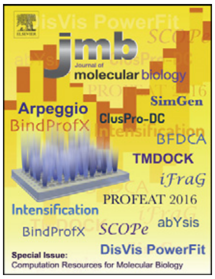

PROFEAT Update: A Protein Features Web Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks

P. Zhang [1,†], L. Tao [2,†], X. Zeng [1], C. Qin [1], S.Y. Chen [1], F. Zhu [3], S.Y. Yang [4], Z.R. Li [5,6], W.P. Chen [7] and Y.Z. Chen [1]

*1 - **Bioinformatics and Drug Design Group,** Department of Pharmacy, National University of Singapore, 117543, Singapore*
*2 - **School of Medicine,** Hangzhou Normal University, Hangzhou 310012, P.R. China*
*3 - **Innovative Drug Research Centre and College of Chemistry and Chemical Engineering,** Chongqing University, Chongqing 401331, P.R. China*
*4 - **State Key Laboratory of Biotherapy and Cancer Center,** West China Hospital, West China Medical School, Sichuan University, Chengdu 610041, P.R. China*
*5 - **College of Chemistry,** Sichuan University, Chengdu 610064, P.R. China*
*6 - **Molecular Medicine Research Center,** State Key Laboratory of Biotherapy, West China Hospital, West China School of Medicine, Sichuan University, Chengdu 610041, P.R. China*
*7 - **Key Lab of Agricultural Products Processing and Quality Control of Nanchang City,** Jiangxi Agricultural University, Nanchang 330045, P.R. China*

***Correspondence to Z.R. Li, W.P. Chen and Y.Z. Chen:** Z. R. Li is to be contacted at: Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, 117543, Singapore. lizerong@scu.edu.cn; iaochen@163.com*

http://dx.doi.org/10.1016/j.jmb.2016.10.013
Edited by Michael Sternberg

Abstract

The studies of biological, disease, and pharmacological networks are facilitated by the systems-level investigations using computational tools. In particular, the network descriptors developed in other disciplines have found increasing applications in the study of the protein, gene regulatory, metabolic, disease, and drug-targeted networks. Facilities are provided by the public web servers for computing network descriptors, but many descriptors are not covered, including those used or useful for biological studies. We upgraded the PROFEAT web server http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi for computing up to 329 network descriptors and protein–protein interaction descriptors. PROFEAT network descriptors comprehensively describe the topological and connectivity characteristics of unweighted (uniform binding constants and molecular levels), edge-weighted (varying binding constants), node-weighted (varying molecular levels), edge-node-weighted (varying binding constants and molecular levels), and directed (oriented processes) networks. The usefulness of the network descriptors is illustrated by the literature-reported studies of the biological networks derived from the genome, interactome, transcriptome, metabolome, and diseasome profiles.

Introduction

The investigations of biological [1–3], disease [4–7] and pharmacological [8–11] processes are facilitated by the study of the relevant protein, gene regulatory, metabolic, and drug-targeted networks. In particular, network descriptors, initially developed for describing the architectures of communication networks, have recently been applied for studying biological networks [7,12]. For instance, the neighborhood connectivity has been applied for measuring the specificity and stability of the protein networks [13]. The clustering coefficient has been used for analyzing the organizational properties of the human protein network [14]. Moreover, some network descriptors have not yet been used but are potentially useful for the analysis of biological networks. For instance, the topological robustness measurement for the social networks [15] is potentially useful for describing the robustness or the alternative signaling capability of biological networks.

Useful facilities have been provided by a number of publicly accessible tools for computing network descriptors. These are Cytoscape [16], NAViGaTOR [17], Gephi [18], VANESA [19], Pajek [20], SpectralNET [21], PINA [22], Hubba [23], GraphWeb [24], tYNA [25] and VisANT [26] for calculating 3–23 network descriptors. Moreover, users with programming expertise may use NetworkX [27], igraph [28] and QuACN [29] for computing no more than 100 network descriptors. Nonetheless, many literature-reported network properties are not covered in these tools (Table 1), some of which are useful for biological studies. For instance, the weighted clustering coefficient has been utilized to predict the gene modules in gene co-expression network [30,31], the interconnectivity has been applied to prioritize the disease-associated genes [32,33], and the PageRank centrality from Google search has been used for analyzing metabolic networks and gene regulatory networks [34,35].

To cater the need for computing a more comprehensive set of network descriptors, we added a new network descriptor module in PROFEAT[‡], previously introduced [36] and updated [37] as a web server for computing the structural and physicochemical descriptors of proteins and protein pairs. The new module provides 177 descriptors (31 node level, 145 network level, and 1 edge level) for an undirected unweighted network (unoriented network with uniform binding constants and molecular levels), 317 descriptors (85 node level, 227 network level, and 5 edge level) for an undirected edge-weighted network (unoriented network with varying binding constants and uniform molecular levels), 189 descriptors (39 node level, 149 network level, and 1 edge level) for an undirected node-weighted network (unoriented network with uniform binding constants and varying molecular levels), 329 descriptors (93 node level, 231 network level, and 5 edge level) for an undirected edge-node-weighted network (unoriented network with varying binding constants and varying molecular levels), and 23 descriptors (11 node level and 12 network level) for a directed unweighted network (oriented process with uniform binding constants and molecular levels; Table 2). Apart from the full set of network descriptors, a subgroup of the network descriptors, which have been extensively used in studying biological networks [13,14] or applied for probing specific biological or therapeutic questions [38], were selected into a slim set of network descriptors. The typical interpretations and biological implications of the network descriptors in the slim set are summarized in Table 3.

Although the network descriptors outside the slim set have been less frequently or are not yet used in

Table 1. The list of supported network types, number of network descriptors, required programming skills, auto-split of multiple networks, and network visualization of PROFEAT and other public tools

Tool name	Number of descriptors	Network types					Auto-detect-&-split multiple networks?	Program skills required?	Network visualization	Graphlet decomposition
		Unweighted	Edge-weighted	Node-weighted	Edgenode-weighted	Directed unweighted				
PROFEAT	up to 329	✓	✓	✓	✓	✓	✓	×	×	×
NetworkX	~100	✓	✓	×	×	✓	×	✓	×	×
igraph	~100	✓	✓	×	×	✓	×	✓	×	✓
QuACN	~100	✓	×	×	×	×	×	✓	×	×
Cytoscape	~23	✓	×	×	×	✓	×	×	✓	×
NAViGaTOR	~13	✓	✓	×	×	×	×	×	✓	×
Gephi	~10	✓	×	×	×	×	×	×	✓	×
VANESA	~10	✓	✓	×	×	✓	×	×	✓	×
Pajek	~9	✓	✓	×	×	×	×	×	✓	×
SpectralNET	~9	✓	✓	✓	✓	×	×	×	✓	×
PINA	~8	✓	×	×	×	×	×	×	✓	×
Hubba	~6	✓	✓	×	×	×	×	×	✓	×
GraphWeb	~4	✓	✓	×	×	✓	×	×	✓	×
tYNA	~4	✓	×	×	×	✓	×	×	✓	×
VisANT	~3	✓	×	×	×	×	×	×	✓	×

Table 2. The numbers of the network descriptors (both the full set and the slim set) for different network types computed by PROFEAT, and the biological representations of each network type

Network type	Biological representations	Full set of network descriptors				Slim set of network descriptors			
		Node level	Network level	Edge level	Total	Node level	Network level	Edge level	Total
Undirected unweighted network	unoriented network with uniform binding constants and uniform molecular levels	31	145	1	177	19	28	1	48
Undirected edge-weighted network	unoriented network with varying binding constants and uniform molecular levels	85	227	5	317	41	44	5	90
Undirected node-weighted network	unoriented network with uniform binding constants and varying molecular levels	39	149	1	189	23	28	1	52
Undirected edgenode-Weighted Network	unoriented network with varying binding constants and varying molecular levels	93	231	5	329	45	44	5	94
Directed unweighted network	oriented network with uniform binding constants and uniform molecular levels	11	12	0	23	5	11	0	16

the study of biological networks, some of them might be possibly useful in describing certain biological network properties not covered by the slim set. In particular, the machine learning method [39,40] and the newly emerged, more advanced deep learning [41] method, which have been used or are starting to be used in biological studies, may be potentially used to classify biological networks in a similar manner as the machine learning classification of protein sequences [39] and to select the marker network descriptors of specific biological processes or populations like the machine learning biomarker selection from the patient gene expression data [40]. These machine learning or deep learning methods typically conduct classifications and marker selections from a large pool of descriptors [39–41], and the noise brought by the high number of descriptors is reduced by means of the established feature selection methods [40]. Given their possible usefulness in biological networks studies and the availability of the methods for using them and for noise reduction, these network descriptors were computed by the PROFEAT to facilitate the relevant studies.

Results

PROFEAT network module structure and access

The PROFEAT procedures for computing the network descriptors are outlined in Supplementary Fig. S1, the network descriptor indices at node/network/edge-level are listed in Table S1–S3, the typical applications of network descriptors in systems biology are summarized in Table S4, the input and output formats/examples of different network types are provided in Table S6–S11, and the detailed descriptor algorithms are given in Supplementary Section F. The PROFEAT network descriptor module consists of five data input fields, each for computing one of the five network types (undirected unweighted, undirected edge-weighted, undirected node-weighted, undirected edge-node-weighted, and directed unweighted networks). Given an input network file, the network descriptors could be computed by uploading the network file in a particular input field and selecting the full set or the slim set of network descriptors, followed by clicking the “Submit” button at the bottom of the input fields. The output file for each inputted network is stored at such URL “http://bidd2.nus.edu.sg/cgi-bin/profeat2016/network/profeat-result.cgi?uid=net-x”, where the numerical “x” is a unique network id for each individual job. For a smaller-sized network input, the output will be immediately displayed on the result window. For a larger-sized network input, users could access the URL later to retrieve their results.

The input network file adapts two formats, simple interaction file (SIF) and nested network (NET), both extensively used for storing biological interaction and network data in databases such as Pathway Commons [42] and in softwares such as Cytoscape [16] Pajek [20]. The SIF format has two interacting nodes (in the first and third columns) in each line and their relationship in the second column, with an optional edge weight in the fourth column. The NET format has three sections, where vertices list all the nodes, edges list all the undirected links between two vertices separated by a white space, and arcs list all the directed links separated by a white space. The output file in a tab-delimited text file format is composed of (1) the header section with each row starting with “!”, followed by the network name, total number of networks, nodes, and edges respectively, (2) the node-level descriptors section with each column showing the descriptor index, name, and value for each node in the network (the node label is provided in the first row), and (3) the network-level descriptors section with each column showing the descriptor index, name, and value. Supplementary Section D is provided for case studies with all types of sample inputs/outputs.

For computing the network descriptors, the following information is required: an undirected unweighted

Table 3. Typical interpretations and biological implications of the slim set of network descriptors

Network Descriptor	Level	Typical interpretation and biological implication
Connectivity/adjacency-based properties		
Degree	Node	Number of interacting partners
Number of selfloops	Node	Number of homodimers formed by two identical molecules
Number of triangles	Node	Number of the smallest unit of molecular interaction clusters
Clustering coefficient	Node	Tendency of each molecule to form groups in the network
Neighborhood connectivity	Node	Indicate if a molecule is near the high-degree hubs of the network
Topological coefficient	Node	Extent of a molecule in sharing its partners in the network
Interconnectivity	Node	How close of a molecule is connected with its neighbors, reflecting the alternative signaling capacity
Bridging coefficient	Node	How well the molecule is linked between high-degree hubs
Degree centrality	Node	Prioritize the molecules by their number of interactions
Number of nodes and edges	Network	Number of molecules and interactions in the biological network
Number of selfloops	Network	Total number of homodimers formed in the network
Maximum/minimum connectivity	Network	The highest/lowest number of interactions for a molecule
Average number of neighbors	Network	The average number of interactions for all molecules
Network density	Network	Efficiency of the information transmitting in the biological network
Average clustering coefficient	Network	Overall tendency of all molecules to form groups in the network
Transitivity	Network	Another measure of tendency of forming groups in the network
Heterogeneity	Network	Reflect the tendency of a biological network to have molecular hubs
Degree centralization	Network	Indicate the biological network is highly connected or decentralized
Shortest path length-based properties		
Average shortest path length	Node	A measure of signal transmission distances or reaction steps from one molecule to all other molecules in the network
Eccentricity	Node	Identify the peripheral or marginal molecules in the network
Radiality	Node	Another indicator for peripheral molecules in the network
Closeness centrality	Node	A measure of how fast the signaling information or reaction spreads from one molecule to all other molecules
Eccentricity centrality	Node	A similar measure as closeness centrality
Load centrality	Node	The extent of a molecule involved in efficient signal transmission
Betweenness centrality	Node	The importance of a molecule to offer efficient alternative signaling
Bridging centrality	Node	A measure of how much information flowing through the molecule
Network diameter	Network	The longest signal transmission or reaction distance in the network
Network radius	Network	The shortest signal transmission or reaction distance in the network
Characteristic path length	Network	The average signal transmission or reaction distance in the network
Average eccentricity	Network	The overall peripherality of all molecules in the network
Global efficiency	Network	The efficiency of information exchange, signaling transmission, or chemical reaction across the biological network
Topological indices		
Hierarchy	Network	Index for power-law distribution of molecular interactions
Robustness	Network	Stability of a biological network for studying diseases and variations
Wiener index, BalabanJ index, Randic connectivity index	Network	Well-known topological properties for molecular characterization
Eigenvector-based complexity indices		
Eigenvector centrality	Node	The iteratively converged importance of a molecule by considering the importance of its interaction partners
Page rank centrality	Node	The iteratively converged importance of a molecule by considering the importance of its interaction partners and its number of partners
Graph energy, Laplacian energy	Network	Well-known eigenvalue-derived properties in mathematical chemistry
Entropy-based complexity indices		
Info content on degree equality	Network	Entropy of probability distribution of the molecular interactions
Radial centric information index	Network	Entropy of probability distribution of the peripheral molecules
Bonchev information index	Network	Entropy of probability distribution of the efficient signaling transmission distances
Edge-weighted properties		
Strength	Node	Indicate if a molecule having strong interactions with its partners
Assortativity	Node	Indicate if a molecule having strong interactions with its partners and also near the high-degree hubs in the network
Edge-weighted interconnectivity	Node	A complexity measure of how close and how strong a molecule is interacting with its partners
Edge-weighted transitivity	Network	A measure of tendency of forming groups in the weighted network
Edge weight	Edge	Interaction kinetic constants, binding affinity, correlation coefficient between molecular levels, interaction confidence score, etc.
Edge betweenness	Edge	Prioritize the important interactions in the biological network, and facilitate the identification of key modules or clusters

Table 3. (*continued*)

Network Descriptor	Level	Typical interpretation and biological implication
Node-weighted properties		
Node weight	Node	Molecular level, expression level, expression fold change, etc.
Node-weighted neighborhood score	Node	Identify the regions with high molecular abundance if the node weight is molecular level, or the regions with high differentially expressed genes if the node weight is gene expression fold change.
Directed properties		
In-degree	Node	The number of molecules that control or regulate a specific molecule
Out-degree	Node	The number of molecules that are controlled or regulated by a specific molecule
Directed local clustering coefficient	Node	Tendency of each molecule to form circulated groups in the network
In-degree (avg, max, min)	Network	The average/highest/lowest number of molecules that control or regulate other molecules in the network
Out-degree (avg, max, min)	Network	The average/highest/lowest number of molecules that are controlled or regulated by other molecules in the network
Directed global clustering coefficient	Network	Overall tendency of all molecules to form circulated groups in the network

network only needs the binary interactions. To compute an undirected edge-weighted network, we need the edge weight, which could be kinetic constant, binding affinity, gene co-expression level, interaction confidence level, or other measurements of the strength of the interacting nodes. Note that the edge length is inversely related to edge weight, as the higher edge weight is typically representing the stronger interaction or the closer relation [43], such that the weighted-distance descriptors are calculated based on the reciprocal of edge weights. The undirected node-weighted network requires an additional node-weighted file, where the node label should be correctly matched to the network file, and the node weight could represent the gene expression level, protein/metabolite level, etc. The undirected edge-node-weighted network requires both the edge-weighted network file and the node-weighted file together for the computation. For all weighted networks, the weight normalization is carried out, such that weighted features will be calculated based on both the original and the normalized weight. Lastly, for a directed unweighted network, the SIF format defines the earlier node points to the latter one, and the NET format lists the directed links in the arc section. Additionally, if there are multiple disconnected networks included in a single input, PROFEAT enables the automatic detection of each connected network, ranks them by size, and computes the network descriptors for each one, respectively.

Discussion

Applications of the network descriptors in studying biological networks derived from the genome, interactome, transcriptome, metabolome, and diseasome profiles

The usefulness of the network descriptors in characterizing the connectivity, organizational, robustness, and stability properties of the biological networks is illustrated in the cases of literature-reported studies of the biological networks built from the genome (e.g., genetic interaction network established by the genome-wide analysis of functionally cooperative double mutants [44]), interactome (e.g., the protein–protein interactions [13], and drug–target interactions [45]), transcriptome (e.g., gene co-expression network based on the exhaustive pairwise profile similarity comparison [46], and gene regulatory network derived from the regulatory interactions between transcription factors and target genes [47]), metabolome (e.g., metabolomics correlation network constructed based on significant correlations among metabolite levels [48]), and diseasome (e.g., the human disease gene network generated from the OMIM (Online Mendelian Inheritance in Man)-based disorder–disease gene associations [9]) profiles, respectively.

A yeast genetic interaction network of ~4000 cooperative gene pairs among ~1000 genes has been constructed by the systematic analysis of functionally cooperative double mutants, which has been subsequently analyzed by using the network descriptor degree (the number of mutant genes cooperative with a mutant gene) to show that the network follows a power-law degree distribution containing many genes with few interactions and a few genes with many interactions, and these few genes are more important for fitness than the less connected genes [49]. In another study of the yeast protein–protein interaction network of 4549 physical interactions among 3278 proteins, based on the analysis of the network descriptor degree (the number of proteins interacting with a protein), it has been found that the links between high degree proteins are systematically suppressed, whereas those between a high and a low degree protein are favored, which decreases the likelihood of crosstalk between different functional modules of the cell and increases the overall robustness of a network [13].

Drug–target networks have been constructed, such that a node represents a drug and that two nodes are connected if they share a common target [45]. In the analysis of a drug–target network derived by docking 1000 FDA-approved drugs to 2500 protein pockets of the human genome [45], three network descriptors degree (the number of drugs sharing the same target with a drug), betweenness (the number of times a drug serves as a linking bridge along the shortest path between two drugs) and a clustering coefficient (the tendency of a drug to form clusters with other drugs in the network) have been used for the comparative analysis of this network with respect to a compound–protein network derived by docking 1592 compounds from the NCI diversity set to 1918 protein pockets, which showed that the drug–target network has a significantly lower degree, comparable betweenness, and slightly lower clustering coefficient, suggesting that the drugs share less number of targets and are more loosely connected than the NCI active compounds. In particular, anticancer drugs are among the drugs with the highest degree and betweenness, and most anticancer compounds are also the most selective compounds in the network.

Based on the exhaustive pairwise gene expression profile similarity comparisons, a yeast gene co-expression network has been constructed and analyzed by using two network descriptors, degree (the number of genes co-expressed with a gene) and clustering coefficient (the level of the clustering of co-expressed genes) [46]. The analysis has indicated that the network follows a clear power-law degree distribution not correlated with the mean expression levels, and the average clustering coefficient of the network is several orders of magnitude greater than that predicted by a pure scale-free growth model, indicative of an underlying hierarchical organization of modularity in the network. The degree descriptor has also been used to derive a co-expressed protein–protein interaction degree and measure as a robust predictor of protein evolutionary rate irrespective of experimental method [50].

In another study [47], the gene regulatory network of the yeast has been constructed from 7074 regulatory interactions between 142 transcription factors and 3420 target genes (interactions can be between transcription factors and non-transcription factor targets or between two transcription factors). The global topological properties of that network have been studied by using four network descriptors in-degree (the number of transcription factors regulating a target), out-degree (the number of target genes for each transcription factor), path length (the number of intermediate regulators between a transcription factor and a terminating target gene), and clustering coefficient (the level of transcription factor inter-regulation). The small in-degrees indicate that transcription factors regulate in simpler combinations, and the large out-degrees imply that each transcription factor has greater regulatory influence by targeting more genes simultaneously. The short paths signify faster propagation of the regulatory signal, while long paths suggest slower action arising from the formation of regulatory chains to control intermediate phases. High clustering coefficients indicate greater inter-regulation between transcription factors. The analysis of the two subnetworks in the endogenous processes (cell cycle and sporulation) and the three subnetworks of the exogenous states (diauxic shift, DNA damage, and stress response) has suggested that these subnetworks have evolved to produce rapid, large-scale responses in exogenous states and carefully coordinated processes in endogenous conditions.

A metabolomic correlation network in *Arabidopsis* has been constructed based on the significant correlations among the metabolite levels in the root tissues and the aerial parts obtained by the gas chromatography–time-of-flight mass spectrometry and published information, respectively [48]. Six network descriptors degree (the number of metabolites significantly correlated to a metabolite), clustering coefficient (the level of the clustering of significantly correlated metabolites), network density (existing metabolite correlations divided by the number of possible correlations), average path length, number of connected components (number of metabolites correlated with another metabolite), and the number of edges (number of metabolite correlations) have been used to assess the threshold-dependent changes in the network topology, which revealed that the network contains tissue- and/or genotype-dependent metabolomics clusters, and some of these clusters are related to the respective biochemical pathways [48].

A human disease gene network of 1284 distinct disorders and 1777 disease-related genes has been generated from the OMIM-based disorder–disease gene associations such that a link is established between a disorder and a disease gene if a mutation in that gene leads to the disorder [51]. The distribution behavior of the drug targets in this network has also been studied [9] by using the network descriptor degree (the number of genes connected to a disorder or the number of disorders connected to a gene), which showed that for both the disorder nodes connected to a drug target and the disease gene nodes encoding a drug target, their average degrees are higher than random cases. Moreover, the distribution of the drug targets in this network exhibits a clustered pattern with the targets primarily enriched in some regions of the network. Specifically, starting from a node in the network, the ratio of drug targets with respect to the distance from the node was measured, which showed a strong enrichment in the first and the second neighbors and thus a bias toward the clustering of drug targets in the network.

Perspectives

Recent investigations have demonstrated the usefulness of network descriptors in facilitating the studies of the genome [44], interactome [13,45], transcriptome [46,47], metabolome [48], and diseasome [9] from the perspectives of biological networks. The progress toward more extensive and more reliable network-based studies of the biological, disease, and therapeutically relevant processes may be constrained by such factors as the incomplete knowledge of the biological networks and regulatory components, limited capability of the currently available network analysis and modeling tools, and the limited computer powers for more in-depth analysis and modeling of the properties and the dynamics of the biological networks. By providing the network descriptor computational facility, PROFEAT complements other resources to collectively provide the information [42], modeling tools [52] and parameters [53] of biological networks. These and the more enhanced ability in generating and analyzing various biological networks from the genome [44], drug–target interaction [45], transcriptome [46,47], metabolome [48], and diseasome [9] profiles will enable more comprehensive and in-depth investigations of the functional roles and the dynamics of the biological networks in regulating biological and cellular systems [1], disease processes [4] and therapeutic actions [8].

Materials and Methods

Network descriptor computational methods

The PROFEAT computed network descriptors are broadly grouped into two groups (local/global). The first group (Supplementary Table S1 and Section F.1) consists of the node-level descriptors that are based on the connectivity/adjacency matrix (degree, selfloop, triangle, and clustering coefficient) and on the shortest-path-length matrix (closeness centrality, betweenness centrality, and eccentricity). These descriptors are illustrated in Fig. 1. Degree deg_i is the number of edges directly linked to the studied node [1]. The Number of Selfloops is the number of edges linking to itself. The Number of Triangles "$tri_i = \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} A_{ij}A_{ik}A_{jk}$" implies the level of segregation, and it is the basis for the global transitivity [43]. Clustering Coefficient is locally defined as "$cluster_i = \frac{2e_i}{deg_i(deg_i-1)}$" and globally defined as "$cluster_G = \frac{1}{N}\sum_{i=1}^{N} cluster_i$", where N is number of nodes, and e_i is the number of links among all neighbors of node i ($e_i = 0$ if $deg_i < 2$) [54]. The global clustering coefficient characterizes the overall tendency of the nodes to form groups or clusters in the network [1]. Closeness Centrality is defined as the reciprocal of the average shortest path length, a measure of information spreading speed from a given node to the other reachable nodes in the network [55]: "$closeness_i = \left(\frac{1}{N}\sum_{j=1}^{N} D_{ij}\right)^{-1}$", where D_{ij} represents the shortest path length between i and j [56]. Betweenness Centrality "$betweenness_i = \frac{\sum_{s \neq i \neq t}\sigma_{st}(i)}{\sigma_{st}}$" indicates the number of times a node is serving as a bridge along the shortest path between any other two nodes, where s and t are different from i; σ_{st} (i) is the number of the shortest paths from s to t passing through i; and σ_{st} is the number of all the shortest paths from s to t [57]. Eccentricity "$eccentricity_i = \max\{D_{ij}\}$" is the largest shortest path length between node i and all the others, identifying the peripheral nodes in the network.

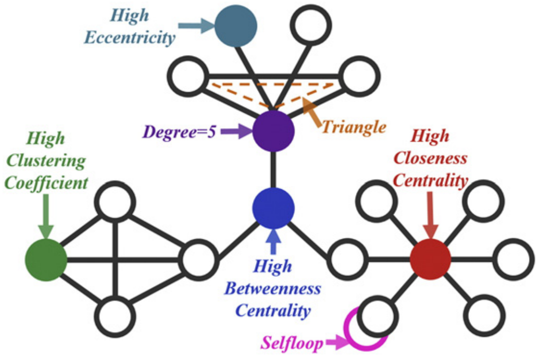
High Eccentricity
Triangle
Degree=5
High Clustering Coefficient
High Closeness Centrality
High Betweenness Centrality
Selfloop

Fig. 1. Graphic illustration of the network descriptors (degree, selfloop, triangle, clustering coefficient, closeness centrality, betweenness centrality, and eccentricity) in a hypothetic network.

The second group (Supplementary Table S2 and Section F.2) is for the network-level features, including the descriptors based on adjacency matrix (degree centralization and heterogeneity), the eigenvalue-based complexity indices (graph energy), and the entropy-based complexity indices (information content of degree equality). Connectivity/Degree Centralization is useful for differentiating the highly connected (e.g., star-shaped) networks from the decentralized networks [58]. Heterogeneity measures the variation of the degree distribution, implying the tendency of a network to have hubs. Biological networks have high heterogeneity as they usually have some central nodes that are highly connected, with the rest having few connections. These two descriptors are computed by calculating first the network density as "$density_G = 2 \cdot E/N(N-1)$" where E is the number of edges; then, the connectivity/degree centralization is defined by "$centralization_G = \frac{N}{N-2}\left(\frac{\max(deg_G)}{N-1} - density_G\right)$" and the heterogeneity is defined by "$heterogeneity_G = \sqrt{N \cdot \sum_{i=1}^{N}\left(deg_i^{\,2}\right) \Big/ \left(\sum_{i=1}^{N} deg_i\right)^2 - 1}$" [59]. The Graph Energy of a network is the summation of all non-zero eigenvalues $\{\lambda_1, \lambda_2 \dots \lambda_k\}$ based on the adjacency matrix "$Energy_G = \sum_{i=1}^{k} |\lambda_i|$" [60]. Information Content of Degree Equality measures the Shannon's entropy of vertex degree distribution "$I_{vertexDegree} = -\sum_{i=1}^{k^d} \frac{N^d{}_i}{N} \cdot log_2\left(\frac{N^d{}_i}{N}\right)$", where $N^d{}_i$ is the number of nodes having the same degree, and k^d is the maximum of degree [61].

To facilitate the studies of biological networks of varying molecular levels and/or binding constants, PROFEAT also provides the edge/node-weighted descriptors. For instances, the edge-weighted clustering coefficient has been applied to the prediction of the gene modules in gene co-expression network [30,31] and is given by "$cluster^{EW}{}_i = \frac{\sum_{j=1}^{N}\sum_{k=1}^{N}\widehat{W_{ij}}\widehat{W_{ik}}\widehat{W_{jk}}}{\left(\sum_{k=1}^{N}\widehat{W_{ij}}\right)^2 - \sum_{k=1}^{N}\widehat{W_{ij}}^2}$" . Node-weighted cross degree and node-weighted local clustering coefficient [62] have been used to analyze the networks with heterogeneous node weights, in the study of Earth's spatial network and international trade network. These descriptors are computed by the following procedure: first, the extended adjacency matrix “$ExtA_{ij} = A_{ij} + \delta_{ij}$” is computed, where A_{ij} is the adjacency matrix, and δ_{ij} is Kronecker's delta constant. The Node-Weighted Cross Degree is defined by “$crossdeg^{NW}{}_i = \sum_{j=1}^{N} ExtA_{ij} \cdot NW_i$”, where the NW_i is the node weight of node i; and the Node-Weighted Local Clustering Coefficient is defined by “$cluster^{NW}{}_i = \frac{1}{crossdeg^{NW}{}_i{}^2} \sum_{j=1}^{N} \sum_{k=1}^{N} ExtA_{ij} \cdot NW_j \cdot ExtA_{ik} \cdot NW_k \cdot ExtA_{jk}$”, which is assumed to zero if the node-weighted cross degree is zero. Directed Local Clustering Coefficient was introduced to measure the brain connectivity, as the neuroconnection is considered as directed [43]. It gives “$cluster^{D}{}_i = \frac{\frac{1}{2}\sum_{j,h \in N}(A_{ij}+A_{ji})(A_{ih}+A_{hi})(A_{jh}+A_{hj})}{(deg_i{}^{+}+deg_i{}^{-})(deg_i{}^{+}+deg_i{}^{-}-1)-2\sum_{j \in N} A_{ij} \cdot A_{ji}}$”, where deg_i^{+} and deg_i^{-} are the in/out-degree of node i.

Comparative evaluation of the PROFEAT computed network descriptors

Performance evaluation of CPU (Central Processing Unit) time was carried on PROFEAT slim set of network descriptors by running 10 different-scaled human-tissue-specific PPI networks of 5 different network types, where the details are given in Supplementary Section E, Table S12, and Fig. S2. PROFEAT computed network descriptor values and the job execution times were evaluated against those from the three popular tools NetworkX, Cytoscape, and Gephi. The evaluated descriptors are those covered by these tools, including degree, number of triangles, closeness/betweenness centrality, local/global clustering coefficient, connectivity centralization, and heterogeneity. These descriptors were computed for three undirected unweighted networks, which are human-tissue-specific PPI networks for hippocampus, muscle, and ovary (with 107, 315, and 1165 nodes and 146, 632, and 2230 edges, respectively). As the CPU times on the public tools cannot be directly obtained, we used the job execution times (from the time of file input to the time of file output, roughly CPU time plus 5 s on PROFEAT) for measuring the time cost in computing and obtaining these descriptors. The comparative results are summarized in Supplementary Table S13–S14. The PROFEAT computed values of all the evaluated descriptors for the three networks are in good agreement with those computed from the popular tools. The slim set of PROFEAT network descriptors was selected, and the job execution times of PROFEAT for the first two networks were faster than those of the public tools (5 s *versus* 10–15 s, and 8 s *versus* 15–20 s), and PROFEAT took comparable time cost to the other tools in the third network (45 s *versus* 30 s). The longer job execution times of PROFEAT arose from its computation of the larger number of network descriptors in contrast to the computation of the smaller set of user-selected descriptors by the other tools.

Acknowledgments

This work supported by Singapore Academic Research Fund [R-148-000-208-112], National Natural Science Foundation of China [81202459], Fundamental Research Funds for the Central Universities [CQDXWL-2012-Z003, CDJZR14-46-88-01], the 973 Program (2013CB967204), and National Natural Science Foundation of China (81325021).

Conflict of interest: The authors declare that they have no conflict of interest.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at doi:10.1016/j.jmb.2016.10.013.

Received 30 June 2016;
Received in revised form 25 September 2016;
Accepted 6 October 2016
Available online 12 October 2016

Keywords:
biological network;
biological pathway;
network descriptor;
protein descriptor;
Web server

†P.Z. and L.T. contributed equally to this work.
‡http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi

Abbreviations used:
SIF, simple interaction file; NET, nested network.

References

[1] A.L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet. 5 (2004) 101–113.

[2] S.H. Yook, Z.N. Oltvai, A.L. Barabasi, Functional and topological characterization of protein interaction networks, Proteomics 4 (2004) 928–942.

[3] A. Ma'ayan, Introduction to network analysis in systems biology, Sci. Signal. 4 (2011) tr5.

[4] A.L. Barabasi, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (2011) 56–68.

[5] D.Y. Cho, Y.A. Kim, T.M. Przytycka, Chapter 5: Network biology approach to complex diseases, PLoS Comput. Biol. 8 (12) (2012) e1002820.

[6] L.I. Furlong, Human diseases through the lens of network biology, Trends Genet. 29 (2013) 150–159.

[7] B. Zhang, Y. Tian, Z. Zhang, Network biology in medicine and beyond, Circ. Cardiovasc. Genet. 7 (2014) 536–547.

[8] A.L. Hopkins, Network pharmacology: the next paradigm in drug discovery, Nat. Chem. Biol. 4 (2008) 682–690.

[9] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, M. Vidal, Drug-target network, Nat. Biotechnol. 25 (2007) 1119–1126.

[10] A. Pujol, R. Mosca, J. Farres, P. Aloy, Unveiling the role of network and systems biology in drug discovery, Trends Pharmacol. Sci. 31 (2010) 115–123.

[11] E. Guney, J. Menche, M. Vidal, A.L. Barabasi, Network-based in silico drug efficacy screening, Nat. Commun. 7 (2016) 10,331.

[12] V. Marx, Cancer: smoother journeys for molecular data, Nat. Methods 12 (2015) 299–302.

[13] S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks, Science 296 (2002) 910–913.

[14] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, W. Birchmeier, H. Lehrach, et al., A human protein–protein interaction network: a resource for annotating the proteome, Cell 122 (2005) 957–968.

[15] M. Piraveenan, S. Uddin, K.S.K. Chung, Measuring robustness of networks under sustained targeted attacks, International Conference on Advances in Social Networks Analysis and Mining, IEEE, Computer Society Washington, DC, USA, IEEE, Istanbul, 2012, ISBN 978-0-7695-4799-2, 38–45.

[16] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.

[17] A. Djebbari, M. Ali, D. Otasek, M. Kotlyar, K. Fortney, S. Wong, et al., NAViGaTOR: large scalable and interactive navigation and analysis of large graphs, Internet Math. 7 (2011) 314–347.

[18] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, 3rd International AAAI Conference on Weblogs and Social Media, San Jose, California, USA, 2009.

[19] C. Brinkrolf, S.J. Janowski, B. Kormeier, M. Lewinski, K. Hippe, D. Borck, et al., VANESA—a software application for the visualization and analysis of networks in system biology applications, J. Integr. Bioinform. 11 (2014) 239.

[20] V. Batagelj, A. Mrvar, Pajek—program for large network analysis, Connections 21 (1998) 47–57.

[21] J.J. Forman, P.A. Clemons, S.L. Schreiber, S.J. Haggarty, SpectralNET—an application for spectral graph analysis and visualization, BMC Bioinformatics 6 (2005) 260.

[22] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T.P. Makela, S. Hautaniemi, Integrated network analysis platform for protein–protein interactions, Nat. Methods 6 (2009) 75–77.

[23] C.Y. Lin, C.H. Chin, H.H. Wu, S.H. Chen, C.W. Ho, M.T. Ko, Hubba: hub objects analyzer—a framework of interactome hubs identification for network biology, Nucleic Acids Res. 36 (2008) W438–W443.

[24] J. Reimand, L. Tooming, H. Peterson, P. Adler, J. Vilo, GraphWeb: mining heterogeneous biological networks for gene modules with functional significance, Nucleic Acids Res. 36 (2008) W452–W459.

[25] K.Y. Yip, H. Yu, P.M. Kim, M. Schultz, M. Gerstein, The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks, Bioinformatics 22 (2006) 2968–2970.

[26] Z. Hu, J. Mellor, J. Wu, T. Yamada, D. Holloway, C. Delisi, VisANT: data-integrating visual framework for biological networks and modules, Nucleic Acids Res. 33 (2005) W352–W357.

[27] A.A. Hagberg, D.A. Schult, P.J. Swart, Exploring network structure, dynamics, and function using networkX, Proceedings of 7th Python in Science Conference, Pasadena, CA, USA 2008, pp. 11–15.

[28] T.N. Gabor Csardi, The igraph software package for complex network research, Int. J. Complex. Syst. 1695 (2006).

[29] L.A. Mueller, K.G. Kugler, A. Dander, A. Graber, M. Dehmer, QuACN: an R package for analyzing complex biological networks quantitatively, Bioinformatics 27 (2011) 140–141.

[30] B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis, Stat. Appl. Genet. Mol. Biol. 4 (2005) (Article17).

[31] J. Saramäki, M. Kivelä, J.P. Onnela, K. Kaski, J. Kertész, Generalizations of the clustering coefficient to weighted complex networks, Phys. Rev. E 75 (2007) 027105.

[32] D. Emig, A. Ivliev, O. Pustovalova, L. Lancashire, S. Bureeva, Y. Nikolsky, et al., Drug target prediction and repositioning using an integrated network-based approach, PLoS One 8 (2013) e60618.

[33] C.L. Hsu, Y.H. Huang, C.T. Hsu, U.C. Yang, Prioritizing disease candidate genes by a gene interconnectedness-based approach, BMC Genomics 12 (2011) S25.

[34] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knosel, et al., Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes, PLoS Comput. Biol. 8 (2012) e1002511.

[35] D. Banky, G. Ivan, V. Grolmusz, Equal opportunity for low-degree network nodes: a PageRank-based method for protein target identification in metabolic graphs, PLoS One 8 (2013) e54204.

[36] Z.R. Li, H.H. Lin, L.Y. Han, L. Jiang, X. Chen, Y.Z. Chen, PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Nucleic Acids Res. 34 (2006) W32–W37.

[37] H.B. Rao, F. Zhu, G.B. Yang, Z.R. Li, Y.Z. Chen, Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Nucleic Acids Res. 39 (2011) W385–W390.

[38] L. Yao, A. Rzhetsky, Quantitative systems-level determinants of human genes targeted by successful drugs, Genome Res. 18 (2008) 206–213.

[39] C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen, SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Res. 31 (2003) 3692–3697.

[40] Z.Q. Tang, L.Y. Han, H.H. Lin, J. Cui, J. Jia, B.C. Low, et al., Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation, Cancer Res. 67 (2007) 9996–10,003.

[41] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (2015) 831–838.

[42] E.G. Cerami, B.E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, et al., Pathway commons, a web resource for biological pathway data, Nucleic Acids Res. 39 (2011) D685–D690.

[43] M. Rubinov, O. Sporns, Complex network measures of brain connectivity: uses and interpretations, NeuroImage 52 (2010) 1059–1069.

[44] A.H. Tong, M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, N. Page, et al., Systematic genetic analysis with ordered arrays of yeast deletion mutants, Science 294 (2001) 2364–2368.

[45] X. Peng, F. Wang, L. Li, K. Bum-Erdene, D. Xu, B. Wang, et al., Exploring a structural protein-drug interactome for new therapeutics in lung cancer, Mol. BioSyst. 10 (2014) 581–591.

[46] S.L. Carter, C.M. Brechbuhler, M. Griffin, A.T. Bond, Gene co-expression network topology provides a framework for molecular characterization of cellular state, Bioinformatics 20 (2004) 2242–2250.

[47] N.M. Luscombe, M.M. Babu, H. Yu, M. Snyder, S.A. Teichmann, M. Gerstein, Genomic analysis of regulatory

network dynamics reveals large topological changes, Nature 431 (2004) 308–312.

[48] A. Fukushima, M. Kusano, H. Redestig, M. Arita, K. Saito, Metabolomic correlation-network modules in *Arabidopsis* based on a graph-clustering approach, BMC Syst. Biol. 5 (2011) 1.

[49] A.H. Tong, G. Lesage, G.D. Bader, H. Ding, H. Xu, X. Xin, et al., Global mapping of the yeast genetic interaction network, Science 303 (2004) 808–813.

[50] K. Pang, C. Cheng, Z. Xuan, H. Sheng, X. Ma, Understanding protein evolutionary rate by integrating gene co-expression with protein interactions, BMC Syst. Biol. 4 (2010) 179.

[51] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabasi, The human disease network, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 8685–8690.

[52] V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, M. Glont, et al., BioModels: ten-year anniversary, Nucleic Acids Res. 43 (2015) D542–D548.

[53] P. Kumar, B.C. Han, Z. Shi, J. Jia, Y.P. Wang, Y.T. Zhang, et al., Update of KDBI: kinetic data of bio-molecular interaction database, Nucleic Acids Res. 37 (2009) D636–D641.

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

[55] M.E.J. Newman, A measure of betweenness centrality based on random walks, Soc. Networks 27 (2003) 39–54.

[56] E.W. Dijkstra, A note on two problems in connexion with graphs, Numer. Math. 1 (1959) 269–271.

[57] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2001) 163–177.

[58] H.W. Ma, A.P. Zeng, The connectivity structure, giant strong component and centrality of metabolic networks, Bioinformatics 19 (2003) 1423–1430.

[59] J. Dong, S. Horvath, Understanding network concepts in modules, BMC Syst. Biol. 1 (2007) 24.

[60] I. Gutman, B. Zhou, Laplacian energy of a graph, Linear Algebra Appl. 414 (2006) 29–37.

[61] D.G. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, John Wiley & Sons Ltd, 1983 264, ISBN-10: 0471900877 ISBN-13: 978-0471900870.

[62] M. Wiedermann, J.F. Donges, J. Heitzig, J. Kurths, Node-weighted interacting network measures improve the representation of real-world complex systems, EPL Europhys. Lett. 102 (2013) 28,007.