Biased random walk with restart for essential proteins prediction

Pengli Lu(卢鹏丽), Yuntian Chen(陈云天), Teng Zhang(张腾), and Yonggang Liao(廖永刚)

Citation: Chin. Phys. B, 2022, 31 (**11**): 118901. DOI: 10.1088/1674-1056/ac7a17

What follows is a list of articles you may be interested in

Passenger management strategy and evacuation in subway station under Covid-19

Xiao-Xia Yang(杨晓霞), Hai-Long Jiang(蒋海龙), Yuan-Lei Kang(康元磊), Yi Yang(杨毅), Yong-Xing Li(李永行), and Chang Yu(蔚畅)

Chin. Phys. B, 2022, 31 (**7**): 078901. DOI: 10.1088/1674-1056/ac43b3

Advantage of populous countries in the trends of innovation efficiency

Dan-Dan Hu(胡淡淡), Xue-Jin Fang(方学进), and Xiao-Pu Han(韩筱璞)

Chin. Phys. B, 2022, 31 (**6**): 068903. DOI: 10.1088/1674-1056/ac5614

Synchronization in multilayer networks through different coupling mechanisms

Xiang Ling(凌翔), Bo Hua(华博), Ning Guo(郭宁), Kong-Jin Zhu(朱孔金), Jia-Jia Chen(陈佳佳), Chao-Yun Wu(吴超云), and Qing-Yi Hao(郝庆一)

Chin. Phys. B, 2022, 31 (**4**): 048901. DOI: 10.1088/1674-1056/ac2b19

Collective behavior of cortico-thalamic circuits: Logic gates as the thalamus and a dynamical neuronal network as the cortex

Alireza Bahramian, Sajjad Shaukat Jamal, Fatemeh Parastesh, Kartikeyan Rajagopal, and Sajad Jafari

Chin. Phys. B, 2022, 31 (**2**): 028901. DOI: 10.1088/1674-1056/ac0eeb

Cascading failures of overload behaviors using a new coupled network model between edges

Yu-Wei Yan(严玉为), Yuan Jiang(蒋沅), Rong-Bin Yu(余荣斌), Song-Qing Yang(杨松青), and Cheng Hong(洪成)

Chin. Phys. B, 2022, 31 (**1**): 018901. DOI: 10.1088/1674-1056/ac1337

Biased random walk with restart for essential proteins prediction

Pengli Lu(卢鹏丽)[1,†], Yuntian Chen(陈云天)[1], Teng Zhang(张腾)[1], and Yonggang Liao(廖永刚)[2]

[1] *School of Computer and Communication, Lanzhou University of Technology, Lanzhou* 730050, *China*
[2] *China Mobile Communications Group Gansu Co., Ltd., Lanzhou* 730070, *China*

(Received 23 April 2022; revised manuscript received 6 June 2022; accepted manuscript online 18 June 2022)

Predicting essential proteins is crucial for discovering the process of cellular organization and viability. We propose biased random walk with restart algorithm for essential proteins prediction, called BRWR. Firstly, the common process of practice walk often sets the probability of particles transferring to adjacent nodes to be equal, neglecting the influence of the similarity structure on the transition probability. To address this problem, we redefine a novel transition probability matrix by integrating the gene express similarity and subcellular location similarity. The particles can obtain biased transferring probabilities to perform random walk so as to further exploit biological properties embedded in the network structure. Secondly, we use gene ontology (GO) terms score and subcellular score to calculate the initial probability vector of the random walk with restart. Finally, when the biased random walk with restart process reaches steady state, the protein importance score is obtained. In order to demonstrate superiority of BRWR, we conduct experiments on the YHQ, BioGRID, Krogan and Gavin PPI networks. The results show that the method BRWR is superior to other state-of-the-art methods in essential proteins recognition performance. Especially, compared with the contrast methods, the improvements of BRWR in terms of the ACC results range in 1.4%–5.7%, 1.3%–11.9%, 2.4%–8.8%, and 0.8%–14.2%, respectively. Therefore, BRWR is effective and reasonable.

Keywords: PPI network, essential proteins, random walk with restart, gene expression

PACS: 89.75.–k **DOI:** 10.1088/1674-1056/ac7a17

1. Introduction

Protein is an important component of all cells and tissues of the human body.[1] Essential proteins are indispensable for organisms survival and evolution.[2] Essential proteins prediction not only sheds light on revealing the structure and function of genes, but also provides importance guidance in the study of disease diagnosis and drug targets.[3–5] In biomedicine field, many methods based on biological experiments have been proposed to predict essential protein, such as single-gene knockout,[6] RNA interference,[7] and conditional gene knockout.[8] These methods have made great contributions in helping people understand cells and in research of new drugs.[9] These traditional experimental procedures can provide an accurate prediction but suffered from huge cost and time consuming.

In recent years, a variety of biological datasets, such as genomics, proteomics, transcriptomics, and gene ontology (GO) data have been obtained by high-throughput experiments, for instance, two-hybrid systems, mass spectrometry, and protein micro-arrays. At the same time, centrality-lethality[10] rule shows that proteins with highly connected neighbors tend to be essential. Thus, many complex network centrality methods have been successfully devoted to essential protein prediction problem, such as degree centrality (DC),[11] betweenness centrality (BC),[12] subgraph centrality (SC),[13] eigenvector centrality (EC),[14] local average center (LAC),[15] network centrality (NC),[16] and multi-order K-shell vector (MKV).[17] However, the above-mentioned methods only consider the topology features[18] of the network and ignore the inherent biological significance of essential proteins,[19] which results in undesirable performance. Actually, biological information is critical for the identification of essential proteins, and there are a variety of biological dadasets that can be used. Therefore, how to integrate the PPI networks and some kinds of biological dadasets to improve the efficiency and accuracy of essential proteins prediction is also a challenge. Currently, some researchers try to use the biological information to essential protein prediction algorithm, such as GO terms,[20] subcellular localization,[21] gene expression sequence,[22,23] protein complexes information,[24] and other biological information. The GEG[25] method is based on semantic similarity and gene expression sequence. The united complex centrality (UC)[24] method considers the number of protein appearances in the complexes. The LIDC[26] method combines network local action with protein complexes information. PeC[22] and Wdc[27] integrate network topology characteristics and gene expression sequence. UDoNC[28] by integrating the protein domain data with PPI data improves the efficiency distinctly. However, it still has large development space for accurately predicting essential proteins.

As an optimization algorithm, random walk (RW) has been widely used in link prediction,[29] recommender systems,[30] ranking,[31] community detection,[32–34] and transmission dynamics.[35] The essential proteins prediction

[†]Corresponding author. E-mail: lupengli88@163.com

method EssRank[36] was developed based on the PageRank[37] model, which is a web ranking algorithm based on RW.[38] The DEP-MSB[39] method was designed also based on PageRank, which integrates a variety of biological information and six centrality methods. In the traditional random walk, the transition probability of a walker from the current vertex to the next vertex is equal. However, due to the complex diversity of the protein network structure and the protein itself contains complex biological characteristics, the walker in the transition process will be affected by the neighbors' proteins, and not necessarily an equal probability of movement.[40] Based on the above description, we propose a biased random walk with restart method named BRWR for the prediction of essential proteins. In algorithm BRWR, first we define a novel similarity adjacency matrix and reconstruct the transition probability matrix, which makes the walker moved towards similar neighbors with a high probability from the initial vertex when walker moves to the adjacent vertices. In addition, we use GO terms score and subcellular score to calculate the initial probability vector of the random walk with restart. Finally, when BRWR process reaches steady state, we can obtain a score vector for proteins and the top prioritized proteins are regarded as the candidate essential proteins. To assess the performance of our algorithm BRWR, we compare our algorithms with other previous algorithms on protein data set BioGRID, YHQ, Krogan and Gavin. The results through different evaluation measures indicate that BRWR outperforms the state-of-the-art approaches with stable performance for identifying the essential proteins.

2. Related work

2.1. PPI network

The PPI network can be abstracted and formularized as an unweighted and undirected graph $G = (V, E)$, where V is the vertex set corresponding to proteins, E is the edges set denoting the interactions between proteins. There are $n = |V|$ proteins and $m = |E|$ edges in the PPI network. The adjacency matrix of the PPI network, denoted by $\boldsymbol{A}$, is the $n \times n$ matrix whose (i, j)-entry is 1 if $v_i \sim v_j$, and it is 0 otherwise. Let $d(v_i)$ be the degree of vertex v_i. $\boldsymbol{D}$ denotes the degree diagonal matrix with diagonal entries $d(v_1), d(v_2), \ldots, d(v_n)$. Let $\Gamma(v_i)$ be the neighbor set of vertex v_i, and $|\Gamma(v_i)| = d(v_i)$.

2.2. Transition probability matrix

Consider a random walk on G: start at a vertex v_i; if at the t-th step the walker is at a vertex v_i^t, it moves to a neighbor of v_i^t with probability $1/d(v_i^t)$. Clearly, the sequence of random vertices $(v_i^t : t = 0, 1, \ldots)$ is a Markov chain. The transmission matrix is defined as $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{D}^{-1}$.[38] Give a distribution $\boldsymbol{P}^t$, the rule of the walk can be expressed by $\boldsymbol{P}^{t+1} = \boldsymbol{M} \times \boldsymbol{P}^t$. The transition matrix of other type random walk, like lazy random walk,[41] can be described as $\boldsymbol{M} = (1/2)(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{D}^{-1})$. In Ref. [34] to calculate $\boldsymbol{M}$, the authors used normalization of the matrix $(\boldsymbol{A} + \boldsymbol{I})$ such that the sum of probability in each column is 1 and defined $\boldsymbol{M} = (\boldsymbol{I} + \boldsymbol{A})(\boldsymbol{I} + \boldsymbol{D})^{-1}$. Adding an identity matrix to the adjacency matrix can avoid self-loops in graph. Therefore, the degree of each vertex is incremented by 1 to provide aperiodicity in the graph. Based on the above description, we propose a new biased transition probability matrix $(\boldsymbol{BM})$ in the next section.

3. Method

The random walk with restart (RWR)-based method fully uses the global topological information of PPI network. In our paper, first we process a variety of protein biological information, use it to judge the similarity between proteins in the network, and reconstruct the transition probability matrix, which makes the random walk biased, so as to better mine the global topological properties of proteins. Second, we use GO terms score and subcellular score to construct the initial probability vector of RWR. The overall process of the BRWR algorithm is shown in Fig. 1.

3.1. Similarity transition probability matrix

The traditional adjacency matrix is determined by considering whether there is an edge between vertices in the network and if a walker selects the neighbors randomly, which ignores the influence of functionally similar neighbors on vertex. Therefore, we use gene expression sequence and subcellular location information to redefine the transition probability matrix, which makes the walker tend to its more similar neighbors during random walks.

(i) Gene express similarity

The gene expression data of proteins were divided into three cycles, each characterized by 12 time points, which is denoted as $T = \{g_1, g_2, \ldots, g_{12}, \ldots, g_{24}, \ldots, g_{36}\}$, where $T(i)$ is the gene expression of a protein at time i ($i \in [1, 36]$). We use 3-σ principle to calculate the threshold, which can be used to determine whether a protein is active. The 3-σ formula for the threshold can be written as

$$SP(v) = \mu(v) + 3\sigma(v)\left(1 - \frac{1}{1 + \sigma^3(v)}\right), \tag{1}$$

where $\mu(v)$ and $\sigma(v)$ are the mean and variance of proteins gene expression value from time 1 to 36, respectively.

In order to compare the gene expression value of each time with the threshold, we calculate the mean value of the gene expression value at 36 time points in 3 cycles. According to the periodicity of gene expression data, the gene expression average value AT_I at each time point is of the three cycles and

can be calculated as follows:

$$AT_I = \frac{T(I)+T(I+12)+T(I+24)}{3} \quad (I \in [1,12]). \quad (2)$$

Then, if the average value AT_I of gene expression at time I is greater than the threshold SP, it is considered to be active at time I. For two adjacent proteins to be active at least at one same time point during the 12 time points, we assume that they are similar.

(ii) Subcellular location similarity

Moreover, in subcellular location information, which contains 11 subcellular information, including: cytoskeleton, plasma, nucleus, endosome, extracellular, golgi, mitochondrion, peroxisome, endoplasmic, vacuole, cytosol. If two adjacent proteins appear in the same subcellular, we assume that they have the same function. Thus they are similar.

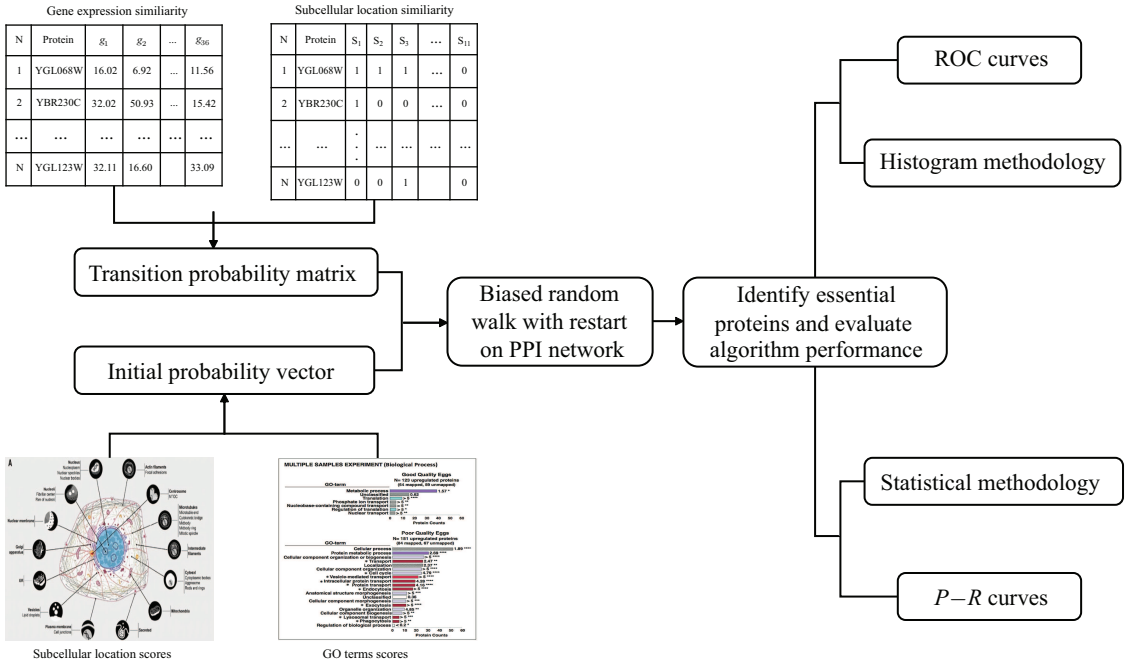

Fig. 1. The overall flow of BRWR.

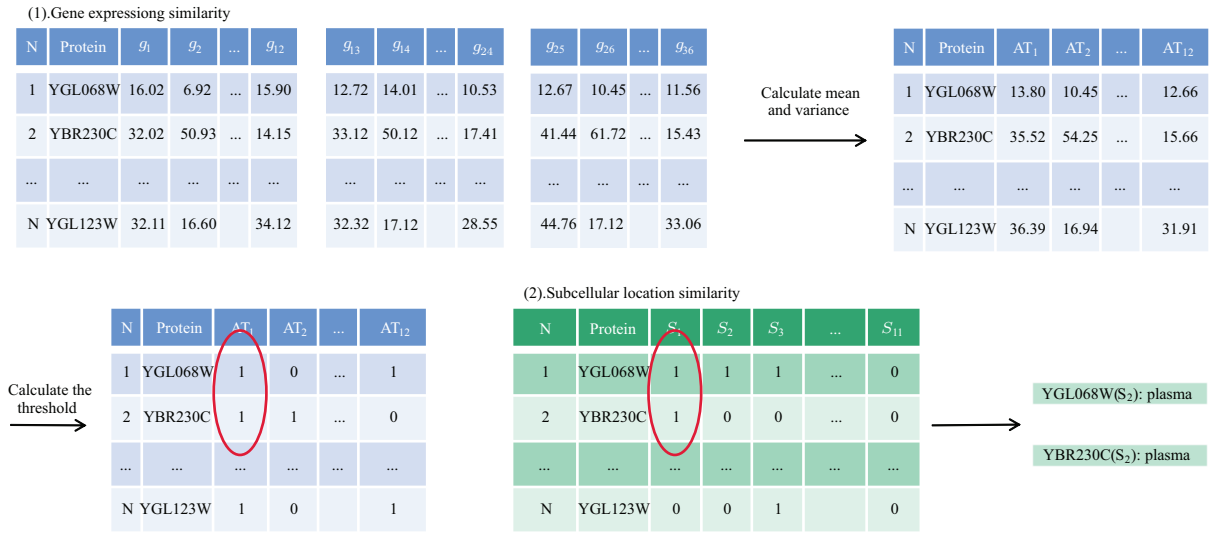

Fig. 2. The similarity of gene expression and subcellular location.

For example, as shown in Fig. 2, we assume that YGL068W and YBR230C are two interacting proteins. Both YGL068W and YBR230C are active at AT_3. Meanwhile, YGL068W and YBR230C are located in the same subcellular plasma. Thus YGL068W is similar to YBR230C. In this paper, we use the value of GS (gene-similar) to indicate whether two adjacent proteins have gene expression similarity. If $GS_{ij} = 1$, we assume that i and j have gene expression

similarity, and if $GS_{ij} = 0$, we assume that i and j have no gene expression similarity. Like gene expression similarity, we use SS (subcellular-similar) to denote the functional similarity of subcellular location. If $SS_{ij} = 1$, we assume that i and j have functional similarity, and if $SS_{ij} = 0$, we assume that i and j have no functional similarity. Finally, if two adjacent proteins are active at least one same point and have the same function, we will think that they are strong similar. If two adjacent proteins are active at least at one same point or have the same function, we will think that they are similar, If two adjacent proteins are active at different point and do not have same function, we will think that they are dissimilar. Based on the above assumptions, we can calculate the $\boldsymbol{SA}$ (similar adjacency matrix), $\boldsymbol{SD}$ (similar degree matrix) and $\boldsymbol{BM}$ (biased transition matrix). For instance, the similarities of gene expression and subcellular location are shown in Fig. 2.

Definition 1 The similar adjacency matrix of the network is denoted by $\boldsymbol{SA} \in \boldsymbol{R}^{n\times n}$, where sa_{ij} is defined as follows:

$$sa_{ij} = \begin{cases} 3a_{ij}, & \text{if GS = 1 and SS =1,} \\ 2a_{ij}, & \text{if (GS = 1 and SS = 0) or (GS = 0 and SS =1),} \\ a_{ij}, & \text{otherwise.} \end{cases} \tag{3}$$

Definition 2 The similar degree matrix is denoted by $\boldsymbol{SD} \in \boldsymbol{R}^{n\times n}$, where sd_{ij} is defined as follows:

$$sd_{ij} = \begin{cases} \sum_{j=1}^{n} sa_{ij}, & i = j, \\ 0, & i \neq j. \end{cases} \tag{4}$$

Definition 3 The transition probability matrix is denoted by $\boldsymbol{BM} \in \boldsymbol{R}^{n\times n}$, and the equation as follows:

$$\boldsymbol{BM} = (\boldsymbol{I} + \boldsymbol{SA})(\boldsymbol{I} + \boldsymbol{SD})^{-1}, \tag{5}$$

where each element of $\boldsymbol{BM}$ represents the probability of remaining in the current vertex or transiting to its neighbors.

3.2. Initial protein score vector

In this section, we take advantage of the initial protein scores to denote the initial probability vector of the random walk with restart. We use GO terms score and subcellular scores to calculate the initial protein scores.

(i) GO terms score (GOS)

GOS is a biological resource which describes the functional properties of genes. The more the same GO terms the two interacting proteins have, the more similar their functions will be. We obtain the relationship between proteins essentiality and GO terms by analyzing the correlation of GO terms between two interacting proteins. GOS is defined as the sum of the protein and its neighbors belong to the same cluster.

$$\mathrm{GOS}(v) = \sum_{j\in\Gamma(i)} \left(\frac{|GO_i \cap GO_j|^2}{|GO_i| \times |GO_j|} \right), \tag{6}$$

where GO_i and GO_j are the GO terms of proteins v_i and v_j, respectively; $|GO_i|$ and $|GO_j|$ are the numbers of GO terms annotating of proteins v_i and v_j, respectively; $|GO_i \cap GO_j|$ denotes the number of GO terms intersection of GO_i and GO_j.

(ii) Subcellular scores (SC)

There are 11 subcellular locations, and proteins in different subcellular locations have different functions. The number of proteins in different compartments is different, and the essential proteins in different compartments are also different. In addition, a protein can present in different compartments at the same time.[42] The value of SC is calculated by the ratio of $Sc(I)$ to Sc_{total},

$$SC(v_i) = \sum_{v\in I} \frac{Sc(I)}{Sc_{\text{total}}}, \tag{7}$$

where $Sc(I)$ ($I \in \{1,2,3,\ldots,11\}$) denotes the number of proteins appearing in the I_{th} subcellular location and I represents the 11 subcellular, Sc_{total} denotes the total number of proteins appearing in all subcellular locations. In this scores vector, since proteins in different networks are distributed differently in subcellulars, each network corresponds to one different SC.

Definition 4 The initial scores of proteins is denoted as $\boldsymbol{IP} \in \boldsymbol{R}^{n\times 1}$, each element represents the initial scores of a protein, and ip_{v_i} can be calculated by

$$ip_{v_i} = SC(v_i) \cdot \mathrm{GOS}(v_i). \tag{8}$$

3.3. Essential proteins prediction based on biased random walk with restart

RWR is a significant prioritization algorithm,[43] which has been used for gene as well as protein complex prioritization in previous studies.[44,45] In the RWR algorithm, random walks start from the seed vertex and move to the direct neighbors or get back to the seed vertex randomly. RWR can be denoted as the following formula:

$$\boldsymbol{P}^{t+1} = (1-\alpha) \cdot \boldsymbol{M} \cdot \boldsymbol{P}^{t} + \alpha \cdot \boldsymbol{P}^{0}, \tag{9}$$

where $\boldsymbol{P}^0$ is the initial probability vector; $\boldsymbol{P}^t$ is a probability vector to reach each vertex at step t ($\boldsymbol{P} \in R^{n\times 1}$); α is the restart probability; and $\boldsymbol{M}$ ($\boldsymbol{M} \in \boldsymbol{R}^{n\times n}$) is the transition matrix, in which M_{ij} denotes the probability from vertex v_i transit to v_j. Eventually, the process gets to steady state until condition $||P^{t+1} - P^t|| < \varepsilon$ holds.

In a random walk, the transition probability of walker from the current vertex to the next vertex is equal. Due to

the fact the interaction relationship between two proteins in the protein network is not generated randomly, the transition process will be affected to a certain extent. The protein itself has complex biological characteristics, therefore transiting between vertices is not necessarily an equal probability movement, but a biased movement. Based on the above problems, we use the proposed $\boldsymbol{BM}$ as the transition probability matrix. The global algorithm of the RWR model is used to identify essential proteins, and the ranking scores of proteins can be calculated by

$$\boldsymbol{BR}^{t+1} = (1-\alpha)\cdot \boldsymbol{BM}\cdot \boldsymbol{BR}^{t} + \alpha \cdot \boldsymbol{IP}, \tag{10}$$

where $\boldsymbol{BR}^{\mathrm{T}} = (BR(v_1), BR(v_2), \ldots BR(v_n))$, until ε meets the preset conditions, the BRWR process reaches a steady state. In this paper, we assume $\varepsilon = 10^{-6}$. Algorithm 1 gives the description of BRWR.

Algorithm 1 The algorithm of the BRWR.

Input:
The data of PPI network $G=(V,E)$;
The data of protein complexes $C=C_i(V(C_i),E(C_i))|C_i \subset G$;
The data of GO terms $GT=(V,g)$;
The data of subcellular location information $SC=(V,s)$.
Output: The sorted value of **BR** after reaching steady state;
for each $e(v_i,v_j)\in E$ in PPI **do**
    Calculate the value of $a_{i,j}$ by **A**;
    Calculate the value of $sa_{i,j}$ by Eq. (3);
    Calculate the value of $sd_{i,j}$ by Eq. (4);
end for
Calculate the similarity transition probability matrix **BM** by Eq. (5);
for each $v_i \in G$ **do**
    Calculate the GOS scores of each protein by Eq. (6);
    Calculate the SC scores of each protein by Eq. (7);
    Calculate the vector **IP** by Eq. (8).
end for
Initialize the vector $\mathbf{IP}=(ip_{v1},ip_{v2},\ldots,ip_{vn})$, set $\alpha = 0$;
Initialize the vector $\mathbf{BR}^t=(1,1,\ldots,1)$;
while do$||\mathbf{BR}^{t+1}-\mathbf{BR}^t|| \geq \epsilon$
    Compute $\mathbf{BR}^{t+1}$ by Eq. (10);
end while
repeat
    step 12, set $\alpha = \alpha + 0.1$;
until $\alpha = 1$
Sequence proteins according to each elements of **BR** values that reached steady state;
return **BR**;

4. Datasets and evaluation settings

4.1. Datasets

In order to evaluate the performance of the algorithm BRWR, we consider the PPI data of saccharomyces cerevisiae (yeast) protein as one of experimental materials, because this organism has relative complete, reliable PPI and essential proteins data. We use four sets of PPI network data, including YHQ,[46] BioGRID,[47] Kroagn,[48] and Gavin.[49] The Kroagn, BioGRID and Gavin data are gathered from the BioGRID database,[47] the YHQ data was constructed by Yu *et al.*[46] After removing the multiple edges and self-interactions, the properties of the network are given in Table 1. The standard essential proteins data were gathered from four different databases: MIPS,[50] SGD,[51] DEG,[52] and SGDP.[2] The gene expression data were downloaded from GEO (gene expression omnibus) database.[53] The subcellular location data were downloaded from COMPARTMENTS database.[54] This data contains 11 subcellular location information. The GO terms data were cut-down version of the GO ontologies, available at (https://www.yeastgenome.org/).[55]

Table 1. Data details of the three protein networks: YDIP, YHQ, and Krogan, from BioGRID.

Dataset	Proteins	Interactions	Essential proteins
BioGRID	5616	52833	1199
YHQ	4743	23294	1108
Krogan	2674	7075	784
Gavin	1430	6531	617

4.2. Evaluation settings

We compare BRWR with a number of existing methods. The proteins are ranked by the essentiality predicted by each method. Then, we select the top 25 percent proteins in the obtained sequence as candidate essential proteins and the remaining 75 percent were selected as candidates non-essential proteins. By comparing the selected top 25 percent proteins with the standard essential proteins dataset, we can get the number of candidate essential proteins that are truly identified as the

essential proteins. True positive (TP) is the number of candidate essential proteins correctly identified as essential proteins. False negative (FN) is the number of candidate essential proteins that were incorrectly identified as non-essential proteins. False positive (FP) is the number of candidate non-essential proteins that were misidentified as essential proteins. True negative (TN) denotes the number of candidate non-essential proteins correctly identified as non-essential proteins.

5. Results

5.1. Parameter analysis

In our method of BRWR, the adjustment of parameter will affect the performance of BRWR, there are three parameters: α, k and t, where α is related to the accuracy of the prediction results. To investigate the effect of parameter α, we evaluate the prediction accuracy by setting values of α range from 0.1 to 1; k is the top k percent of ranked proteins; t is the iterations times of biased random walk in which the process reaches steady state. As shown in Table 2, the iterations time t decreases when α increases. Through comparison, when α is 0.5, 0.8, 0.6 and 0.7, respectively, the algorithm in networks YHQ, BioGRID, Kroagn and Gavin has best performance.

Table 2. Number of essential proteins predicted by BRWR for different α.

Dataset	α	k						t
		1%	5%	10%	15%	20%	25%	
BioGRID	0	37	156	308	426	530	630	630
	0.1	37	170	310	447	550	660	117
	0.2	36	169	319	460	561	670	67
	0.3	36	166	324	463	571	671	45
	0.4	37	167	332	465	581	671	33
	0.5	37	165	337	470	582	667	26
	0.6	38	167	340	475	588	665	20
	0.7	38	168	341	476	590	665	16
	0.8	**37**	**169**	**344**	**475**	**593**	**669**	**12**
	0.9	38	173	347	477	590	664	9
	1	38	175	350	480	588	665	1
YHQ	0	8	107	264	377	469	555	920
	0.1	14	111	254	387	494	576	134
	0.2	16	115	258	391	498	582	73
	0.3	22	115	261	391	505	578	48
	0.4	26	117	262	394	506	582	35
	0.5	**27**	**122**	**266**	**395**	**508**	**581**	**27**
	0.6	26	127	267	397	505	578	21
	0.7	26	127	272	398	503	577	17
	0.8	26	131	273	404	503	577	13
	0.9	25	132	274	405	504	577	10
	1	28	156	272	396	502	567	1
Krogan	0	18	84	165	229	279	324	262
	0.1	25	95	177	249	312	353	90
	0.2	25	99	183	255	313	359	51
	0.3	25	102	184	257	314	362	37
	0.4	25	103	186	257	317	365	28
	0.5	24	106	187	262	316	364	23
	0.6	**24**	**105**	**186**	**261**	**315**	**365**	**18**
	0.7	24	105	185	261	315	365	15
	0.8	24	105	186	258	313	363	12
	0.9	24	104	188	255	312	363	9
	1	24	104	190	258	316	362	1
Gavin	0	13	61	105	152	196	237	366
	0.1	13	64	113	165	205	252	104
	0.2	13	64	117	164	212	256	61
	0.3	13	65	118	165	214	254	42
	0.4	13	65	118	167	219	257	32
	0.5	14	64	121	167	220	258	24
	0.6	14	64	121	169	220	259	19
	0.7	**14**	**64**	**122**	**171**	**221**	**259**	**15**
	0.8	14	64	122	171	220	258	12
	0.9	14	64	122	173	222	257	9
	1	14	64	122	172	222	256	1

5.2. Validated by histograms

To verify the performance of BRWR, we compare it with other algorithms (BC, DC, SC, EC, LAC, UC, NC, PeC, Wdc) by histograms. First, we calculate the score according to each method, and rank the proteins in descending order. Then the top 1%, 5%, 10%, 15%, 20% and 25% proteins are selected as candidate proteins. Finally, the number of essential proteins in these candidate essential proteins is determined based on the standard data set of known essential proteins. The comparison results are shown in Figs. 3–6.

Figure 3 shows the prediction results of each method on the BioGRID dataset. DC, BC, EC, SC, LAC, NC, UC, PeC and Wdc find out 598, 539, 324, 525, 621, 636, 532, 558 and 621 essential proteins at the top 25% candidates, respectively. Our algorithm BRWR discovers 669 essential proteins. Compared with other algorithms, the number of essential proteins predicted by BRWR on BioGRID dataset is significantly higher.

Figure 4 shows the prediction results of each method on the YHQ dataset. DC, BC, EC, SC, LAC, NC, UC, PeC and Wdc find out 539, 548, 511, 479, 542, 538, 518, 491 and 538 essential proteins at the top 25% candidates, respectively. Our algorithm BRWR discovers 581 essential proteins. Compared with other algorithms, the number of essential proteins predicted by BRWR on YHQ dataset is significantly higher.

Figure 5 shows the prediction results of each method on the Krogan dataset. DC, BC, EC, SC, LAC, NC, UC, PeC and Wdc find out 318, 248, 285, 284, 326, 325, 319, 321 and 333 essential proteins at the top 25% candidates, respectively. Our algorithm BRWR discovers 365 essential proteins. Compared with other algorithms, the number of essential proteins predicted by BRWR on Krogan dataset is significantly higher.

Figure 6 shows the prediction results of each method on the Gavin dataset. DC, BC, EC, SC, LAC, NC, UC, PeC and Wdc find out 221, 172, 157, 210, 254, 252, 232, 234 and 247 essential proteins at the top 25% candidates, respectively. Our algorithm BRWR discovers 259 essential proteins. Compared with other algorithms, the number of essential proteins predicted by BRWR on Gavin dataset is significantly higher.

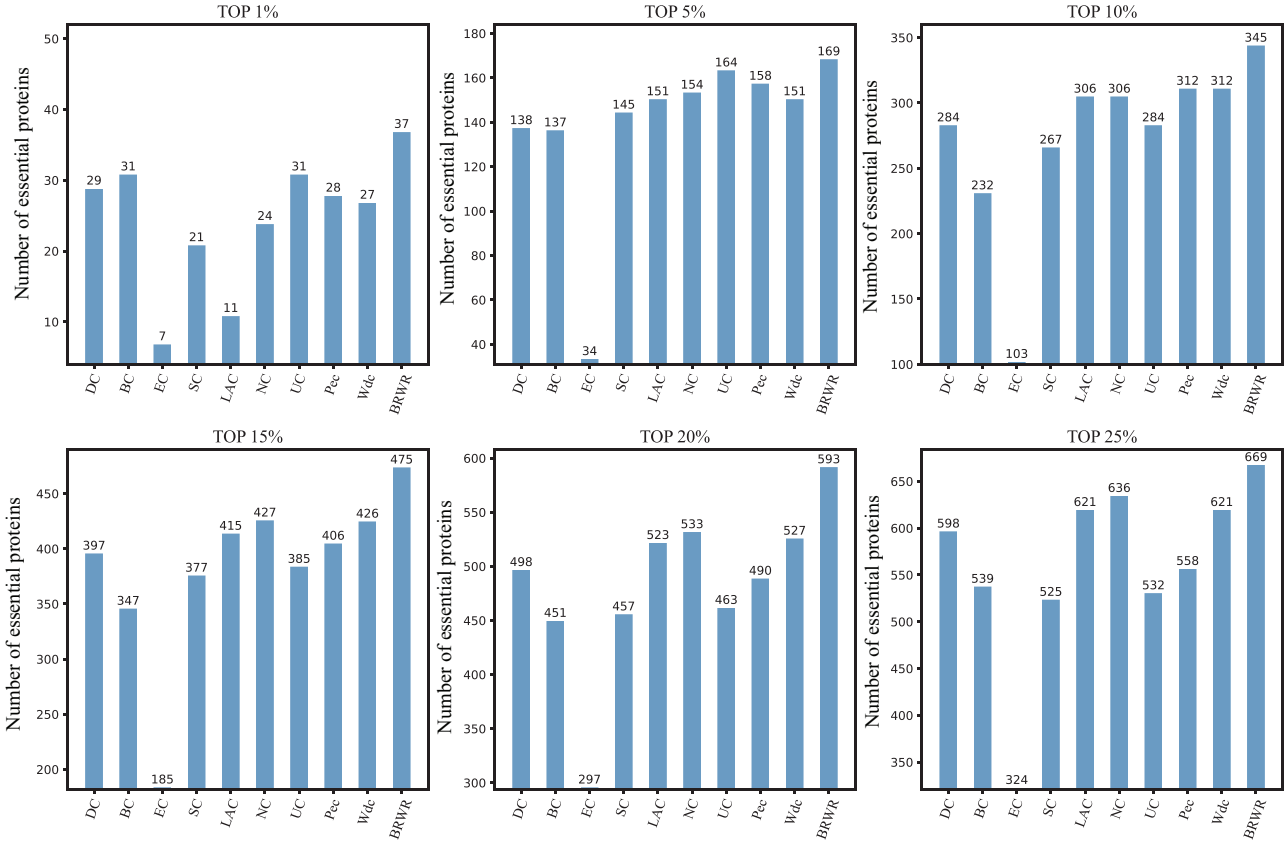

Fig. 3. Comparison of the number of essential proteins detected by BRWR and other nine previous centrality measures from the BioGRID network.

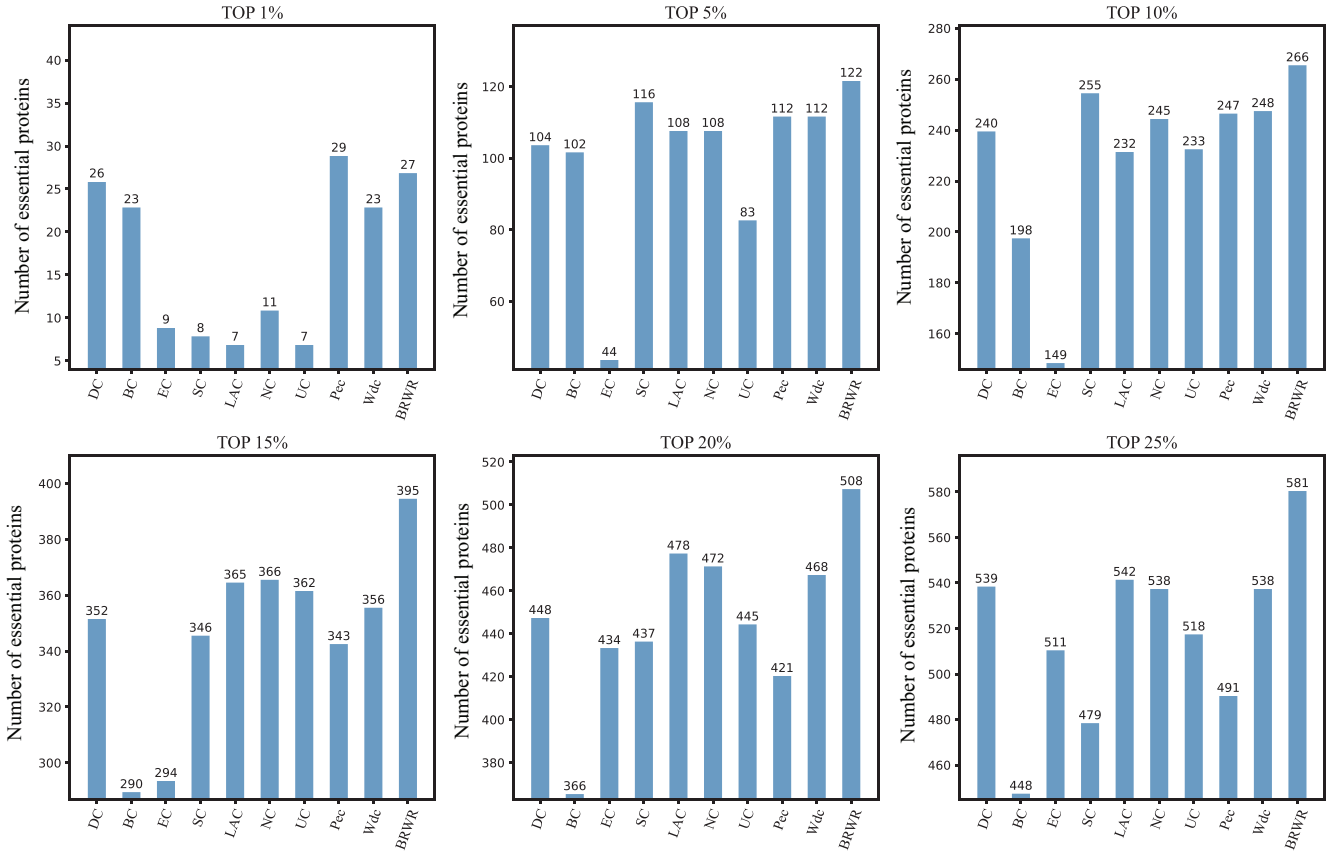

Fig. 4. Comparison of the number of essential proteins detected by BRWR and other nine previous centrality measures from the YHQ network.

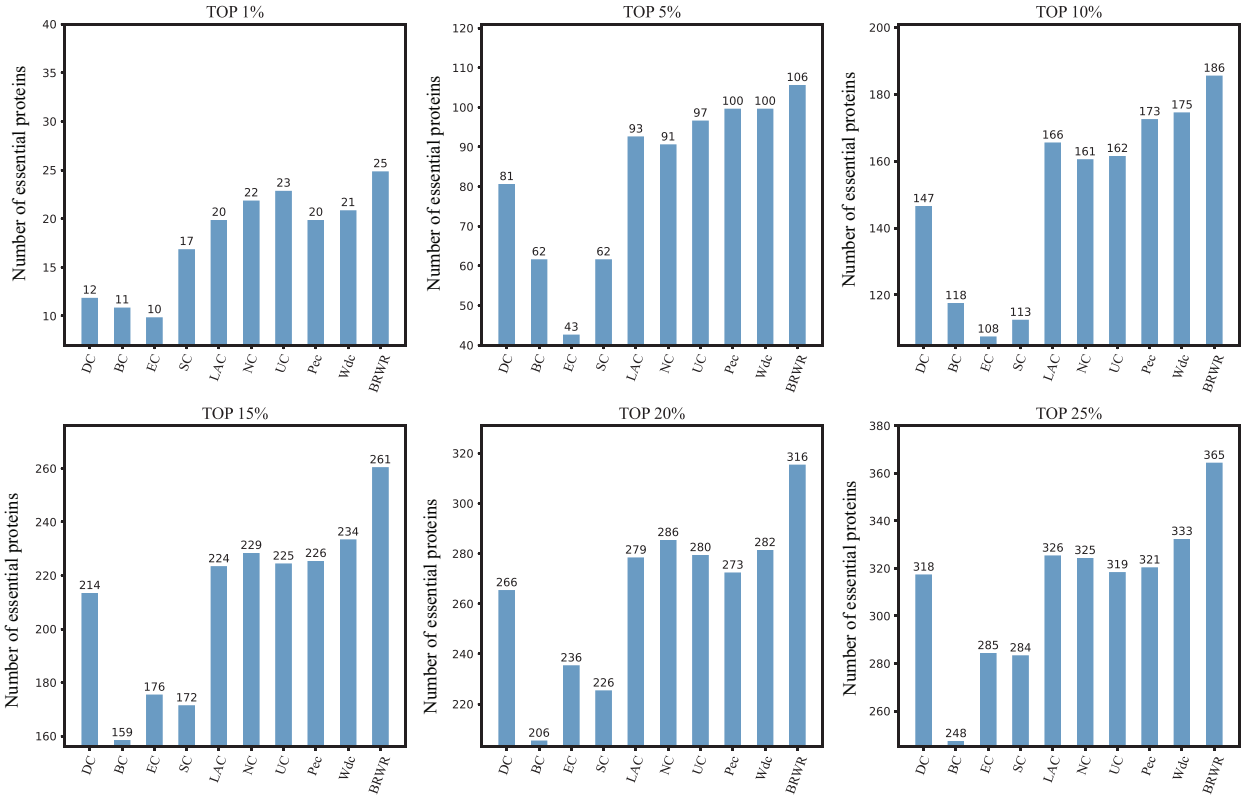

Fig. 5. Comparison of the number of essential proteins detected by BRWR and other nine previous centrality measures from the Krogan network.

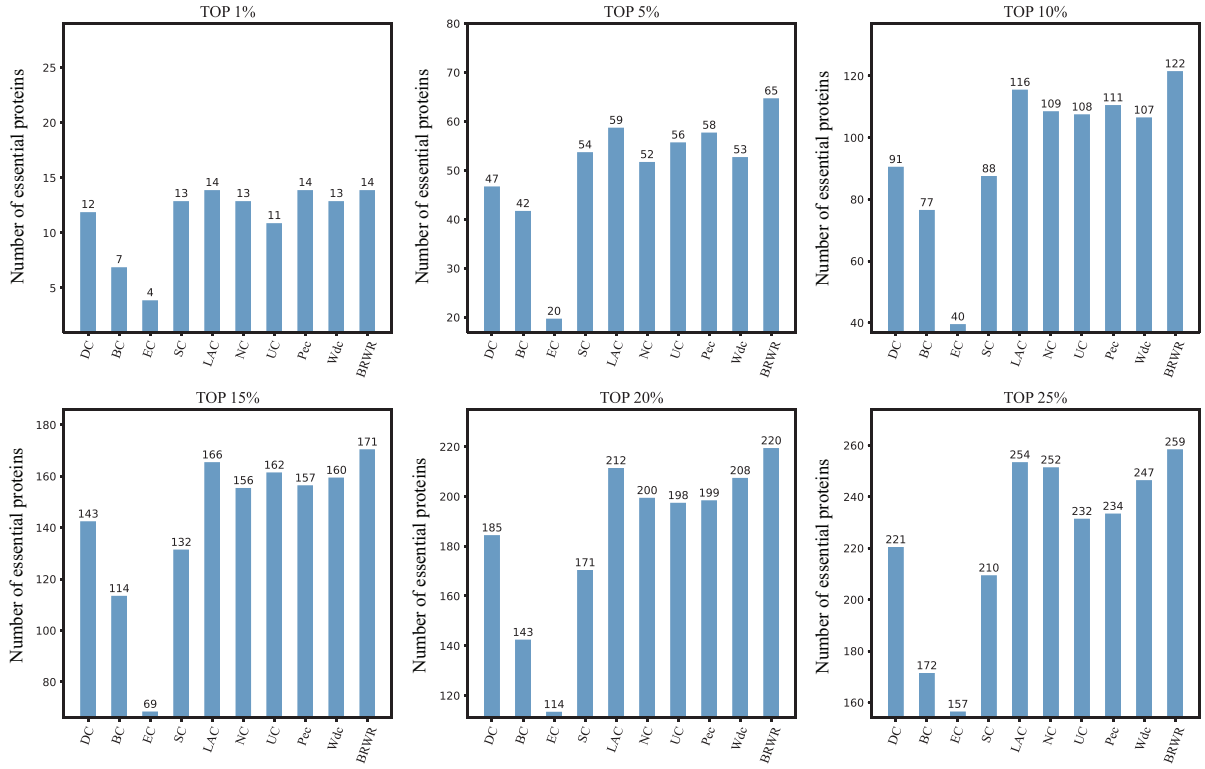

Fig. 6. Comparison of the number of essential proteins detected by BRWR and other nine previous centrality measures from the Gavin network.

5.3. Validated by six statistical methods

Six statistical methods are a more comprehensive performance evaluation method, which are commonly used in classification tasks. To evaluate the performance of BRWR, we compare each method by using six statistical measures including accuracy: $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$, F-measure: $F = \frac{2\mathrm{SN}\cdot\mathrm{PPV}}{\mathrm{SN}+\mathrm{PPV}}$, positive predictive value: $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$, negative predictive value: $\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$, specificity: $\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$, and sensitivity:

$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$.

Table 3. Comparison of results of SN, SP, PPV, NPV, *F*-measure and ACC.

Dataset	Methods	SN	SP	PPV	NPV	*F*-measure	ACC
BioGRID	DC	0.499	0.804	0.431	0.843	0.427	0.734
	BC	0.450	0.790	0.389	0.828	0.417	0.712
	EC	0.270	0.753	0.246	0.776	0.257	0.642
	SC	0.438	0.788	0.381	0.825	0.407	0.708
	LAC	0.518	0.816	0.456	0.850	0.485	0.748
	NC	0.530	0.814	0.460	0.853	0.493	0.749
	UC	0.443	0.798	0.395	0.828	0.418	0.717
	Pec	0.465	0.794	0.402	0.833	0.431	0.718
	Wdc	0.518	0.810	0.448	0.849	0.480	0.743
	BRWR	**0.558**	**0.822**	**0.482**	**0.862**	**0.517**	**0.761**
YHQ	DC	0.486	0.812	0.469	0.823	0.477	0.729
	BC	0.404	0.786	0.392	0.795	0.398	0.689
	EC	0.461	0.812	0.456	0.816	0.459	0.723
	SC	0.432	0.808	0.435	0.807	0.433	0.713
	LAC	0.489	0.815	0.474	0.824	0.482	0.732
	NC	0.486	0.813	0.470	0.823	0.478	0.730
	UC	0.467	0.810	0.456	0.817	0.461	0.723
	Pec	0.442	0.790	0.418	0.806	0.429	0.702
	Wdc	0.486	0.812	0.468	0.823	0.477	0.729
	BRWR	**0.524**	**0.822**	**0.500**	**0.835**	**0.512**	**0.746**
Krogan	DC	0.406	0.814	0.480	0.763	0.440	0.692
	BC	0.316	0.777	0.376	0.728	0.344	0.640
	EC	0.364	0.796	0.431	0.747	0.394	0.667
	SC	0.362	0.795	0.429	0.746	0.393	0.666
	LAC	0.416	0.817	0.491	0.767	0.450	0.697
	NC	0.415	0.817	0.490	0.768	0.449	0.697
	UC	0.407	0.813	0.480	0.763	0.440	0.692
	Pec	0.409	0.816	0.486	0.765	0.444	0.695
	Wdc	0.425	0.822	0.504	0.771	0.461	0.704
	BRWR	**0.466**	**0.839**	**0.551**	**0.787**	**0.504**	**0.728**
Gavin	DC	0.357	0.834	0.623	0.628	0.454	0.626
	BC	0.277	0.773	0.484	0.582	0.353	0.557
	EC	0.254	0.756	0.445	0.568	0.324	0.538
	SC	0.339	0.820	0.592	0.617	0.431	0.611
	LAC	0.412	0.873	0.713	0.659	0.522	0.672
	NC	0.407	0.869	0.705	0.656	0.516	0.668
	UC	0.374	0.845	0.651	0.637	0.475	0.641
	Pec	0.399	0.863	0.691	0.651	0.506	0.661
	Wdc	0.337	0.903	0.730	0.639	0.461	0.658
	BRWR	**0.420**	**0.880**	**0.730**	**0.664**	**0.533**	**0.680**

The six statistical indicators of each method are calculated on BioGRID, YHQ, Krogan and Gavin datasets. As shown in Table 3, the six index values of our algorithm BRWR are all higher than those of the compared algorithms. Especially, the ACC values on BioGRID, YHQ, Krogan and Gavin datastes are 0.761, 0.746, 0.728, 0.680, respectively.

5.4. Validated by the *P*–*R* curve

In essence, our study is a classic unbalanced dichotomy problem, that is, the proteins in the PPI network are divided into essential proteins and non-essential proteins. The *P*–*R* curve is a performance evaluation method for binary-classification problems in machine learning, which can easily assess the performance of the classification ability of BRWR. Thus we use it to evaluate our algorithm. The *x*-axis denotes recall ($\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$) and the *y*-axis denotes Precision ($\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$). The result shown in Fig. 7, the *P*–*R* curves show that there is a superior performance for essential proteins in comparison of BRWR with other algorithms.

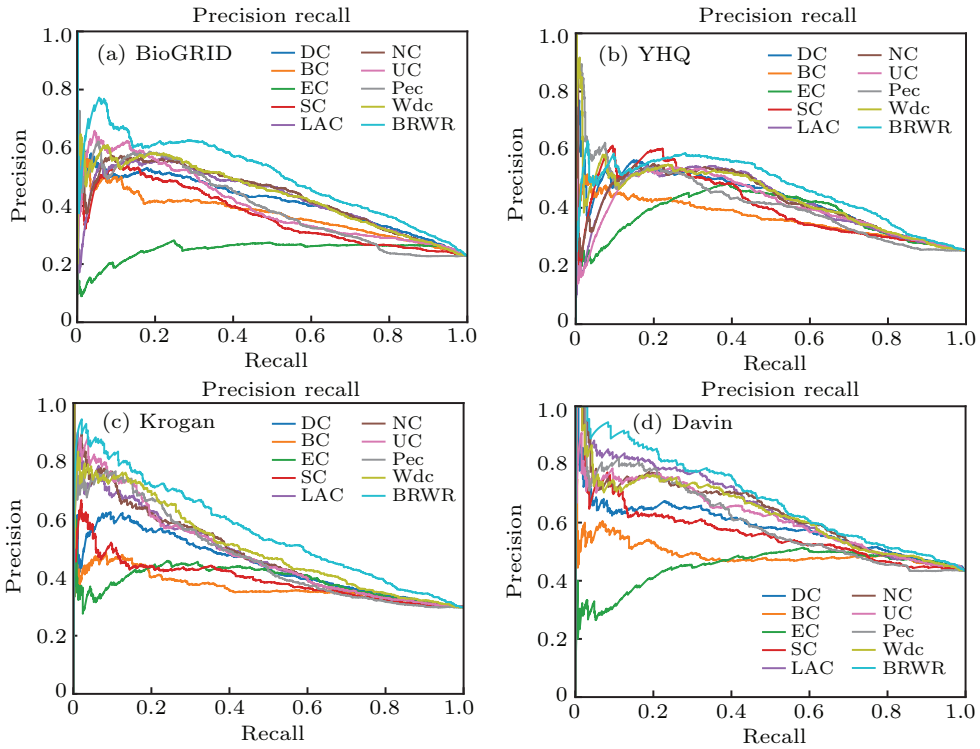

Fig. 7. The performances of BRWR and other nine centrality measures on the BioGRID, YHQ, Krogan and Gavin datasets, validated by *P*–*R* curves.

5.5. Validated by the ROC curve

In addition, the ROC curve is a classic evaluation method. The *x*-axis represents the false positive rate, the *y*-axis represents the true positive rate, and the area of ROC reflects the quality of the algorithm. We describe the ROC curve, which also shows that our algorithm is better than several other algorithms. The result is shown in Fig. 8.

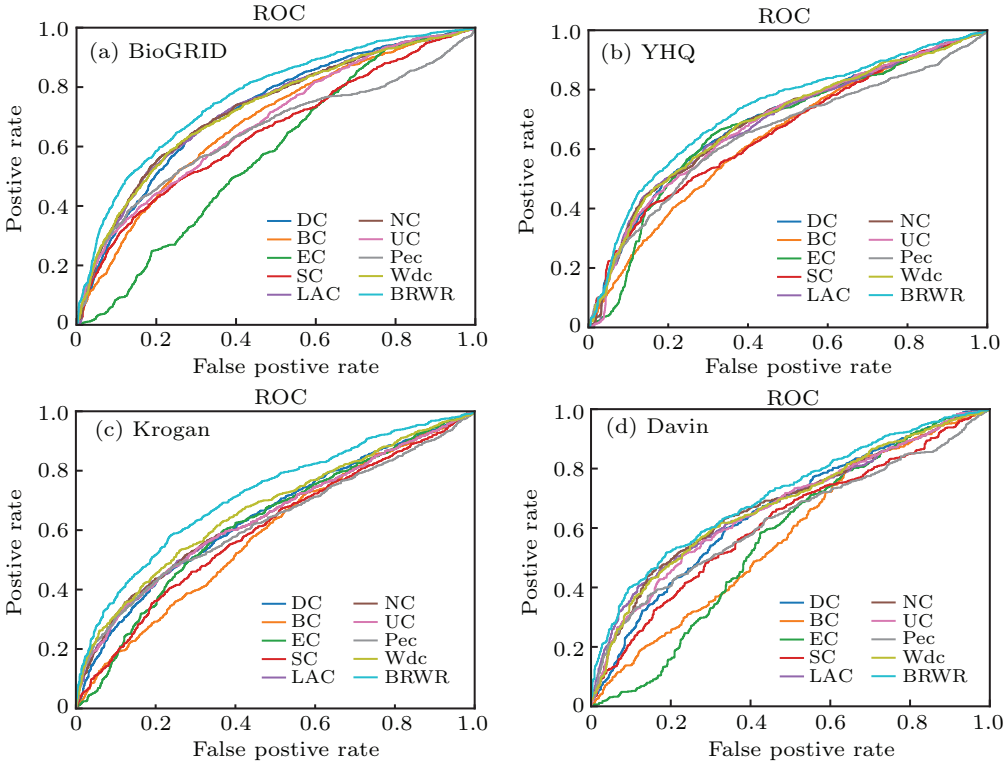

Fig. 8. The performances of BRWR and other nine centrality measures on the BioGRID, YHQ, Krogan and Gavin datasets, validated by ROC curves.

6. Conclusion

The prediction of essential proteins is an indispensable research for us to know the organisms survival and evolution. Up to date, many methods have been proposed for predicting essential proteins. However, it is still of challenge to improve the prediction accuracy. In this paper, we propose a new algorithm BRWR based on the RWR model for prediction of essential proteins. Firstly, the adjacency matrix is reconstructed by using gene expression sequence and subcellular location information, and named as similarity adjacency matrix. The similarity adjacency matrix is used to construct a biased transition probability matrix, which makes the process of random walk biased. In addition, the subcellular scores are fused with the GO terms information to construct the initialization probability vector in the BRWR. Experimental results show that our proposed method has higher accuracy and stable performance in predicting essential proteins. The improvements of BRWR in terms of the average ACC results range in 1.4%–5.7%, 1.3%–11.9%, 2.4%–8.8%, and 0.8%–14.2%, respectively.

Acknowledgement

Project supported by the National Natural Science Foundation of China (Grant Nos. 11861045 and 62162040).

References

[1] Kamath R, Fraser A, Dong Y, *et al*. 2003 *Nature* **421** 231
[2] Winzeler E, Shoemaker D, Astromoff A, *et al*. 1999 *Science* **285** 901
[3] Jeong H and Mason S 2001 *Nature* **411** 41
[4] Gerardo J and Childs B 2001 *Nature* **409** 853
[5] Nivit G, Shailendra S, Trilok A, *et al*. 2014 *J. Comput. Biol.* **21** 456
[6] Giaever G, Chu A, Ni L, *et al*. 2002 *Nature* **418** 387
[7] Cullen L, Arndt G, *et al*. 2005 *Immunol. Cell Biol.* **83** 217
[8] Roemer T, Jiang B, Davison J, *et al*. 2003 *Mol. Microbiol.* **50** 1
[9] Acencio M L and Lemke N 2009 *BMC Bioinform.* **10** 290
[10] Karthik R, Nandita D and Govind K J 2014 *Syst. Synth. Biol.* **8** 73
[11] Freeman L C 1978 *Soc. Networks* **1** 215
[12] Joy M, Brock A, Ingber D, *et al*. 2005 *J. Biotechnol.* **2005** 594674
[13] Estrada E and Juan A 2005 *Physica A* **364** 581
[14] Bonacich P 1987 *Am. J. Sociol.* **92** 1170
[15] Li M, Wang J X, Chen X, *et al*. 2011 *Comput. Biol. Chem.* **35** 143
[16] Li M, Wang J X, Wang H, *et al*. 2012 *IEEE ACM Trans. Comput. Biol. Bioinform.* **9** 1070
[17] Wang K L, Wu C X, Ai J, *et al*. 2019 *Acta Phys. Sin.* **68** 196402 (in Chinese)
[18] Huang L Y, Huo Y L, Wang Q, *et al*. 2019 *Acta Phys. Sin.* **68** 018901 (in Chinese)
[19] Wuchty S and Stadler P 2003 *J. Theor. Biol.* **223** 45
[20] Hsing M, Byler K G and Cherkasov A 2008 *BMC Syst. Biol.* **2** 80

[21] Li M, Ni P and Chen X 2017 *IEEE Trans. Comput. Biol. Bioinformat.* **16** 1386

[22] Li M, Zhang H and Wang J X 2012 *BMC Syst. Biol.* **6** 15

[23] Xiao Q H, Wang J X and Peng X, *et al*. 2015 *BMC Genom.* **16** (Suppl. 3) S1

[24] Li M and Lu Y 2017 *IEEE Trans. Comput. Biol. Bioinformat.* **14** 380

[25] Zhang W, Xu J, Li X, *et al*. 2016 *IEEE Trans. Nanobiosci.* **15** 939

[26] Luo J W and Qi Y 2015 *PLoS One* **10** e0131418

[27] Tang X, Wang J, Zhong J and Pan Y 2014 *IEEE Trans. Comput. Biol. Bioinformat.* **11** 407

[28] Peng X, Wang J, Wu F X and Pan Y 2015 *PLoS One* **10** e0130743

[29] Zhou Y, Wu C and Tan L 2021 *Physica A* **570** 125783

[30] Park H, Jung J and Kang U 2017 *IEEE International Conference on Big Data*, March 10–12, 2017, Beijing, China

[31] Jung J, Jin W, Sael L and Kang U 2016 *IEEE 16th International Conference on Data Mining (ICDM)*, December 12–15, 2016, Barcelona, Spain, p. 973

[32] Zhou H J 2003 *Phys. Rev. E* **67** 061901

[33] Zhou H J and Lipowsky R 2004 *Lecture Notes in Computer Science* (Berlin: Springer) Vol. 3038 pp. 1062–1069

[34] Bahadori S, Moradi P and Zare H 2020 *Appl. Intell.* **51** 3561

[35] Bestehorn M, Riascos P and Michelitsch M 2021 *Continuum Mechanics and Thermodynamics* **33** 1027

[36] Xu B, Guan J H, Wang Y and Wang Z W 2017 *IEEE Trans. Comput. Biol. Bioinformat.* **16** 377

[37] Lv L S, Bardou D, Hu P, Liu Y Q and Yu G H 2022 *Chaos Solitons Fractals* **159**

[38] Lovász L 2004 *Lecture Notes in Mathematics* (Berlin: Springer) Vol. 2 pp. 1–46

[39] Liu W, Ma L and Chen L 2020 *J. Theor. Biol.* **504** 110414

[40] Zhu Z Q, Jin X L and Huang Z L 2012 *Chin. Phys. Lett.* **29** 038901

[41] Lin L, Xu X, Ping H, *et al*. 2013 *IEEE 10th Web Information System and Application Conference*, November 10–15, 2013, Yangzhou, China, p. 281

[42] Lei X, Zhao J, Fujita H, *et al*. 2018 *Knowl. Based Syst.* **151** 136

[43] Tong H H, Faloutsos C and Pan J Y 2007 *IEEE Sixth International Conference on Data Mining (ICDM'06)*, June 25–28, 2006, Las Vegas, USA, pp. 613–622

[44] Razaghi-Moghadam Z, Abdollahi R, Goliaei S, *et al*. 2016 *J. Biomed. Inform.* **64** 139

[45] Liu Z and Luo J 2017 *Comput. Biol. Chem.* **69** 41

[46] Yu H, Greenbaum D, Lu H, Zhu X and Gerstein M 2004 *Trends Genet.* **20** 227

[47] Stark C, Breitkreutz B J, Chatr-aryamontri A, *et al*. 2011 *Nucleic Acids Res.* **39** 698

[48] Krogan N, Cagney G, Yu H, *et al*. 2006 *Nature* **440** 637

[49] Gavin A C, Aloy P, Grandi P, *et al*. 2006 *Nature* **440** 631

[50] Mewes H, Frishman D, Mayer K, *et al*. 2006 *Nucleic Acids Res.* **34** 169

[51] Cherry J, Adler C, Ball C, *et al*. 1998 *Nucleic Acids Res.* **26** 73

[52] Zhang R and Lin Y 2009 *Nucleic Acids Res.* **37** 455

[53] Tu B P, Kudlicki A, Rowicka M and McKnight S L 2005 *Science* **310** 1152

[54] Binder J X, Sune P F, Kalliopi T, *et al*. 2014 *J. Biol. Databases Curation* **2014** bau012

[55] Lei X, Yang X Q and Schreiber G 2018 *PLoS One* **13** e0198998