BMC Medical Genomics

RESEARCH Open Access

Measuring disease similarity and predicting disease-related ncRNAs by a novel method

Yang Hu[1†], Meng Zhou[2†], Hongbo Shi[2], Hong Ju[3], Qinghua Jiang[1*] and Liang Cheng[2*]

From IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2016
Shenzhen, China. 15-18 December 2016

Abstract

Background: Similar diseases are always caused by similar molecular origins, such as diasease-related protein-coding genes (PCGs). And the molecular associations reflect their similarity. Therefore, current methods for calculating disease similarity often utilized functional interactions of PCGs. Besides, the existing methods have neglected a fact that genes could also be associated in the gene functional network (GFN) based on intermediate nodes.

Methods: Here we presented a novel method, InfDisSim, to deduce the similarity of diseases. InfDisSim utilized the whole network based on random walk with damping to model the information flow. A benchmark set of similar disease pairs was employed to evaluate the performance of InfDisSim.

Results: The region beneath the receiver operating characteristic curve (AUC) was calculated to assess the performance. As a result, InfDisSim reaches a high AUC (0.9786) which indicates a very good performance. Furthermore, after calculating the disease similarity by the InfDisSim, we reconfirmed that similar diseases tend to have common therapeutic drugs (Pearson correlation $\gamma^2 = 0.1315$, $p = 2.2\text{e-}16$). Finally, the disease similarity computed by infDisSim was employed to construct a miRNA similarity network (MSN) and lncRNA similarity network (LSN), which were further exploited to predict potential associations of lncRNA-disease pairs and miRNA-disease pairs, respectively. High AUC (0.9893, 0.9007) based on leave-one-out cross validation shows that the LSN and MSN is very appropriate for predicting novel disease-related lncRNAs and miRNAs, respectively.

Conclusions: The high AUC based on benchmark data indicates the method performs well. The method is valuable in the prediction of disease-related lncRNAs and miRNAs.

Keywords: Information flow, Disease similarity, Gene functional network, lncRNA similarity network

Background

One way to indicate the associations between pair-wise diseases in quantitatively is their similarity. In comparison with the associations, disease similarity can indicate the relationships between diseases of multiple categories more clearly and easily, for instance, cancers [1]. In the previous studies, disease similarity was exploited to compute similarities between protein-coding RNA genes (PCGs), which can help to disclose the complex pathogenesis of diseases [1]. Moreover, disease similarity was also employed to calculate similarities between microRNA genes (miRNAs) [2, 3], and long non-coding RNA genes (lncRNAs) [4–8], respectively, which could be applied for constructing functional network of non-coding RNA genes (ncRNAs). Recently, similarity between diseases was even utilized to predict potential therapeutic drugs for diseases [9–12].

Semantic associations and disease gene associations are often considered to be quantitative for evaluating disease similarity. Semantic associations between diseases were documented in the ontology around disease terms. The most widely used ontology for calculating disease similarity is Disease Ontology (DO) [13], which

* Correspondence: qhjiang@hit.edu.cn; liangcheng@hrbmu.edu.cn
†Equal contributors
[1]School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China
[2]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150001, China
Full list of author information is available at the end of the article

is the first ontology to be established around disease terms. DO defines a type of semantic association named 'IS_A' relationship, which reflects set inclusion relationships between disease terms [14]. Disease terms of DO could build a directed acyclic graph (DAG) based on the 'IS_A' relationship. Disease-related genes were distributed in different sources, such as Comparative Toxicogenomics Database (CTD) [15], Online Mendelian Inheritance in Man (OMIM) [16], Gene Reference into Functions (GeneRIFs) [17], Genetic Association Database (GAD) [18], and so on.

Three widely used methods for computing the similarity of terms of ontology were presented by Resnik [19], Lin [20], and Wang et al. [21] repectively. All of these three methods were utilized for computing disease similarity by DOSim [1]. Resnik presented Information content (IC) of terms of ontology [19], and in this method, IC of the most informative common ancestor (MICA) of pair-wise diseases was served as the similarity of them. Due to the IC of the pair-wise terms and the IC of the MICA could contribute to the similarity of them, Lin [20] improved Resnik's method. By the contrast of Resnik's and Lin's method, Wang et al. [21] computed the similarity between terms fully based on semantic associations of terms in ontology.

In recent years, three methods for calculating similarity of terms of DO were presented. Disease-related genes have been the focus of all these methods. In another word, the similarity of two diseases was converted to the similarity of the two gene sets of diseases. Mathur and Dinakarpandian first presented to utilize the figure of overlapping genes to calculate disease similarity [22]. Even though two gene sets have no shared genes, these two sets could also be connected by their presence during the same or similar biological process. Therefore, Mathur and Dinakarpandian designed a process-similarity based (PSB) method to compute disease similarity based on biological process terms of Gene Ontology [23, 24]. Besides biological process, co-expression [25] and protein-protein interaction [26] could also be employed to similarity of disease-related gene sets [27, 28]. Hence, Cheng et al. combined semantic association and the comprehensive gene functional network to compute disease similarity (SemFunSim) [11], which performs very well.

Improved knowledge has suggested that semantic associations and disease gene associations are two types of significant associations, which were widely exploited to measure disease similarity. Recent studies focused on incorporating disease gene associations from different views. Eventually, comprehensive gene functional network (GFN) was incorporated in SemFunSim method [11], in which functional interactions of pair-wise genes were considered. Obviously, it is straightforward to consider that whether the entire network could be completely utilized to measure disease similarity. For this purpose, we designed a novel method, called *InfDisSim*, to figure out disease similarity by modeling the information flow in the comprehensive GFN in this study.

Methods

Date source

Disease ontology

Disease terms and semantic associations were originated from DO [13] (Table 1), which is manually curated for diseases names. As for now, it includes 7124 'IS_A' relationships between 6920 terms.

Disease gene association network

Disease-related genes are derived from the latest version of diversed open source sources involving CTD [15], GAD [18], GeneRIFs [17], and OMIM [16]. Disease terms in these databases were distributed to DO according to SIDD [29]. After integrating all of these four widely used sources, 130,144 associations between 3178 disease terms and 11,717 genes were obtained as disease gene association network (Additional file 1).

Comprehensive gene functional network

Comprehensive GFN was estimated from HumanNet [30], which is built around *Homo sapiens*. Multiple interactions spanning human mRNA co-expression, protein-protein interaction, protein complex, and comparative genomics data sets, combining with alike lines of evidence from orthologs in yeast, fly and worm are comprehensively analyzed for the network utilizing a probabilistic method. Currently, it contains 476,399 interactions among 16,243 genes [30].

Disease-related drugs

Disease-related drugs were derived from robust, publicly accessible databases CTD [15], which elucidates the process that chemicals affect human health. Disease terms in CTD were distributed to DO according to SIDD [29]. As a result, 16,639 associations between 1093 diseases and 3887 drugs were obtained.

Table 1 Data sources

Data source	Web site (Date of download)
DO	http://disease-ontology.org/ (Jun 2016)
CTD	http://ctdbase.org/ (Jun 2016)
GeneRIF	http://www.ncbi.nlm.nih.gov/gene/about-generif (Jun 2016)
GAD	https://geneticassociationdb.nih.gov/ (Jun 2016)
OMIM	http://www.omim.org/ (Jun 2016)
HumanNet	http://www.functionalnet.org/humannet/download.html (Jun 2016)
LncRNADisease	http://www.cuilab.cn/lncrnadisease (Jun 2016)

Disease-related lncRNAs

Human lncRNA-disease associations [31–36] were incorporated into the lncRNA similarity network (LSN), which was constructed based on disease similarity, to predict potential relationships between diseases and lncRNAs. These associations were derived from a manually curated database LncRNADisease [37], which provided experimentally supported disease-lncRNA associations. After removing disease terminologies not in DO and deploying of duplicate associations, 602 associations between 167 diseases and 338 lncRNAs were obtained (Additional file 2).

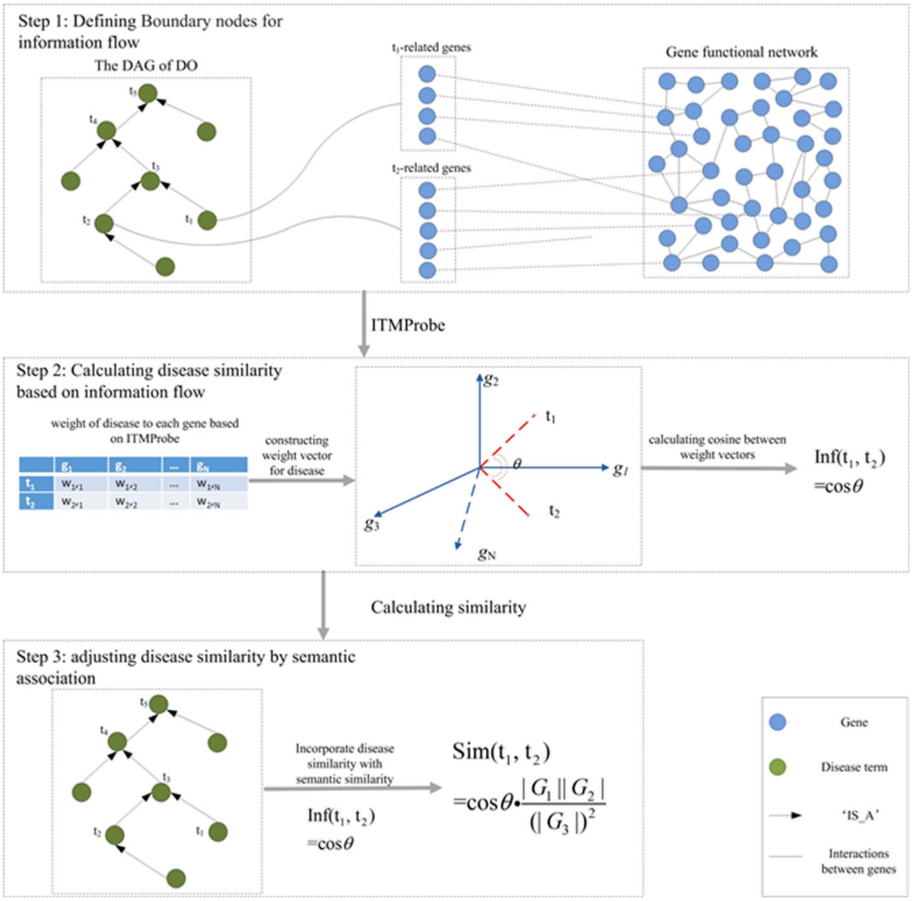
Disease-related miRNAs

Disease-related human miRNAs were extracted from the Human microRNA Disease Database (HMDD) v2.0 [3]. After manually mapping disease terms of HMDD to DO, we got 5710 associations between 556 miRNAs and 265 diseases (Additional file 3).

Method for calculating disease similarity

In this study, we designed a novel method to compute disease similarity by modelling the information flow in the comprehensive GFN. In the previous study, a tool called ITM Probe [38] was created for analyzing information flow in the network based on random walk with damping. Currently, three models involving absorbing, emitting, and channel were employed in ITM Probe. According to these three models [39], the initial nodes which are the starting points of the random walk and the sink nodes which are the ending points of the random walk are regarded as boundary nodes, and the rest of the nodes in the network are regarded as transient nodes. Channel model [39] was designed for directed information flow, which extends absorbing model that specify the source of the information flow and emitting model that distributes end of information flow.

Here, channel model was employed to the network involving disease gene association network and the comprehensive GFN. In this network, disease terms couldn't be directly linked to each other, however, they could be associated based on their related genes. According to Fig. 1, diseases in the network were considered as boundary nodes, and all the genes were considered as transient nodes. To distribute a weight to each transient


Step 1: Defining Boundary nodes for information flow
The DAG of DO
t5
t4
t3
t2
t1
t1-related genes
t2-related genes
Gene functional network
ITMProbe
Step 2: Calculating disease similarity based on information flow
weight of disease to each gene based on ITMProbe
constructing weight vector for disease
g2
t1
θ
g1
t2
g3
gN
calculating cosine between weight vectors
Inf(t1, t2)
=cosθ
Calculating similarity
Step 3: adjusting disease similarity by semantic association
Incorporate disease similarity with semantic similarity
Inf(t1, t2)
=cosθ
Gene
Disease term
'IS_A'
Interactions between genes

	g_1	g_2	...	g_N
t_1	$w_{1\times1}$	$w_{1\times2}$	...	$w_{1\times N}$
t_2	$w_{2\times1}$	$w_{2\times2}$	...	$w_{2\times N}$

$$\mathrm{Sim}(t_1, t_2) = \cos\theta \cdot \frac{|G_1||G_2|}{(|G_3|)^2}$$

Fig. 1 Workflow of InfDisSim to demonstrate the basic ideas of measuring disease similarity

nodes for disease, a given disease was considered as both the source node and the sink node in the information flow, and damping factor was distributed as 0.85 based on previous study [39]. Assuming N genes exist in the integrative network. Each disease can be represented as N-dimension vector based on the ITM Probe. For a give disease t_1, the weight vector can be described as:

$$WV_{t_1} = \{w_{1,1}, w_{1,2}, \dots, w_{1,i}, \dots, w_{1,N}\}, \tag{1}$$

where WV_{t_1} indicates a weight vector of t_1, and $w_{1,\ i}$ indicates the weight score of t_1 on the ith dimension. Then, disease similarity based on the information flow could be defined as the cosine of their vectors as following:

$$\mathrm{Inf}(t_1, t_2) = \frac{\sum_{i=1}^{N} w_{1,i} \cdot w_{2,i}}{\sqrt{\sum_{i=1}^{N} {w_{1,i}}^2}\sqrt{\sum_{j=1}^{N} {w_{2,j}}^2}}. \tag{2}$$

Because disease similarity could be reflected by semantic associations and the disease gene associations, the disease similarity is defined as following:

$$\mathrm{InfDisSim}(t_1, t_2) = \mathrm{Inf}(t_1, t_2)\frac{|G_1||G_2|}{(|G_{MICA}|)^2}, \tag{3}$$

where G_1, G_2 indicates gene set of t_1 and t_2, respectively. G_{MICA} is the gene set of t_3, which is the most informative common ancestor of t_1 and t_2. And $|.|$ represents the number of terms in the specified set.

According to Lin's research, the definition of similarity between pair of terms of DO is as following:

$$Sim(t_1, t_2) = \frac{2 \times IC(t_{MICA})}{IC(t_1) + IC(t_2)}, \tag{4}$$

or

$$Sim(t_1, t_2) = \frac{\log \frac{|G_{root}|^2}{|G_{MICA}|^2}}{\log \frac{|G_{root}|^2}{|G_1| \cdot |G_2|}}, \tag{5}$$

where G_{root} represents gene sets of the root node of the DAG of DO. According to the eq. 5, the semantic similarity between t_1 and t_2 is proportional to $|G_1|$ and $|G_2|$, and is inversely proportional to $|G_{MICA}|$. Therefore, the proportional relation of Eq. 3 is consistent with the proportional relation of Lin's method.

Assuming T_1 and T_2 are two disease sets, which includes n, and m diseases, respectively. Similarity between two disease sets (Fig. 2) was defined in the eq. 6 as following:

$$sim(T_1, T_2) = \frac{\sum_{1 \le i \le n} Sim(t_{1,i} -> T_2) + \sum_{1 \le j \le m} Sim(t_{2,j} -> T_1)}{n + m}, \tag{6}$$

where $t_{1,i}$, and $t_{2,j}$ represent the ith and jth diseases of T_1 and T_2, respectively. $Sim(t_{1,\ i} -> T_2)$ represents similarity from a disease term of T_1 to T_2. Taken $t_{1,1}$ for example, the eq. 7 gives the definition as following:

$$Sim(t_{1,1} -> T_2) = \max_{1 \le j \le m} sim(t_{1,1}, t_{2,j}). \tag{7}$$

Method for predicting disease-related lncRNAs and miRNAs

Disease-related lncRNAs and miRNAs were indicated applying a global network ranking algorithm called random walk with restart (RWR) [40]. The random walker starts from one or several seed nodes and then randomly transits to neighboring nodes considering the probabilities of the edges connected the two nodes. And the probability of returning to the seed node is supposed as γ. Then, RWR algorithm can be defined as following:

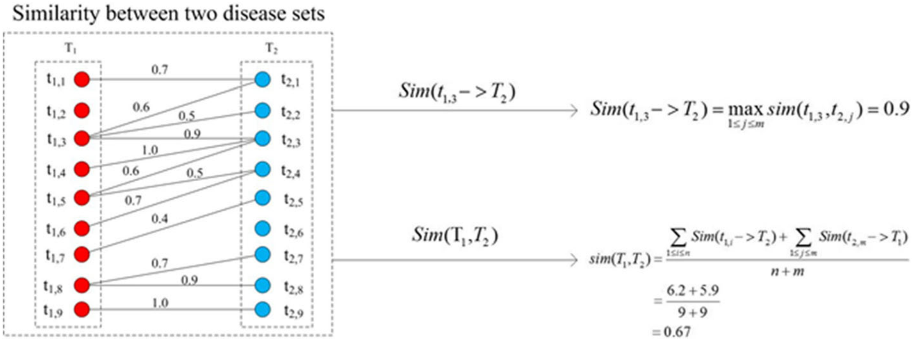


Fig. 2 Shows an example of calculating similarity between disease sets T1 and T2

$$\mathrm{P}_{t+1} = \gamma \mathrm{P}_0 + (1-\gamma)\mathrm{AP}_t, \qquad (8)$$

where P_0 represents the initial probability vector, which changes with the step t and the probability γ, P_t is a vector in which the ith element represents the probability of finding the walker at node i and step t, A indicates the column-normalized adjacency matrix of the network. The algorithm was implemented until the difference between P_t and P_{t+1} falling below 10^{-10}, which indicates all the nodes' status become stable.

Based on our method, researchers can predict novel lncRNA-disease and miRNA-disease associations based on RWR. Firstly, a LSN (MSN) could be constructed for RWR. A lncRNA (miRNA) has associations with a set of diseases. Hence, similarity between two lncRNAs (miRNAs) could be computed based on their related disease sets, which promotes to construct a LSN (MSN). Then, lncRNAs (miRNAs) could be scored for each disease based on RWR, in which the known lncRNAs (miRNAs) of a disease are considered as seed nodes. For each disease, the unknown lncRNAs (miRNAs) of it could be scored. After ranking the lncRNAs (miRNAs) based on the scores, disease-related lncRNAs (miRNAs) are finally predicted.

Method for validating the performance of *InfDisSim*

Figure 3 shows the process of performance validation. At the beginning, a benchmark set including 70 pairs between 47 diseases was derived from two public articles respectively(Additional file 4). One of them is Suthram et al.'s study [41], by which similar pairs of diseases were recognized according to the disease-related mRNA expression data and the human protein interaction network. The other is Pakhomov et al.'s study [42], in which similar pairs of diseases were manually checked by experts in related fields. Then, a random set involving ten times of the benchmark set was obtained from DO. After that, the similarities of benchmark set and random set were calculated by the state-of-art methods including Resnik's, Lin's, Wang's, PSB, SemFunSim, and *InfDisSim*. Finally, the receiver operating characteristic (ROC) curve was drew for assessing the performance of these methods. Furthermore, the experiment was iterated 100 times, and the average of the region under the ROC curve (AUC) for each method was obtained.

Results

Performance evaluation based on benchmark set

ROC curves of the state-of-art methods based on a benchmark set and a random set are shown in Fig. 4a. The figure indicates that the AUCs of Resnik's, Lin's, Wang's, PSB, SemFunSim and *InfDisSim* are 0.6283, 0.6586, 0.6837, 0.8807, 0.9843, and 0.9786, respectively. Obviously, the performances of three typical methods involving Resnik's, Lin's, and Wang's methods are almost the same. And all of these three methods perform generally. By the contrast, three novel methods that predicted more disease gene associations and gene interactions perform superior, of which the performances of SemFunSim and *InfDisSim* are the best and nearly the same.

Resnik's, Lin's, and Wang's methods concentrated on sematic associations. Few of disease gene associations were employed by these three methods. With more and more disease gene associations and gene interactions identified, it is easier to study similarity between diseases in molecular level. Fortunately, three methods including PSB, SemFunSim, and *InfDisSim* have intergrated these associations into semantic associations. It is easy to find the interactions between genes including mRNA coexpression, protein-protein interaction, protein complex, and so on. Although PSB method only applied cooccurrenced biological process of genes, its performance has already been improved. To enhance the performance, SemFunSim and *InfDisSim* methods employed comprehensive gene functional associations from two different views. And both of these two methods perform excellently.

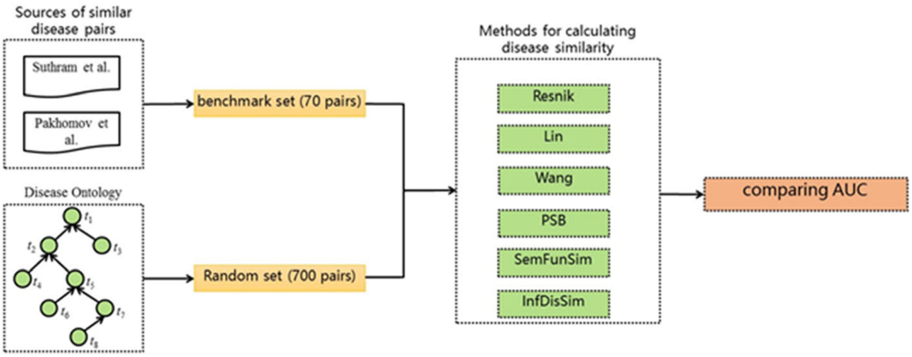


Fig. 3 The process of performance evaluation. AUC represents the area under the receiver operating characteristic curve

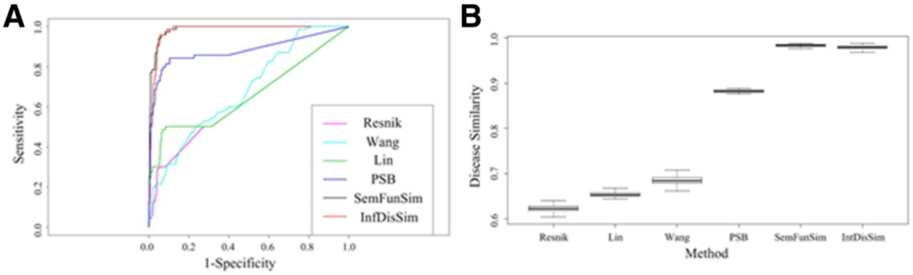


Fig. 4 AUC analysis based on the benchmark set. **a** ROC curves of the state-of-art methods. **b** AUCs of 100 iterators

Figure 4b shows the AUCs of the 100 iterators, which are consistent with the Fig. 4a. From this figure, the average AUCs of the 100 iterators are 0.6223, 0.6538, 0.6851, 0.8824, 0.9832, and 0.9788, respectively.

Relationship between disease similarity by *InfDisSim* and co-occurrence drugs

Previous studies have indicated that similar diseases could have common therapeutic drugs [9, 10]. Therefore, it is possible that similar diseases tend to have more co-occurrence drugs. To prove this, we discuss the relationship of disease similarity by *InfDisSim* with co-occurrence drugs. In this study, we employed the Jaccard index as the measure for disease similarity by drugs. As a consequence, *InfDisSim* disease similarity showed significant positively correlated with the co-occurrence drugs (Pearson correlation $\gamma^2 = 0.1315$, $p = 2.2\text{e-}16$; Fig. 5). Results demonstrate that disease similarity detected by our method is correlated with co-occurrence drugs, which have a very strong correlation with disease similarity.

Application of disease similarity to the prediction of disease-related lncRNAs

For the sake of showing the usefulness of disease similarity computed by our InfDisSim, we firstly constructed a lncRNA similarity network (LSN) based on disease similarity, and then identified disease-related lncRNAs based on LSN. The similarity of each pair of 111 lncRNAs was computed using the eq. 6. After that, the z-score of each pair of lncRNAs was computed based on these scores. Then, each similarity score gained a one-sided *P*-value. Finally, all of these lncRNA similarity scores were appiled to construct LSN (Additional file 5).

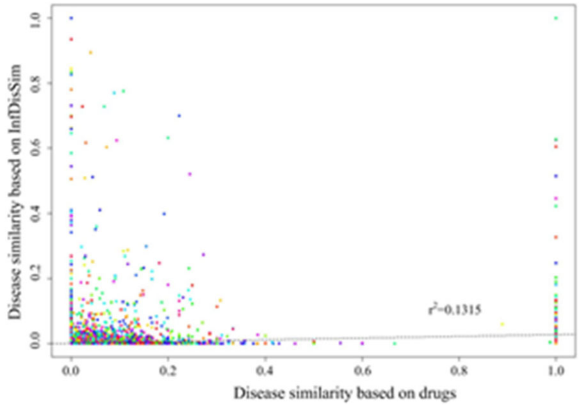


Fig. 5 The relationship between disease similarity based on InfDisSim and co-occurrence drugs

LSN was further employed to predict disease-related lncRNAs employing RWR algorithm. According to the known 331 associations between 125 diseases and 111 lncRNAs, the performance of the LSN was assessed by leave-one-out cross validation. Finally, an AUC of 0.9893 was obtained (Fig. 6).

Application of disease similarity to the prediction of disease-related miRNAs

We also utilized the disease similarity to construct a MSN and predict disease-related miRNAs based on the network. Here, we calculated similarity of each pair of 265 miRNAs and corresponding one-sided P-value. All of these miRNA similarity scores were employed to construct MSN (Additional file 6) for predicting disease-

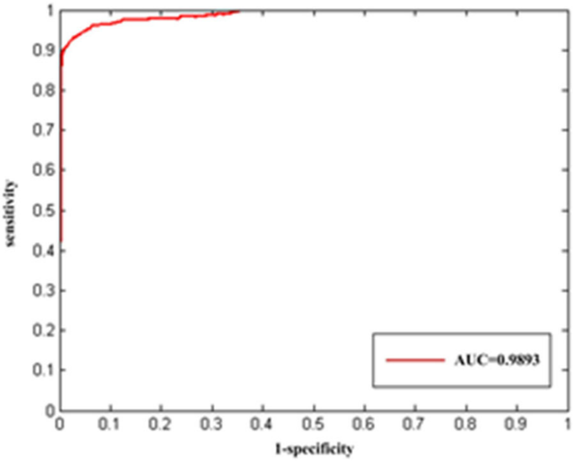


Fig. 6 The ROC curve of our method based on leave-one-out cross validation on experimentally verified lncRNA-disease associations

related miRNAs. The performance of the MSN was assessed by leave-one-out cross validation. As a result, we got an AUC of 0.9007.

Discussion

To identify the disease-related ncRNAs, including lncRNAs and miRNAs, we presented a novel method based on disease similarity using a random walk. With the high AUC performance of predicting disease-related miRNAs and lncRNAs (0.9893, 0.9007), the proposed methods in this paper may also be applied to predict other disease-related modules, e.g. SNP and risk pathways [43, 44].

Conclusions

In this study, we presented a novel method, *InfDisSim*, to figure out disease similarity by semantic association and disease-related genes. In time of computing similarity based on genes, information flow was modelled into a comprehensive GFN, which is constructed by integrating multiple interactions involving mRNA co-expression, protein-protein interaction, protein complex, and so on. In the precious study, SemFunSim has introduced the interactions of pair-wise genes between different gene set. Here, the whole network was fully employed based on information flow. It introduced a novel view to compute disease similarity.

The performance of *InfDisSim* was validated employing the benchmark set. The high AUC (0.9786) indicates its excellent performance. Then, we assessed the observation that similar diseases could have common therapeutic drugs. Finally, *InfDisSim* disease similarity was significant positively correlated with the co-occurrence drugs (Pearson correlation $\gamma^2 = 0.1315$, $p = 2.2e\text{-}16$; Fig. 5). Therefore, *InfDisSim* disease similarity could be utilized to predict potential associations between diseases and drugs.

lncRNA similarity and miRNA similarity could be computed based on *InfDisSim* disease similarity. Here, for all the pairs of lncRNAs (miRNAs), which was applied to construct a LSN (MSN), we calculated their similarities. The network was further used to predicate disease-related lncRNAs (miRNAs). As a result, the high AUC (0.9893, 0.9007) illustrates that the LSN (MSN) is very appropriate for predicting potential associations between diseases and lncRNAs (miRNAs) based on RWR.

Additional files

Additional file 1: Disease-gene associations. (TXT 2080 kb)

Additional file 2: Disease-lncRNA associations. (TXT 3 kb)

Additional file 3: Disease-miRNA associations. (TXT 132 kb)

Additional file 4: Benchmark set of similar disease pairs. (TXT 11 kb)

Additional file 5: lncRNA functional similarity scores. (TXT 3310 kb)

Additional file 6: miRNA functional similarity scores. (TXT 9539 kb)

Acknowledgements

The authors thank for editors and reviewers for the improvement of this manuscript.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. HIT NSRIF 201856), the National Science and Technology Major Project [Nos: 2016YFC1202302], the National Natural Science Foundation of China (No: 61502125 and 61571152), the National High-tech R&D Program of China (863 Program) [Nos: 2014AA021505, 2015AA020101, 2015AA020108], Heilongjiang Postdoctoral Fund (Grant No. LBH-Z6064, LBH-Z15179), and China Postdoctoral Science Foundation (Grant No. 2016M590291). Publication costs were funded by the Fundamental Research Funds for the Central Universities (Grant No. HIT NSRIF 201856), the National Natural Science Foundation of China (No: 61502125 and 61571152), Heilongjiang Postdoctoral Fund (Grant No. LBH-Z6064, LBH-Z15179), and China Postdoctoral Science Foundation (Grant No. 2016M590291).

Availability of data and materials

All data generated or analyzed during this study are included in this published article.

About this supplement

This article has been published as part of *BMC Medical Genomics* Volume 10 Supplement 5, 2017: Selected articles from the IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2016: medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-10-supplement-5.

Authors' contributions

LC, and YH conceived and designed the experiments. LC, MZ, HS, HJ, QJ analysed data. LC wrote this manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China. [2]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150001, China. [3]Department of information engineering, Heilongjiang biological science and technology Career Academy, Harbin 150001, China.

Published: 28 December 2017

References

1. Li J, Gong B, Chen X, Liu T, Wu C, Zhang F, Li C, Li X, Rao S, Li X. DOSim: an R package for similarity between diseases based on disease ontology. BMC bioinformatics. 2011;12:266.
2. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics. 2010;26(13):1644–50.
3. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. 2014;42(Database issue):D1070–4.
4. Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci Rep. 2015;5:13186.

5. Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Mol BioSyst. 2014;10(8):2074–81.
6. Chen X, Yan CC, Luo C, Ji W, Zhang Y, Dai Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. Sci Rep. 2015;5:11338.
7. Cheng L, Shi H, Wang Z, Hu Y, Yang H, Zhou C, Sun J, Zhou M. IntNetLncSim: an integrative network analysis method to infer human lncRNA functional similarity. Oncotarget. 2016;7:47864–74.
8. Zhou M, Wang X, Li J, Hao D, Wang Z, Shi H, Han L, Zhou H, Sun J. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Mol BioSyst. 2015;11(3): 760–9.
9. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol. 2011;7:496.
10. Cheng L, Jiang Y, Wang Z, Shi H, Sun J, Yang H, Zhang S, Hu Y, Zhou M. DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. Sci Rep. 2016;6:30024.
11. Cheng L, Li J, Ju P, Peng J, Wang Y. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. PLoS One. 2014;9(6):e99415.
12. Cheng L, Sun J, Xu W, Dong L, Hu Y, Zhou M. OAHG: an integrated resource for annotating human genes with multi-level ontologies. Sci Rep. 2016;6: 34820.
13. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, et al. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015;43(Database issue):D1071–8.
14. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46.
15. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, et al. The comparative Toxicogenomics database: update 2013. Nucleic Acids Res. 2013;41(Database issue):D1104–14.
16. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for online Mendelian inheritance in man (OMIM®). Hum Mutat. 2011;32(5):564–7.
17. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM. Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc. 2003;2003:460–4.
18. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004;36(5):431–2.
19. Resnik P: Using information content to evaluate semantic similarity in a taxonomy. 1995 *arXiv preprint cmp-lg/9511007*.
20. Lin D. An information-theoretic definition of similarity. ICML. 1998;1998: 296–304.
21. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
22. Mathur S, Dinakarpandian D. Automated ontological gene annotation for computing disease similarity. AMIA Summits Transl Sci Proc. 2010;2010:12–6.
23. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Inform. 2012;45(2):363–71.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25(1):25–9.
25. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302(5643):249–55.
26. Ortutay C, Vihinen M. Identification of candidate disease genes by integrating gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res. 2009;37(2):622–8.
27. Peng J, Wang T, Wang J, Wang Y, Chen J. Extending gene ontology with gene association networks. Bioinformatics. 2016;32(8):1185–94.
28. Peng J, Uygun S, Kim T, Wang Y, Rhee SY, Chen J. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. BMC Bioinformatics. 2015;16:44.
29. Cheng L, Wang G, Li J, Zhang T, Xu P, Wang Y. SIDD: a semantically integrated database towards a global view of human disease. PLoS One. 2013;8(10):e75504.
30. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21(7):1109–21.
31. Sun J, Chen X, Wang Z, Guo M, Shi H, Wang X, Cheng L, Zhou M. A potential prognostic long non-coding RNA signature to predict metastasis-free survival of breast cancer patients. Sci Rep. 2015;5:16553.
32. Zhou M, Guo M, He D, Wang X, Cui Y, Yang H, Hao D, Sun J. A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. J Transl Med. 2015;13:231.
33. Zhou M, Sun Y, Sun Y, Xu W, Zhang Z, Zhao H, Zhong Z, Sun J. Comprehensive analysis of lncRNA expression profiles reveals a novel lncRNA signature to discriminate nonequivalent outcomes in patients with ovarian cancer. Oncotarget. 2016;7:32433–48.
34. Zhou M, Wang X, Shi H, Cheng L, Wang Z, Zhao H, Yang L, Sun J. Characterization of long non-coding RNA-associated ceRNA network to reveal potential prognostic lncRNA biomarkers in human ovarian cancer. Oncotarget. 2016;7(11):12598–611.
35. Zhou M, Xu W, Yue X, Zhao H, Wang Z, Shi H, Cheng L, Sun J. Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma. Oncotarget. 2016;7:29720–38.
36. Zhou M, Zhao H, Wang Z, Cheng L, Yang L, Shi H, Yang H, Sun J. Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. J Exp Clin Cancer Res. 2015;34:102.
37. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013;41(Database issue):D983–6.
38. Stojmirovic A, Yu YK. ITM probe: analyzing information flow in protein networks. Bioinformatics. 2009;25(18):2447–9.
39. Stojmirovic A, Yu YK. Information flow in interaction networks II: channels, path lengths, and potentials. J Comput Biol. 2012;19(4):379–403.
40. Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82(4):949–58.
41. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol. 2010;6(2): e1000662.
42. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. AMIA Ann Symp Proc. 2010;2010:572.
43. Jiang Q, Jin S, Jiang Y, Liao M, Feng R, Zhang L, Liu G, Hao J. Alzheimer's disease variants with the genome-wide significance are significantly enriched in immune pathways and active in immune cells. Molecular Neurobiology. 2017;54(1):594–600.
44. Liu G, Zhang F, Jiang Y, Hu Y, Gong Z, Liu S, Chen X, Jiang Q, Hao J. Integrating genome-wide association studies and gene expression data highlights dysregulated multiple sclerosis risk pathways. Multiple Sclerosis Journal. 2017;23(2):205–212.