PLOS ONE

RESEARCH ARTICLE

The integration of weighted gene association networks based on information entropy

Fan Yang[1], Duzhi Wu[2]*, Limei Lin[1], Jian Yang[3], Tinghong Yang[1], Jing Zhao[4]*

1 Department of Mathematics, Army Logistics University of PLA, Chongqing, China, **2** Rongzhi College of Chongqing Technology and Business, Chongqing, China, **3** School of Pharmacy, Second Military Medical University, Shanghai, China, **4** Institute of Interdisciplinary Complex Research, Shanghai University of Traditional Chinese Medicine, Shanghai, China

* wwwdz63@163.com (DW); zhaojanne@gmail.com (JZ)

Check for updates

OPEN ACCESS

Citation: Yang F, Wu D, Lin L, Yang J, Yang T, Zhao J (2017) The integration of weighted gene association networks based on information entropy. PLoS ONE 12(12): e0190029. https://doi.org/10.1371/journal.pone.0190029

Editor: Lars Kaderali, Universitatsmedizin Greifswald, GERMANY

Received: April 27, 2017

Accepted: December 6, 2017

Published: December 22, 2017

Copyright: © 2017 Yang et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by National Natural Science Foundation of China under Grant Nos. 61372194, 81260672 to JZ, http://www.nsfc.gov.cn/; Chongqing Education Reform Project of Graduate (yjg152017) to JZ, http://www.cqjw.gov.cn/; National Science and Technology Major Project of China (2015ZX09101043-008) to JZ, http://www.nmp.gov.cn/. The funders had no role in study design, data collection and analysis,

Abstract

Constructing genome scale weighted gene association networks (WGAN) from multiple data sources is one of research hot spots in systems biology. In this paper, we employ information entropy to describe the uncertain degree of gene-gene links and propose a strategy for data integration of weighted networks. We use this method to integrate four existing human weighted gene association networks and construct a much larger WGAN, which includes richer biology information while still keeps high functional relevance between linked gene pairs. The new WGAN shows satisfactory performance in disease gene prediction, which suggests the reliability of our integration strategy. Compared with existing integration methods, our method takes the advantage of the inherent characteristics of the component networks and pays less attention to the biology background of the data. It can make full use of existing biological networks with low computational effort.

Introduction

In recent years, high-throughput biological experimental techniques[1, 2] have generated massive omic data sources at the molecular level, such as protein-protein interaction data[3], gene co-expression data[4], and transcriptional regulation data[5]. Arduous efforts have been dedicated to unravel the interplays between all genes in organisms by integrating these data into interaction networks [6–10]. In these networks, nodes represent genes, edges represent interactions between genes, and edge weights are evidence scores of the interactions fused from various biological data sources[11, 12]. Network-based approaches not only have provided more convenient platform for discovering more abundant interactions of genes, but also been employed to infer new disease genes based on links with known disease genes.

In previous studies, there are two main methods to integrate various biological functional data into a comprehensive network. One is subjective scoring integration method, and the other is statistical inference scoring algorithm.

The subjective scoring integration method first rationally evaluates the likelihood of functional coupling between genes by analyzing the objective information, and then scores the genes' interaction by a mathematical function. Various kinds of information are used to judge

decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

the confidence of interactions between genes, such as the reliability of experimental techniques, number of validation studies and orthologous appearance in model organisms. In principle, interactions identified by low-throughput experiments such as *in vitro* X-ray crystallography get higher scores than those obtained from high-throughput experiments such as affinity based technologies. Schaefer *et al* generated a scored human PPI network HIPPIE (Human Integrated Protein-Protein Interaction rEference) from multiple sources, and developed an expertly curated scoring scheme that takes into account three types of information: superiority of experimental techniques to identify the PPI, numbers of studies that have found the PPI, and orthologs of the interacting proteins that have been found experimentally to interact in other model organisms[13]. The interactions are mainly retrieved from several public databases, such as BioGrid[14], IntAct[15], MINT[16], DIP[17], and BIND[18].

Statistical inference scoring method rescored the possibility of associations between genes based on prior information, which provides some basic evidence for the interaction of gene pairs [19–22]. We classify this statistical scoring method into two categories. One is based on probability likelihood ratio scoring, and the other is based on Bayesian method. Lee *et al* used a log likelihood scoring (*LLS*) scheme to construct a genome-scale functional network of human genes called HumanNet, which incorporated diverse data sources, such as gene expression, protein interaction, genetic interaction, sequence, literature, and comparative genomics data. This network includes data directly collected from human genes, as well as those from orthologous genes of yeast, worm, and fly [23]. They utilized HumanNet to predict disease genes and the results showed a favorable performance. Mering *et al* generated a human gene association network STRING using a naive Bayesian model, which integrated known associations collected from many public databases and predicted associations via gene neighborhood, gene fusion, gene co-occurrence, and gene co-expression[20, 24, 25]. They scored each gene pair based on several evidences, such as the co-membership in KEGG pathways, which support their functional coupling. Additionally, Alexeyenko *et al* mixed information derived from proteomics and genomics pipelines and generated a human gene association network FunCoup by an optimized Bayesian framework[19]. Then, they used the network FunCoup to construct a protein-protein interaction network associated with the disease Alzheimer[26]. STRING and FunCoup are scored by similar method which integrated various evidences supporting the associations between genes.

The existing human genome gene association networks, such as HumanNet, STRING and FunCoup, have successful applications in biological research and disease gene prediction [27–29]. However, there is a huge difference between these networks. We find that they share more than 80% common genes, but the common edges are very limited (the proportion of common edges to total number is less than 10%). Therefore, it is necessary to integrate the existing networks to obtain a weighted gene association network, which contains more abundant information by making best use of the information of previous networks. Such work could be valuable to understand cellular processes and study the pathology of complex diseases[30, 31].

In this paper, we propose an algorithm based on information entropy [32–36] to integrate multiple weighted gene association networks (WGANs). Using this algorithm, we integrate four existing human gene association networks to construct a much larger network. To verify the reliability of this integrated network, we use it as background network to predict disease genes and compare its performance with the original networks.

Materials and methods

Data resources

In this paper, an integration strategy is designed to integrate different weighted gene association networks, and an example is provided to show the integration performance. In our

example, four human gene association networks are constructed from different databases, and a network called GO [37–39] is used as a test network for the selection of parameters in the integration model. Detailed information of the networks is presented as follows.

1. HIPPIE: a scored human PPI network integrated from multiple sources, which used an expertly curated scoring scheme that takes into account the reliability of three types of information. The authors aimed to map many gene pairs in different public databases and give them a score calculated by analyzing the reliability of three types of information.
2. HumanNet: a genome-scale functional association network of human genes which were integrated from 21 large-scale genomics and proteomics datasets. In this network, the edge weights are structured by calculating the log likelihood scoring (*LLS*) for each pair and integrating these *LLS* to a final weight.
3. FunCoup: a genome-wide functional association network constructed from the version 3.0 of FunCoup database, which integrates large amounts of genomic data by an optimized Bayesian approach.
4. STRING: a gene association network constructed from the version 9.1 of SRING database which aims to collect and predict many types of gene-gene associations, including physical and functional interactions from diverse sources. In the network, the weight of each link represents a probabilistic confidence score.
5. GO: a human gene association network constructed from the Gene Ontology database which provides structured, controlled vocabularies and classifications that contain several domains of molecular and cellular biology. In this network, there is a link if two genes share at least three GO terms, while the number of shared terms is assigned as weight of current link.

Basic information of above networks is listed in Table 1. Since genes in these databases are presented in multiple identifiers and are obtained by distinct algorithms, we first map the identifiers into the Entrez gene codes and then normalize the weight of each edge into the area (0, 1].

The workflow for the integration of networks

In this section, we outline the workflow that integrates these four networks (see Fig 1) The steps are as follows:

Construct the four networks from the databases and normalize the edge weight to the area (0, 1].

1. Combine all the nodes and edges of the four networks to construct a union network denoted by UNet.
2. For each edge (i, j) in the union network UNet, rescore the edge weight by the model of information entropy.

Table 1. Basic information of the four original networks (HIPPIE, HumanNet, FunCoup and STRING) and the GO network.

Network	HIPPIE	HumanNet	FunCoup	STRING	GO
Nodes	16,514	16,243	16,626	16,213	18,386
Edges	235,184	476,399	4,044,929	3,180,982	45,449,515
Average degree	28.48	58.66	486.58	392.40	4943.93
Average clustering coefficient	0.129	0.246	0.438	0.232	0.786

https://doi.org/10.1371/journal.pone.0190029.t001

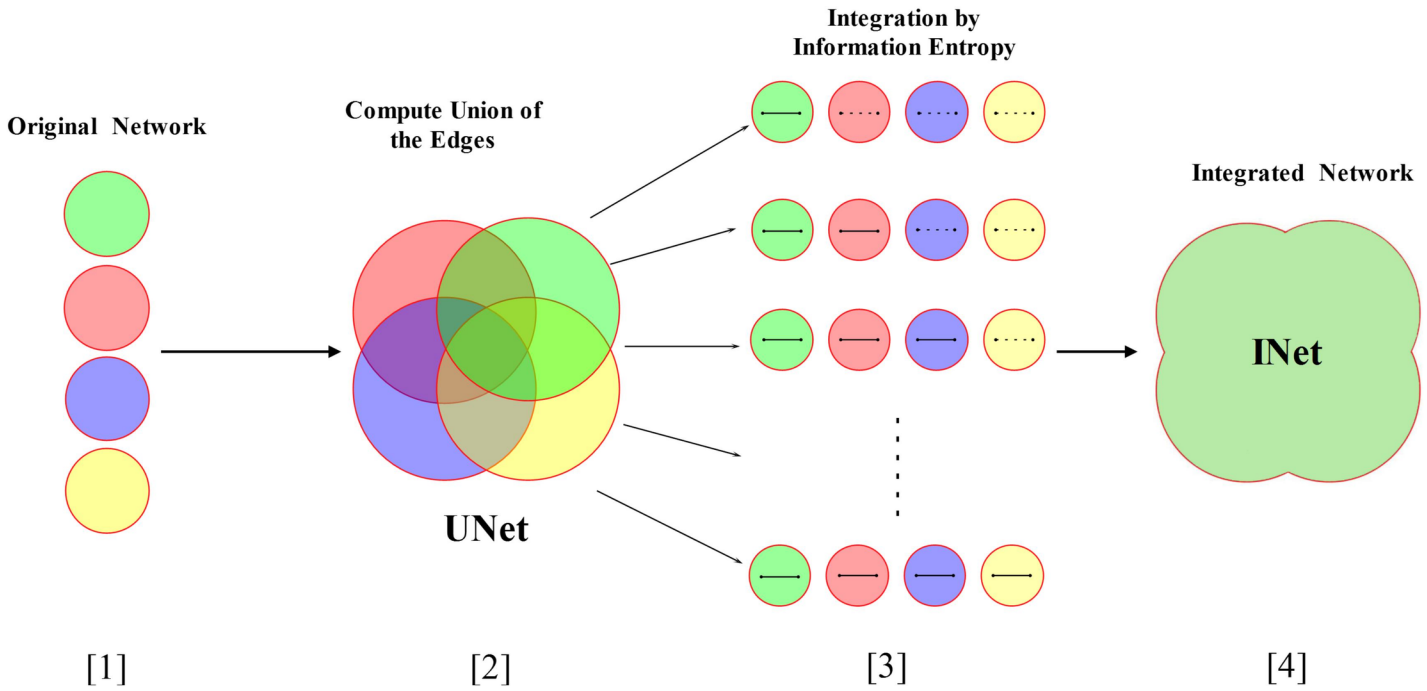

Fig 1. The workflow for the integration of networks. Circles with different colors denote different networks. The solid line in circle means an edge existing in the network and the dashed line means an edge absent in the network. In Step [3], each row denotes the same pair of genes which is connected in at least one of the networks.

https://doi.org/10.1371/journal.pone.0190029.g001

3. Construct the integrated network (denoted as INet), which has the same nodes and edges as the union network while the edge weights are obtained in step 3.

Integration model based on information entropy

We employ information entropy to integrate the weights of the four networks. Generally speaking, entropy is a variable used to measure the disorder degree of a system. It has specific explanations in different areas. Considering a random event, entropy describes the mean uncertain degree of the random variable. This measure has also been widely used in the field of information theory, which is known as information entropy.

For a random variable X, if X contains n possible instances denoted by $x_i (i = 1,2...n)$ whose occurrence possibility is $p_i = P(x_i)$, the uncertain degree of the occurrence of X can be defined by information entropy as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) * \log_2 p(x_i) \quad (1)$$

Suppose there are m WGANs to be integrated, which are denoted by $Net_1, \cdots, Net_m$. We combine all edges in the m WGANs to obtain an edge union set E_U and construct a union network UNet using all the edges and nodes in the set E_U. Then our task is to calculate a weight for each edge in the network UNet from weights of the m original networks. For each gene pair (i, j) in set E_U, $W_k^{(ij)}$ is the edge weight of the gene pair (i, j) in the kth network. We further integrate the edge weight of gene pair (i, j) in m WGANs into a new evidence score of the integrated network as follows,

$$W^{(ij)} = \alpha_1^{(ij)} W_1^{(ij)} + \ldots + \alpha_m^{(ij)} W_m^{(ij)} \quad (2)$$

where $\alpha_1^{(ij)}, \alpha_2^{(ij)} \ldots \alpha_m^{(ij)}$ are positive numbers which represent the integration parameters of gene pair (i, j) in corresponding networks. Larger $\alpha_k^{(ij)}$ suggests larger contribution of the corresponding network Net_k in the integration ($k = 1,2,\ldots m$), $\alpha_1^{(ij)} + \ldots + \alpha_m^{(ij)} = 1$.

To design reasonable integration parameters, we describe the average uncertain degree of an edge existing between a gene pair (i, j) by information entropy and use it to define the integration parameters. As we know, $W^{(ij)}$ can be explained as the probability that the edge exists between gene i and j in a weighted gene association network. For the sake of convenience, we define the following random variable Y,

$$Y = \begin{cases} 1 & i \text{ and } j \text{ are linked} \\ 0 & i \text{ and } j \text{ are not linked} \end{cases} \tag{3}$$

$$P(Y = 1) = W^{(ij)}, P(Y = 0) = 1 - W^{(ij)}.$$

Therefore, the random variable Y describes whether there are interactions between node i and j. According to information entropy described in Eq (1), we can describe the uncertain degree of the interactions between node i and j as following:

$$H(W^{(ij)}) = H(Y) = -W^{(ij)}\log_2 W^{(ij)} - (1 - W^{(ij)})\log_2(1 - W^{(ij)}) \tag{4}$$

In practice, one pair of genes may have interaction in some of the original networks, and have none interaction in the rest of networks. For the latter, it is mathematically meaningless when computing information entropy to depict the uncertain degree of gene pairs. Thus it is necessary to preprocess such pair of genes. In this case, we assume that there is an interaction between the gene pair (i, j) in the network and give it a rather small weight ε. Similarly, for a gene pair (i, j) with edge weight 1, we also change its weight as $1-\varepsilon$. In this paper, we set ε = 0.001. In fact, we choose ε = 0.001 just because it is far smaller than each weight of the existing edges in the original networks.). In this way, each edge (i, j) in the union set EU owns m weights $W_k^{(ij)} (1 \leq k \leq m, 0 < W_k^{(ij)} < 1)$.

For the kth WGAN, the edge weight of gene pair (i, j) is $W_k^{(ij)}$, thus the entropy that an edge exists between the gene pair (i, j) is

$$H(W_k^{(ij)}) = -W_k^{(ij)}\log_2 W_k^{(ij)} - (1 - W_k^{(ij)})\log_2(1 - W_k^{(ij)})$$

The larger the entropy $H(W_k^{(ij)})$, the greater the uncertain degree of existing edge between the gene pair (i, j). Therefore, the entropy $H(W_k^{(ij)})$ can be used to design the integration parameters $\alpha_k^{(ij)} (1 \leq k \leq m)$ in Eq (2). It is reasonable to give a relatively small value to $\alpha_k^{(ij)}$ if the entropy $H(W_k^{(ij)})$ is large. We define a function which decreases with the increasing of $H(W_k^{(ij)})$ as follows,

$$C_k^{(ij)} = C(W_k^{(ij)}) = 1 - e^{-\frac{1}{[H(W_k^{(ij)})]^\theta}} \tag{5}$$

where $\theta > 0$ is an adjustment parameter, which can be properly selected by training with real data. Our function is specifically designed to restrict the integration parameter $C_k^{(ij)}$ in the area (0, 1).Then we design the integration parameters $\alpha_k^{(ij)}$ by normalizing $C_k^{(ij)}$ as follows:

$$\alpha_k^{(ij)} = \frac{C_k^{(ij)}}{\sum_{s=1}^{m} C_s^{(ij)}}, k = 1, \ldots, m \tag{6}$$

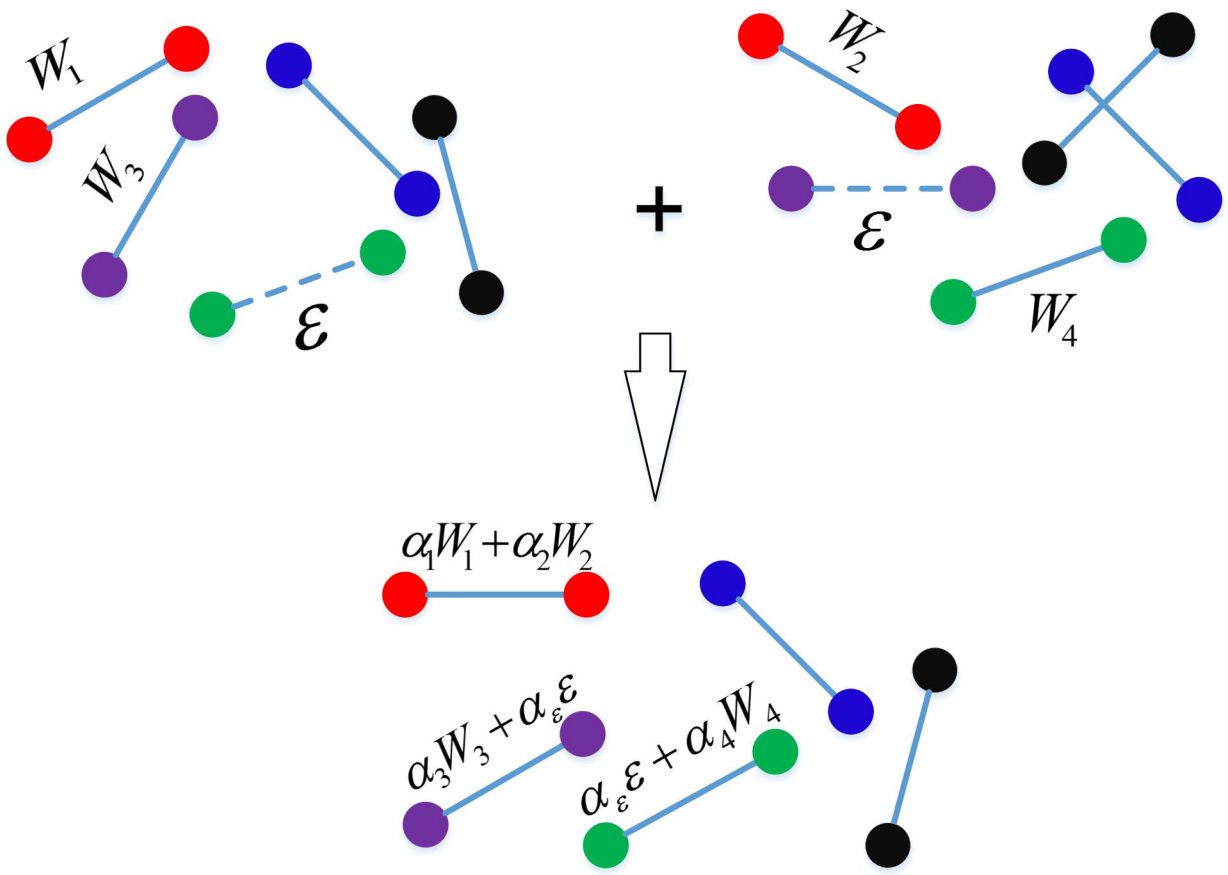

Fig 2. A simple example of integrating two networks.

https://doi.org/10.1371/journal.pone.0190029.g002

Lastly, for any edge (i, j) in the union set E_U, by integrating the edge weights $W_k^{(ij)}(1 \leq k \leq m)$ given in the m networks, the new edge weight $W^{(ij)}$ of the gene pair (i, j) in the integrated network is derived according to Eq (2).

To sum up, our algorithm mainly includes the computation of information entropy for each edge in the original networks and edge weights in the integration process. Firstly, computations of information entropy for each edge in the UNet are polynomial whose computational complexity is $CCE < O(e)$, where e is the number of edges in the four original networks. Secondly, computations of edge weights in the integration process are polynomial. Thus total computational complexity is $CC = CCW \times ES \times CCE$. Here, CCW is the computational complexity of edge weights, ES is the number of edges in the four original networks, CCE is the computational complexity of information entropy. Therefore this algorithm is polynomial. Fig 2 shows a simple example of integrating two networks.

The selection of the adjustment parameter in the integration model

In this section, we select the adjustment parameter θ in Eq (5) for a more accurate integration.

We construct the GO network according to the functional information of all human genes in the gene ontology database (GO), and use it as a test network to determine the adjustment parameter θ of the integration model. GO database is established by Gene Onotology Consortium, which describes almost all known genes in multiple species and annotates them with biological function by standard lexical items (GO term). A GO term represents a specific

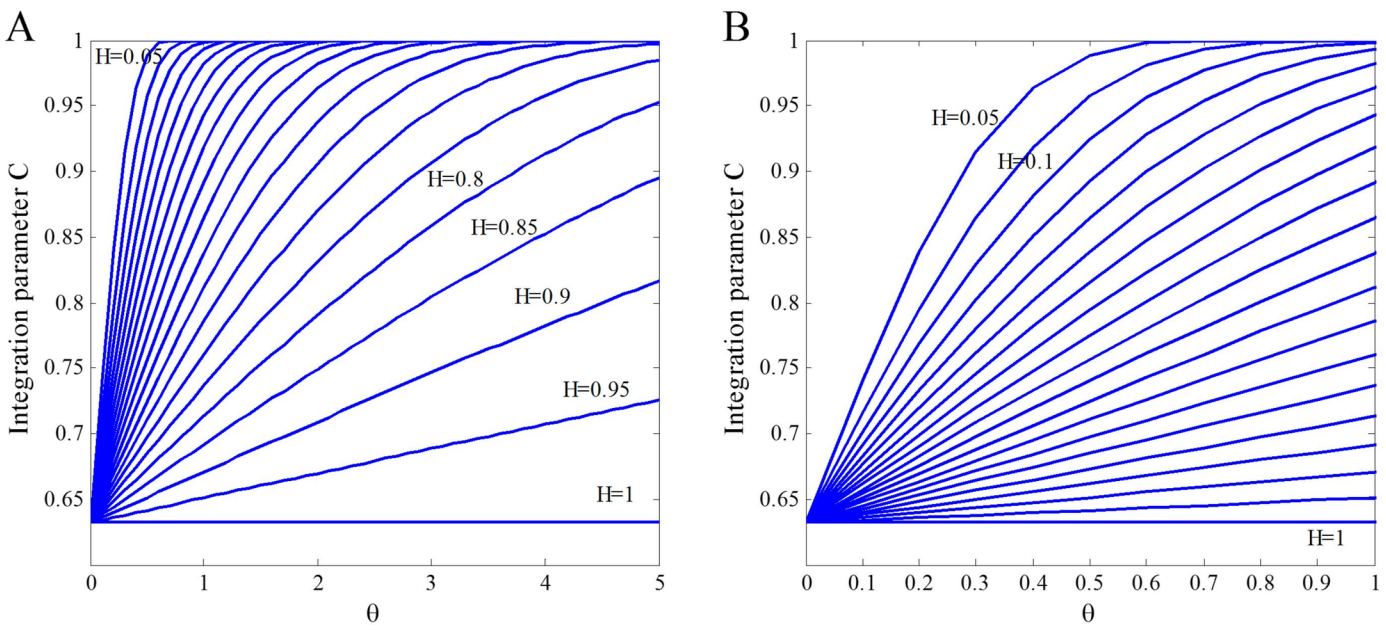

Fig 3. Variation trend of the integration parameter *C* with the change of the adjustment parameter *θ* and information entropy *H* based on the Eq (5). (A) *θ* changes in the area [0, 5]; (B) *θ* is limited to the area [0, 1].

https://doi.org/10.1371/journal.pone.0190029.g003

biological meaning the gene owns. The more GO terms two genes shared, the higher biological similarity they have. Generally, two genes can be linked in the network if they share at least one common GO term. In order to enrich our selection, we finally take gene pairs which share at least three common GO terms, and the number of shared terms is assigned as weight of the current link. Afterwards, we normalized the edge weight into the interval of (0, 1]. Finally, a normalized GO network is constructed with its edge weight proportional to the biological similarity. Since the GO database is established entirely from biological knowledge without using any computational data, this network is different from the networks we used as sources for integration. Therefore, we use it as the comparing background to adjust the parameters.

θ is the adjustment parameter which is used to adjust the influence of information entropy of the pair (i, j) on the integration parameter. In order to get a reasonable adjusting interval of θ, we need to observe that how the integration parameter C changes with θ and H in Eq (5). In Fig 3A we set θ in the area [0, 5] and plot the variation trend of C with the change of θ and H. It shows that, for small values of H (such as $H = 0.05, 0.1$), when the value of θ is larger than 1, the value of C has no significant difference. Thus in Fig 3B we limit θ in the area [0, 1]. We can see that, for different values of H, the area [0.2, 0.6] of θ makes C a relatively significant difference. Therefore, we set the area for adjusting the parameter θ as [0.2, 0.6].

Finally, θ can be determined by solving the following optimization problem:

$$\min_{\theta} f(\theta) = \min_{\theta} \sum (W_{\theta} - W_{GONet})^2 \quad \theta \in [0.2, 0.6] \tag{7}$$

where W_{θ} represents the weights of the common edges with GO network in the integrated network, W_{GONet} is the weight of the corresponding edge in the GO network.

Network-based disease gene prediction

Weighted gene association networks provide a new platform for the study of complex diseases in the context of molecular networks [40–45]. As pointed out in previous works, the

functionally related genes tend to cluster in the network and result in same or similar disease. Thus it is worthwhile to infer potential disease genes by network modeling. Based on such observation, we take genes known to be associated with a particular disease as network 'seeds', and then rank the candidate genes according to their proximity with the seeds by neighborhood weighing rule[46]. Specifically, for a given disease, each gene i in the network is prioritized according to the sum of the weights of its links to the known disease (seed) genes:

$$S_i = \sum_{j \in seed} W_{ij} \tag{8}$$

where W_{ij} is the weight of the link between genes i and j. If gene i has no links with any seed genes, S_i is 0.

To test the quality of the final integrated network, we use it and the four original networks as background network respectively. Disease gene prediction is conducted in each background network respectively and the performances are compared. We conduct two kinds of experiments as follows.

Evaluating the power of the networks in disease gene prediction using leave-one-out cross validation. We extract 609 distinct disease genes from the Online Mendelian Inheritance in Man (OMIM) database. These genes are assembled into 113 seed gene sets corresponding to 113 disorders of human disease phenotypes, in which each seed set contains at least 3 genes. Next, we evaluate the integrated network by leave-one-out cross validation[47, 48] based on the neighborhood weighing rule. This evaluation treats each known gene-disease set as a test case, and assesses how well each known disease gene ranks against a background set of genes when the remaining disease genes are used as seeds. Then, all test cases are pooled together to evaluate the overall performance.

Evaluating the power of the networks in predicting new disease genes. DisGeNET is a discovery platform which provides open access to one of the largest collections of genes and variants associated with human diseases. We extracted 519 genes associated with 49 diseases from the DisGeNET [49] database, in which the 49 diseases are also included in the OMIM database. Then, we took genes in OMIM as seeds and genes in DisGeNET but not in OMIM as test genes to predict new disease genes using the final integrated network and four original networks as background network, respectively. All candidates (including the 519 disease genes) are ranked based on their connectivity with the seed set.

Afterwards, we extracted 24 known disease genes of obesity[50] from the OMIM database and 367 disease genes searched from the literature by Hancock *et al* [51]. We take 24 known disease genes of obesity as seeds and 367 related disease genes as test genes, and conduct disease gene prediction using the final integrated network and four original networks as background network, respectively. For each gene i in the background network, we first calculate S_i by Eq (8), and rank them in descending order.

Performance assessment. We evaluate the performance of the networks in disease gene prediction through following two criterions.

Accuracy: All test cases are pooled together, and the performance is evaluated by calculating the percentage of tested disease genes by varying rank cutoff in the interval [0, 100]. Consequently, the higher percentage the test disease genes, the better performance the background network in disease gene prediction.

AUC value: We plot ROC curves for prediction results and compute their AUC values. ROC (false positive rate vs true positive rate) curve is plotted by changing the rank cutoff from 1 to the number of all genes in the background network by turns. In detail, false positive rate is the fraction of non-seed genes ranked above the threshold, while true positive rate is the

proportion of seed genes ranked above the threshold. AUC is the area under the ROC curve, which lies in the interval [0.5, 1]. It will be 0.5 if all disease genes are distributed at random in the rank, and larger area indicates better performance.

Both of the criterions are used to evaluate the performance of different networks in disease gene prediction. Percentage of test genes shows the accuracy in certain rank cutoff. AUC value shows the predicting accuracy which synthesizing all rank cutoffs in gene association network. The latter criterion suggests an overall performance and it is more robust than the former.

Results and Discussion

Comparison of the four original weighted human gene association networks

Although the four networks under study are all weighted association networks of human genes, they were built by different research teams. HIPPIE is a scored human PPI network integrated from multiple sources. Links in this network are weighed by an expertly curated scoring scheme, which takes into account three types of information including experimental techniques, evidence numbers, and orthologs. Networks HumanNet, FunCoup and STRING come mainly from varieties of databases constructed by fusing physical interaction data and functional association data by log likelihood scoring methods or naive Bayesian framework. The number of nodes in the four networks is almost the same, but the number of edges in each network is quite different. For example, as shown in Table 1, Networks HIPPIE and HumanNet have rather fewer edges than networks FunCoup and STRING.

Fig 4 shows the information of overlapped nodes and edges between the four original weighted human gene association networks. It can be seen that they share 12,127 common gene nodes and 27,382 edges. As listed in Table 2, over 70% nodes of the four networks are common (Fig 5A). However, the numbers of common edges are rather few (Fig 5B). For example, only 0.65% edges of FunCoup appear in the other 3 networks. That is, the proportion of overlapped nodes is far larger than that of the overlapped edges.

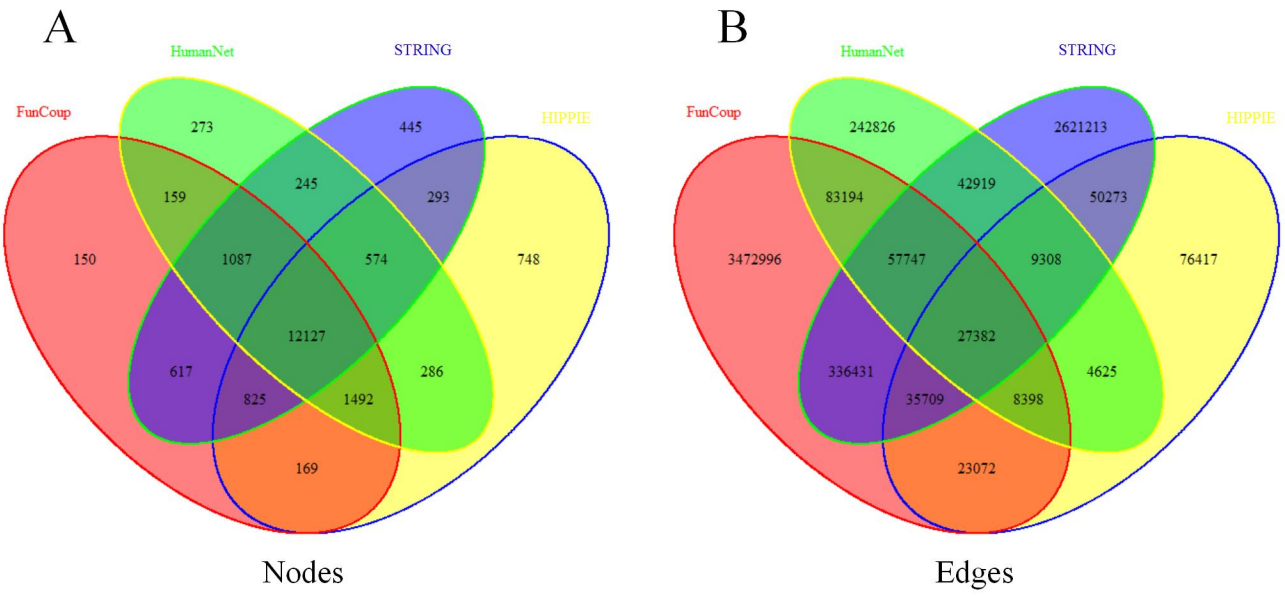

Fig 4. The number of overlapped nodes and edges of the four original networks under study. (A)Common nodes of the four original networks (HIPPIE, HumanNet, FunCoup and STRING). (B)Common edges of the four original networks.

https://doi.org/10.1371/journal.pone.0190029.g004

Table 2. The common nodes and edges information of four networks.

Network	HIPPIE	HumanNet	FunCoup	STRING
Proportion of common nodes occupied in networks	73.43%	74.66%	72.94%	74.80%
Proportion of common edges occupied in networks	11.24%	4.05%	0.65%	0.83%

https://doi.org/10.1371/journal.pone.0190029.t002

In Fig 5C and 5D, we present the fraction of nodes and edges of the four networks in the union network UNet. Each of the four networks takes over eighty percentage of nodes in the union network, while edges of each network only take very small fraction of the union, which is proportional to the numbers of edges of these networks.

The comparisons suggest that it is necessary to make full use of these networks so as to create a network that covers more information about interplays between genes.

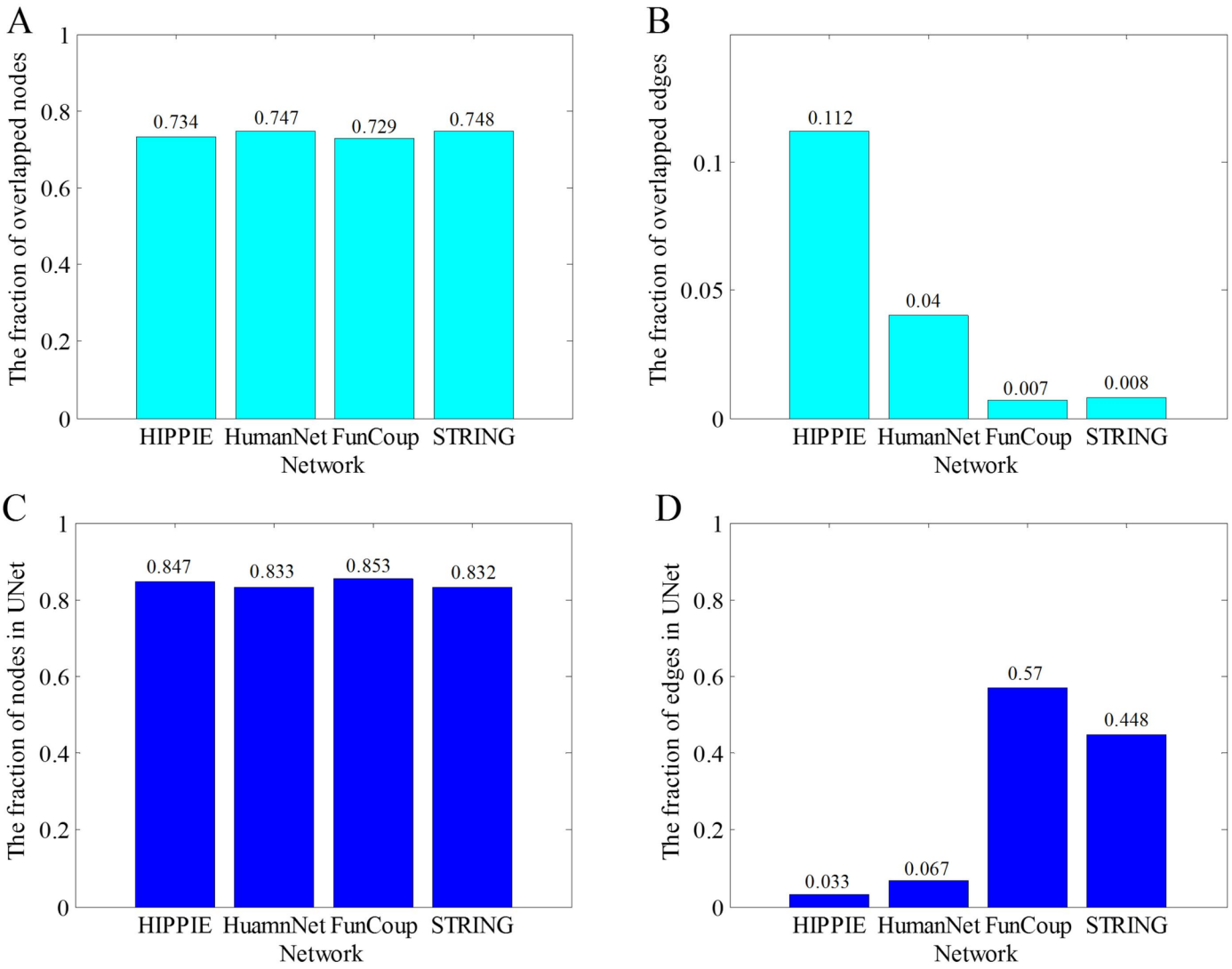

Fig 5. The comparison of nodes and edges in the four original networks and the union network UNet. (A) The fraction of overlapped nodes of the four original networks in their own network. (B) The fraction of overlapped edges of the four original networks in their own network. (C) The fraction of nodes of the four original networks in the union network UNet. (D) The fraction of edges of the four original networks in the union network UNet.

https://doi.org/10.1371/journal.pone.0190029.g005

Construction of an integrated human gene association network by data integration

Now we integrate the four weighted human gene association networks, HIPPIE, HumanNet, FunCoup and STRING according to the proposed integration model. We first combine all nodes and edges in these four networks to build the union network UNet which consists of 19,490 nodes and 7,092,510 links. Then we calculate the weight of each edge in this union network by Eq (2), in which the integration parameters are obtained by Eqs (4–6). In order to choose a proper adjustment parameter θ in Eq (5), we take the GO network as the training network to solve the optimization problem (7). The training result is given in Fig 6.

In Fig 6, $f(\theta)$ reaches minimum when θ is 0.3. Thus the adjustment parameter θ in Eq (5) is selected as 0.3. Then we get the weight of every edge in the union network. In this way, the integrated network (INet), a weighted network based on the union network, is obtained. See S1 File for data of this network and S2 File for the Matlab code of the integration algorithm.

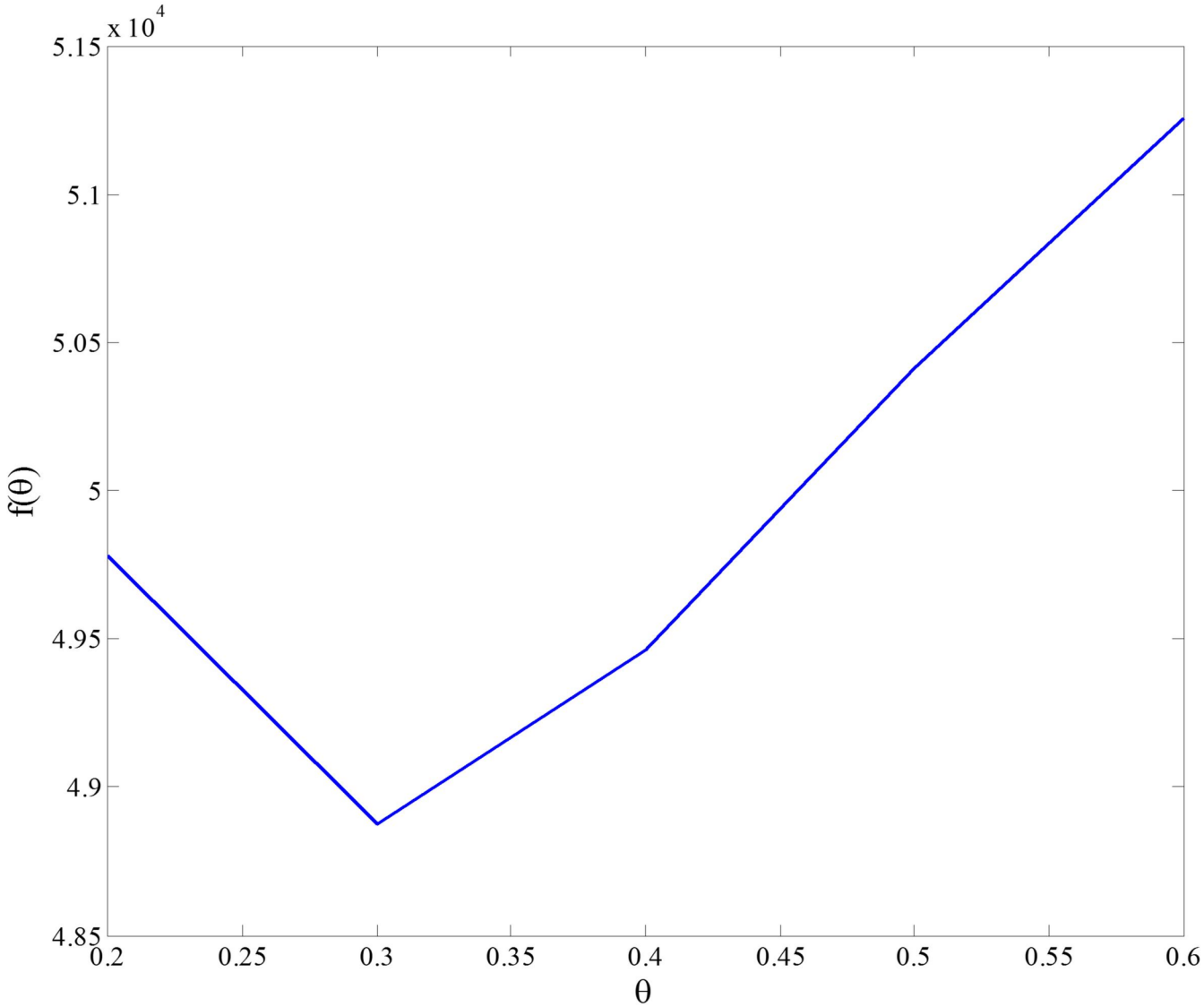

Fig 6. Training process to determine adjustment parameter θ.

https://doi.org/10.1371/journal.pone.0190029.g006

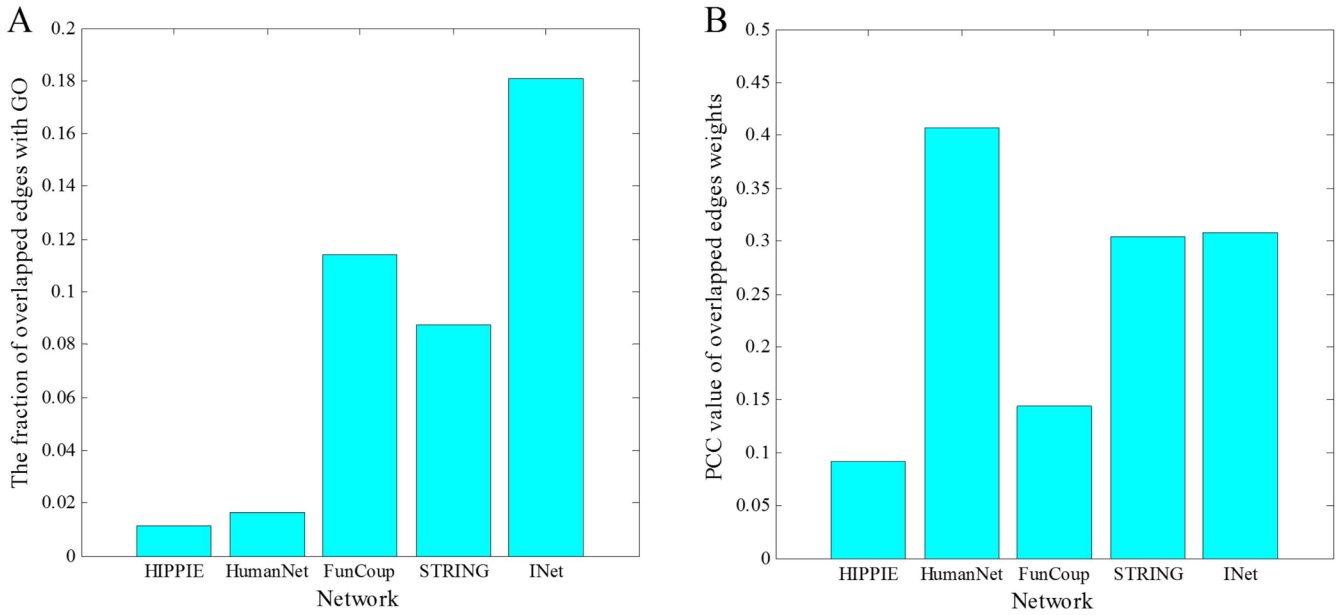

Fig 7. Comparisons of overlapped edges and weights information of the networks with that of the GO network. (A)The fraction of overlapped edges of the INet network and the four original networks with the GO network. (B) The Pearson correlation coefficient between the vectors for the overlapped edge weights of the network and the GO network.

https://doi.org/10.1371/journal.pone.0190029.g007

Now we verify the reliability of our information entropy method for determining the edge weights. For the network INet and each of the four original networks, we first identify its overlapped edges with the GO network. Then, restricting to these overlapped edges, we calculate the Pearson correlation coefficient (PCC) between the vectors for the edge weights of the network and the GO network. All the Pearson correlation coefficients are larger than zero and all the p-values are much smaller than 0.05, implying the statistically significant positive linear correlation between the weights in all the five cases. In Fig 7A we show the fraction of overlapped edges of the INet network and the four original networks with the GO network. Since INet includes all edges of the four original networks, it has the largest fraction of overlapped edges with the GO network. Fig 7B shows that the weight of network INet has a relatively higher positive correlation with that of the GO network than most networks (higher than HIPPIE, STRING and FunCoup, but lower than HumanNet), suggesting a high functional consistency of the weight determined by our model.

The optimal θ is 0.3 and corresponding Pearson correlation coefficient between edges in INet and GO is 0.3075. To test the sensitivity of the parameter, we respectively set θ as 0.1 and 0.4 and calculated the corresponding Pearson correlation coefficient value. The results are 0.3062 and 0.3066 respectively. This result suggests that functional correlation of gene pairs in the INet is robust with the change of the parameter.

In summary, on the one hand, the integrated network provides a much larger network which includes more interaction information between genes. On the other hand, the higher correlation of its weights with those of GO implies a high functional consistency between the connected genes.

Assessment of the integrated network in disease gene prediction

It has been known that genes associated with same disease phenotype tend to be functional related and are clustered together in gene association networks. Thus network-based methods

are widely used in disease gene prediction, in which a gene association network is usually used as a background network. Here we assess the quality of the networks under study by using them as background networks for disease gene prediction.

Assessment by leave-one-out validation. We extract 609 disease genes from OMIM database, and assemble them into 113 seed gene sets corresponding to 113 disorders of human disease phenotypes. For each disease, we first extract one disease gene as a test gene. Then, we use the remaining genes as seed genes to predict the test gene. Afterwards, we get the score of each gene in the network based on the neighborhood weighing rule in Eq (8) and rank them in descending order. Last, we pool all test cases together and calculate the percentage of tested disease genes above various rank cutoffs. Besides, we plotted the ROC curve and computed the AUC value for the prediction results based on each background network. In Fig 8, we show performance comparison and ROC curves of disease gene prediction based on the integrated network INet, the four original networks and the GO network. From Fig 8A we see that INet has a much higher precision of disease gene prediction than that of HumanNet, GONet, HIPPIE and FunCoup, but poorer precision than STRING in the top 100 ranks. In Fig 8B, it is observed that the integrated network INet has the highest AUC value compared with the four original networks and the GONet. In this case, the integrated network has the highest AUC value, but it performs not better than STRING in precision of the top 100 ranks. Comparing to the precision in the top 100 ranks, AUC value shows the precision on the overall ranks. Therefore, in this leave-one-out validation, the integrated network does well in overall performance than other networks.

Considering that three of the original networks (HumanNet, HIPPIE and FunCoup) have much poorer performance than STRING, we guess that the poorer performance of INet than STRING in the top 100 precision is probably caused by them. To verify our inference, we integrate STRING with one of these three networks, respectively. In Fig 9A–9C we compare the performance of the integrated network with the component networks. In all the three cases, the performance of the integrated network is between that of the better and the poorer original

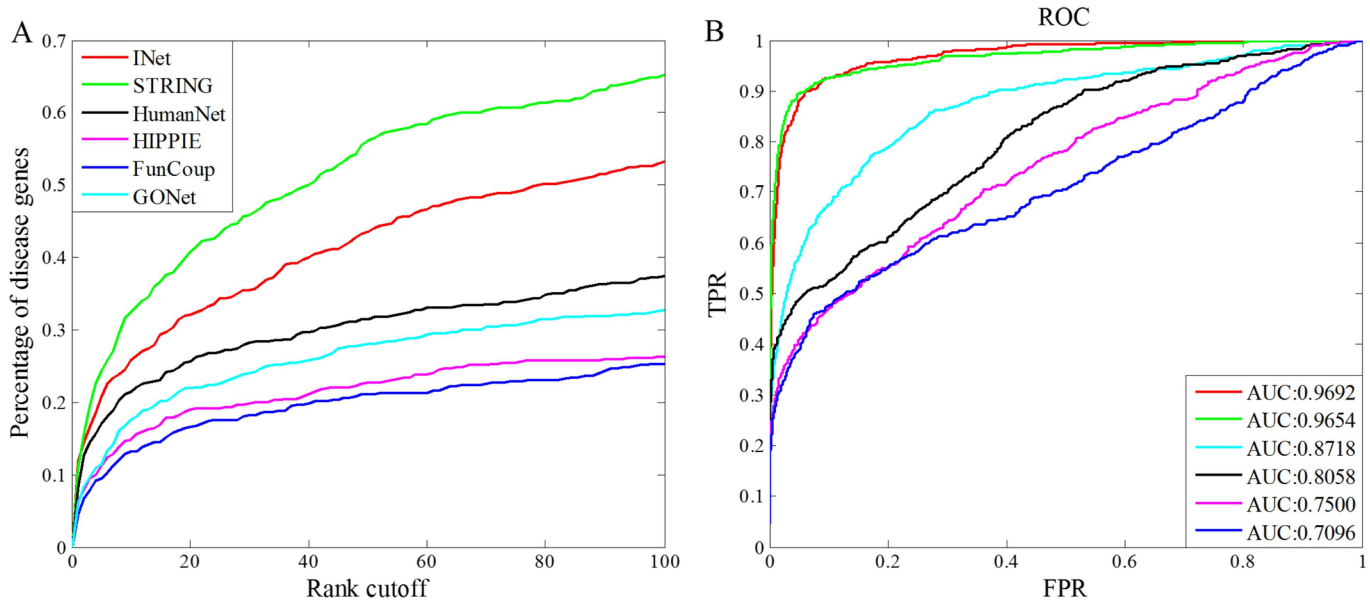

Fig 8. Performance comparison of disease gene prediction based on the integrated network INet, the four original networks and GO network. (A)Percentage of the test genes ranked within top 100. (B) ROC curves and AUC values for the prediction results.

https://doi.org/10.1371/journal.pone.0190029.g008

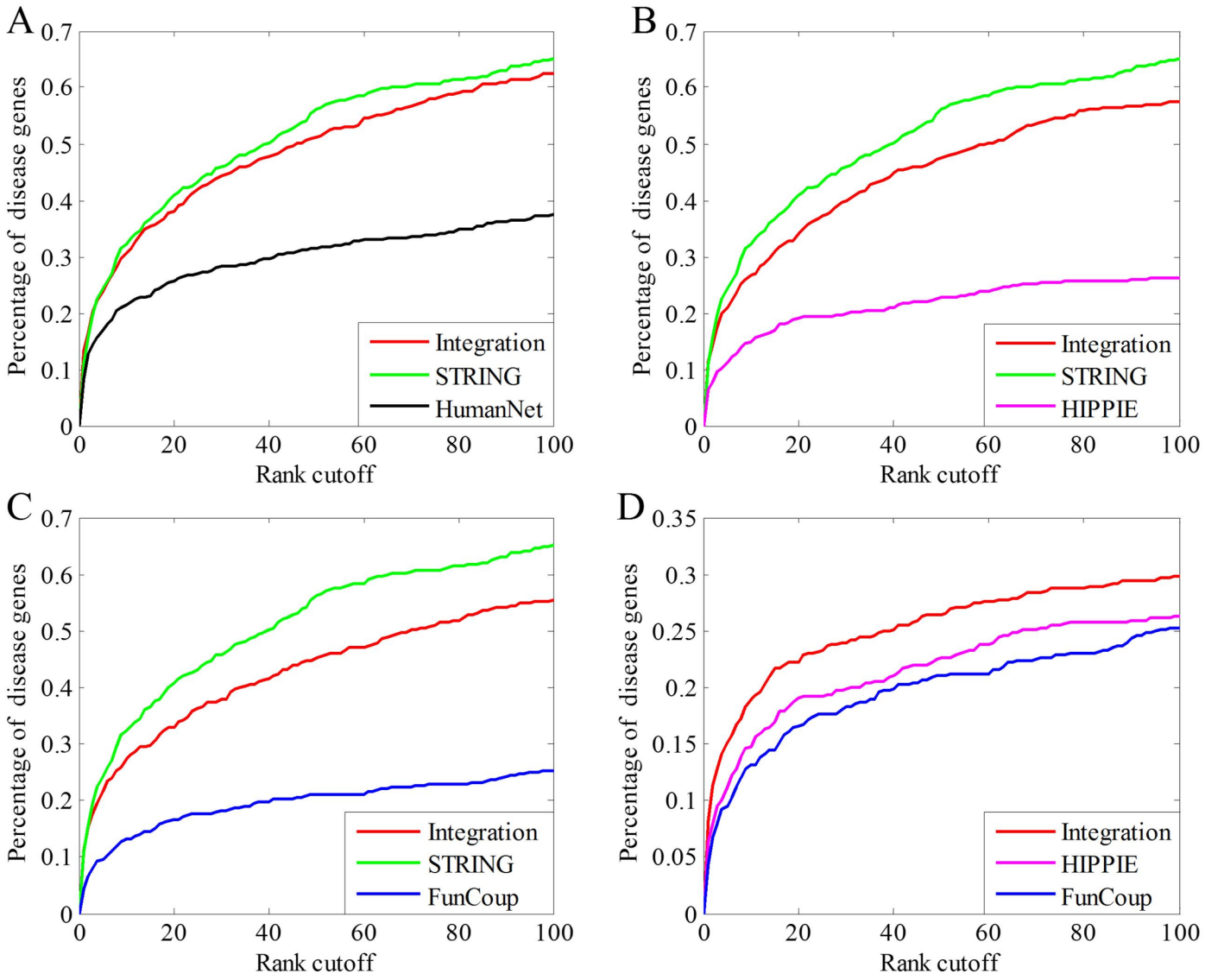

Fig 9. Performance comparisons of different networks in disease gene prediction by leave-one-out cross validation. (A) The integrated network is constructed from String and HumanNet; (B) The integrated network is constructed from STRING and HIPPIE. (C) The integrated network is constructed from STRING and FunCoup. (D)The integrated network is constructed from HIPPIE and FunCoup.

https://doi.org/10.1371/journal.pone.0190029.g009

networks and much closer to that of STRING. These results verify our conjecture and indicate that the performance of the integrated network could be weakened by networks which have much poorer performance. We also integrate the two networks that have the poorest performance, HIPPIE and FunCoup. As shown in Fig 9D, the integrated network exhibits better prediction precision than both of the original networks, implying that the integration has the potential to reach a "one plus one makes more than two" effect.

Assessment by predicting new disease genes. We extracted 519 genes associated with 49 diseases from the DisGeNET database, in which the 49 diseases are also included in the OMIM database. Then, we took genes in OMIM as seeds and genes in DisGeNET as test genes to conduct disease gene prediction based on the INet and the four original networks as background network, respectively. Note that, there are no overlapped genes in seeds and test genes. As

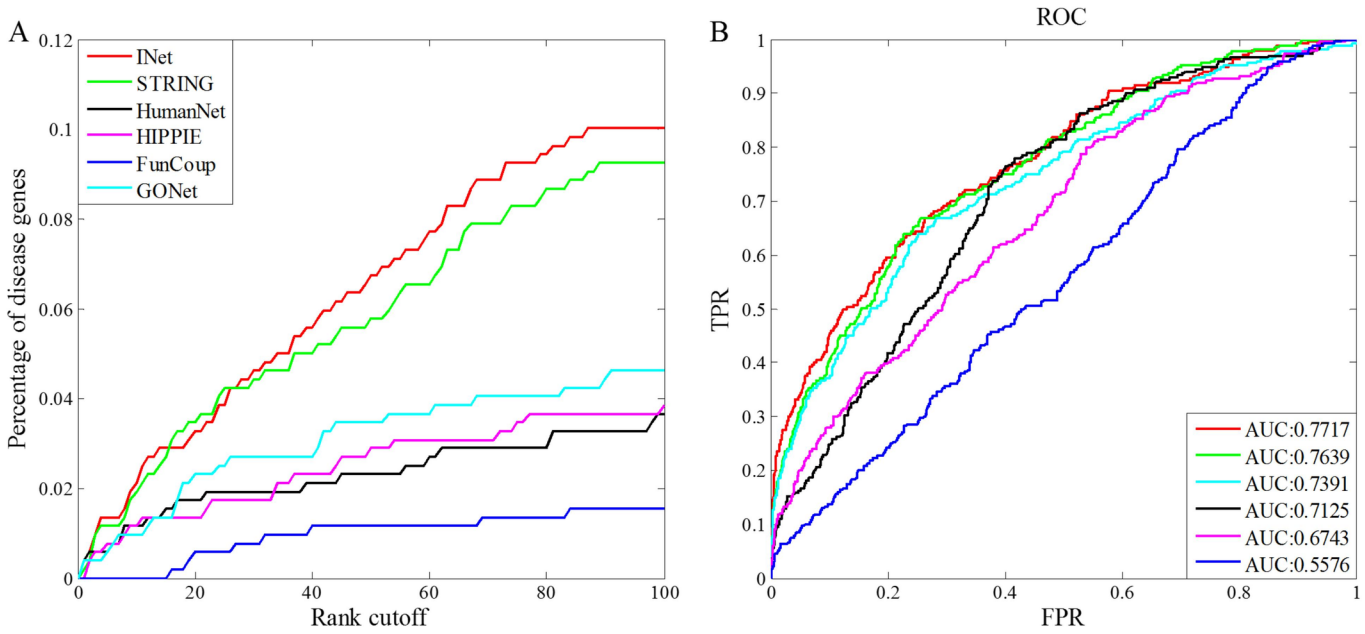

Fig 10. Performance comparisons for the power of different networks in predicting new disease genes. (A)Percentage of the test genes ranked within top 100. (B) ROC curves and AUC values for the prediction results.

https://doi.org/10.1371/journal.pone.0190029.g010

Fig 10A and 10B indicates, the integrated network not only presents the best prediction accuracy for the top 100 genes, but also has the highest AUC value than other networks. On the whole, the results of this case suggest that the integrated network could have a preferable performance than the original networks in predicting new disease genes.

Prediction of obesity associated genes using the integrated network

To further test the reliability of the integrated network in the study of complex diseases, we mimic the search for new disease genes for the polygenic disease obesity. We use the 24 disease genes of obesity included in OMIM database as known disease genes to explore other new disease genes. We treat the 367 obesity associated genes collected from literature as unknown disease genes. In this case, the 24 genes from OMIM and the 367 genes from literature are seed and test genes, respectively. The integrated network INet, the four original networks (HIPPIE, HumanNet, FunCoup, STRING) and the GO network are background networks. We score all genes in the background network by Eq (8) and rank their scores decreasingly. Then we calculate the percentage of test genes under different rank cutoffs based on the results obtained from these networks. Also, ROC curves are plotted to compare the performance of these networks.

Fig 11A presents the comparison between the results by INet, the four original networks and the GO network in the top 100 ranks, respectively. As shown in Fig 11A, INet and STRING exhibit the best performance among the six networks in comparison. When focusing on the top 20 ranks, INet performs more outstanding than STRING. Specifically, the top 12 genes predicted by INet are all correct, i.e., they are all included in the test set. This result further verifies good performance of INet in predicting new disease genes. Fig 11B presents the ROC curves of these networks in predicting obesity disease genes. It indicates that the GO network has higher AUC value than the other networks, although its performance is not good enough in top 100 ranks. In this case, AUC value of the integrated network is lower than GONet and STRING.

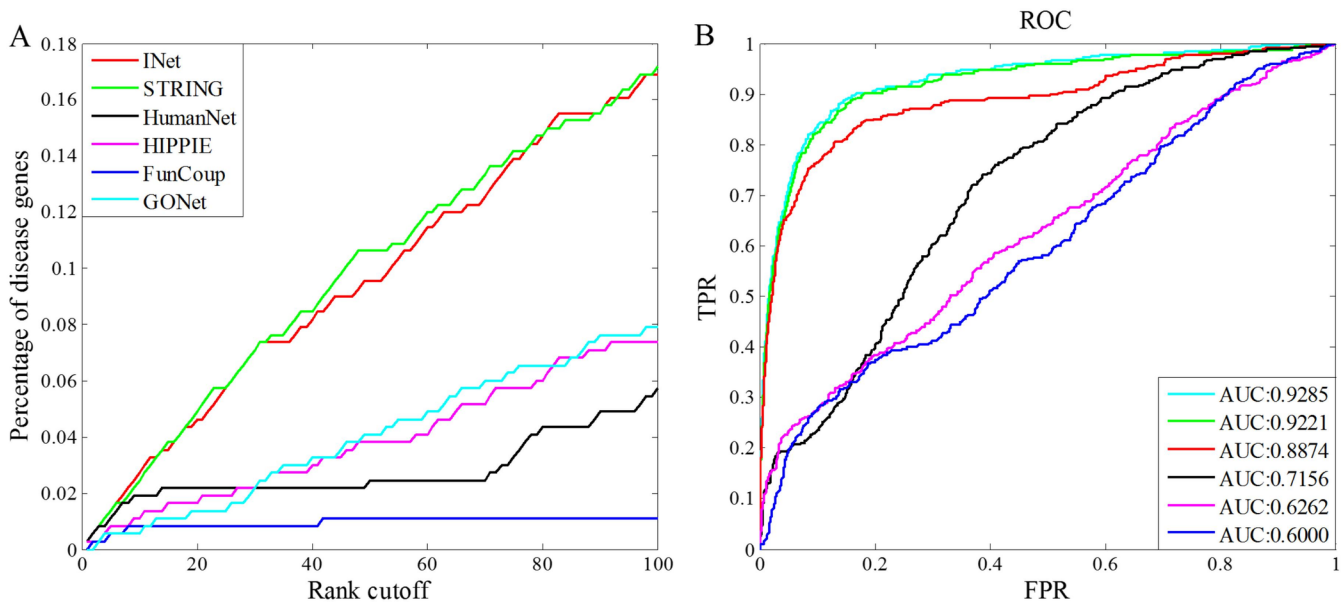

Fig 11. Performance comparison of obesity disease gene prediction based on different networks. (A) Percentage of the test genes ranked within top 100. (B) ROC curves and AUC values for the prediction of test genes.

https://doi.org/10.1371/journal.pone.0190029.g011

Conclusion

In this paper, we propose a novel method based on information entropy for the integration of weighted gene association networks. We use this method to construct a weighted human gene association network from four existing networks, STRING, FunCoup, HumanNet and HIPPIE. The constructed network (named INet) includes all nodes and edges from the four original networks, while its edge weights are calculated by our information entropy algorithm using the weights in the four original networks. Edge weights of the network INet show quite high positive correlation with that of the GO network. Thus the edge weights determined by our algorithm are highly correlated with functional consistency between corresponding gene pairs. In addition, this integrated network exhibits satisfactory performance in disease gene prediction, which indicates its reliability and application significance. In summary, the network constructed from the proposed integration method includes more abundant biological information, which plays a satisfactory effect in predicting disease genes compared with previous methods. This method is insightful for the information integration of multiple weighted genomic scale gene association networks.

Supporting information

S1 File. Data of the integrated network INet.txt.
(RAR)

S2 File. MATLAB program for the integration of multiple weighted network.
(RAR)

Author Contributions

Conceptualization: Duzhi Wu, Jing Zhao.

Data curation: Fan Yang, Limei Lin, Jian Yang.

Formal analysis: Fan Yang, Jian Yang.

Funding acquisition: Jing Zhao.

Investigation: Fan Yang.

Methodology: Duzhi Wu, Tinghong Yang.

Resources: Limei Lin.

Supervision: Duzhi Wu, Tinghong Yang, Jing Zhao.

Validation: Limei Lin, Jian Yang.

Writing – original draft: Fan Yang, Limei Lin.

Writing – review & editing: Duzhi Wu, Jing Zhao.

References

1. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nature Reviews Genetics. 2015; 16(5):299–311. https://doi.org/10.1038/nrg3899 PMID: 25854182
2. Gupta S, De Puysseleyr V, Van der Heyden J, Maddelein D, Lemmens I, Lievens S, et al. MAPPI-DAT: data management and analysis for protein-protein interaction data from the high-throughput MAPPIT cell microarray platform. Bioinformatics (Oxford, England). 2017.
3. Gromiha MM, Yugandhar K, Jemimah S. Protein–protein interactions: scoring schemes and binding affinity. Current Opinion in Structural Biology. 2017; 44:31–8. https://doi.org/10.1016/j.sbi.2016.10.016 PMID: 27866112
4. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC bioinformatics. 2004; 5(1):18.
5. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296(5568):752–5. https://doi.org/10.1126/science.1069516 PMID: 11923494
6. Sun Y, Weng Y, Zhang Y, Yan X, Guo L, Wang J, et al. Systematic expression profiling analysis mines dys-regulated modules in active tuberculosis based on re-weighted protein-protein interaction network and attract algorithm. Microbial Pathogenesis. 2017.
7. Gao B, Shao Q, Choudhry H, Marcus V, Dong K, Ragoussis J, et al. Weighted gene co-expression network analysis of colorectal cancer liver metastasis genome sequencing data and screening of anti-metastasis drugs. International journal of oncology. 2016; 49(3):1108–18. https://doi.org/10.3892/ijo.2016.3591 PMID: 27571956
8. Hobbs ET, Pereira T, O'Neill PK, Erill I. A Bayesian inference method for the analysis of transcriptional regulatory networks in metagenomic data. Algorithms for Molecular Biology. 2016; 11(1):19.
9. Specht AT, Li J. LEAP: Constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics. 2016:btw729.
10. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology. 2005; 4(1):1128.
11. Yang J, Yang T, Wu D, Lin L, Yang F, Zhao J. The integration of weighted human gene association networks based on link prediction. BMC Systems Biology. 2017; 11(1):12. https://doi.org/10.1186/s12918-017-0398-0 PMID: 28137253
12. Peng J, Wang T, Wang J, Wang Y, Chen J. Extending gene ontology with gene association networks. Bioinformatics. 2016; 32(8):1185–94. https://doi.org/10.1093/bioinformatics/btv712 PMID: 26644414
13. Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: Integrating protein interaction networks with experiment based quality scores. PloS one. 2012; 7(2): e31826. https://doi.org/10.1371/journal.pone.0031826 PMID: 22348130
14. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic acids research. 2006; 34(suppl 1):D535–D9.
15. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic acids research. 2011:gkr1088.
16. Chatr-Aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: the Molecular INTeraction database. Nucleic acids research. 2007; 35(suppl 1):D572–D4.

17. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic acids research. 2004; 32(suppl 1):D449–D51.
18. Bader GD, Betel D, Hogue CW. BIND: the biomolecular interaction network database. Nucleic acids research. 2003; 31(1):248–50. PMID: 12519993
19. Alexeyenko A, Sonnhammer EL. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome research. 2009; 19(6):1107–16. https://doi.org/10.1101/gr.087528.108 PMID: 19246318
20. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research. 2013; 41 (D1):D808–D15.
21. Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T, et al. Refined phylogenetic profiles method for predicting protein–protein interactions. Bioinformatics. 2005; 21(16):3409–15. https://doi.org/10.1093/bioinformatics/bti532 PMID: 15947018
22. Yu J, Finley RL. Combining multiple positive training sets to generate confidence scores for protein–protein interactions. Bioinformatics. 2009; 25(1):105–11. https://doi.org/10.1093/bioinformatics/btn597 PMID: 19010802
23. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research. 2011; 21(7):1109–21. https://doi.org/10.1101/gr.118992.110 PMID: 21536720
24. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic acids research. 2005; 33(suppl 1):D433–D7.
25. Emmert-Streib F, Dehmer M, Shi Y. Fifty years of graph matching, network alignment and network comparison: Elsevier Science Inc.; 2016. 180–97 p.
26. Association As. 2017 Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2017.
27. Le D-H, Dang V-T. Ontology-based disease similarity network for disease gene prediction. Vietnam Journal of Computer Science. 2016; 3(3):197–205.
28. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nature biotechnology. 2006; 24(5):537–44. https://doi.org/10.1038/nbt1203 PMID: 16680138
29. Peng J, Bai K, Shang X, Wang G, Xue H, Jin S, et al. Predicting disease-related genes using integrated biomedical networks. BMC genomics. 2017; 18(1):1043.
30. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nature chemical biology. 2008; 4(11):682–90. https://doi.org/10.1038/nchembio.118 PMID: 18936753
31. Mattick JS. The Future of Molecular Pathology. Molecular Pathology in Cancer Research: Springer; 2016. p. 349–57.
32. Gray RM. Entropy and information theory: Springer Science & Business Media; 2011.
33. Dehmer M, Li X, Chen Z, Emmertstreib F, Shi Y. Mathematical foundations and applications of graph entropy. 2016.
34. Cao S, Dehmer M, Shi Y. Extremality of degree-based graph entropies. Information Sciences. 2014; 278(10):22–33.
35. Cao S, Kang Z, Kang Z. Network Entropies Based on Independent Sets and Matchings: Elsevier Science Inc.; 2017. 265–70 p.
36. Chen Z, Dehmer M, Emmert-Streib F, Shi Y. Entropy bounds for dendrimers. Applied Mathematics & Computation. 2014; 242(2):462–72.
37. Consortium GO. The gene ontology (GO) project in 2006. Nucleic acids research. 2006; 34(suppl 1): D322–D6.
38. Consortium GO. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004; 32(suppl 1):D258–D61.
39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000; 25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651
40. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011; 12(1):56–68. https://doi.org/10.1038/nrg2918 PMID: 21164525
41. Liu Z-P, Wang Y, Zhang X-S, Chen L-n. Network-based analysis of complex diseases. IET Systems Biology. 2012; 6(1):22–33. https://doi.org/10.1049/iet-syb.2010.0052 PMID: 22360268
42. Yin T, Chen S, Wu X, Tian W. GenePANDA—a novel network-based gene prioritizing tool for complex diseases. Scientific Reports. 2017;7.

43. Zhang Q, Li J, Xue H, Kong L, Wang Y. Network-based methods for identifying critical pathways of complex diseases: a survey. Molecular BioSystems. 2016; 12(4):1082–9. https://doi.org/10.1039/c5mb00815h PMID: 26888073

44. Al-Harazi O, Al Insaif S, Al-Ajlan MA, Kaya N, Dzimiri N, Colak D. Integrated genomic and network-based analyses of complex diseases and human disease network. Journal of Genetics and Genomics. 2016; 43(6):349–67. https://doi.org/10.1016/j.jgg.2015.11.002 PMID: 27318646

45. Chasman D, Siahpirani AF, Roy S. Network-based approaches for analysis of complex biological systems. Current opinion in biotechnology. 2016; 39:157–66. https://doi.org/10.1016/j.copbio.2016.04.007 PMID: 27115495

46. Linghu B, Snitkin ES, Hu Z, Xia Y, DeLisi C. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome biology. 2009; 10(9):1.

47. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. nature. 2002; 415(6871):530–6. https://doi.org/10.1038/415530a PMID: 11823860

48. Kohavi R, editor A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai; 1995: Stanford, CA.

49. Piñero J, Bravo À, Queraltrosinach N, Gutiérrezsacristán A, Deupons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2017; 45(Database issue):D833–D9. https://doi.org/10.1093/nar/gkw943 PMID: 27924018

50. Grundy SM. Obesity, metabolic syndrome, and cardiovascular disease. The Journal of Clinical Endocrinology & Metabolism. 2004; 89(6):2595–600.

51. Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, et al. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet. 2008; 4(2):e32. https://doi.org/10.1371/journal.pgen.0040032 PMID: 18282109