OXFORD

Briefings in Bioinformatics, 18(6), 2017, 1057–1070

doi: 10.1093/bib/bbw071
Advance Access Publication Date: 18 August 2016
Paper

A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks

Peng Zhang,* Lin Tao,* Xian Zeng, Chu Qin, Shangying Chen, Feng Zhu, Zerong Li, Yuyang Jiang, Weiping Chen and Yu-Zong Chen

Corresponding authors: Yuyang Jiang, The Ministry-Province Jointly Constructed Base for State Key Lab, Shenzhen Technology and Engineering Lab for Personalized Cancer Diagnostics and Therapeutics, and Shenzhen Kivita Innovative Drug Discovery Institute, Tsinghua University Shenzhen Graduate School, Shenzhen, P.R. China, 518055. Tel.: +86-755-26036087; Fax: +86-755-26036029; Email: jiangyy@sz.tsinghua.edu.cn; Weiping Chen, Key Lab of Agricultural Products Processing and Quality Control of Nanchang City, Jiangxi Agricultural University, Nanchang, P. R. China 330045. Tel.:+86-791-83813420; Fax: +86-791-83813655; E-mail: iaochen@163.com; Yu-Zong Chen, Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Singapore 117543. Tel.:+65-6516 6877; Fax: +65-6774 6756; Email: phacyz@nus.edu.sg
*These authors contributed equally to this work.

Abstract

The genetic, proteomic, disease and pharmacological studies have generated rich data in protein interaction, disease regulation and drug activities useful for systems-level study of the biological, disease and drug therapeutic processes. These studies are facilitated by the established and the emerging computational methods. More recently, the network descriptors developed in other disciplines have become more increasingly used for studying the protein–protein, gene regulation, metabolic, disease networks. There is an inadequate coverage of these useful network features in the public web servers. We therefore introduced upto 313 literature-reported network descriptors in PROFEAT web server, for describing the topological, connectivity and complexity characteristics of undirected unweighted (uniform binding constants and molecular levels), undirected edge-weighted (varying binding constants), undirected node-weighted (varying molecular levels), undirected edge-node-weighted (varying binding constants and molecular levels) and directed unweighted (oriented process) networks. The usefulness of the PROFEAT computed network descriptors is illustrated by their literature-reported applications in studying the protein–protein, gene regulatory, gene co-expression, protein–drug and metabolic networks. PROFEAT is accessible free of charge at http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi.

Key words: network descriptor; network feature; biological network; web server

Peng Zhang is a PhD candidate in BioInformatics and Drug Design Group (BIDD), Department of Pharmacy, National University of Singapore (NUS), Singapore, and currently a visiting scholar in Tsinghua University Shenzhen Graduate School, China. His research interests include bioinformatics, computational biology and network-based protein/gene data mining.
Lin Tao is a postdoc in Tsinghua University Shenzhen Graduate School. His research focuses on bioinformatics and natural product study.
Xian Zeng is a PhD candidate in BIDD, Department of Pharmacy, NUS, Singapore. His research interests lie in pharmacogenomics and drug design.
Chu Qin is a postdoc in BIDD, Department of Pharmacy, NUS, Singapore. Her research interest is computer-aided drug design.
Shangying Chen is a PhD candidate in BIDD, Department of Pharmacy, NUS, Singapore. Her research interests are cheminformatics and drug discovery.
Feng Zhu is a professor in Innovative Drug Research Center, Chongqing University, China.
Zerong Li is a professor in College of Chemistry, Sichuan University, China.
Yuyang Jiang is a Professor in Tsinghua University Shenzhen Graduate School, China.
Weiping Chen is a professor in Key Lab of Agricultural Products Processing and Quality Control of Nanchang City, Jiangxi Agricultural University, China.
Yuzong Chen is a professor in in Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Singapore. His research interests include bioinformatics, computational biology, computer-aided drug design and herbal medicine.
Submitted: 15 April 2016; **Received (in revised form):** 14 June 2016

Introduction

Quantitative analysis of biological networks is needed for more extensive study of biological [1], disease [2] and pharmacological [3] processes. These analyses can be facilitated by the knowledge of the network descriptors that characterize the connectivity, organizational, robustness and stability properties of the biological [1, 4, 5], disease [2, 6–8] and drug-targeted networks [3, 9–11]. A number of network descriptors have initially been developed for describing network architectures and communication phenomena in such areas as sociology and communications. For instance, the centrality indices and betweenness represent the degree of centralization of a network and the points of influence in the network of sociological studies [12]; the clustering coefficient models the structural trends and relations in interpersonal relationships [13]; and the compactness and stress indices characterize the internal structure of a communication network [14].

Some of these network descriptors have been applied for studying biological, disease and drug-targeted networks, which to a large extent share the same architectural features with other complex networks [1]. So far, the established network theory or graph theory, from the field of mathematics and computer science, has facilitated to reveal enrichment patterns, systematic understandings, high-level relationships and network-based clues in biological networks [8, 15]. For instance, the betweenness centrality has been used for the modularity analysis of interaction information in a liver central carbon metabolism network [16] and for the assessment of protein druggability based on the profiles of the drug targets in the human protein network [17]. The clustering coefficient and topological coefficient have been used for analyzing the organizational properties of the human protein network [18]. The neighborhood connectivity has been applied for measuring the specificity and stability of the protein networks [19]. Nonetheless, a substantial number of the network descriptors have not yet been used but are potentially useful for the analysis of a more variety of features of biological networks. For instance, the geographical indices for transportation systems [20] are potentially useful for describing the spatial and structural features of a biological network, and the topological robustness measurement for the social networks [21] can be potentially used for measuring the robustness or the alternative signaling capability of biological networks.

A number of resources are available for computing network descriptors, particularly Cytoscape [22], NAViGaTOR [23], Gephi [24], VANESA [25], Pajek [26], SpectralNET [27], PINA [28], Hubba [29], GraphWeb [30], tYNA [31] and VisANT [32] that enable the computation of approximately 23, 13, 10, 10, 9, 9, 8, 6, 4, 4 and 3 network descriptors, respectively (Table 1). Moreover, users knowledgeable of the respective programming languages can use Python library NetworkX [33], R packages igraph [34] and QuACN [35] for computing ~100 network properties. However, these programming-based tools are hardly applicable for the users without computation expertise [36]. Compared with the literature-reported network descriptors (Tables 2 and 3), these resources cover a limited number of network descriptors, and some of the uncovered network descriptors have been applied in systems biology studies. For instance, the PageRank centrality from Google search algorithm has been used for analyzing the metabolic networks and gene regulatory networks [51, 52, 68]; the interconnectivity has been applied to prioritize the disease-associated genes [41–43], and the weighted clustering coefficient has been used to predict the significant genes in gene co-expression network [56, 69].

Table 1. List of the network descriptors provided by the existing publically accessible tools that do not require programing skill

Tool name (no. of provided descriptors) [Ref]	List of descriptors	
	Node level	Network level
Cytoscape (23) [22]	Degree, in/out-degree, number of self-loops, clustering coefficient, topological coefficient, neighborhood connectivity, avg shortest path length, eccentricity, radiality, closeness centrality, betweenness centrality, stress	Number of nodes/edges/self-loops, density, diameter, radius, centralization, heterogeneity, avg number of neighbors, characteristic path length
NAViGaTOR (13) [23]	Clustering coefficient, degree centrality, betweenness centrality	Number of nodes/edges, density, min/avg/max degree, diameter, avg clustering coefficient, characteristic path length
Gephi (10) [24]	Degree, clustering coefficient, betweenness centrality, closeness centrality, eigenvector centrality, PageRank centrality, HITS	Diameter, density, avg clustering coefficient, avg shortest path length
VANESA (10) [25]	Degree, avg/max shortest path length	Min/avg/max degree, avg shortest path length, density, centralization, clustering coefficient
Pajek (9) [26]	Degree, avg shortest path length, degree centrality, closeness centrality, betweenness centrality	Diameter, degree centralization, closeness centralization, betweenness centralization
SpectralNET (9) [27]	Degree, clustering coefficient, min/avg/max shortest path length	Number of nodes, diameter, avg clustering coefficient, avg shortest path length
PINA (8) [28]	Degree, shortest path length, clustering coefficient, closeness centrality, betweenness centrality, degree centrality, eigenvector centrality	Diameter
Hubba (6) [29]	Degree, bottleneck, subgraph centrality, edge percolation component, max neighborhood component, density of max neighborhood	N.A.
GraphWeb (4) [30]	Betweenness centrality	Number of nodes/edges, density
tYNA (4) [31]	Degree, clustering coefficient, eccentricity, betweenness centrality	N.A.
VisANT (3) [32]	Degree, shortest path length, clustering coefficient	N.A.

Table 2. List of the node-level descriptors provided by PROFEAT and their typical applications in systems biology studies

Descriptor group	Node-level descriptors	Typical applications in systems biology
Connectivity to immediate neighbors Connectivity to next immediate neighbors	Degree, scaled connectivity, number of self-loops, number of triangles, Z score Clustering coefficient, neighborhood connectivity, topological coefficient, interconnectivity, bridging coefficient	Clustering coefficient used to illustrate the hierarchical architecture of metabolism [1, 37], identify the functional modules from genomic associations [38], and predict protein function by network-based methods [39].Degree and clustering coefficient used to validate if the experimental drugs are more associated with existing proteins in the drug–target network [9], and predict candidate genes in coronary artery disease [40].Topological coefficient and clustering coefficient used to identify high-confidence interactions in a large-scale PPI network [18].Neighborhood connectivity applied for measuring the specificity and stability of protein networks [19].Interconnectivity applied to prioritize the disease-associated genes in drug target discovery [41–43].
Distance relationships to all other nodes	Average shortest path length, distance sum, eccentricity, eccentric, deviation, distance deviation, radiality	Eccentricity and distance deviation used to prioritize the metabolic biomarkers in obesity [44].Radiality used to analyze gene regulatory networks [45].
Centrality measure based on distance to all other nodes	Closeness centrality (avg, sum) eccentricity centrality, harmonic centrality, residual centrality	Betweenness centrality, degree centrality, bridging centrality and other centrality measures used to expose the relationship between network topology and system function of proteins [3, 40, 46–48], classify the important nodes in drug discovery [49] and understand genes implicated in disease [50].
Centrality measure based on shortest paths passing thru the studied node	Stress centrality, betweenness centrality, normalized betweenness, bridging centrality	
Centrality measure based on degree or/and neighbors' centrality	Degree centrality, page rank centrality, eigenvector centrality	PageRank centrality used to identify protein target in metabolic networks [51], identify candidate marker genes for prognostic prediction of pancreatic cancer patients [52].Eigenvector centrality, together with other centralities, were applied to predict the synthetic genetic interactions [53, 54].
Edge-weighted descriptor	Strength, assortativity, disparity, geometric mean of triangles, edge-weighted local clustering coeff (Barrat's, Onnela's, Zhang's, Holme's)	The strength of the associations between genes was used as the edge weight in gene co-expression analysis [55].Edge-weighted clustering coefficient was used to predict the significant genes in gene co-expression network [56], or a general biology system [57].
Node-weighted descriptor	Node weight, node-weighted cross degree, node-weighted local clustering coeff.	
Directed and unweighted descriptor	In-degree, out-degree, directed local clustering coefficient, neighborhood connectivity (only-in), neighborhood connectivity (only-out), neighborhood connectivity (in and out), average directed neighbor degree	In/out-degree have been applied for five directed biological networks, to identify and ranks the regulators in the networks [58].

There is a need for the relevant web servers to provide more comprehensive coverage and more user-friendly service in computing the network descriptors. We therefore introduced a new network descriptor module in PROFEAT (http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi), which was previously introduced [70] and updated [71] as a web server for computing the structural and physicochemical descriptors of proteins, peptides and protein–protein interaction (PPI) pairs. The new module supports the computation of 173 descriptors (29 node level, 144 network level) for an undirected unweighted network (un-oriented network with uniform binding constants and uniform molecular levels), 303 descriptors (77 node level, 226 network level) for an undirected edge-weighted network (un-oriented network with varying binding constants and uniform molecular levels), 183 descriptors (35 node level, 148 network level) for an undirected node-weighted network (un-oriented network with uniform binding constants and varying molecular levels), 313 descriptors (83 node level, 230 network level) for an undirected edge-node-weighted network (un-oriented network with varying binding constants and varying molecular levels) and 20 descriptors (9 node level, 11 network level) for a directed unweighted network (oriented network with uniform binding constants and uniform molecular levels). The codes for computing these descriptors were tested against the PPI data sets [72–74], and the performance evaluation was conducted. Table 4 summarizes the number of network descriptors, the list of supported network types, the network visualization and some other features of PROFEAT in comparison with the other 14 publically accessible tools. These network descriptors were fully documented in Supplementary Section A and Section E, and their applications in systems biology networks were summarized in Tables 2 and 3.

Materials and methods

The PROFEAT computed network descriptors are broadly grouped into two groups. The first group (Table 2, Supplementary Table S1 and Section E.1) consists of the node-level descriptors that represent the connectivity profiles to the immediate neighbors (degree and triangle) and the next immediate neighbors (clustering coefficient), and the centrality measure based on the distance to all the other nodes (closeness centrality) and the number of shortest paths passing through the studied node (betweenness centrality).

Table 3. List of the network-level descriptors provided by PROFEAT and their typical applications in systems biology studies

Descriptor group	Network-level descriptors	Typical applications in systems biology
Global connectivity profiles	Number of nodes/edges/self-loops, max/min connectivity, avg number of neighbors, total adjacency, network density, average clustering coefficient, transitivity, heterogeneity, degree centralization, central point dominance	Density, heterogeneity, degree centralization and global clustering coefficient used to compare and study the PPI networks between drosophila and yeast [59].
Network measure based on shortest paths	Total distance, diameter, radius, shape coefficient, characterisitc path length, network eccentricity, avg eccentricity, network eccentric, eccentric connectivity, unipolarity, integration, variation, avg distance, mean distance deviation, centralization, global efficiency	Characteristic path length and global efficiency used to describe the brain neuro-connectivity network [60].
Topological index based on connectivity	Edge complexity index, randic connectivity index, atom-bond connectivity index, Zagreb index (1, 2, modified, augmented, variable), Narumi index, Narumi geometric index, Narumi harmonic index, alpha index, beta index, pi index, eta index, hierarchy, robustness, medium articulation	Randic connectivity index and Zagreb indices applied to access the complexity in chemistry and biology [61, 62].Medium articulation evaluated for measuring the graph features of PPI, genetic interaction, and metabolic networks [63].
Topological index based on shortest paths	Complexity index (A, B), Wiener index, hyper-wiener, Harary index (1, 2), Compactness index, Superpendentic index, Hyper-distance-path index, BalabanJ index, BalabanJ-like indices (1, 2, 3), Geometric arithmetic indices (1, 2, 3), product of row sums, Topological index (Schultz, Gutman), Szeged index, efficiency complexity	Wiener index, BalabanJ index and Graph complexity index used to access the complexity in chemistry and biology [61, 62].BalabanJ index used to classify the metabolic networks from three domains of life [64].Efficiency complexity assessed for measuring the graph features of PPI, genetic interaction and metabolic networks [63].
Entropy-based complexity	Shannon's entropy-derived information content of (degree equality/edge equality/edge magnitude/distance degree/distance degree equality), radial centric information index, distance degree compactness, distance degree centric index, graph distance complexity, information layer index, Bonchev information index (1, 2, 3), Balaban-like information index (1, 2)	Radial centric information index used to classify the metabolic networks from three domains of life [64].Bonchev indices, and some other entropy-based measures evaluated for possible use in areas of biology and chemistry [65, 66].
Eigenvalue-based complexity	Graph energy, laplacian energy, spectral radius, Estrada index, Laplacian Estrada index, Quasi-Weiner index, Mohar index (1, 2), graph index complexity, 50 Dehmer's eigenvalue properties based on matrices of (adjacency/Laplacian/distance/distance path/augmented vertex degree/extended adjacency/vertex connectivity/random walk Markov/weighted structural function 1/weighted structural function 2)	Dehmer proposed a set of 50 eigenvalue-based descriptors, which possess high discriminative power to capture structural information of graphs, to predict biological and pharmacological properties [67].Graph index complexity was discussed in measuring the graph features of real-world systems, including PPI network, genetic interaction network and metabolic network [63].
Edge-weighted descriptors	Weighted transitivity, edge-weighted global clustering coeff (Barrat's, Onnela's, Zhang's, Holme's)	Weighted transitivity used to describe the brain neuro-connectivity network [60].
Node-weighted descriptor	Total node weight, node-weighted global clustering coeff	
Directed and unweighted descriptor	In-degree (max, avg, min), out-degree (max, avg, min), directed global clustering coefficient	In/out-degree applied for directed biological networks, to identify and rank the regulators in the networks [58].

These descriptors are illustrated in Figure 1. In detail, **Degree** deg_i is the number of nodes directly connected to the studied node [1]. Number of **Triangle** '$tri_i = \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} A_{ij} A_{ik} A_{jk}$'implies the level of segregation, and it is the basis for measuring the network transitivity [60]. **Clustering Coefficient** $cluster_i$ of node i is locally defined as '$cluster_i = \frac{2e_i}{deg_i(deg_i-1)}$'and the global clustering coefficient $cluster_G$ is '$cluster_G = \frac{1}{N}\sum_{i=1}^{N} cluster_i$', where N is number of nodes, e_i is the number of connected pairs among all neighbors of node i and $e_i = 0$ if a node has less than two neighbors [75]. The global average clustering coefficient characterizes the overall tendency of the nodes to form groups or clusters in the network [1]. **Closeness Centrality** of a node is defined as the reciprocal of its average shortest path length in the graph, which is a measure of how fast information spreads from a given node to the other reachable nodes in the network [76]. '$centralityCloseness_i = \frac{1}{\frac{1}{N}\sum_{j=1}^{N} D_{ij}}$', where D_{ij} represents the shortest path length between node i and node j, computed based on the Dijkstra's algorithm [77]. **Betweenness Centrality** '$centralityBetweenness_i = \frac{\sum_{s\neq i\neq t}\sigma_{st}(i)}{\sigma_{st}}$' indicates the number of times a node i serves as a linking bridge along the shortest path between two nodes s and t, where node s and node t are different from node i in the network, $\sigma_{st}(i)$ is the number of shortest paths from s to t passing through i, and σ_{st} is the number of shortest paths from s to t [16, 78].

The second group (Table 3, Supplementary Table S2 and Section E.2) includes network-level features, including the network connectivity descriptors (degree centralization and heterogeneity) [59], the eigenvalue-based network complexity descriptor (graph energy) [79] and the entropy-based network complexity descriptor (information content of degree equality) [80]. **Connectivity Centralization** (or namely degree centralization) is useful for

Table 4. The number of network descriptors, the list of network types and visualization features of PROFEAT and other publically accessible tools

Tool name	Number of descriptors	Network Types					Auto-split multiple networks?	Program skills required?	Network visual interface
		Unweighted	Edge weighted	Node weighted	EdgeNode weighted	Directed			
PROFEAT	up to 313	✓	✓	✓	✓	✓	✓	x	x
NetworkX [33]	~100	✓	✓	x	x	✓	x	✓	x
igraph [34]	~100	✓	✓	x	x	✓	x	✓	x
QuACN [35]	~100	✓	x	x	x	x	x	✓	x
Cytoscape [22]	~23	✓	x	x	x	✓	x	x	✓
NAViGaTOR [23]	~13	✓	✓	x	x	x	x	x	✓
Gephi [24]	~10	✓	x	x	x	x	x	x	✓
VANESA [25]	~10	✓	✓	x	x	✓	x	x	✓
Pajek [26]	~9	✓	✓	x	x	x	x	x	✓
SpectralNET [27]	~9	✓	✓	✓	✓	x	x	x	✓
PINA [28]	~8	✓	x	x	x	x	x	x	✓
Hubba [29]	~6	✓	✓	x	x	x	x	x	✓
GraphWeb [30]	~4	✓	✓	x	x	✓	x	x	✓
tYNA [31]	~4	✓	x	x	x	✓	x	x	✓
VisANT [32]	~3	✓	x	x	x	x	x	x	✓

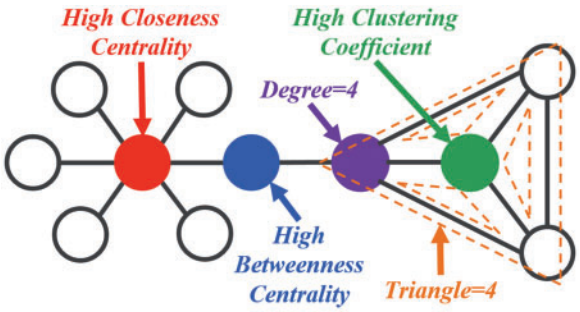


Figure 1. Graphic illustration of the network descriptors degree, triangle, clustering coefficient, closeness centrality and betweenness centrality in a hypothetic network. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

distinguishing such characteristics as highly connected networks (e.g. star shaped) or decentralized networks, which have been used for studying the structural differences of metabolic networks [81]. **Heterogeneity** measures the variation of the connectivity distribution, such that reflecting the tendency of a network to have hubs. This heterogeneity index is biologically meaningful, as biological networks are usually heterogeneous with some central nodes highly connected and the rest of the nodes having few connections in the network. These two descriptors are computed as follows: network density variable is defined as '$density_G = \frac{2 \cdot E}{N(N-1)}$', where E is the number of edges. Then the connectivity centralization is calculated by '$centralization_G = \frac{N}{N-2}\left(\frac{max(deg_G)}{N-1} - density_G\right)$'and the heterogeneity is calculated by '$heterogeneity_G = \sqrt{\frac{N \cdot \sum_{i=1}^{N}(deg_i{}^2)}{\left(\sum_{i=1}^{N} deg_i\right)^2}} - 1$' [59]. **Graph Energy** of a network is defined as the sum of the absolute value of all nonzero eigenvalues $\{\lambda_1, \lambda_2 \ldots \lambda_k\}$ based on the adjacency matrix '$Energy_G = \sum_{i=1}^{k} |\lambda_i|$'[79]. **Information Content of Degree Equality** measures the probability distribution of vertex degree '$I_{vertexDegree} = -\sum_{i=1}^{k^d} \frac{N_i^d}{N} \cdot log_2\left(\frac{N_i^d}{N}\right)$', where N_i^d is the number of nodes having the same degree, and k^d is the maximum of degree [80].

To support the analysis of biological networks with varying binding constants and varying molecular levels, PROFEAT also supports the computation of the edge/node-weighted descriptors. For instances, the edge-weighted clustering coefficient [56, 69], the node-weighted cross degree, the node-weighted local clustering coefficient [82] and the directed local clustering coefficient [60]. **Edge-Weighted Clustering Coefficient** has been applied to the prediction of the significant genes in gene co-expression network [56, 69] and it is given by '$cluster\ EW_i = \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} \widehat{W}_{ij}\widehat{W}_{ik}\widehat{W}_{jk}}{\left(\sum_{k=1}^{N}\widehat{W}_{ij}\right)^2 - \sum_{k=1}^{N}\widehat{W}_{ij}{}^2}$', where the normalized edge weight is defined as '$\widehat{W}_{ij} = \frac{W_{ij}}{max\{W\}}$'. **Node-Weighted Cross Degree** and **Node-Weighted Local Clustering Coefficient** [82] are useful for analyzing the networks with heterogeneous node weights, which has been recently derived for Earth's spatial network and international trade network study. 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 relationship between node i and node j, and δ_{ij} is Kronecker's delta constant. The node-weighted cross degree is calculated 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 then computed by the formula '$cluster^{NW}{}_i = \frac{1}{crossdeg^{NW}{}_i{}^2} \sum_{j=1}^{N}\sum_{k=1}^{N} ExtA_{ij} \cdot NW_j \cdot ExtA_{ik} \cdot NW_k \cdot ExtA_{jk}$', which works only if the node-weighted cross degree is not zero, otherwise the local clustering coefficient will be assumed to be zero. **Directed Local Clustering Coefficient** has been introduced to measure the brain connectivity, as the neuro-connection is considered as a directed link [60]. This descriptor is defined by '$cluster^D{}_i = \frac{\frac{1}{2}\sum_{j,h\in N}(A_{ij}+A_{ji})(A_{ih}+A_{hi})(A_{jh}+A_{hj})}{(deg_i{}^+ + deg_i{}^-)(deg_i{}^+ + deg_i{}^- - 1) - 2\sum_{j\in N} A_{ij} \cdot A_{ji}}$', where deg_i^- and $deg^+{}_i$ represents the in-degree and out-degree of node i, respectively.

Results

To facilitate more extensive use of network descriptors in systems biology studies, we added the biological network descriptor computational module in PROFEAT web server at (http://bidd2.nus.edu.sg/cgi-bin/profeat2016/network/profnew.cgi).

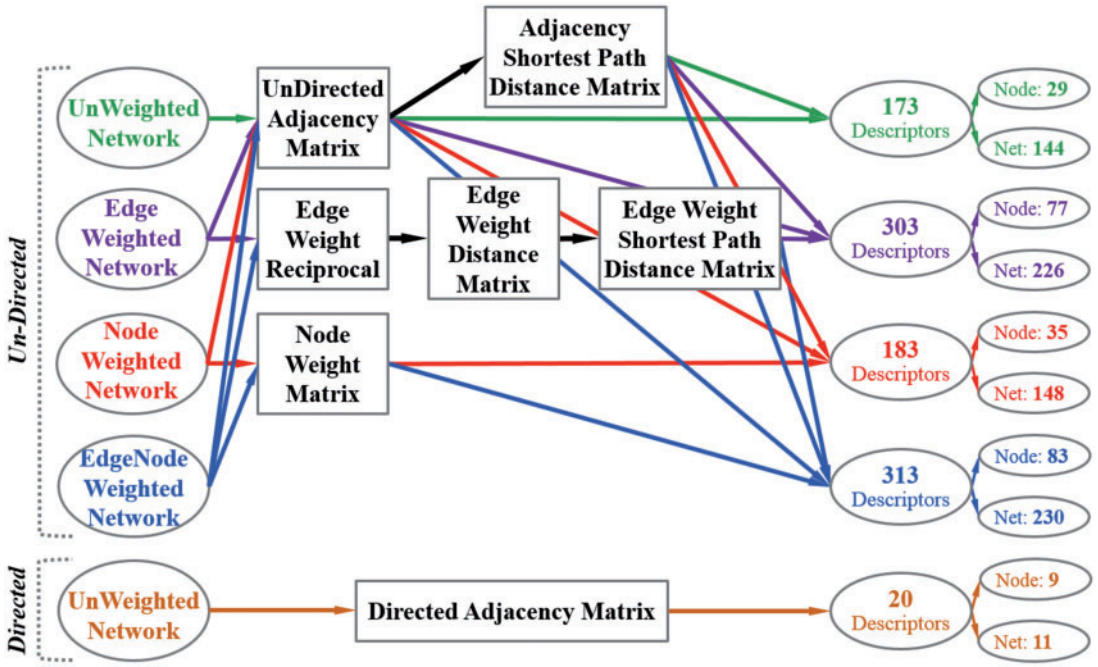


Figure 2. The computational flowchart of PROFEAT network descriptors, where 'node:' gives the number of node-level descriptors and 'net:' gives the number of network-level descriptors. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

PROFEAT network descriptor module is composed 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 un-weighted networks). Given an input network file, each type of network descriptors can be computed by uploading the file in a particular input field followed by the click of the 'Submit' button at the bottom of the input fields. Once the job is submitted, the network is read and the adjacency matrix is stored in hashable dictionary data type for faster data access. The deep-first-search is then carried out to check and split the disconnected networks if any. The adjacency-based shortest path lengths and the edge-weighted shortest path lengths are computed and also stored in hashable data matrices, and followed by the calculation of each descriptor according to its computation algorithm (Supplementary Section E). The output can be retrieved in two ways. For the jobs of smaller network sizes, the output is directly displayed at the automatically popup PROFEAT output window, with a download bottom to save the output file. For the jobs of larger network sizes the output is directed to a file stored at a URL such as http://bidd2.nus.edu.sg/cgi-bin/profeat2016/network/profeat-result.cgi?uid=net-x, where (x = 0, 1, 2,. . .) is a uniquely assigned network id for each individual job, which is accessible to view and save the output descriptors sometime later. The flowchart for computing the network descriptors is given in Figure 2, and the examples of the input and output of different network types are provided in Figure 3 and Supplementary Tables S3–S9.

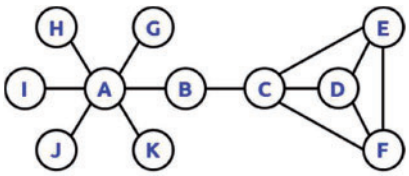
Input files

The input network file can be in either Simple interaction file (SIF) or Nested network file (NET) format, which have been widely used for storing biological interaction and network data in such databases as the Pathway Commons [72] and by such software as Cytoscape [22] and Pajek [26]. Specifically, the SIF format (illustrated in Figure 3) is tab-delimited, specifying the two interacting nodes in each column, with another column of relationship type in between the two node columns, i.e. [node A] tab [relationship] tab [node B]. In the NET format (also illustrated in Figure 3), the input data are in three sections, the *vertices section lists all the nodes, the *edges section contains all the undirected interactions between two nodes with an optional edge weight in the third column and the *arcs section includes all the directed interactions pointing from the earlier node to the later node. The output file is in a tab-delimited text file format composed of (1) the header section with each row starting with the character '!' followed by the network name, total number of networks, nodes and edges respectively, (2) the node-level descriptors section with each row showing the descriptor index, name and value for every node in the network (the node label is provided in the first row) and (3) the network-level descriptors section with each row showing the descriptor index, name and value. The input of a sample network in the SIF and NET format and the typical output are provided in Figure 3. The descriptor indices are described in Supplementary Section A, and the descriptor algorithms are given in Supplementary Section E.

Required information

The required information for computing the network descriptors is as follows: First, for an undirected unweighted network, only adjacency information is needed, and only unweighted descriptors are generated. Second, an undirected edge-weighted network requires a fourth column specifying the edge weight in the input: *[node A] tab [relationship] tab [node B] tab [edge weight]*, where the numerical edge weight can be kinetic constant, binding affinity, gene co-expression level, interaction confidence level or other measurements of the strength of the interacting nodes. Note that, edge length is inversely related to edge weight, as higher weight typically represents stronger

Network



Input in SIF Format

```
A       interact      B
B       interact      C
C       interact      D
C       interact      E
C       interact      F
D       interact      E
D       interact      F
E       interact      F
A       interact      G
A       interact      H
A       interact      I
A       interact      J
A       interact      K
```

Input in NET Format

```
*vertices
A
B
C
D
E
F
G
H
I
J
K
*edges
A       B
B       C
C       D
C       E
C       F
D       E
D       F
E       F
A       G
A       H
A       I
A       J
A       K
*arcs
```

Output

```
! Input Network File Name: sample_network.sif
! Total Number of Networks: 1
! Total Number of Nodes: 11
! Total Number of Edges: 13

# Network File:  sample_network.sif {11 Nodes; 13 Edges}
# # Node-Level Descriptors
[G10.0.0]       Node Label:                   A    B    ...    K
[G10.1]         Un-Weighted Features
[G10.1.1]       Degree:                       6    2    ...    1
      ...                    ...              ...  ...  ...    ...
# # Network-Level Descriptors
[G11.1]         Un-Weighted Features
[G11.1.1]       Number of Nodes:              11
[G11.1.2]       Number of Edges:              13
      ...                    ...              ...
```

Figure 3. The input and output of a sample undirected unweighted network, where (A, B . . . K) are the labels of individual nodes.

interaction and closer relation [60]. Such that the weighted-distance-related descriptors are calculated based on the reciprocal of edge weights. Third, the undirected node-weighted network needs an additional node weight file as [node label] tab [node weight], where the node label should be correctly matched to the network file, and the node weight could be gene expression level, protein/metabolite level, etc. Fourth, the undirected edge-node-weighted network requires the edge-weighted network file and the node weight file together to calculate the network descriptors. For all weighted networks, the weight normalization is carried out, such that weighted features will be calculated based on both the original weight and the normalized weight. Lastly, consider a directed unweighted network, the SIF file is differently defined: for the two interacting nodes in each line, the earlier one points to the latter one: *[node A] tab [relationship] tab [node B]* gives (A→B). Additionally, if there are multiple disconnected networks included in one single input file, PROFEAT is able to automatically detect each connected network, rank them by the number of nodes and output the descriptors for each one, respectively. An illustrative example for such case study is given in the Supplementary Section B.8 and Supplementary Table S9.

Evaluation of the performance of PROFEAT

We evaluated the CPU time of PROFEAT in computing the complete set of network descriptors for 30 GO biological process-specific PPI networks of five different network types and various

Table 5. CPU time in computing the complete set of network descriptors for 30 GO biological process-specific human PPI networks of five different network types

PPI network GO ID	GO biological process	Network size		CPU time (Minutes) for different network types				
		Number of nodes	Number of edges	Unweighted	Edge weighted	Node weighted	EdgeNode-weighted	Directed
GO:0002376	Immune system process	52	51	0.006	0.010	0.007	0.011	0.004
GO:0008219	Cell death	76	81	0.009	0.020	0.011	0.022	0.004
GO:0030705	Cytoskeleton intracellular transport	107	123	0.014	0.044	0.019	0.048	0.005
GO:0006091	Generation of metabolites and energy	125	134	0.022	0.070	0.030	0.077	0.005
GO:0006259	DNA metabolic process	141	152	0.025	0.093	0.035	0.104	0.005
GO:0006913	Nucleocytoplasmic transport	197	219	0.055	0.236	0.084	0.267	0.005
GO:0048646	Anatomical structure formation	242	246	0.092	0.423	0.147	0.446	0.006
GO:0006629	Lipid metabolic process	271	353	0.130	0.602	0.205	0.666	0.006
GO:0000902	Cell morphogenesis	347	377	0.288	1.25	0.402	1.42	0.007
GO:0009790	Embryo development	390	415	0.357	1.74	0.583	1.97	0.007
GO:0007005	Mitochondrion organization	435	513	0.456	2.46	0.768	2.68	0.008
GO:0048870	Cell motility	461	534	0.545	2.86	0.915	3.01	0.009
GO:0006397	mRNA processing	489	681	0.679	3.66	1.12	3.93	0.009
GO:0016192	Vesicle-mediated transport	494	630	0.670	3.49	1.13	3.75	0.009
GO:0034641	Cellular nitrogen metabolic process	563	756	1.01	5.44	1.70	5.57	0.011
GO:0006950	Response to stress	590	772	1.16	6.23	1.92	6.82	0.012
GO:0007010	Cytoskeleton organization	606	799	1.23	6.70	2.09	7.53	0.012
GO:0006464	Cellular protein modification process	627	751	1.40	7.60	2.32	8.24	0.013
GO:0006605	Protein targeting	642	860	1.49	8.22	2.47	9.07	0.014
GO:0006457	Protein folding	670	842	1.69	8.94	2.82	9.80	0.013
GO:0006412	Translation	772	996	2.52	13.77	4.27	15.47	0.017
GO:0006914	Autophagy	825	1001	2.94	16.33	5.03	18.71	0.018
GO:0006810	Transport	872	1089	3.41	19.55	6.01	21.92	0.020
GO:0005975	Carbohydrate metabolic process	1014	1329	5.44	30.78	9.41	35.32	0.026
GO:0007267	Cell–cell signaling	1202	1737	8.99	50.74	15.52	56.93	0.033
GO:0007049	Cell cycle	1513	2262	17.77	102.71	30.99	117.12	0.051
GO:0007568	Aging	1692	2637	24.92	144.16	43.78	158.62	0.062
GO:0030154	Cell differentiation	1752	2742	27.82	163.10	48.25	178.31	0.068
GO:0007155	Cell adhesion	1865	3356	34.26	194.14	58.84	216.88	0.076
GO:0008283	Cell proliferation	2616	4664	78.80	431.63	160.43	491.17	0.146

sizes (Table 5). These networks were constructed as follows: first, the human PPIs were collected from HPRD (Human Protein Reference Database) [74]; second, the GO annotation information for each protein was extracted from UniProt Database [83]; third, the proteins and the PPIs were mapped against the biological process annotations from GOslims Gene Ontology [84]; and finally, the human PPI network associated with each specific GO biological process was constructed. We selected 30 GO biological process networks with varying number of nodes from 52 to 2616 and different number of edges from 51 to 4664. Each of the 30 networks was constructed into five different types. The first four types are undirected unweighted, edge-weighted, node-weighted and edge-node-weighted networks with the edge weights or node weights randomly generated. The fifth type is the directed unweighted network with the direction of each edge tentatively assigned from the left node to the right node in the input of SIF Format.

The CPU time of the tested networks, measured on a Dell OptiPlex9010 Intel Core i7-3770 3.4 GHz CPU, are summarized in Table 5 and Figure 4. Specifically, the CPU time for the unweighted network is within 1 min for a network having no more than 500 nodes or 600 edges, the CPU time is <5 min if the network size is less than 1000 nodes or 1300 edges, while the CPU time increases to >80 min if the network gets larger than 2600 nodes or 4600 edges. On the other hand, the CPU time for the edge-weighted network is <50 min if the network size is no more than 1200 nodes or 1700 edges, and it takes >400 min if the network size is bigger than 2600 nodes or 4600 edges.

We further evaluated the PROFEAT computed values of a selected set of network descriptors and the job execution times in comparison with those of the three popular tools NetworkX, Cytoscape and Gephi. The evaluated network descriptors are degree, number of triangles, local/global clustering coefficient, closeness centrality, betweenness centrality, connectivity centralization and heterogeneity. These descriptors were computed for three undirected unweighted networks, which are (1) PPI network for GO:0030705 cytoskeleton intracellular transport with 107 nodes and 123 edges, (2) PPI network for GO:0009790 embryo development with 390 nodes and 415 edges and (3) PPI network for GO:0007267 cell–cell signaling with 1202 nodes and 1737 edges. Because it was difficult to directly measure the CPU time on the three popular tools, we used the job execution time (from the time of file input to the time of file output, which is roughly CPU time plus 3 s on PROFEAT) for the measurement. The comparative results are presented in Table 6, Supplementary Section C and Supplementary Tables S10–S12. The PROFEAT computed values of all the evaluated network descriptors for the three networks are in good agreement with those computed from the three popular tools. The job execution time of PROFEAT for the first two networks are comparable with those of the three popular tools (5 s versus 5–30 s, and 30 s versus 10–40 s). But the job execution time of PROFEAT for the third

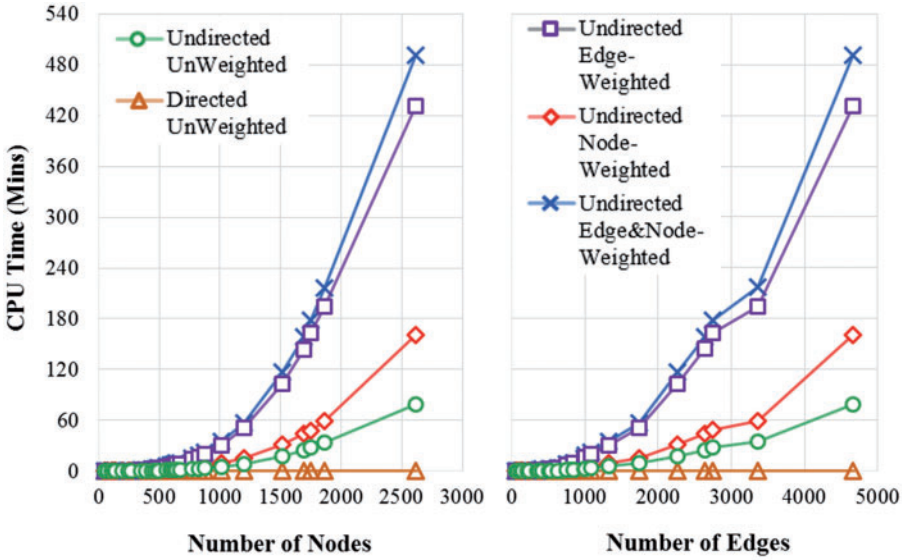


Figure 4. CPU time (mins) in computing the complete set of network descriptors for the networks described in Table 5 with respect to the number of nodes (left) and the number of edges (right). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Table 6. Comparison of the computed network descriptor values and the job execution time for a PPI network GO:0030705 cytoskeleton intracellular transport (107 nodes and 123 edges) by PROFEAT and other popular tools NetworkX, Cytoscape and Gephi

Tool name	PROFEAT	NetworkX	Cytoscape	Gephi
Network descriptor		Computed network descriptor value		
Degree	42 (TUBB)	42 (TUBB)	42 (TUBB)	42 (TUBB)
Number of triangle	4 (KIF5A)	4 (KIF5A)	N.A.	4 (KIF5A)
Local clustering coefficient	1 (SDC3)	1 (SDC3)	1 (SDC3)	1 (SDC3)
Closeness centrality	0.486 (TUBB)	0.485 (TUBB)	0.485 (TUBB)	0.484 (TUBB)
Betweenness centrality	0.816 (TUBB)	0.816 (TUBB)	0.816 (TUBB)	0.816 (TUBB)
Global clustering coefficient	0.025	0.025	0.025	0.110
Connectivity centralization	0.382	N.A.	0.382	N.A.
Heterogeneity	2.046	N.A.	2.045	N.A.
		Job execution time		
	~5 s	~5 s	~30 s	~30 s

The first five descriptors are node-level properties, where the maximum values and the corresponding node's gene names are given, like 'max. Value (gene name)'. The next three descriptors are network-level properties, globally describing the entire network.

network is substantially higher than those of the three popular tools (9 min versus 30–60 s). The longer job execution times of PROFEAT arise from its computation of the full set of network descriptors in contrast to the computation of a smaller subset of descriptors by the other tools.

Discussion

The usefulness of the network descriptors in characterizing the connectivity, organizational, robustness and stability properties of the protein–protein, disease, metabolic and drug targeted networks are illustrated in the following cases of literature-reported studies. These are the study of the topological properties of the human protein–protein network [18], the characteristics of the protein encoding the house-keeping genes and tissue-specific genes in the protein–protein network of specific human tissues [85], the exploration of network descriptors for identifying [86] and analyzing [87] metabolic pathways and for studying evolutional features [88] and phylogenetic relationships [89], the investigation of the target-like characteristics of therapeutic targets in the human protein–protein network [9, 17] and the disease network [9], and the feasibility assessment of using drug target relevant network descriptors for developing machine learning target prediction models [17]. The taxonomy of the biological problems and the use of network indicators discussed here are summarized in Supplementary Section D and Supplementary Table S12.

Study of the topological properties of the human protein–protein network and the characteristics of specific protein classes

The extensive studies of PPI have generated rich knowledge and data for the investigation of the network behavior of proteins. For instance, a PPI map has been constructed as a resource for annotating the proteome, which has been used for probing the topological properties of the human protein–protein network that connects 1705 human proteins via 3186 interactions [18]. Based on the analysis of the network descriptors of this network, it was found that the average clustering coefficient, a

measure of the tendency of the proteins to form groups, diminishes when the number of interactions per protein increases, indicating a hierarchical organization of the network. Moreover, the topological coefficient, a measure of the extent to which a protein shares interaction partners with other proteins, decreases with the number of connections, suggesting that hubs do not have more common neighbors than proteins with fewer connections.

Moreover, a study has been conducted to analyze the topological and organizational properties of the proteins encoded by the house-keeping and tissue-specific genes in the human protein–protein networks of specific human tissues [85]. The tissue-specific networks of 19 tissue types were generated by mapping the tissue-specific gene expression data of the Human Gene Expression Index database [90] to those of the Human Protein Reference Database [91]. There are 149–203 house-keeping genes and 8–318 tissue-specific genes with 464–1933 interactions in the network of each tissue type. Three network descriptors (degree, betweenness centrality, closeness centrality) of the house-keeping and tissue-specific genes were analyzed in comparison with the expected mean values of randomly selected proteins in the human protein–protein network, which showed certain tissue-specific behavior. For instance, for the house-keeping genes, the average degrees in the brain and testes, and the average betweenness centrality in the testes are significantly greater than the expected mean values, which indicate that in these tissues the proteins encoded by house-keeping genes tend to have a greater number of direct neighbors and/or occupy network points incident to higher number of shortest interaction paths. For the tissue-specific genes, the average betweenness centrality in the testes and the average closeness centrality in the ovary are significantly lower than the expected mean, which suggest that the protein products of tissue-specific genes in the testes tend to occupy network positions incident to lower number of shortest interaction paths, and those in the ovary tend to be further away from other genes.

Exploring network descriptors for identifying and analyzing metabolic pathways, and for studying evolutional features and phylogenetic relationships

Metabolic pathways have distinct network topological features shaped by such factors as the evolutionary processes [88, 92]. These topological features are useful and have been explored for identifying [86] and evaluating [87] the elements of metabolic pathways, studying the influence of evolution on metabolic networks [88], and revealing that the phylogenetic relationship of the species are encoded in their metabolic pathways [89]. One strategy in generating metabolic pathways is to search for the shortest reaction paths among the possible connections between the identified metabolites using their structural similarity as the edge weightage in the metabolite graphs [86]. To remove the irrelevant metabolite candidates for each reaction path, two network descriptors (degree and betweenness centrality) have been used to filter out the nodes of lower degree value or lower betweenness centrality value that likely represent side metabolites in a reaction step [86]. In another study, multiple network descriptors such as betweenness centrality and eccentricity have been used in the development of a machine learning model (random forest model) to predict the correct and incorrect enzyme assignments, which successfully distinguished correctly and incorrectly annotated candidates of a missing enzyme (dihydroneopterin aldolase) in the *Plasmodium falciparum* folate biosynthesis pathway [87].

The studies of the evolution of metabolic networks have been based on such models as the Patchwork model [88, 93], which assumes that enzymes refine their substrate specificity after duplication events. In the early stage of evolution, most enzymes have broad substrate specificities for generating multiple metabolic pathways to produce the same metabolites. Eventually, evolution has brought selective advantage to those pathways that generate higher amounts of the key metabolites, and the duplication events lead to the specialization of the enzymes and the respective metabolic pathways. In studying these evolutionary events, five network descriptors (degree, betweenness centrality, clustering coefficient, assortativity and shortest path) have been used for characterizing the topological properties of metabolic networks [88]. The relevant studies have shown that the duplication rate of a metabolic network hubs is relatively low [94], indicative of their central roles, while duplication of genes are mostly localized in the network [88]. Moreover, there is a high retention of duplicates between chemically similar reactions and in closely connected functional modules, and the local connectivity effect of duplications is absent in the interaction networks of nonenzymatic proteins, which suggests that the retention of duplicates results from biochemical rules that govern substrate–enzyme–product relationships [95]. In another study [89], two networks, a network of interacting pathways and a network of interacting metabolites, were constructed from the KEGG and the November 2006 release of the Ma data set for each of the 107 species, which were represented by 35 network descriptors that measure various degree, centrality, distance and cliques-related properties. Then, the metabolic network-based distances derived from the trained regression models were compared with the phylogenetic distances derived from the 16S rRNA sequences, which showed that the metabolic network-based distances reproduce accurately reference phylogenetic distances derived from 16S rRNA sequences. These studies suggest the usefulness of network descriptors in studying the structural, evolutionary and phylogenetic profiles of the metabolic networks.

Study of the target-like characteristics of drug targets in the human protein–protein and disease networks

The targets of approved drugs possess such specific target-like characteristics as the appropriate druggable structures, substantial dissimilarity to human proteins and distinguished systems and tissue distribution profiles [9, 17, 96, 97]. In particular, these targets are distinguished in the human protein–protein and disease–gene networks such that specific network descriptors may be used as the quantitative determinants of drug targets in these networks [9, 17]. In a study of the global relationships between drug targets in the human interactome network [9], a target protein network was first constructed by using 394 targets of 890 approved drugs wherein these targets are connected by their commonly targeted drug(s). In this network, 788 drugs share targets and 305 targets are connected to one another. This network was then overlaid onto the human protein–protein network [98] composed of 7533 proteins and 22 052 non-self-interacting and nonredundant interactions. Overall, 260 targets were mapped onto the human protein–protein network, which on average have a higher degree (with 42% more interacting proteins) than that of the nontarget proteins in the same network. In another study of the druggability properties of 304 targets of approved drugs in the human protein–

protein network [17], a protein–protein network model of 7764 proteins and 28 149 interactions derived from the Human Protein Reference Database [74], in which the drug targets were found to have increased average betweenness, suggesting their tendency to bridge two or more clusters of relatively closely interacting proteins.

The distribution behavior of the drug targets in the human disease network has also been studied [9]. A human disease–gene network of 1284 distinct disorders and 1777 disease-related genes was generated from the OMIM-based disorder–disease gene associations [50], wherein a link is established between a disorder and a disease gene if a mutation in that gene leads to the disorder. Disease genes associated with a single disorder are omitted. In this network, there are 166 genes encoding the targets of approved drugs, 43% of which are associated with two or more disorders. 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 clustering of drug targets in the network.

Evaluation of the feasibility of using drug target-relevant network descriptors for developing machine learning target prediction models

A study has been conducted to further test whether some of the above-described drug target-relevant network descriptors are useful for developing machine learning classification models to predict drug targets based on the available network and other types of descriptors of proteins [17]. In that study, 237 targets (positive samples) and 5037 nontarget proteins (negative samples) were used for constructing and testing the machine learning models (naive Bayesian classifier, logistic regression, radial basis function network and Bayesian network models) in distinguishing targets and nontarget proteins in the human genome. These target and nontarget proteins were represented by two network (degree and betweenness) and three other (tissue expression entropy, SNP-based C_{ratio} and functional family assignment) descriptors. The machine learning models were each trained and tested by using the 10-fold cross-validation method [99]. The performance of these machine leaning models was evaluated by the receiver operating characteristic (ROC) score that measures the overall performance of each model integrated over the entire range of false-positive and false-negative rates. The ROCs of these four models were found to be significantly higher than the value of 0.5 of a random classification model, indicating the usefulness of the network descriptors as part of the ingredients in developing drug target prediction tools based on the sequence, structure and systems profiles of proteins.

Perspectives

The functional studies of proteins frequently require the use of multiple approaches from the perspectives of genetic sequences, protein structures, molecular interactions and biological networks. Protein functional studies particularly at the biological systems and cellular levels can be greatly enhanced by the exploration of the network theories, descriptors and models developed in other fields [12–14, 20, 21, 100–102] and in the study of biological systems [1–3, 6, 9, 103], which offers much more expanded perspectives and avenues to the understanding of biological systems and cellular internal organization, evolution and dynamic behavior than the studies based on the concept of individual molecule or independent group of molecules [1]. The progress toward reliable network-based studies of the biological and disease processes may be constrained by the insufficient information about biological networks, limited capability of the available network analysis and modeling methods and the inadequate computational resources for facilitating the analysis and modeling of biological networks. By providing the facility of the computation of comprehensive network descriptors useful for studying biological systems, PROFEAT complements the other resources in the information [72, 104], modeling tools [105], parameters [106] of biological pathways and the data of PPIs [107–109] for collectively facilitating the investigation of the protein functions and the network dynamics and their roles in biological and cellular systems [1], disease processes [2] and therapeutic regulations [3].

Key Points

- PROFEAT web server computes the currently most comprehensive (upto 313) network descriptors to characterize the node/network-level topology, connectivity and complexity properties of a network.
- It supports different types of networks with different biological representations in terms of binding constants, molecular levels and directed processes.
- It is user friendly with simple input/output and easy operation.
- This article reviewed the usefulness of network descriptors in systems biology applications (e.g. protein–protein interaction network, metabolic network, disease network, drug–target network and gene regulatory network)
- PROFEAT network descriptors could facilitate the functional biological investigations by providing the systematic properties of molecular interaction networks, offering the expanded understandings of biological complex systems and revealing the higher-level clues of what the mechanisms could be.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This work was supported by Singapore Academic Research Fund [R-148-000-208-112], giving to PI: Yuzong Chen; National Natural Science Foundation of China [81202459], giving to PI: Feng Zhu; Fundamental Research Funds for the Central Universities [CQDXWL-2012-Z003, CDJZR14-46-88-01], giving to PI: Feng Zhu and Shenzhen Sci & Tech Bureau [CXB201104210014A], [CXZZ20150529165045064], [CXZZ20150529165045064] giving to PI: Yuyang Jiang; the 973 Program [2013CB967204] and National Natural Science Foundation of China [81325021] giving to PI: Zerong Li.

References

1. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
2. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68.
3. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* 2008;**4**:682–90.
4. Yook SH, Oltvai ZN, Barabasi AL. Functional and topological characterization of protein interaction networks. *Proteomics* 2004;**4**:928–42.
5. Ma'ayan A. Introduction to network analysis in systems biology. *Sci Signal* 2011;**4**:tr5:22–32.
6. Cho DY, Kim YA, Przytycka TM. Network biology approach to complex diseases. *PLoS Comput Biol* 2012;**8**:e1002820.
7. Furlong LI. Human diseases through the lens of network biology. *Trends Genet* 2013;**29**:150–9.
8. Zhang B, Tian Y, Zhang Z. Network biology in medicine and beyond. *Circ Cardiovasc Genet* 2014;**7**:536–47.
9. Yildirim MA, Goh KI, Cusick ME, *et al*. Drug-target network. *Nat Biotechnol* 2007;**25**:1119–26.
10. Pujol A, Mosca R, Farres J, *et al*. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol Sci* 2010;**31**:115–23.
11. Guney E, Menche J, Vidal M, *et al*. Network-based in silico drug efficacy screening. *Nat Commun* 2016;**7**:10331.
12. Freeman LC. A set of measures of centrality based on betweenness. *Sociometry* 1977;**40**:35–41.
13. Holland PW, Leinhardt S. Transtivity in structural models of small groups. *Small Group Res* 1971;**2**:107–24.
14. Shimbel A. Structural parameters of communication networks. *Bull Math Biophys* 1953;**15**:501–7.
15. Marx V. Cancer: smoother journeys for molecular data. *Nat Methods* 2015;**12**:299–302.
16. Yoon J, Blumer A, Lee K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics* 2006;**22**:3106–8.
17. Yao L, Rzhetsky A. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res* 2008;**18**:206–13.
18. Stelzl U, Worm U, Lalowski M, *et al*. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005;**122**:957–68.
19. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science* 2002;**296**:910–13.
20. Rodrigue JP. *The Geography of Transport Systems*. New York, NY: Routledge, 2013.
21. Piraveenan M, Uddin S, Chung KSK. Measuring robustness of networks under sustained targeted attacks. In IEEE International Conference Advances in Social Networks Analysis and Mining, Istanbul, 2012, 38–45.
22. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
23. Djebbari A, Ali M, Otasek D, *et al*. NAViGaTOR: large scalable and interactive navigation and analysis of large graphs. *Internet Math* 2011;**7**:314–47.
24. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks, In 3rd International AAAI Conference on Weblogs and Social Media, San Jose, CA, USA, 2009.
25. Brinkrolf C, Janowski SJ, Kormeier B, *et al*. VANESA - a software application for the visualization and analysis of networks in system biology applications. *J Integr Bioinform* 2014;**11**:239.
26. Batagelj V, Mrvar A. Pajek – program for large network analysis. *Connections* 1998;**21**:47–57.
27. Forman JJ, Clemons PA, Schreiber SL, *et al*. SpectralNET–an application for spectral graph analysis and visualization. *BMC Bioinformatics* 2005;**6**:260.
28. Wu J, Vallenius T, Ovaska K, *et al*. Integrated network analysis platform for protein-protein interactions. *Nat Methods* 2009;**6**:75–7.
29. Lin CY, Chin CH, Wu HH, *et al*. Hubba: hub objects analyzer–a framework of interactome hubs identification for network biology. *Nucleic Acids Res* 2008;**36**:W438–43.
30. Reimand J, Tooming L, Peterson H, *et al*. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res* 2008;**36**:W452–9.
31. Yip KY, Yu H, Kim PM, *et al*. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* 2006;**22**:2968–70.
32. Hu Z, Mellor J, Wu J, *et al*. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res* 2005;**33**:W352–7.
33. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using networkX, In Proceedings of 7th Python in Science Conference, 2008:11–15.
34. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJ Complex Syst* 2006:1695.
35. Mueller LA, Kugler KG, Dander A, *et al*. QuACN: an R package for analyzing complex biological networks quantitatively. *Bioinformatics* 2011;**27**:140–1.
36. Doncheva NT, Assenov Y, Domingues FS, *et al*. Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc* 2012;**7**:670–85.
37. Ravasz E, Somera AL, Mongru DA, *et al*. Hierarchical organization of modularity in metabolic networks. *Science* 2002;**297**:1551–5.
38. Snel B, Bork P, Huynen MA. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci USA* 2002;**99**:5890–5.
39. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007;**3**:88.
40. Zhang L, Li X, Tai J, *et al*. Predicting candidate genes based on combined network topological features: a case study in coronary artery disease. *PLoS One* 2012;**7**:e39542.
41. Emig D, Ivliev A, Pustovalova O, *et al*. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One* 2013;**8**:e60618.
42. Hsu CL, Huang YH, Hsu CT, *et al*. Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics* 2011;**12(Suppl 3)**:S25.
43. Zhu C, Kushwaha A, Berman K, *et al*. A vertex similarity-based framework to discover and rank orphan disease-related genes. *BMC Syst Biol* 2012;**6 (Suppl 3)**:S8.
44. Netzer M, Kugler KG, Muller LA, *et al*. A network-based feature selection approach to identify metabolic signatures in disease. *J Theor Biol* 2012;**310**:216–22.
45. Dirk Koschützki FS. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio* 2008;**2**:193–201.
46. Jacunski A, Tatonetti NP. Connecting the dots: applications of network medicine in pharmacology and disease. *Clin Pharmacol Ther* 2013;**94**:659–69.

47. Joy MP, Brock A, Ingber DE, *et al*. High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol* 2005;**2005**:96–103.
48. Yu H, Kim PM, Sprecher E, *et al*. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 2007;**3**:e59.
49. Harrold JM, Ramanathan M, Mager DE. Network-based approaches in drug discovery and early development. *Clin Pharmacol Ther* 2013;**94**:651–8.
50. Goh KI, Cusick ME, Valle D, *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**:8685–90.
51. Banky D, Ivan G, Grolmusz V. Equal opportunity for low-degree network nodes: a PageRank-based method for protein target identification in metabolic graphs. *PLoS One* 2013;**8**:e54204.
52. Winter C, Kristiansen G, Kersting S, *et al*. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* 2012;**8**:e1002511.
53. Paladugu SR, Zhao S, Ray A, *et al*. Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics* 2008;**9**:426.
54. You ZH, Yin Z, Han K, *et al*. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *BMC Bioinformatics* 2010;**11**:343.
55. Schadt EE, Björkegren JL. NEW: network-enabled wisdom in biology, medicine and healthcare. *Sci Transl Med* 2012;**4**:115rv111.
56. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;**4**:Article17.
57. Barrat A, Barthelemy M, Pastor-Satorras R, *et al*. The architecture of complex weighted networks. *Proc Natl Acad Sci USA* 2004;**101**:3747–52.
58. Pei Wang JL, Yu X. Identification of important nodes in directed biological networks: a betwork motif approach. *PLoS One* 2014;**9**:e106132.
59. Dong J, Horvath S. Understanding network concepts in modules. *BMC Syst Biol* 2007;**1**:24.
60. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 2010;**52**:1059–69.
61. Emmert-Streib F, Dehmer M. Networks for systems biology: conceptual connection of data and function. *IET Syst Biol* 2011;**5**:185–207.
62. Bonchev D. *Complexity in Chemistry, Biology, and Ecology*. New York, USA: Springer US, 2007.
63. Kim J, Wilhelm T. What is a complex graph? *Physica A* 2008;**387**:2637–52.
64. Mueller LA, Kugler KG, Netzer M, *et al*. A network-based approach to classify the three domains of life. *Biol Direct* 2011;**6**:53.
65. Dehmer M, Mowshowitz A. A history of graph entropy measures. *Inf Sci* 2011;**181**:57–78.
66. Dehmer M, Grabner M, Varmuza K. Information indices with high discriminative power for graphs. *PLoS One* 2012;**7**:e31214.
67. Dehmer M. Uniquely discriminating molecular structures using novel eigenvalue-based descriptors. *MATCH Commun Math Comput Chem* 2012;**67**:147–72.
68. Ivan G, Grolmusz V. When the web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics* 2011;**27**:405–7.
69. Saramäki J, Kivelä M, Onnela JP, *et al*. Generalizations of the clustering coefficient to weighted complex networks. *Phys Rev E* 2007;**75**:027105.
70. Li ZR, Lin HH, Han LY, *et al*. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 2006;**34**:W32–7.
71. Rao HB, Zhu F, Yang GB, *et al*. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 2011;**39**:W385–90.
72. Cerami EG, Gross BE, Demir E, *et al*. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011;**39**:D685–90.
73. Kandasamy K, Mohan SS, Raju R, *et al*. NetPath: a public resource of curated signal transduction pathways. *Genome Biol* 2010;**11**:R3.
74. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database–2009 update. *Nucleic Acids Res* 2009;**37**:D767–72.
75. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998;**393**:440–2.
76. Newman MEJ. A measure of betweenness centrality based on random walks. *Soc Networks* 2003;**27**:39–54.
77. Dijkstra EW. A note on two problems in connexion with graphs. *Numer Math* 1959;**1**:269–71.
78. Brandes U. A faster algorithm for betweenness centrality. *J Math Sociol* 2001;**25**:163–77.
79. Gutman I, Zhou B. Laplacian energy of a graph. *Linear Algebra Appl* 2006;**414**:29–37.
80. Leroy G. *Information Theoretic Indices for Characterization of Chemical Structures*. New York, NY : Wiley, 1985.
81. Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 2003;**19**:1423–30.
82. Wiedermann M, Donges JF, Heitzig J, *et al*. Node-weighted interacting network measures improve the representation of real-world complex systems. *EPL* 2013;**102**:28007.
83. UniProt C. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
84. Gene Ontology C. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2015;**43**:D1049–56.
85. Lin WH, Liu WC, Hwang MJ. Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. *BMC Syst Biol* 2009;**3**:32.
86. Frainay C, Jourdan F. Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Brief Bioinform* 2016;pii: bbv115
87. Liberal R, Pinney JW. Simple topological properties predict functional misannotations in a metabolic network. *Bioinformatics* 2013;**29**:i154–61.
88. Yamada T, Bork P. Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol* 2009;**10**:791–803.
89. Mazurie A, Bonchev D, Schwikowski B, *et al*. Phylogenetic distances are encoded in networks of interacting pathways. *Bioinformatics* 2008;**24**:2579–85.
90. Hsiao LL, Dangond F, Yoshida T, *et al*. A compendium of gene expression in normal human tissues. *Physiol Genomics* 2001;**7**:97–104.
91. Peri S, Navarro JD, Amanchy R, *et al*. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003;**13**:2363–71.

92. Raymond J, Segre D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* 2006;**311**:1764–7.
93. Ycas M. On earlier states of the biochemical system. *J Theor Biol* 1974;**44**:145–60.
94. Lu C, Zhang Z, Leach L, *et al*. Impacts of yeast metabolic network structure on enzyme evolution. *Genome Biol* 2007;**8**:407.
95. Diaz-Mejia JJ, Perez-Rueda E, Segovia L. A network perspective on the evolution of metabolism by gene duplication. *Genome Biol* 2007;**8**:R26.
96. Zheng CJ, Han LY, Yap CW, *et al*. Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev* 2006;**58**:259–79.
97. Hopkins AL, CR G. The druggable genome. *Nat Rev Drug Discov* 2002;**1**:727–30.
98. Rual JF, Venkatesan K, Hao T, *et al*. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;**437**:1173–8.
99. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.
100. Latora V, Marchiori M. Efficient behavior of small-world networks. *Phys Rev Lett* 2001;**87**:198701.
101. Sabidussi G. The centrality index of a graph. *Psychometrika* 1966;**31**:581–603.
102. Davis JA. Clustering and hierarchy in interpersonal relations: testing two graph theoretical models on 742 sociomatrices. *Am Sociol Rev* 1970;**35**:843–51.
103. Zhang J, Jia J, Zhu F, *et al*. Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors. *Mol Biosyst* 2012;**8**:2645–56.
104. Kanehisa M, Goto S, Sato Y, *et al*. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;**42**:D199–205.
105. Chelliah V, Juty N, Ajmera I, *et al*. BioModels: ten-year anniversary. *Nucleic Acids Res* 2015;**43**:D542–8.
106. Kumar P, Han BC, Shi Z, *et al*. Update of KDBI: kinetic data of bio-molecular interaction database. *Nucleic Acids Res* 2009;**37**:D636–41.
107. Szklarczyk D, Franceschini A, Wyder S, *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;**43**:D447–52.
108. Orchard S, Ammari M, Aranda B, *et al*. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 2014;**42**:D358–63.
109. Hoffmann R, Valencia A. A gene network for navigating the literature. *Nat Genet* 2004;**36**:664.