a modified centrality measure called *m*th-order degree mass. They observed a strong linear correlation between the Degree, Betweenness and Leverage centrality indices within both real and random networks [26].

Similar in methodology but different in networks, all these studies attempted to quantify correlations between different well-known centrality measures. In contrast, in this study, we develop a formal methodology (besides the correlation analysis) for centrality comparison and usage within network analysis. We comprehensively compared 27 distinct centrality measures applied to 14 small to large biological and random networks. All biological networks are PPIN among same set of proteins which are reconstructed using a variety of computational and experimental methods. We demonstrate how the ranking of nodes (or edges) using these measures depends on the network structure (topology) and why this network concept i.e. centrality requires renewed attention.

Materials and methods

The workflow of this study is schematically presented in Fig. 1. Our workflow starts by constructing and retrieving networks, followed by network global analysis. The centrality analysis and comparing the centrality measures using machine learning methods were the next main steps.

Reconstruction of the networks

In this study, a UniProtKB reviewed dataset [27] was used to retrieve proteins in *Saccharomyces cerevisiae* (6721 proteins). UniProtKB accessions were converted to STRING using the STRINGdb R package, which resulted in 6603 protein identifiers (3rd Sep2016). Interactions among proteins were extracted based on the STRING IDs . In the 2017 edition of the STRING database the results of these interactions are structured in a way to provide maximum coverage; this is achieved by including indirect and predicted interactions on the top of the set. Each result is assigned a score which represents to what extent the interactions are biological, meaningful, specific and reproducible according to the supporting evidence. The STRING database is comprised of channels that are known as "evidence channels", which are formed according to the origin and type of evidence [28]. In this study, 13 evidence channels indicating PPIN of yeast are present: co-expression, co-expression-transferred, co-occurrence, database, database-transferred, experiments, experiments-transferred, fusion, homology, neighborhood-transferred, textmining, textmining-transferred and combined-score.

For the purpose of comparison with real network behavior, a null model network was generated. The null network is the Erdős–Rényi model [29] and was generated using the igraph R package [30]. The generated null network was created with a size similar to the yeast reconstructed PPIN in order to have a more fair comparison.

Fundamental network concepts analysis

To understand the network structure, we reviewed various network features using several R packages [31-33]. First of all, we inquired the score distribution of different channels and whether there is any significant correlation among the scores for each PPI. Then the networks were made readable in the R environment. The network density, clustering coefficient, network heterogeneity, and network centralization properties of the network were calculated. The number of connected components and graph diameter for each network were also computed. Then, the power-law distribution was determined by computing α and r values.

Centrality analysis

For this research study, we are only considering undirected, loop-free connected graphs according to the PPIN topology. For centrality analysis, the following 27 centrality measures were selected: Average Distance [34], Barycenter [35], Closeness (Freeman) [7], Closeness (Latora) [36], Residual closeness [37], ClusterRank [38], Decay [39], Diffusion degree [40], Density of Maximum Neighborhood Component (DMNC) [41], Geodesic K-Path [42, 43], Katz [44, 45], Laplacian [46], Leverage [47], Lin [48], Lobby [49], Markov [50], Maximum Neighborhood Component (MNC) [41], Radiality [12], Eigenvector [51], Subgraph scores [52], Shortest-Paths betweenness [7], Eccentricity [53], Degree, Kleinberg's authority scores [54], Kleinberg's hub scores [54], Harary graph [53] and Information [55]. All these measures can be calculated for undirected networks in a reasonable time. These measures were established using the centiserver [8], igraph [30] and sna [56] R packages. We assorted the centrality measures into five distinct parts including Distance-, Degree-, Eigen-, Neighborhood-based and miscellaneous groups depend on their logic and formulas (Table 1).

Unsupervised machine learning analysis

We used principal components analysis (PCA) [57] as a key step to understanding the influence of a centrality measure within a network. PCA was performed on normalized computed centrality measures . In order to visualize the result of the PCA and clustering analysis, we applied some R packages [58-60]. After normalization, we examined whether the centrality measures in all networks can be clustered by performing the clustering tendency procedure. This required calculating the Hopkins' statistic values and using VAT (Visual Assessment of cluster Tendency) plot. We applied the clustering validation measures to get the best number of clusters using some other R packages [61, 62]. This will provide silhouette scores in different kinds of clustering methods such as hierarchical, k-means, and PAM (Partitioning Around Medoids).. Finally, distance metrics could be formed using the Pearson correlation in order to find any relationships between two centrality measures in each PPIN. To be able to compare the clustering results in the various PPINs, the Jaccard similarity index [63] was used relying on the similarity metrics of the clustering results [64].

Results

Evaluation of network global properties and centrality analysis

By importing the same set of protein names, the 13 different PPINs were extracted from the STRING database using different predictive channels. Note that the PPI scores derived from the neighborhood channel of yeast were all zero. The distribution of STRING links and edge scores are presented for each channel (Fig. 2). As shown in the figure, all these channels independently identify an interaction for each protein pair with specified score or weight. Therefore, we have 13 dissimilar network prediction methods and the following results are independent from each other. The independency between evidence channels is also shown in Fig. 3 by a pairwise scatterplot and Pearson's r correlation coefficient. The scores are not correlated and most of the coefficients are around the zero.

In the following, we consider the evidence channel scores as edge weights in each PPIN. So, along with the Erdos-Renyi null network, the 14 networks were utilized in order to establish an examination of centrality measures. Note that the giant component of each network was calculated to compute corresponding measures. To make sense of the networks and their giant component's structure, some fundamental network concepts were computed as shown in Table 2. The homology, fusion, co-occurrence and database networks contain high numbers of unconnected components. Except the homology network which has the smallest giant component, the density of all networks are between 0.01-0.05, as was expected as real network are typically sparse. The network diameter of the fusion, co-occurrence, database and co-expression are one order of magnitude greater than the others. All of the PPINs except homology network have high r correlation values to the power-law distribution with the diverse alpha coefficients. The high value of the average clustering coefficients of the database and homology indicates the modular structure of these networks. Most of the PPINs have the high value of the heterogeneity and network centralization compared with the null network. The degree distribution and clustering coefficients for the networks are also plotted in Fig. 4 and Fig. 5 respectively. Except the homology network, all the degree distribution are left-skewed similar to the scale-free networks. The scale-free property of each network was visualized (See Supplementary file 1).

In the next step, the 27 centrality measures of nodes were computed in all of the 14 networks (Data is available upon request). As shown in Fig. 6, beside the observed correlation between centrality measures in both combined-score and Erdos-Renyi networks, their distribution and the level of associations showed the vast diversity among all five centrality groups especially in Distance-, Neighborhood-based and Miscellaneous groups. The Eccentricity and Leverage centralities which comes from Distance- and Degree-based groups respectively showed the independent shape of distribution compared to the others. This pattern is repeated in all PPINs to some extent (See Supplementary file 2). However, the multimodal distributions

of centrality measures are presented in the random network but not in the real networks. Also, the value of the associations on average are interestingly higher in the null network than the PPINs.

Dimension reduction and clustering analysis

In the next step, PCA-based dimensionality reduction was used not only to minimize information loss, but to reveal which centrality measures contain the most relevant information and it can effectively identify important or influential nodes in networks. As illustrated in Fig. 7, the profile of the distance to the center of the plot and their directions are mostly similar except for the homology which is similar to the random network. The contribution values of each centrality measure are shown according to rank order in Table 3, based on their corresponding principal components. A similar pattern of the contribution of centrality measures could be observed in all real networks even in homology networks compared with the random null network (See Supplementary file 3). The Closeness centrality Latora is the main contributor of the principal components on average and it showed that it contains more information compared with the other centrality measures in PPINs. In contrast, other well-known centralities i.e. Betweenness and Eccentricity revealed a low contribution value in all PPINs similar to the null network and lower than random threshold. On the contrary, the Degree displayed moderate levels of contribution in all real networks whilst it is the fourth rank of random network contributors. Although, the overall patterns of contributions are similar, all PPINs exhibited this special fingerprint of centrality importance.

Finally, by performing unsupervised categorization, we aim to cluster centrality measures found on the computed values in PPINs. First, by performing the clustering tendency procedure, we find that this data set is clusterable. The distance metrics formed with Pearson correlation coefficients using the normalized centrality measures can be seen in Fig. 8. Then, we validate the measures to obtain the silhouette scores for different clustering methods such as hierarchical, k-means, and PAM (Supplementary files 4 & 5). Considering these scores, the hierarchical clustering algorithm was applied to categorize the standardized centrality measure results in each PPIN. The output of applying the clustering algorithms and the corresponding number of clusters is shown in Table 4. Using the hierarchical algorithm based on Ward's method [65], which relies on determining the adequate number of clusters, the centrality measures are clustered in each PPINs (Fig. 9). The number of clusters, the distance between centralities and the centrality composition in all 13 PPINs displayed independent outcomes and inconstant behavior of each centrality to rank nodes within networks. For better visualization, Table 5 displays the pairwise Jaccard similarity index values for each network pair. The lowest values are related to the homology, neighborhood-transferred and co-occurrence PPINs and among the genome context prediction methods, fusion PPIN is more associated to the others. The high similarity between co-expression and co-expression-transferred was expected but

the similar centrality clustering of the database with both aforementioned PPINs and combined-score with textmining-transferred are remarkable.

Interestingly, it is also illustrated that silhouette scores (calculated from the centralities in each cluster) are related to the contribution value of the corresponding centrality measure. Where there is a high silhouette value, a high contribution value is observed, however, a high contribution value will not always mean a high silhouette value. Fig. 10 shows the relationship between the silhouette scores and contribution values of each centrality measure in the combined-score PPIN. In panel A, each color represents a cluster of centrality measures. As shown in the figure, Latora, Radiality, Residual, Decay, Lin, Leverage, Freeman and Barycenter centralities have been gathered together in the same cluster where the corresponding silhouette scores are all at a high level except the Leverage's score, while the average silhouette score is around 0.66 in that cluster. On the other hand in panel B of Figure 8, we can find the Leverage's contribution value that is under the threshold line and placed in the group with the least amount of contribution. Centralities named Lobby index, ClusterRank, Laplacian, MNC, Degree, Markov, Diffusion degree, Kleinberg's hub, Eigen vector, Authority score, Katz centralities fell in the same group where all the silhouette scores are more than the average equal to 0.61 and in the same way, their corresponding contribution values are at high levels, too. On the other hand, we observe that Shortest path Betweenness (which is in a separated cluster) and Geodesic k path, Subgraph and DMNC (which are all in one cluster) had silhouette values lower than the average 0.03. The silhouette values influence the contribution values, which are the lowest. In all other PPINs, the same relationship between silhouette scores and contribution values was observed as shown in Supplementary files 3 & 5.

Discussion

Network biology is a major part of the relatively young biological research approach called systems biology. Topological analysis of biological networks is a basic approach for understanding complex behaviors of a cell, which can be classified into three categories, which are collective behaviors, subnetwork behaviors and individual behaviors (prioritizing of important nodes by centrality index). It consists of various computational methods and algorithms in order to infer biological conclusions such as biomarker discovery or drug design and repurposing which are mostly at the stage of introduction or development. However, there is no suitable benchmark for network biologists and this could result in inconsistent and non-reproducible outcomes. Therefore, reconsidering the computational methods towards recommending a standard protocol is needed and inevitable.

One of the unsolved problems in network biology and PPIN analysis is sorting the proteins with respect to their influence based on the centrality-lethality rule [66]. Commonly, high Degree value of a protein in PPIN, i.e., a hub is translated to the extent of its corresponding impact on biological functions [67]. However, dependency of Degree and vitality role of proteins has been riddled with some controversy. In a pioneering work, the authors claimed that the high value of degree centrality in the yeast PPIN is likely to be the main protein for the yeast survival [66]. In another study, this rule was re-examined in three distinct PPINs of three species which confirmed the essentiality of central proteins for survival [11]. Similar results were reported for gene co-expression networks of three different species [68] and for metabolic network of *Escherichia coli* [69, 70]. Ernesto Estrada generalized this rule to six other centrality measures. They showed that the subgraph centrality is the best to find important proteins, and using these measures performed significantly better than a random selection [13]. However, He and Zhang proposed that the relationship between hub nodes and essentiality is not related to network architecture. This finding is explained with the involvement of probability theory [67]. Also, regarding modular structure of PPIN, Joy et al. concluded that Betweenness centrality is more likely to be essential than Degree [71]. The representative power of Betweenness as a topological characteristic was also mentioned in mammalian transcriptional regulatory networks which was clearly correlated to Degree [72]. Recently, it has been shown that existing hub nodes in a network do not have a direct relationship with prognostic genes across cancer types [73].

Some works revisited the centrality notion and focused on repairing this rule by adding new definitions and measures [74]. Tew and Li demonstrated functional centrality and showed that it correlates more strongly than pure topological centrality [75]. Recently, Peng et al. introduced localization-specific centrality and claimed that their measure is more likely to be essential in different species [76]. Khuri and

Wuchty introduced minimum dominating sets of PPIN which are enriched by essential proteins. They described that there is a positive correlation between degree and lethality [10]. One of the proposed solutions is to apply functional methods in this context according to the type of biological networks to be analyzed.

In this study, we worked on updated versions of PPINs which were built on different approaches and as it is shown, while the approaches have different properties, all of them were undirected in a same biological category. In addition to improve our network set compared to previous studies, a large number of centrality measures was used to explore the proximity and preference of the measures. We demonstrated that a similar profile of centrality ranking in PPINs i.e. Latora, Barycenter, Diffusion degree, Freeman, Residual, Average, Radiality centralities plays an important role in detecting the center of PPIN models. We inferred that the rationale and logic of networking in each network dictates which centrality indices should be taken. Also, our results demonstrated the relationship between contribution value derived from PCA and silhouette width as a cluster validity index. We are currently working on producing several built-in functions for various networks to recognize the most appropriate network-dependent centrality measure.

However, from this research, we first reasserted that the architecture and global properties of a network can impact on the central component detection and that the center of network would be differed according to network inherent topology. However, the main step before applying centrality measures for biological inference is identifying their ability to demonstrate the distinction between nodes or edges. In other words, we consider whether a centrality measure has enough information to be used for ranking the network components. Similar to other data mining studies, data reduction and low-dimensional projection help to extract interesting features i.e. centralities and corresponding relationships. Thus, in order to quantify connectivity in biological networks, we recommended that before calculating any centrality measures, an analysis of principal components of these measures should be carried out. The analysis of principal components is necessary to find out which measures have the highest contribution values, i.e., which measures carry the most important information.

Figure legends

Fig. 1. A workflow for studying the centrality measures. This follows the construction of the yeast PPIN relying on different kinds of evidence channels as well as the generation of a null network. The workflow contains a comparison of the behavior of several centrality measures using machine learning methods such as principal components analysis and clustering procedures.

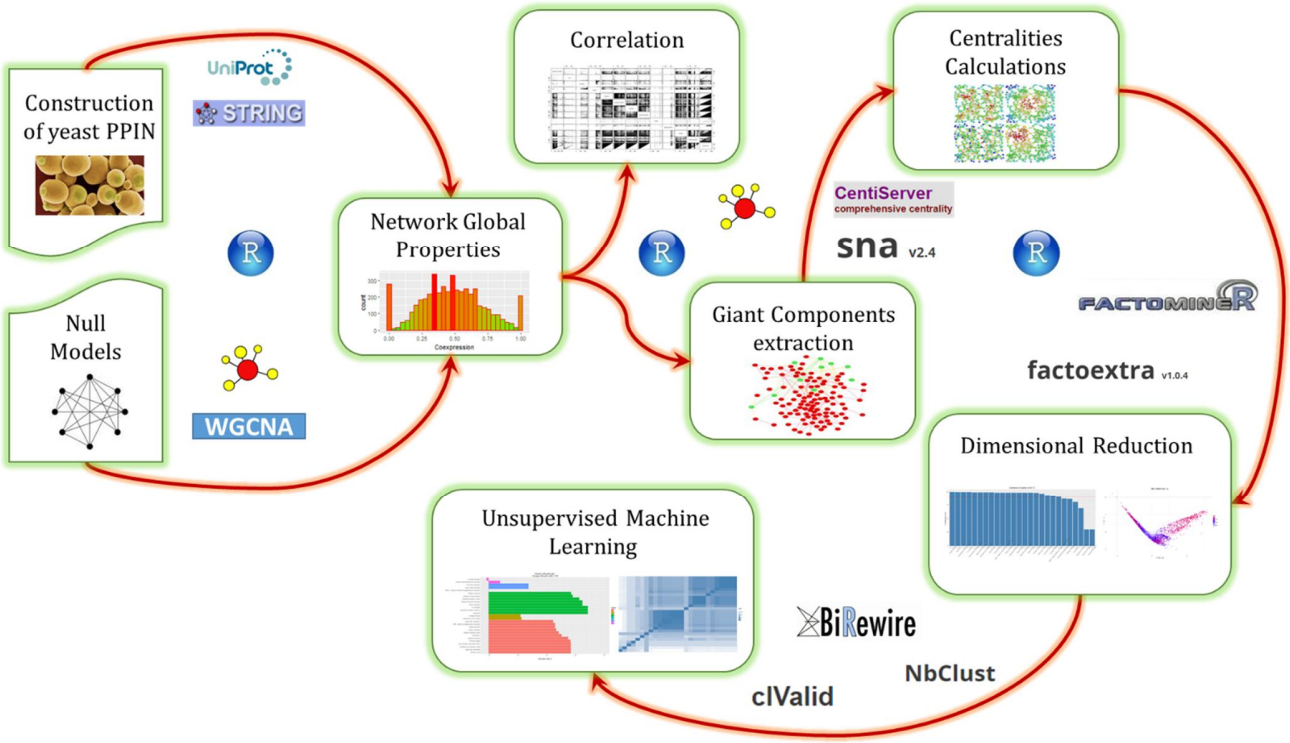

Fig. 2. A graphical representation of the distribution of the evidence channel scores presented in STRING database for the yeast PPIN. The color spectra from green to red indicates low to high values.

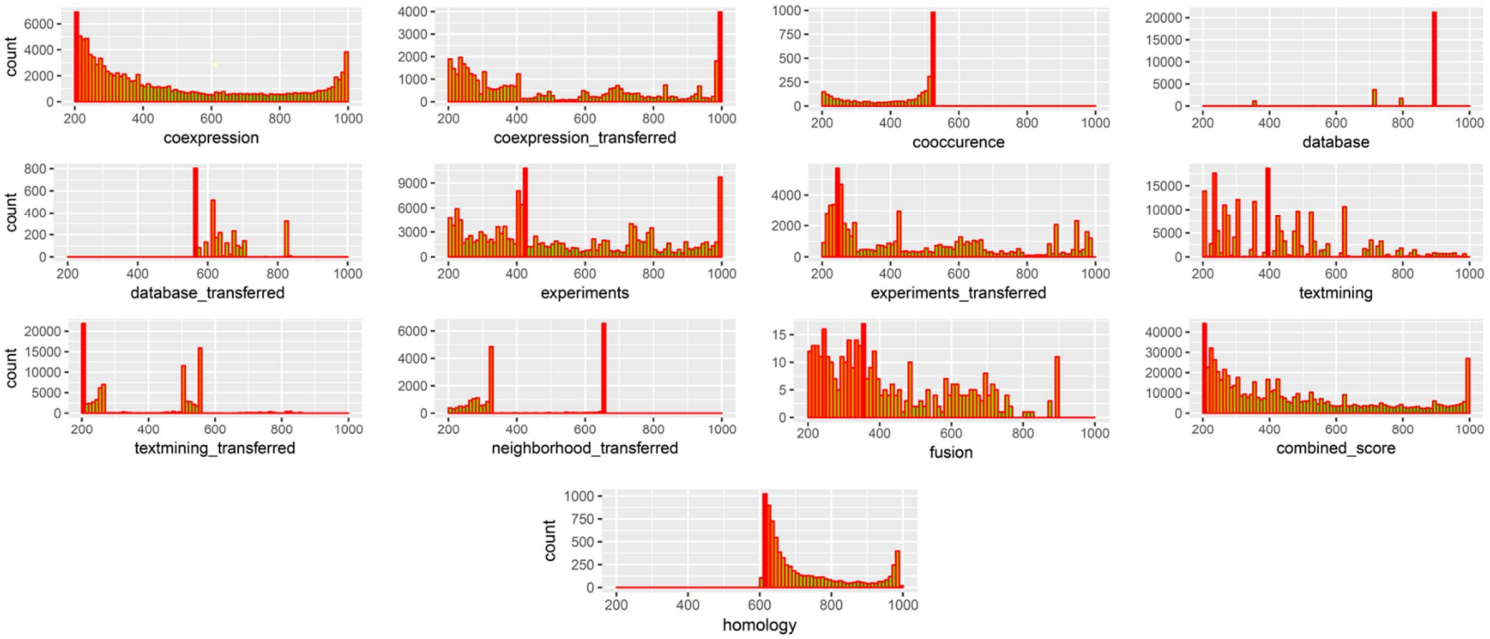

Fig. 3. Pairwise scatterplot between the evidence channel scores. The Pearson's r correlation coefficients are also indicated between the evidence channels. The distributions of scores in each evidence are presented at the diameters of the figure.

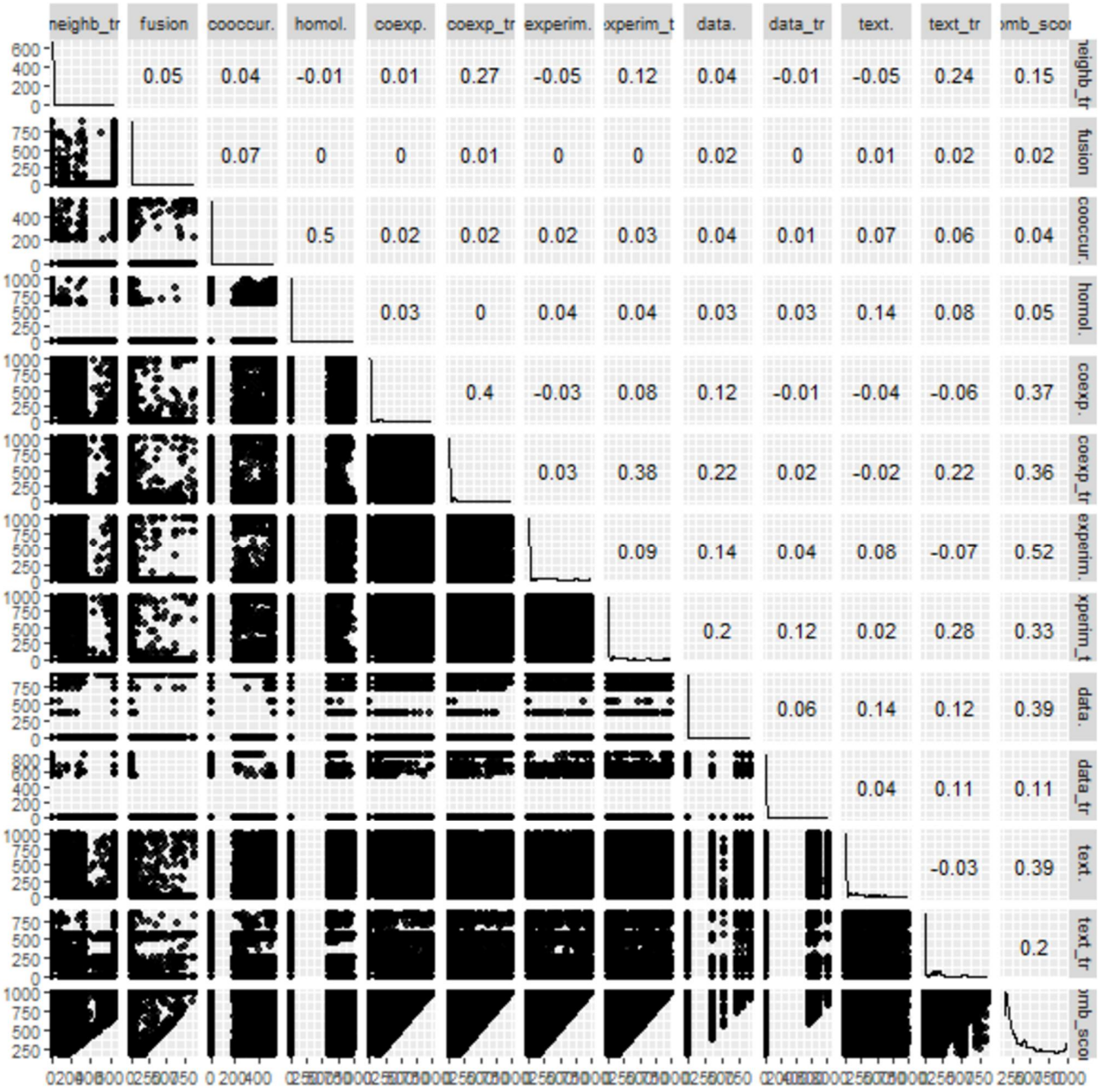

Fig. 4. A graphical representation of the degree distributions in each reconstructed PPIN and the generated null network. The color spectra from green to red indicates low to high values.

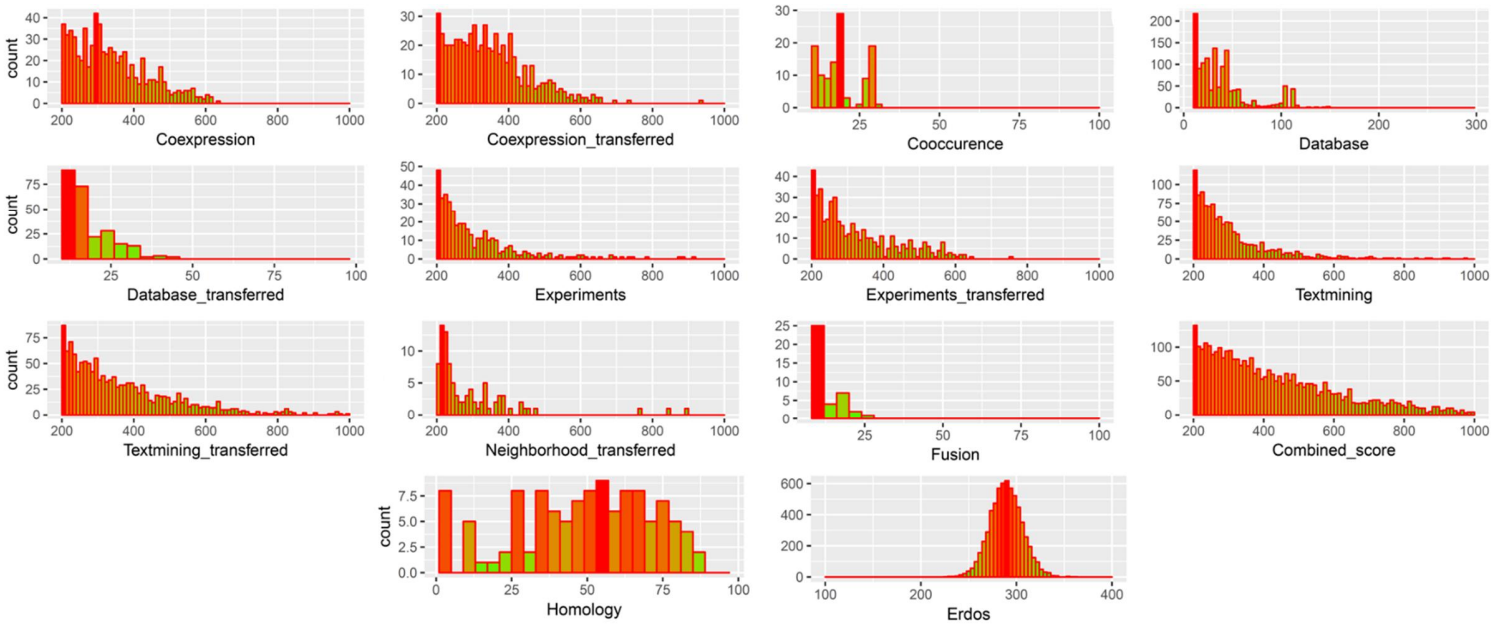

Fig. 5. A graphical representation of the clustering coefficient distribution in each reconstructed PPIN and the generated null network. The color spectra from green to red indicates low to high values.

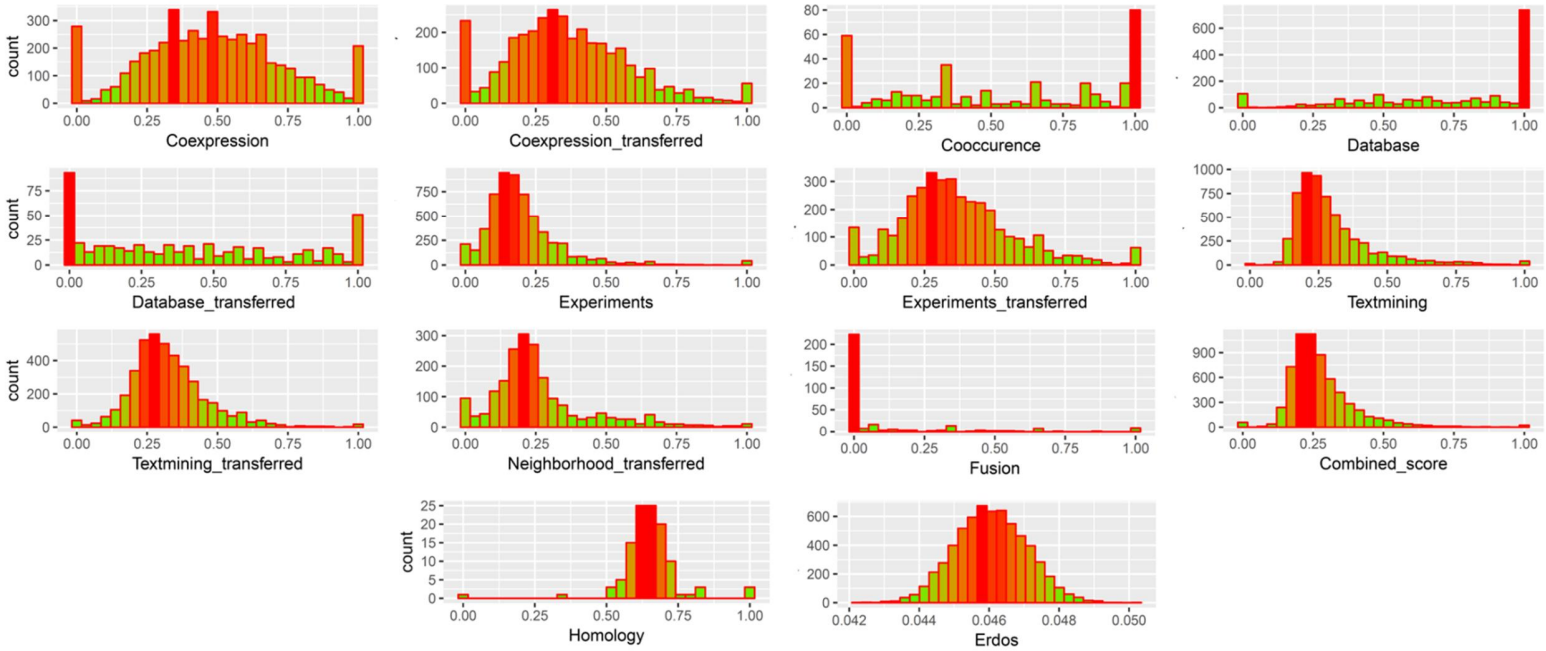

Fig. 6. Pairwise scatterplot between the centrality measures. This figure contains combined-score PPIN and the null network. In this figure, the r Pearson correlation coefficients between centralities beside the centralities distribution are also presented in both networks. For better representation, the scatterplot was divided into three parts corresponding to Table 1 groups. For all networks please refer to the supplementary file 2.

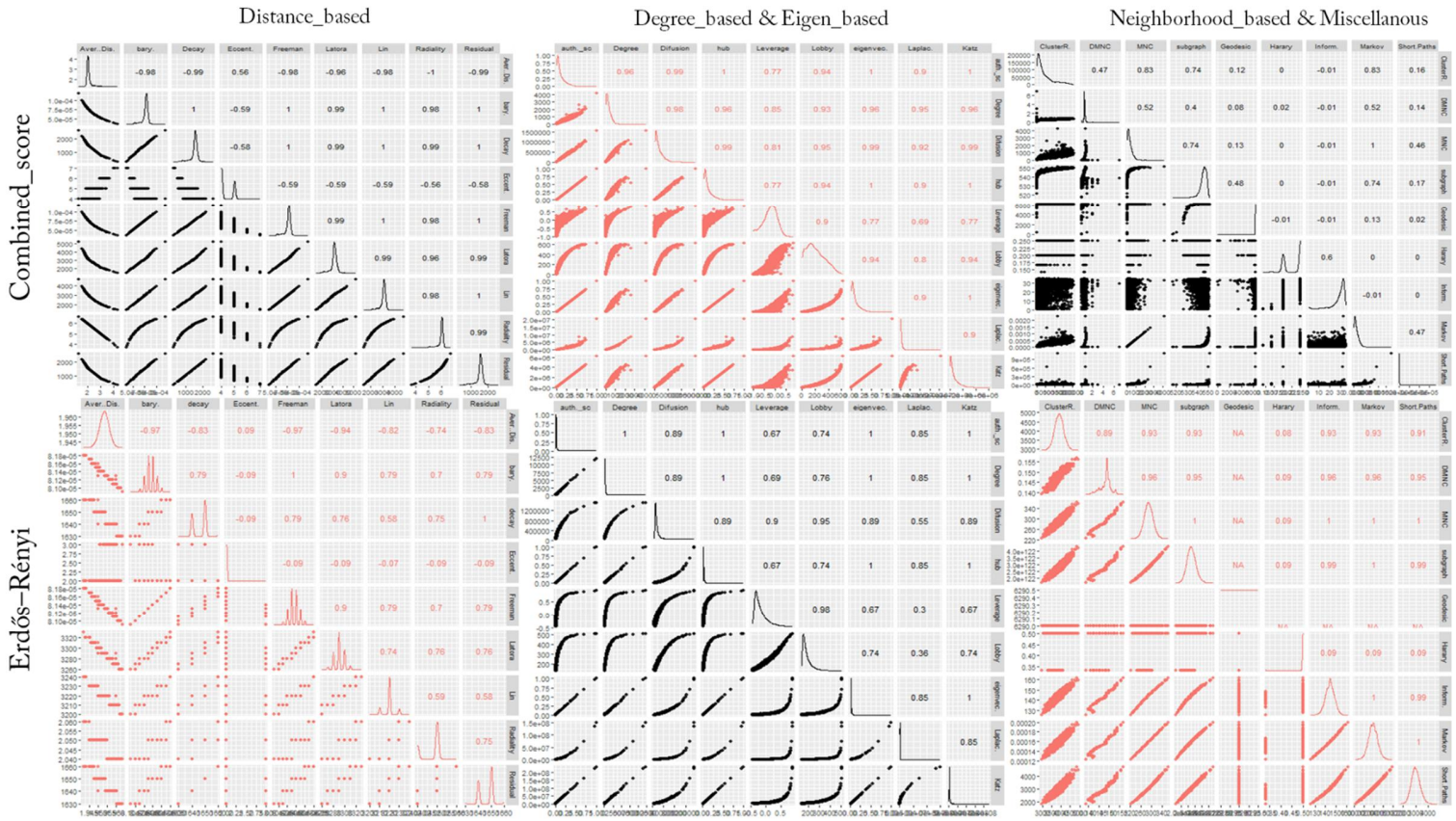

Fig. 7. Biplot representation of the centrality measures in each network. The first two dimensions of multivariate data which can be visualized graphically using PCA. In each plot, nodes are shown as points and centrality measures as vectors.

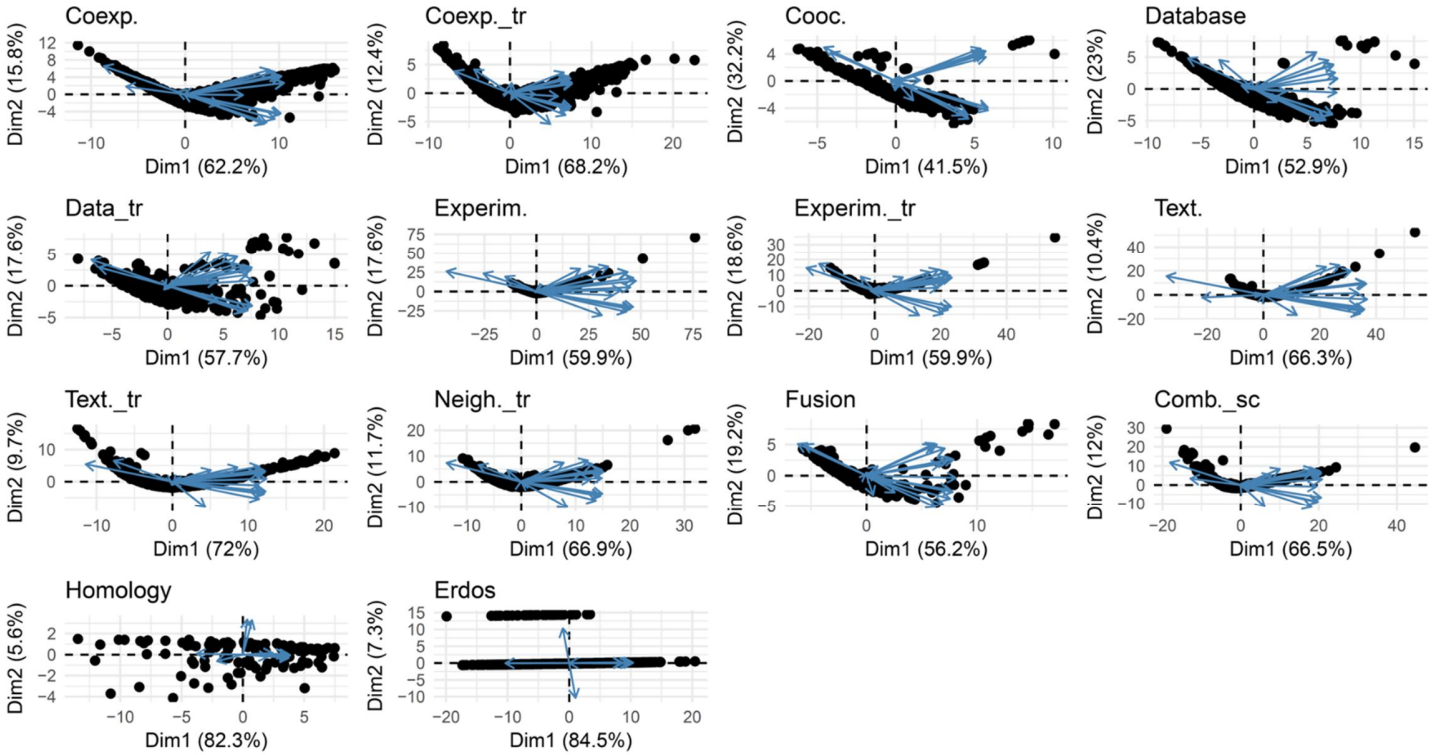

Fig. 8. Distance metrics. The dissimilarities between centrality measures in each PPIN are shown using colored distance measures found on r Pearson correlation coefficients. Colors from blue to red represents distances which vary from 0 to 2.

Fig. 9. The clustering dendrograms. In each dendrogram, the colored boxes show ensued clusters of centrality measures in each PPIN based on predefined distance threshold.

Fig. 10. (A) Clustering silhouette plot of the combined-score PPIN. The colors represent the six clusters of the centrality measures in this PPIN. The average silhouette width was equaled to 0.49. (B) Contribution values of centrality measures according to their corresponding principal components in this PPIN. The number of principal components stand on the network architecture was equal to 3. The dashed line indicates the random threshold of contribution. (C) Line plot between silhouette and contribution values. R value shows the result of regression coefficient analysis and p value has been computed found on Pearson correlation test.

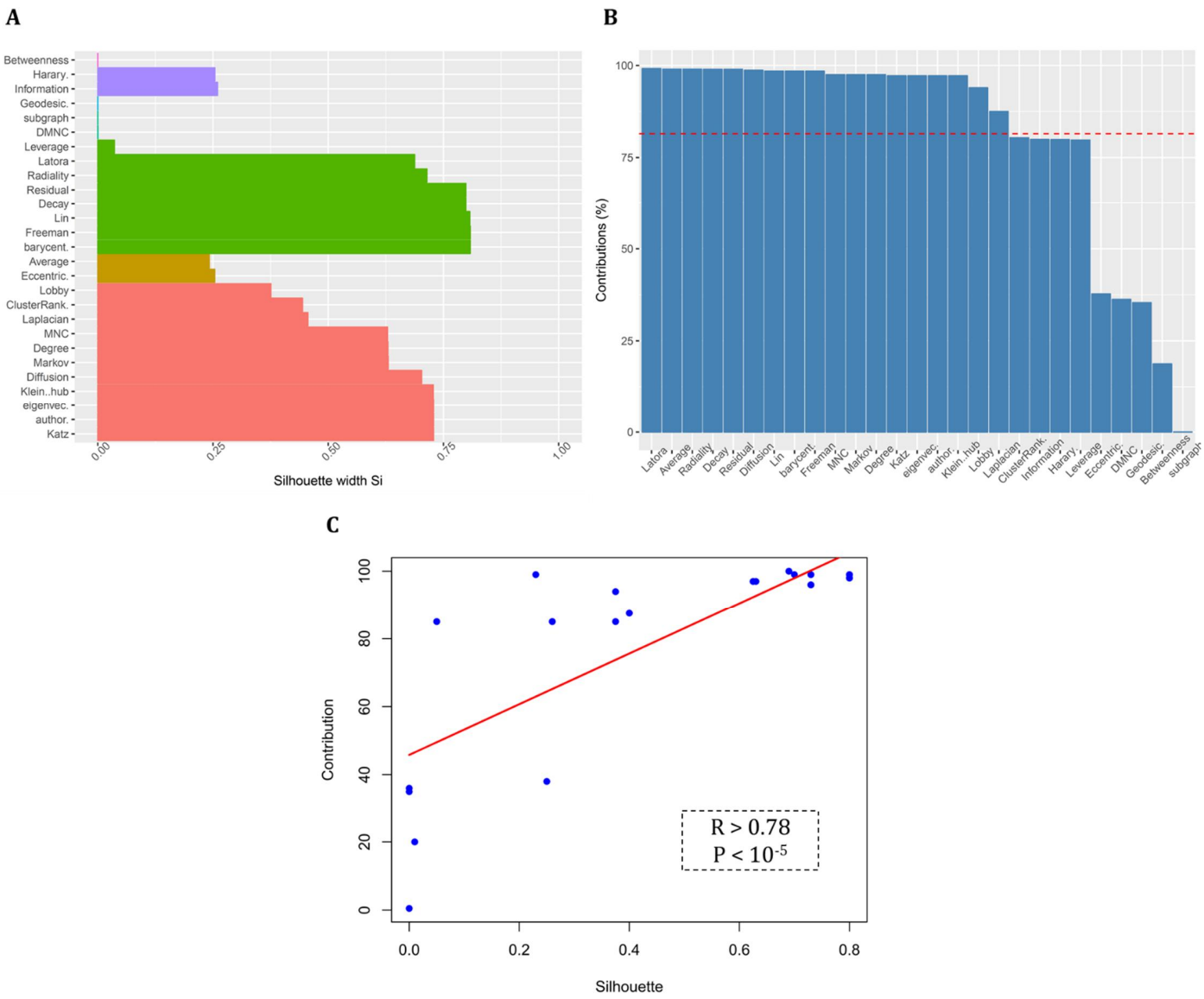

Table legends

Table 1. Centrality measures. These measures represented in five groups depend on their logic and formulas.

Distance_based	Degree-based	Eigen-based	Neighborhood-based	Miscellanous
Average Distance	Authority_score	Eigenvector centralities	ClusterRank	Geodesic K-Path Centrality
Barycenter	Degree Centrality	Katz Centrality (Katz Status Index)	Density of Maximum Neighborhood Component (DMNC)	Harary Graph Centrality
Closeness Centrality (Freeman)	Diffusion Degree	Laplacian Centrality	Maximum Neighborhood Component (MNC)	Information Centrality
Closeness centrality (Latora)	Kleinberg's hub centrality scores		Subgraph centrality scores	Markov Centrality
Decay Centrality	Leverage Centrality			Shortest-Paths Betweenness Centrality
Eccentricity of the vertices	Lobby Index (Centrality)			
Lin Centrality				
Radiality Centrality				
Residual Closeness Centrality				

Table 2. Network global properties of all PPINs and the null network.

Networks / Properties	Nodes	Edges	Connected Components	Nodes/giant component	Edges/Giant Component	Density	Diameter	α value (Power Law)	r value (Power Law)	Average Clustering Coefficent	Heterogeneity	Network Centralization
Homology	2479	7545	648	115	2813	0.43	5	0.01	0.00	0.64	0.47	0.34
Cooccurence	1221	3275	209	425	1653	0.02	22	1.07	0.61	0.47	1.02	0.06
Fusion	1187	1408	222	437	789	0.01	26	1.76	0.91	0.07	1.00	0.05
Database_transferred	622	2975	10	583	2930	0.02	9	1.23	0.69	0.36	0.86	0.06
Neighborhood_transferred	2004	71236	1	2004	71236	0.04	6	0.91	0.72	0.26	1.04	0.41
Database	2496	27766	100	2058	26574	0.01	20	1.18	0.56	0.69	1.07	0.06
Coexpression_transferred	3614	168368	4	3607	168364	0.03	8	0.96	0.79	0.35	1.39	0.26
Experiments_transferred	3870	163403	1	3870	163403	0.02	6	1.03	0.81	0.35	1.59	0.77
Textmining_transferred	4207	364816	1	4207	364816	0.04	6	0.88	0.75	0.32	1.09	0.30
Coexpression	5310	195676	27	5254	195643	0.01	15	1.08	0.82	0.44	1.56	0.11
Textmining	5896	379341	1	5896	379341	0.02	5	1.01	0.66	0.30	0.92	0.35
Experiments	6026	211613	2	6024	211612	0.01	6	1.22	0.82	0.19	1.54	0.56
Combined_score	6294	911414	2	6292	911413	0.05	7	0.76	0.63	0.27	0.90	0.62
Erdős_Rényi	6292	911413	1	6292	911413	0.05	3	_1.095	0.01	0.05	0.06	0.01

Table 3. Ranking of the contribution values based on the PCA for each network. The red to green highlighted cells represent the top to bottom ranked centrality measures in each network. The underlined ranking values are the centrality measure's contribution value which are less than random threshold of contribution.

Centralities / Networks	coexpression_transferred	coexpression	cooccurence	database	database_transferred	experiments	experiments_transferred	fussion	homology	neighborhood_transferred	textmining	textmining_transferred	combined_score	Erdos
Closeness centrality Latora	2	1	1	1	4	3	3	1	1	1	1	1	1	20
Decay Centrality	3	6	9	8	5	5	4	9	2	5	5	5	4	23
barycenter	6	4	3	4	1	15	7	4	13	7	2	9	9	18
Lin Centrality	5	2	4	3	3	16	8	2	12	9	4	7	7	24
Diffusion Degree	1	5	7	2	11	6	8	9	4	6	11	4	6	6
Closeness Centrality Freeman	7	3	2	5	2	14	6	3	14	8	3	8	8	19
Residual Closeness Centrality	4	7	8	9	6	4	5	11	3	4	6	6	5	3
Average Distance	8	14	17	18	13	1	2	14	16	2	7	3	2	2
Radiality Centrality	9	15	18	17	14	2	1	15	17	3	8	2	3	1
Katz Centrality Katz Status Index	10	9	10	13	7	11	15	5	5	12	14	16	13	7
authority_score	11	12	13	11	8	9	14	6	8	13	12	13	14	8
Eigenvector centralities	12	10	11	12	9	12	12	7	6	10	15	14	15	9
Kleinberg s hub centrality scores	13	11	12	10	10	10	13	8	7	11	13	15	16	10
Degree Centrality	16	19	20	15	12	13	11	13	9	15	10	12	12	4
Lapladan Centrality	17	8	6	6	16	18	17	12	15	19	17	18	18	12
Markov Centrality	14	17	23	20	18	8	10	16	11	16	9	11	11	11
Maximum Neighborhood Component (MNC)	15	20	21	16	20	7	16	18	10	14	26	10	10	5
ClusterRank	20	18	5	7	23	23	22	19	19	24	20	19	19	21
Lobby Index Centrality	18	16	15	19	19	19	18	20	18	17	16	17	17	22
subgraph centrality scores	19	21	14	14	17	17	19	16	21	18	18	27	27	13
Geodesick Path Centrality	21	13	22	21	15	21	24	20	23	22	25	23	25	27
Leverage Centrality	23	25	27	22	21	22	23	23	20	23	19	20	22	15
Eccentricity of the vertices	22	24	24	22	22	24	20	22	24	26	24	21	23	26
Shortest Paths Betweenness Centrality	24	26	25	24	24	20	21	25	25	25	21	22	26	16
Harary Graph Centrality	26	23	16	27	26	26	26	27	27	20	23	26	21	25
Information Centrality	27	22	19	26	27	27	27	26	26	21	22	25	20	14
Density of Maximum Neighborhood Component (DMNC)	25	27	26	25	25	25	25	24	22	27	27	25	24	17

Table 4. Clustering information values for PPINs. The Hopkin's statistics threshold for clusterability is 0.05.

Network	Hopkins Statistic	Number of Clusters	Silhouette Average Value
Coexpression	0.25	6	0.36
Coexpression_transferred	0.21	7	0.33
Cooccurence	0.18	6	0.55
Database	0.24	6	0.33
Database_transferred	0.20	9	0.32
Experiments	0.21	9	0.31
Experiments_transferred	0.16	6	0.43
Textmining	0.24	8	0.28
Textmining_transferred	0.20	6	0.35
Neighborhood_transferred	0.26	2	0.39
Fussion	0.16	5	0.48
Combined_score	0.30	7	0.27
Homology	0.23	2	0.46

Table 5. Jaccard index coefficient values for PPINs. The values represent how similar the networks are, in terms of their clustering results. Where a value of 1.00 indicates an exact match, and values equal to 0 show dissimilarity.

	coexp.	coexp_tr	coocc.	comb.	dat_tr	dat.	exp.	exp_tr	fus.	hom.	nei_tr	tex.	tex_tr
coexpression		0.99	0.58	0.77	0.62	1.00	0.58	0.80	0.83	0.41	0.43	0.62	0.76
coexpression_transferred			0.57	0.78	0.63	0.99	0.58	0.81	0.82	0.40	0.43	0.62	0.77
cooccurence				0.47	0.75	0.58	0.44	0.50	0.73	0.29	0.30	0.43	0.48
combined_score					0.52	0.77	0.62	0.63	0.64	0.37	0.39	0.78	0.96
database_transferred						0.62	0.55	0.55	0.55	0.25	0.27	0.47	0.51
database							0.58	0.80	0.83	0.41	0.43	0.62	0.76
experiments								0.59	0.49	0.25	0.27	0.67	0.63
experiments_transferred									0.67	0.41	0.43	0.62	0.62
fussion										0.40	0.42	0.52	0.64
homology											0.91	0.30	0.37
neighborhood_transferred												0.32	0.39
textmining													0.78
textmining_transferred													

Supplementary data

Supplementary file 1: Fitted power law distribution. The degree distribution of each network has been compared to the power law distribution in order to visualize the scale free property in the structure of each network.

Supplementary file 2: Scatterplots between groups of centralities. Each panel indicates scatterplots between centralities groups of two networks.

Supplementary file 3: Contribution values of centralities in each network. These values were computed based on the principal components. The red line shows the threshold used for identifying effective centralities.

Supplementary file 4: Clustering properties results. These properties include connectivity, Dunn and Silhouette scores. These scores suggest the sufficient clustering method by a specific number of clusters.

Supplementary file 5: Clusters silhouette plots. Each color represents a cluster and each bar with specific color indicates a centrality.

Supplementary file 6: Optimal number of clusters. The suitable number of clusters for hierarchical clustering method was computed using the average silhouette values.

Supplementary file 7: Visual assessment of cluster tendency plots. Each rectangular represents the clusters of the calculated results of the centrality measures.

References

1. Giuliani A, Filippi S, Bertolaso M. Why network approach can promote a new way of thinking in biology. Frontiers in Genetics. 2014;5(APR):1-5. doi: 10.3389/fgene.2014.00083.
2. Sobie Ea, Lee Y-S, Jenkins SL, Iyengar R. Systems biology--biomedical modeling. Science signaling. 2011;4(190):tr2-tr. doi: 10.1126/scisignal.2001989.
3. Bruggeman FJ, Westerhoff HV. The nature of systems biology. Trends in microbiology. 2007;15(1):45-50. doi: 10.1016/j.tim.2006.11.003.
4. Jalili M, Salehzadeh-Yazdi A, Gupta S, Wolkenhauer O, Yaghmaie M, Resendis-Antonio O, et al. Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. Frontiers in Physiology. 2016;7(August):375-. doi: 10.3389/fphys.2016.00375.
5. Freeman LC. Going the Wrong Way on a One-Way Street: Centrality in Physics and Biology. Journal of Social Structure. 2008;9(2):1-15.
6. Christakis NA, Fowler JH. The Spread of Obesity in a Large Social Network over 32 Years. New England Journal of Medicine. 2007;357(4):370-9. doi: 10.1056/NEJMsa066082.
7. Freeman LC. Centrality in social networks conceptual clarification. Social networks. 1978;1(3):215-39.
8. Jalili M, Salehzadeh-Yazdi A, Asgari Y, Arab SS, Yaghmaie M, Ghavamzadeh A, et al. CentiServer: A Comprehensive Resource, Web-Based Application and R Package for Centrality Analysis. PloS one. 2015;10(11):e0143111.
9. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411(6833):41-2.
10. Khuri S, Wuchty S. Essentiality and centrality in protein interaction networks revisited. BMC bioinformatics. 2015;16(1):1.
11. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular Biology and Evolution. 2005;22(4):803-6. doi: 10.1093/molbev/msi072.
12. Zotenko E, Mestre J, O'Leary DP, Przytycka TM. 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(8):e1000140.
13. Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics. 2006;6(1):35-40.
14. Park K, Kim D. Localized network centrality and essentiality in the yeast–protein interaction network. Proteomics. 2009;9(22):5143-54.
15. Altaf-Ul-Amin M, Chandra DF, Wijaya SH, Kanaya S. Relation of essentiality and functionality of yeast proteins with their centrality values in a PPI network. Yeast. 2015;32:S64-S. PubMed PMID: WOS:000361466200070.
16. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications. 2014;5:3231-. doi: 10.1038/ncomms4231.
17. Landherr A, Friedl B, Heidemann J. A Critical Review of Centrality Measures in Social Networks. Bus Inform Syst Eng+. 2010;2(6):371-85. doi: 10.1007/s12599-010-0127-3. PubMed PMID: WOS:000285112400005.
18. Kumar A, Mehrotra KG, Mohan CK. Neural Networks for Fast Estimation of Social Network Centrality Measures. Adv Intell Syst. 2015;415:175-84. doi: 10.1007/978-3-319-27212-2_14. PubMed PMID: WOS:000369539900014.
19. Tong JW, Luo JT. Correlation and Stability Study of Centrality in Social Network. Aer Adv Eng Res. 2015;22:664-7. PubMed PMID: WOS:000365403500157.

20. Wu Q, Qi XQ, Fuller E, Zhang CQ. "Follow the Leader": A Centrality Guided Clustering and Its Application to Social Network Analysis. Sci World J. 2013. doi: Artn 368568

10.1155/2013/368568. PubMed PMID: WOS:000326535800001.
21. Buttner K, Scheffler K, Czycholl I, Krieter J. Social network analysis - centrality parameters and individual network positions of agonistic behavior in pigs over three different age levels. Springerplus. 2015;4:185. doi: 10.1186/s40064-015-0963-1. PubMed PMID: 25932371; PubMed Central PMCID: PMCPMC4409614.
22. Koschützki D, Schreiber F, editors. Comparison of Centralities for Biological Networks. German Conference on Bioinformatics; 2004: Citeseer.
23. Dwyer T, Hong S-H, Koschützki D, Schreiber F, Xu K, editors. Visual analysis of network centralities. Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60; 2006: Australian Computer Society, Inc.
24. Valente TW, Coronges K, Lakon C, Costenbader E. How correlated are network centrality measures? Connections (Toronto, Ont). 2008;28(1):16.
25. Batool K, Niazi MA. Towards a Methodology for Validation of Centrality Measures in Complex Networks (vol 9, e90283, 2014). Plos One. 2014;9(5). doi: ARTN e98379

10.1371/journal.pone.0098379. PubMed PMID: WOS:000339614800108.
26. Li C, Li Q, Van Mieghem P, Stanley HE, Wang H. Correlation between centrality metrics and their application to the opinion model. The European Physical Journal B. 2015;88(3):65.
27. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Plant Bioinformatics: Methods and Protocols. 2016:23-54.
28. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Research. 2016:gkw937.
29. Erdos P, Rényi A. On the evolution of random graphs. Publ Math Inst Hung Acad Sci. 1960;5(1):17-60.
30. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006;1695(5):1-9.
31. Butts CT. network: a Package for Managing Relational Data in R. Journal of Statistical Software. 2008;24(2):1-36.
32. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9(1):1.
33. Horvath S. Weighted network analysis: applications in genomics and systems biology: Springer Science & Business Media; 2011.
34. del Rio G, Koschützki D, Coello G. How to identify essential genes from molecular networks? BMC Systems Biology. 2009;3(1):1.
35. Viswanath M. Ontology-based automatic text summarization. University of Georgia. 2009.
36. Latora V, Marchiori M. Efficient behavior of small-world networks. Physical review letters. 2001;87(19):198701.
37. Dangalchev C. Residual closeness in networks. Physica A: Statistical Mechanics and its Applications. 2006;365(2):556-64.
38. Chen D-B, Gao H, Lü L, Zhou T. Identifying influential nodes in large-scale directed networks: the role of clustering. PloS one. 2013;8(10):e77455.
39. Hurajová J, Gago S, Madaras T. Decay Centrality. 15th Conference of Košice Mathematicians; 2.–5. apríla; Herl'any2014.
40. Kundu S, Murthy CA, Pal SK. A New Centrality Measure for Influence Maximization in Social Networks. Lect Notes Comput Sc. 2011;6744:242-7. PubMed PMID: WOS:000306290700040.

41. Lin C-Y, Chin C-H, Wu H-H, Chen S-H, Ho C-W, Ko M-T. Hubba: hub objects analyzer—a framework of interactome hubs identification for network biology. Nucleic acids research. 2008;36(suppl 2):W438-W43.
42. Borgatti SP, Everett MG. A graph-theoretic perspective on centrality. Social networks. 2006;28(4):466-84.
43. De Meo P, Ferrara E, Fiumara G, Ricciardello A. A novel measure of edge centrality in social networks. Knowledge-based systems. 2012;30:136-50.
44. Grassler J, Koschutzki D, Schreiber F. CentiLib: comprehensive analysis and exploration of network centralities. Bioinformatics. 2012;28(8):1178-9. doi: 10.1093/bioinformatics/bts106. PubMed PMID: 22390940.
45. Junker BH, Koschützki D, Schreiber F. Exploration of biological network centralities with CentiBiN. BMC bioinformatics. 2006;7(1):1.
46. Qi X, Fuller E, Wu Q, Wu Y, Zhang C-Q. Laplacian centrality: A new centrality measure for weighted networks. Information Sciences. 2012;194:240-53.
47. Joyce KE, Laurienti PJ, Burdette JH, Hayasaka S. A new measure of centrality for brain networks. PLoS One. 2010;5(8):e12200.
48. Hoffman AN, Stearns TM, Shrader CB. Structure, context, and centrality in interorganizational networks. Journal of Business Research. 1990;20(4):333-47.
49. Korn A, Schubert A, Telcs A. Lobby index in networks. Physica A: Statistical Mechanics and its Applications. 2009;388(11):2221-6.
50. White S, Smyth P, editors. Algorithms for estimating relative importance in networks. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining; 2003: ACM.
51. Bonacich P. Power and centrality: A family of measures. American journal of sociology. 1987:1170-82.
52. Estrada E, Rodriguez-Velazquez JA. Subgraph centrality in complex networks. Physical Review E. 2005;71(5):056103.
53. Hage P, Harary F. Eccentricity and centrality in networks. Social networks. 1995;17(1):57-63.
54. Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM). 1999;46(5):604-32.
55. Stephenson K, Zelen M. Rethinking centrality: Methods and examples. Social Networks. 1989;11(1):1-37.
56. Butts CT. sna: Tools for social network analysis. R package version. 2010;2(2).
57. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(4):433-59.
58. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. Journal of statistical software. 2008;25(1):1-18.
59. Kassambara A. factoextra: Visualization of the outputs of a multivariate analysis. R Package version 1.0. 1. 2015.
60. Maechler MR, Struyf P, Hubert A, Hornik M. K. 2012 cluster: Cluster Analysis Basics and Extensions. R package version. 1(3).
61. Charrad M, Ghazzali N, Boiteau V, Niknafs A, Charrad MM. Package 'NbClust'. J Stat Soft. 2014;61:1-36.
62. Brock G, Pihur V, Datta S, Datta S. clValid, an R package for cluster validation. Journal of Statistical Software (Brock et al, March 2008). 2011.
63. Campello RJ. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters. 2007;28(7):833-41.
64. Gobbi A, Albanese D, Iorio F. Package 'BiRewire'. 2016.

65. Ward Jr JH. Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 1963;58(301):236-44.
66. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411(6833):41-2. doi: Doi 10.1038/35075138. PubMed PMID: WOS:000168432800033.
67. He X, Zhang J. Why do hubs tend to be essential in protein networks? PLoS genetics. 2006;2(6):e88-e. doi: 10.1371/journal.pgen.0020088.
68. Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organisms. PLoS biology. 2004;2(1):E9-E. doi: 10.1371/journal.pbio.0020009.
69. Wagner A, Fell DA. The small world inside large metabolic networks. Proceedings Biological sciences / The Royal Society. 2001;268(1478):1803-10. doi: 10.1098/rspb.2001.1711.
70. Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 2003;19(11):1423-30. doi: 10.1093/bioinformatics/btg177.
71. Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. Journal of Biomedicine and Biotechnology. 2005;2005(2):96-103. doi: 10.1155/JBB.2005.96.
72. Potapov AP, Voss N, Sasse N, Wingender E. Topology of mammalian transcription networks. Genome informatics International Conference on Genome Informatics. 2005;16(2):270-8. doi: 162270 [pii].
73. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature communications. 2014;5.
74. Zhang X, Xiao W, Acencio ML, Lemke N, Wang X. An ensemble framework for identifying essential proteins. BMC Bioinformatics. 2016;17(1):322. doi: 10.1186/s12859-016-1166-7.
75. Tew KL, Li XL, Tan SH. Functional centrality: Detecting lethality of proteins in protein interaction networks. Genome Inform Ser. 2007;19:166-77. PubMed PMID: WOS:000253521200015.
76. Peng X, Wang J, Wang J, Wu FX, Pan Y. Rechecking the Centrality-Lethality Rule in the Scope of Protein Subcellular Localization Interaction Networks. PLoS One. 2015;10(6):e0130743. doi: 10.1371/journal.pone.0130743. PubMed PMID: 26115027; PubMed Central PMCID: PMCPMC4482623.