Inferring Candidate CircRNA-Disease Associations by Bi-random Walk Based on CircRNA Regulatory Similarity

Chunyan Fan[1], Xiujuan Lei[1(✉)], and Ying Tan[2]

[1] School of Computer Science, Shaanxi Normal University, Xi'an 710119, China
xjlei@snnu.edu.cn

[2] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China

Abstract. Identification of associations between circular RNAs (circRNA) and diseases has become a hot topic, which is beneficial for researchers to understand the disease mechanism. However, traditional biological experiments are expensive and time-consuming. In this study, we proposed a novel method named BWHCDA, which applied bi-random walk algorithm on the heterogeneous network for predicting circRNA-disease associations. First, circRNA regulatory similarity is measured based on circRNA-miRNA interactions, and circRNA similarity is calculated by the average of circRNA regulatory similarity and Gaussian interaction profiles (GIP) kernel similarity for circRNAs. Similarly, disease similarity is the mean of disease semantic similarity and GIP kernel similarity for diseases. Then, the heterogeneous network is constructed by integrating circRNA network, disease network via circRNA-disease associations. Subsequently, the bi-random walk algorithm is implemented on the heterogeneous network to predict circRNA-disease associations. Finally, we utilize leave-one-out cross validation and 10-fold cross validation frameworks to evaluate the prediction performance of BWHCDA method and obtain AUC of 0.9334 and 0.8764 ± 0.0038, respectively. Moreover, the predicted hsa_circ_0000519-gastric cancer association is analyzed. Results show that BWHCDA could be an effective resource for clinical experimental guidance.

Keywords: CircRNA-disease associations · Bi-random walk · CircRNA regulatory similarity

1 Introduction

Circular RNAs (CircRNAs) are a type of non-coding RNAs with closed loop structures formed by back splicing [1]. Recently, large number of circRNAs are widely found in various livings [2], and they could regulate gene expression at transcriptional or post-transcriptional levels by titrating microRNAs (miRNAs) [3], regulating transcription and splicing [4, 5], even several circRNAs could translate to produce polypeptides [6]. Increasing researches have demonstrated that the mis-regulation of circRNAs may cause

Y. Tan et al. (Eds.): ICSI 2020, LNCS 12145, pp. 485–494, 2020.
https://doi.org/10.1007/978-3-030-53956-6_44

abnormal cellular functions and associated with various diseases [7]. Thus, disease-associated circRNAs are becoming a class of promising biomarkers for disease diagnosis and treatment.

However, it is costly and laborious to identify the disease-related circRNAs with biomedical experiments. Recently, several computational approaches have been developed. Lei *et al*. [8] firstly designed a path weighted approach named PWCDA to predict circRNA-disease associations. Likewise, KATZHCDA [9] is developed based on KATZ model to measure the probability for each pair of circRNA-disease associations, in which the circRNA expression similarity and disease phenotype similarity matrices are used as priori knowledge to establish the circRNA-disease heterogeneous network. DWNN-RLS [10] is designed based on Kronecker regularized least squares to predict the associations between circRNAs and diseases. iCircDA-MF [11] is developed based on non-negative matrix factorization by integrating the circRNA-gene, gene-disease and circRNA-disease relationships. Wang *et al*. [12] utilized a recommendation algorithm PersonalRank to measure the relevance between circRNAs and diseases based on circRNA expression profiles and functional similarity. Although several methods have developed for the circRNA-disease association prediction, it is still a challenge to obtain sufficiently accurate results.

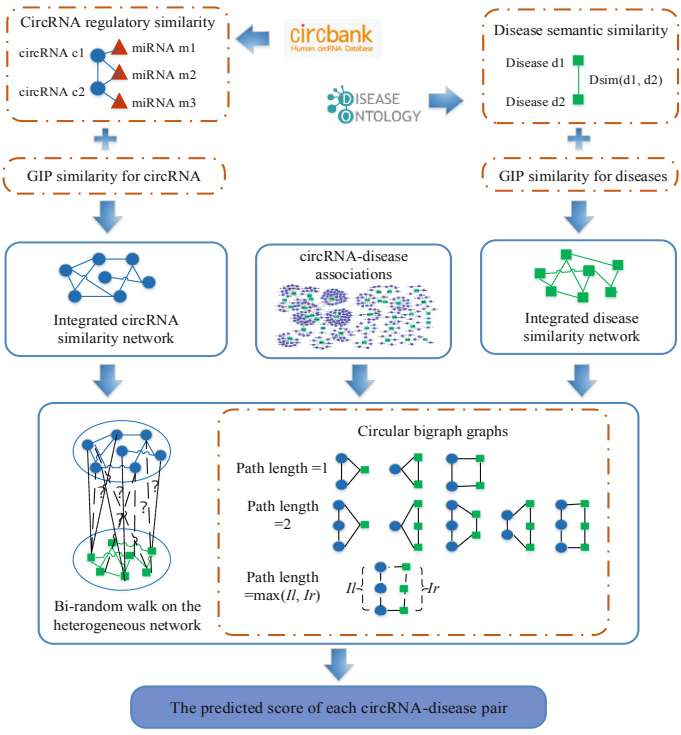

Fig. 1. The flowchart of BWHCDA method.

In this study, we developed a novel framework for forecasting circRNA-disease associations named BWHCDA, which integrated multiple similarity measures and implemented bi-random walk algorithm (Fig. 1). First, circRNA regulatory similarity is effective measured based on circRNAs may play essential roles in regulating miRNA function in disease occurrence and progression. Moreover, combined with Gaussian interaction profiles (GIP) kernel similarity for circRNAs, the integrated circRNA similarity is effectively measured. Similarly, disease similarity is denoted as the average of disease sematic similarity and GIP kernel similarity for diseases. Subsequently, the heterogeneous network is constructed by combing the circRNA network, disease network and circRNA-disease associations. Then, circular bigraph (CBG) patterns are introduced in bi-random walk algorithm to predict the missing associations based on the heterogeneous network. The results show that BWHCDA could be considered as a powerful tool for predicting circRNA-disease associations.

2 Methods

2.1 Human CircRNA-Disease Associations

The experimentally validated human circRNA-disease associations are extracted from the CircR2Disease database [13]. Then, we choose the associations that circRNAs have been recorded in circBase database [14] and disease name recorded in disease ontology database [15]. Finally, we retained 371 circRNA-disease associations between 325 circRNAs and 53 diseases as the gold standard dataset. The circRNA-disease adjacency matrix $A(i,j)$ is established, if there is an association between circRNA and disease, $A(i,j)$ is set as 1, otherwise 0.

2.2 CircRNA Regulatory Similarity

The miRNA-circRNA interactions are downloaded from the CircBank database [16], and the interactions overlapped with disease-related circRNAs are selected to measure the regulatory similarity of circRNAs. It is measured as follows:

$$SC\_RG(c_i, c_j) = \frac{card(M_i \cap M_j)}{\sqrt{card(M_i)} \cdot \sqrt{card(M_j)}} \quad (1)$$

where the set of M_i have relationship with circRNA c_i and the set of miRNA M_j have relationship with circRNA c_j.

2.3 Disease Semantic Similarity

The disease names are described as hierarchical directed acyclic graph (DAG) based on the Medical Subject Headings (MeSH) descriptions for diseases. And disease semantic similarity is calculated by the DOSE [17] tool with Wang *et al.* method.

2.4 GIP Kernel Similarity

Based on the assumption that similar circRNAs (diseases) are tend to have similar interaction or non-interaction pattern with diseases (circRNAs) [18], the GIP kernel similarity for circRNAs and diseases are respectively calculated as follows:

$$SC\_cGIP(c(i), c(j)) = exp(-\gamma_c \|c(i) - c(j)\|^2)$$
$$\gamma_c = \frac{1}{(\frac{1}{n_c}\sum_{i=1}^{n_c} \|c(i)\|^2)} \tag{2}$$

$$SD\_dGIP(d(i), d(j)) = exp(-\gamma_d \|d(i) - d(j)\|^2)$$
$$\gamma_d = \frac{1}{(\frac{1}{n_d}\sum_{i=1}^{n_d} \|d(i)\|^2)} \tag{3}$$

where $c(i)$ (or $d(i)$) denotes the circRNA (disease) interaction profiles, which is the i-th row (column) of the adjacency matrix A. The parameters γ_c and γ_d are used to control the kernel bandwidth. n_c (or n_d) is the number of circRNAs (diseases).

2.5 Integrated Similarity for CircRNAs and Diseases

The new circRNA similarity scores (SC) are calculated with the average scores of the circRNA regulatory similarity and GIP kernel similarity for circRNAs. Similarly, the integrated disease similarity (SD) is denoted as the mean of the disease semantic similarity and GIP kernel similarity for diseases. Then, the integrated circRNA similarity and integrated disease similarity are adjusted with the logistic function [19].

$$S(x) = \frac{1}{1 + e^{cx+d}} \tag{4}$$

where x is the value of element of matrix SC or SD. Parameters c and d control the adjustment effects, and we set c as -15 and set d as log(9999), respectively.

2.6 The Construction of Heterogeneous Network

According to the circRNA similarity and disease similarity measures, the circRNA network and disease network can be constructed. Next, the weighted heterogeneous circRNA-disease network is constructed based on the circRNA network, disease network via gold standard circRNA-disease associations. The heterogeneous network could be considered as a bipartite graph, the nodes represent circRNAs or diseases, the edges represent three types of interactions of circRNA-circRNA, disease-disease and circRNA-disease.

2.7 BWHCDA Method

Based on the topology and structure characteristics of circRNA network and disease network, the concept of CBG was introduced. A CBG is described as a subgraph of a circRNA path $\{c_1, c_2, \ldots, c_n\}$ and a disease path $\{d_1, d_2, \ldots, d_m\}$, in which the

two ends connected by circRNA-disease associations (c_1, d_1) and (c_n, d_m). The CBG indicates a vicinity relation between the two association (c_1, d_1) and (c_n, d_m), which is generalized by their distance to other associations in the circRNA network and disease network. The length of CBG patterns (l, r) is determined by the longer length of circRNA path and the disease path. In this study, we hypothesize that most potential associations tend to be covered by many shorter CBGs in the unknown circRNA-disease network. If there are more CBG patterns between circRNAs and diseases, the higher possibility of circRNA-disease associations are.

By iteratively adding the circRNA path and disease path, we calculates the CBGs weighted by decay factor α that ranges from 0 to 1. Because of different structures and topologies in the circRNA network and disease network, disparate optimal number of random walk steps are generated. Therefore, parameters l and r are introduced to restrict the number of random steps in circRNA similarity network and disease similarity network, respectively. The iterative process of bi-random walk is described as follows:

$$\text{On the circRNA network: } Cc = \alpha \cdot SC_L \cdot CD_{t-1} + (1-\alpha)A \tag{5}$$

$$\text{On the disease network: } Dd = \alpha \cdot CD_{t-1} \cdot SD_L + (1-\alpha)A \tag{6}$$

where α is the decay factor that controls the importance of CBG for different paths, SL_L and SD_L represent the normalized matrix by using Laplace regularization.

$$SC_L = Dc^{-1/2}(S_c)Dc^{-1/2} \tag{7}$$

$$SD_L = Dd^{-1/2}(S_d)Dd^{-1/2} \tag{8}$$

where $Dc(i,i)$ (or $Dd(i,i)$) is the diagonal matrix of circRNA similarity matrix S_c (S_d).

By combining the propagation scores of matrices Cc and Dd, the relevance scores of unknown circRNA-disease associations could be obtained. The BWHCDA algorithm is outlined as Table 1.

3 Results

3.1 Prediction Performance

To assess the performance of BWHCDA method, leave-one-out cross validation (LOOCV) and 5-fold cross validation (10-fold CV) framework are performed on the gold standard datasets. For LOOCV, each known circRNA-disease association is removed in turn as testing sample, and the other associations are regarded as training samples. Then, the unknown circRNA-disease associations are considered as candidate associations, and the prediction performance is assessed by the predicted rank of test sample. In the framework of 10-fold CV, circRNA-disease associations are randomly divided into ten subsets, and each subset is utilized in turn as test set and the remaining as the train set on each time. To decrease the sample division bias, we perform 100 times repetitions of 10-fold CV. The receiver operating characteristic (ROC) curves are plotted to show the prediction performance by calculating the true positive rate (TPR) and false positive rate (FPR). Furthermore, the area under the curves (AUCs) are calculated to evaluate the overall performance.

Table 1. The pseudocode of BWHCDA algorithm

Algorithm: BWHCDA
Input: circRNA-disease association interaction adjacency matrix A, circRNA-circRNA regulated similarity matrix SC_RG, disease semantic similarity matrix DS_Dss, parameter α, iteration step I_l and I_r
Output: predicted interaction matrix CD
BWHCDA (A, SC_RG, DS_Dss, α, I_l, I_r)
1. Calculate GIP similarity matrix SC_cGIP and SD_dGIP with A;
2. The integrated circRNA similarity matrix SC is obtained based on SC_RG and SC_cGIP, while the disease similarity SD is calculated based on DS_Dss and SD_dGIP;
3. Obtain the matrix by logistic function S_c and S_d;
4. Obtain the normalized matrix SC_L and SD_L by Laplacian normalization operation.
5. CD_0=A/sun(A); // CD_0 is the initial probability
6. //Iteration process;
7. Max_Iter = max([I_l, I_r])
8. **for** t = 1 to Max_Iter
9. $m = n = 0$;
10. //Random walk in circRNA similarity network
11. **if** (t<= I_l)
12. $m = 1$;
13. $Cc = \alpha * SC_L * CD_{t-1} + (1-\alpha) * A$;
14. **end if**
15. // Random walk in disease similarity network
16. **if** (t<= I_r)
17. $n = 1$;
18. $Dd = \alpha * CD_{t-1} * SD_L + (1-\alpha) * A$;
19. **end if**
20. //combination of results
21. $CD_t = (m * Cc + n * Dd)/(m+n)$
22. **end for**
23. **return** CD;

3.2 Effects of Parameters

There are three parameters in the BWHCDA method, including α, I_l, I_r. To test the effects of the three parameters, we set α value as {0.2, 0.4, 0.6, 0.8}, and I_l, I_r are set from 1 to 5, respectively. Then, we could calculate AUC values based on LOOCV and the effects of these parameters are shown in Tables 2, 3, 4 and 5. The results indicate that α has little effects on prediction performance. When $\alpha = 0.4$, $I_l = 4$ and $I_r = 5$, the AUC value of LOOCV is the highest with step length less than five. When $\alpha = 0.4$, $I_l = 3$, $I_r = 4$, AUC value of LOOCV is the highest within step length less than four. The AUC value of LOOCV is the highest within step length than three steps when $\alpha = 0.4$, $I_l = 2$, $I_r = 3$. And when $\alpha = 0.6$, $I_l = 1$, $I_r = 2$, the AUC value of LOOCV is the highest within two steps. Finally, we set three parameters as $\alpha = 0.4$, $I_l = 2$, $I_r = 3$, respectively.

3.3 Comparison with Other Methods

To further evaluate the prediction performance of BWHCDA, we compare it with other five methods including KATZHCDA [9], PageRank [20], NCP [21], BDSILP [22] and

Table 2. When α is set as 0.2, the effect of parameters I_l and I_r for LOOCV AUC.

$\alpha = 0.2$	$I_r = 1$	$I_r = 2$	$I_r = 3$	$I_r = 4$	$I_r = 5$
$I_l = 1$	0.9084	0.9278	0.9126	0.9046	0.9009
$I_l = 2$	0.6851	0.9097	0.9295	0.9137	0.9050
$I_l = 3$	0.3462	0.7035	0.9099	0.9298	0.9138
$I_l = 4$	0.1645	0.3678	0.7057	0.9099	0.9298
$I_l = 5$	0.1429	0.1666	0.3706	0.7059	0.9099

Table 3. When α is set as 0.4, the effect of parameters I_l and I_r for LOOCV AUC.

$\alpha = 0.4$	$I_r = 1$	$I_r = 2$	$I_r = 3$	$I_r = 4$	$I_r = 5$
$I_l = 1$	0.9084	0.9314	0.9209	0.9116	0.9058
$I_l = 2$	0.7788	0.9109	**0.9334**	0.9243	0.9138
$I_l = 3$	0.5692	0.8018	0.9116	**0.9341**	0.9253
$I_l = 4$	0.3632	0.6156	0.8056	0.9116	**0.9343**
$I_l = 5$	0.2076	0.4049	0.6259	0.8062	0.9115

Table 4. When α is set as 0.6, the effect of parameters I_l and I_r for LOOCV AUC.

$\alpha = 0.6$	$I_r = 1$	$I_r = 2$	$I_r = 3$	$I_r = 4$	$I_r = 5$
$I_l = 1$	0.9084	**0.9322**	0.9254	0.9165	0.9095
$I_l = 2$	0.8194	0.9119	0.9331	0.9285	0.9209
$I_l = 3$	0.6887	0.8428	0.9118	0.9328	0.9296
$I_l = 4$	0.5445	0.7381	0.8476	0.9108	0.9324
$I_l = 5$	0.4094	0.6150	0.7506	0.8480	0.9100

Table 5. When α is set as 0.8, the effect of parameters I_l and I_r for LOOCV AUC.

$\alpha = 0.8$	*$I_r = 1$*	*$I_r = 2$*	*$I_r = 3$*	*$I_r = 4$*	*$I_r = 5$*
$I_l = 1$	0.9084	0.9318	0.9261	0.9180	0.9098
$I_l = 2$	0.8422	0.9123	0.9299	0.9273	0.9203
$I_l = 3$	0.7543	0.8653	0.9109	0.9261	0.9255
$I_l = 4$	0.6670	0.8022	0.8694	0.9078	0.9229
$I_l = 5$	0.5789	0.7359	0.8158	0.8688	0.9046

HeteSim [23]. Consequently, BWHCDA method achieve the best performance among these six approaches based on AUC values of LOOCV and 10-fold CV with the same datasets (Figs. 2 and 3). Therefore, BWHCDA method is better than other five methods.

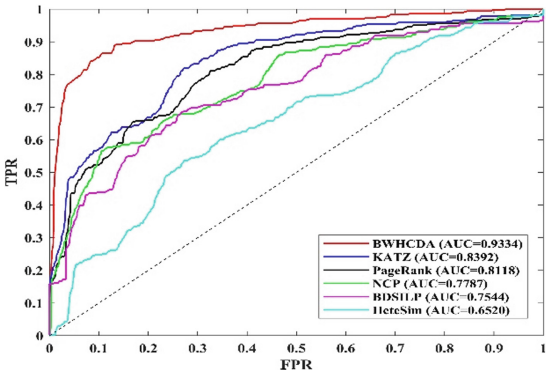

Fig. 2. Comparison of BWHCDA and other methods in terms of ROC curves in LOOCV.

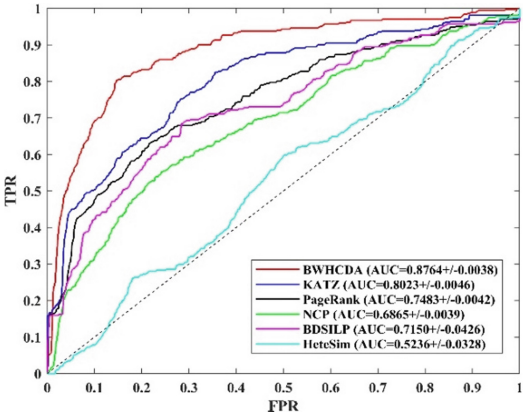

Fig. 3. Comparison of BWHCDA and other methods in terms of ROC curves in 10-fold CV.

3.4 Case Studies

To further assess the prediction performance of BWHCDA method, we analyze the predicted hsa_circ_0000519-gastric cancer association. As shown in Fig. 4, hsa_circ_0000519 may interact with miRNAs including hsa-miR-1233, hsa-miR-1258, hsa-miR-1296, hsa-miR-146b-3p, hsa-miR-521 to play their biological roles. The miRNA targets gene of these miRNAs have been validated related with gastric cancer, including hsa-miR-1258 target HPSE, hsa-miR-146b-3p target PER1 and IRAK1, hsa-miR-521 target ERCC8, hsa-miR-1296-5p target ERBB2. In addition, hsa-miR-1296 has been validated associated with gastric cancer. Therefore, hsa_circ_0000519 may be a potential biomarker for gastric diagnosis and prognosis.

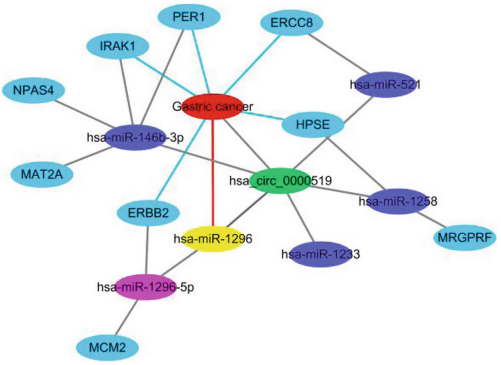

Fig. 4. The hsa_circ_0000519-miRNA-mRNA-gastric cancer interaction network.

4 Conclusion

Prioritizing the potential associations between circRNAs and diseases is benefit to the development of the understanding of the disease mechanism, diagnose and treatment for diseases. The reasons that why BWHCDA method has better performance is shown as following aspects. First, bi-random explored the CBG patterns with iteratively implement random walk on the circRNA similarity network and disease similarity network. In addition, BWHCDA is a multi-task learning method that could forecast potential circRNA-disease associations simultaneously rather than mine candidate circRNAs for specific diseases. Therefore, BWHCDA could be an effective method for biomedical research.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (61972451, 61672334, 61902230) and the Fundamental Research Funds for the Central Universities, Shaanxi Normal University (GK201901010).

References

1. Barrett, S.P., Wang, P.L., Salzman, J.: Circular RNA biogenesis can proceed through an exon-containing lariat precursor. eLife **4**, e07540 (2015)
2. Memczak, S., et al.: Circular RNAs are a large class of animal RNAs with regulatory potency. Nature **495**, 333–338 (2013)
3. Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P.P.: A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell **146**, 353–358 (2011)
4. Zhang, Y., et al.: Circular intronic long noncoding RNAs. Mol. Cell **51**, 792–806 (2013)
5. Conn, V.M., et al.: A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. Nat. Plants **3**, 17053 (2017)
6. Sun, P., Li, G.: CircCode: a powerful tool for identifying circRNA coding ability. Front. Genet. **10**, 981 (2019)
7. Meng, S., et al.: CircRNA: functions and properties of a novel potential biomarker for cancer. 16

8. Lei, X., Fang, Z., Chen, L., Wu, F.X.: PWCDA: path weighted method for predicting circRNA-disease associations. Int. J. Mol. Sci. 19 (2018)
9. Fan, C., Lei, X., Wu, F.X.: Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. Int. J. Biol. Sci. **14**, 1950–1959 (2018)
10. Yan, C., Wang, J., Wu, F.X.: DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. BMC Bioinform. **19**, 520 (2018)
11. Wei, H., Liu, B.: iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. Briefings Bioinform. (2019)
12. Wang, Y., Nie, C., Zang, T., Wang, Y.: Predicting circRNA-disease associations based on circRNA expression similarity and functional similarity. **10** (2019)
13. Fan, C., Lei, X., Fang, Z., Jiang, Q., Wu, F.X.: CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database: J. Biol. Databases Curation **2018** (2018)
14. Glazar, P., Papavasileiou, P., Rajewsky, N.: circBase: a database for circular RNAs. RNA (New York, N.Y.) **20**, 1666–1670 (2014)
15. Schriml, L.M., et al.: Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. **40**, D940–D946 (2012)
16. Liu, M., Wang, Q., Shen, J., Yang, B.B., Ding, X.: Circbank: a comprehensive database for circRNA with standard nomenclature. RNA Biol. **16**, 899–905 (2019)
17. Yu, G., Wang, L.-G., Yan, G.-R., He, Q.-Y.J.B.: DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics **31**, 608–609 (2015)
18. van Laarhoven, T., Nabuurs, S.B., Marchiori, E.: Gaussian interaction profile kernels for predicting drug-target interaction. Bioinform. (Oxford, England) **27**, 3036–3043 (2011)
19. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. **6**, e1000641 (2010)
20. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Stanford InfoLab (1999)
21. Gu, C., Liao, B., Li, X., Li, K.: Network consistency projection for human miRNA-disease associations inference. Sci. Rep. **6**, 36054 (2016)
22. Zhang, W., Yang, W., Lu, X., Huang, F., Luo, F.: The bi-direction similarity integration method for predicting microbe-disease associations. IEEE Access **6**, 38052–38061 (2018)
23. Fan, C., Lei, X., Guo, L., Zhang, A.: Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. Neurocomputing **323**, 76–85 (2019)