Deng *et al. BMC Medical Informatics and Decision Making* 2019, **19**(Suppl 6):269
https://doi.org/10.1186/s12911-019-0968-8

BMC Medical Informatics and Decision Making

RESEARCH Open Access

MultiSourcDSim: an integrated approach for exploring disease similarity

Lei Deng[1], Danyi Ye[1], Junmin Zhao[2] and Jingpu Zhang[2*]

From: IEEE International Conference on Bioinformatics and Biomedicine (2018)
Madrid, Spain 3-6 December 2018

Abstract

Background: A growing collection of disease-associated data contributes to study the association between diseases. Discovering closely related diseases plays a crucial role in revealing their common pathogenic mechanisms. This might further imply treatment that can be appropriated from one disease to another. During the past decades, a number of approaches for calculating disease similarity have been developed. However, most of them are designed to take advantage of single or few data sources, which results in their low accuracy.

Methods: In this paper, we propose a novel method, called MultiSourcDSim, to calculate disease similarity by integrating multiple data sources, namely, gene-disease associations, GO biological process-disease associations and symptom-disease associations. Firstly, we establish three disease similarity networks according to the three disease-related data sources respectively. Secondly, the representation of each node is obtained by integrating the three small disease similarity networks. In the end, the learned representations are applied to calculate the similarity between diseases.

Results: Our approach shows the best performance compared to the other three popular methods. Besides, the similarity network built by MultiSourcDSim suggests that our method can also uncover the latent relationships between diseases.

Conclusions: MultiSourcDSim is an efficient approach to predict similarity between diseases.

Keywords: Disease similarity network, Diffusion component analysis, Integrating multiple data sources

Background

Quantitative measurement of disease similarity is gaining more and more attentions because it helps to reveal common psychophysiology and improve clinical decision-making systems, so as to better understand human diseases status and more accurately classify diseases [1]. It also plays a crucial role in identifying novel drug indications [2], since diseases may have the same or similar therapeutic targets, suggesting that they may be treated with the same or similar drugs [3–6]. In the past few decades, our understanding of human diseases has made remarkable progress [7]. For example, the network-based approaches [8–11] to calculating the similarity between diseases is impressive. Constructing a disease similarity network based on biological data to explore the relationship between diseases has become one of the research hotspots of modern biology and medicine. At present, the measurement of similarity disease research is necessary.

In previous studies, various properties of human genes (such as predicted function or amino-acid sequence length) and Gene Ontology (GO) [12–14] biological processes have been correlated with the chance of causing a disease [15–17]. The calculation approaches of disease similarity can be roughly divided into function-based methods [18, 19] and semantic-based methods [20]. The functional-based approach calculates similarities between diseases by comparing genes associated with diseases [18, 19]. For instance, the BOG [18] method, which was

*Correspondence: zhangjp@csu.edu.cn
[2]School of Computer and Data Science, Henan University of Urban Construction, 467000, Pingdingshan, China
Full list of author information is available at the end of the article

designed by Mathur and Dinakarpandian, calculates the similarity between diseases by comparing gene overlaps of related diseases. Moreover, BOG [18] also considers the self-information of each disease. However, its shortcoming is that it does not consider the functional link between disease-related genes. Further, Mathur and Dinakarpandian proposed a method based on process similarity (PSB [19]). The method provides functions to measure similarity, including the similarity function based on GO terms [12], and the similarity function between entities annotated with terms extracted from the ontology based on both co-occurrence and information content. The semantic-based method is extensively used in biomedical and bioinformatics. For instance, Resnik's method [21] calculates the similarity between diseases according to the information content of the most informative common ancestor. Lin's method [22] incorporates not only the information content of the most informative common ancestor but also the the information content of the two disease terms. Jiang and Conrath et al. [23] represented the similarity between two terms through the semantic distance.

In addition, phenotype similarity plays an important part in a lot of biological similarity and biomedical applications, and it is also the most common way of classifying diseases [24]. For example, the Human phenotype ontology (HPO) is a controlled and standardized vocabulary that describes the abnormal phenotype of human disease. And Medical Subject Headings (MeSH) [25] use this approach to classify diseases.

Although there are many patterns for measuring similarity between diseases, most of them use a single biological data source, and few methods using multiple biological data sources are proposed. For example, some of the previous approaches calculate the similarity according to genes related with diseases. Nevertheless, there exist some diseases which are unrelated or rarely related to genes. Thus, depending solely on individual biological data associated with disease might greatly affects the prediction performance of the methods. In this work, a novel approach named MultiSourcDSim is proposed to compute the similarity between diseases by integrating multiple biological datasets. In MultiSourcDSim, firstly, three disease similarity networks are respectively built by using a variety of biological data such as gene-disease associations, GO biological process-disease associations and symptom-disease associations. Secondly, the high-dimensional vector of each node is extracted by running restart random walks [26] on each network, and low-dimensional vectors that can represent the high-dimensional topological patterns in each network are learned. Finally, the similarity between diseases is obtained by calculating the cosine score between two low-dimensional vectors. The experiments demonstrate that disease similarity predicted by our method is significantly correlated with disease category of MeSH, implying that the network constructed by our method is capable of detecting the latent relationships between diseases. Moreover, the results also show that MultiSourcDSim outperform the other three popular methods.

Methods

Datasets

CTD's MEDIC disease vocabulary which is downloaded in http://ctdbase.org (March 4, 2018) is chosen as criterion for describing diseases. CTD's MEDIC disease vocabulary is a modified subset of descriptors from the Diseases [C] branch of the U.S. National Library of Medicine's MeSH, combined with genetic disorders from the Online Mendelian Inheritance in Man (OMIM) database, and we use MeSH to mark disease terms. Each record in CTD's MEDIC disease vocabulary contains 9 fields, 4 of which are retained for calculating disease similarity. They are respectively DiseaseID, DiseaseName, AltDiseaseIDs (alternative identifiers) and ParentIDs (identifiers of the parent terms).

We have collected three data sets associated with disease, namely gene-disease associations, GO biological process-disease associations, and symptom-disease associations. In the three sets, a great deal of biological information bound up with diseases is included. For instance, each record in the gene-disease associations contains 9 fields (GeneSymbol, GeneID, DiseaseName, DiseaseID, DirectEvidence, InferenceChemicalName, InferenceScore, OmimIDs, PubMedIDs). In the three data sets, 3,125,954 gene-disease associations containing 3254 disease terms and 668,760 GO biological process-disease associations containing 5720 disease terms are pooled from http://ctdbase.org(March 4, 2018), and each record in the two data sets is identified by MeSH markers. The gene terms and the gene ontology biological process terms are labeled with the NCBI gene identifiers and GO identifiers, respectively. The 80,638 symptom-disease associations are collected from paper [27], which describes 4040 diseases. However, the diseases in the symptom-disease associations are marked by the MeSH names. To obtain the Mesh identifiers corresponding to the names, we map the disease names in the symptom-disease associations to the IDs in the CTD's MEDIC disease vocabulary. After screening for the co-occurring diseases term in all associations, 8126 diseases are extracted.

Overview of MultiSourcDSim

In our method, we combine three disease-related data sets to calculate the similarity between diseases more accurately. Specifically, we firstly construct three disease similarity networks through computing the similarity

respectively according to the gene-disease associations, GO biological process-disease associations, and symptom-disease associations. Secondly, the compact low-dimensional feature representations of diseases from the three similarity networks are learned by running Diffusion Component Analysis (DCA) [28–30]. Finally, the disease similarity is calculated according to the learned representations.

Calculate semantic similarity of diseases

MeSH is a vocabulary that gives uniformity and consistency to the indexing and cataloging of biomedical literature. It is organized in a manner of tree structures with 16 main branches. Category C represents diseases. In our approach, the semantic similarity of diseases is measured by using the special structure between MeSH descriptor [25]. We build a directed acyclic graph (DAG) to clarify the associations among various diseases. The nodes in the DAG represent the MeSH descriptor. Child nodes are more specialized (containing more disease information) and parent nodes are more generalized (containing less disease information). In addition to the relationships of the disease itself, we also combine the relationships between disease and other biological entity, namely gene, GO and symptom. The probability of a disease occurs in a disease-related data set is just its frequency in the data set. The frequency of a disease term t is calculated as:

$$f(t) = self(t) + \sum_{tc \in children(t)} f(tc). \tag{1}$$

Here, $self(t)$ represents the number of occurrences of the disease term t in a single data set, and the disease term tc is a direct child of the disease item t, belonging to the $children(t)$ collection. In other words, the frequency of the disease term t in a single disease-related data set is defined as the frequency of its own occurrence plus the frequency of occurrence of all its child nodes. The probability that the disease term t appears in the disease-related data set is as follows:

$$prob(t) = \frac{f(t)}{N}. \tag{2}$$

Here, N indicates the frequency of occurrence of the root node in the corresponding DAG.

Then, the similarity scores are computed according to the probabilities of diseases based on the metric proposed by Lin et al. [22]. In Lin's method, the similarity is measured in terms of information theory. It is believed that the similarity between terms is determined by their generality (information content of common ancestor nodes) and particularity (their respective information content). Therefore, the semantic similarity depends on the maximum ratio of the information content of the common ancestor nodes of the two terms to the sum of the information content of the two terms themselves. Generally, the higher the degree of information sharing between two terms, the higher the semantic similarity score, and on the contrary, the lower the similarity score. This definition is as follows:

$$Score(t1, t2) = \max_{t \in (LCA(t1,t2))} \left(\frac{2 * \log prob(t)}{\log prob(t1) + \log prob(t2)} \right). \tag{3}$$

Here, $LCA(t1, t2)$ is the set of least common ancestors of term $t1$ and $t2$. The similarity scores fall in the range [0, 1].

Integrate multiple networks and learn representations

We construct three disease similarity networks according to the similarity scores. To achieve the compact integration of multiple similarity network, we adopt DCA strategy to capture low-dimensional vectors representing topological patterns of networks. In DCA, the random walk with restart (RWR) method [26] is firstly employed to analyze the structure of each network.

The RWR from a node i is defined as:

$$s_i^{t+1} = (1 - a)s_i^t T + ae_i. \tag{4}$$

Here, T denotes the probability transfer matrix. s_i^t is specified as an n-dimensional vector, where each entry is the probability of visiting a node at t iterations from the initial node i. e_i is the initial probability vector, where $e_i(i)$=1 and $e_i(j)$=0, $\forall j \neq i$. a is the restart probability. After several iterations, a stable distribution is obtained, and s_i is regard as the 'diffusion state' of the node i.

There exists noise in the diffusion states obtained in this manner, and the dimensionality is high. To solve this problem, we utilize fewer dimensions to approximate each diffusion state s_i through a polynomial logistic model based on the potential vector representation of nodes in a network. Specifically, the probability assigned to node j in the diffusion state of node i is as follows:

$$\hat{s}_{ij} = \frac{exp\left\{x_i^T w_j\right\}}{\sum_{j'} exp\left\{x_i^T w_j'\right\}}, \tag{5}$$

where $\forall i, x_i, w_j \in R^d$ for $d \ll n$. x_i and w_j represent the node feature and context feature of node i respectively.

The goal is to find the low-dimensional vector representation of nodes w and x that best approximates a set of observed diffusion states $s = \{s_1, \ldots, s_n\}$ according to the logistic model. To achieve the goal, KL-divergence is used as the objective function to optimize, which is given by:

$$\min_{w,x} C(s, \hat{s}) = \frac{1}{n} \sum_{i=1}^{n} D_{KL}\left(s_i || \hat{s}_i\right), \tag{6}$$

where n is the number of nodes. By writing out the definition of KL-divergence, the formula is written as:

$$C\left(s,\hat{s}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[-H(s_i) - \sum_{j=1}^{n} s_{ij}\left(x_i^T w_j - \log\left(\sum_{j'=1}^{n}\exp\left\{x_i^T w_{j'}\right\}\right)\right)\right], \tag{7}$$

where $H(\cdot)$ denotes the entropy. In order to combine the three disease similarity networks, the formula (6) is modified as follows:

$$\min_{w,x} C(s,\hat{s}) = \frac{1}{n}\sum_{m=1}^{M}\sum_{i=1}^{n} D_{KL}\left(s_i^m||\hat{s}_i^m\right). \tag{8}$$

Here, M represents the number of networks. In this work, M is equal to 3. To minimize the objective function, we compute the gradients with regard to the parameters w and x. The low-dimensional vector representations are obtained by the quasi-Newton L-BFGS method with these gradients.

To improve efficiency, we can employ singular value decomposition (SVD) to optimize the alternative objective function [31].

Calculate the similarity between diseases

After extracting the low-dimensional representations for all nodes which can best explain the connectivity patterns in the networks, we utilize the learned representations as features for calculating the disease similarity. In this study, the number of nodes in the three networks, namely the total number of diseases is 8126, and the dimension of these features is set to 600. The similarity between diseases is measured through cosine score, which is as follows:

$$cosine(d_x, d_y) = \frac{\sum_i d_{x,i} d_{y,i}}{\sqrt{\sum_i d_{x,i}^2 d_{y,i}^2}}. \tag{9}$$

Here, d_x and d_y are two vectors which represent two disease respectively. Obviously, the similarity is between 0 and 1.

Results

The degree distribution of disease similarity networks

We adopt gene-disease associations, GO biological process-disease associations and symptom-disease associations as the sources of disease similarity network, and construct the small similarity networks based on the Lin's measure separately. In order to better understand the topology of these networks, we calculate the degree distribution of nodes in the network. Figures 1, 2 and 3 elucidates the degree distribution of disease node in three small disease similarity networks.

In the disease similarity network based on gene-disease association dataset (GDN), there exist 3254 diseases and 32733 connections. Marfan Syndrome (MeSH: D008382), which is the relation with 178 diseases, has the maximum degree. There are 225 diseases with degree 1 (Fig. 1). 5720 diseases and 249490 relationships make up the disease similarity network based on GO biological process-disease association dataset (BPDN). The disease with the maximum degree is Martin-Probst Deafness-Mental Retardation Syndrome (MeSH: C564495), the degree is 1024. As shown in Fig. 2, nearly half of the disease nodes have margins with about 100 other disease nodes. And similarity values of all disease pairs are computed in the disease similarity network based on symptom-disease association dataset (SDN), and the distribution of 48279 similarity values (between 4040 diseases) is acquired.

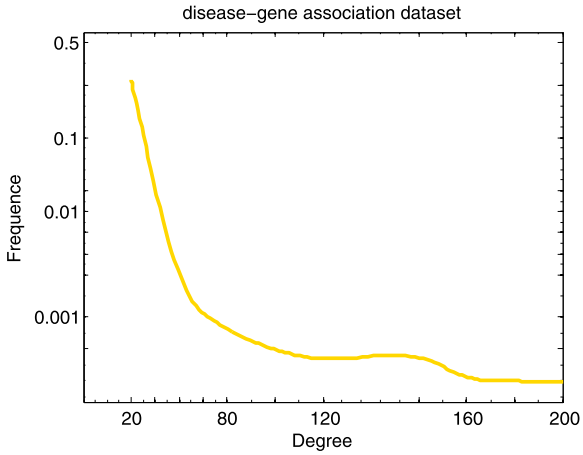

Fig. 1 Degree distribution of disease node in the small similarity network built based on disease-gene association dataset

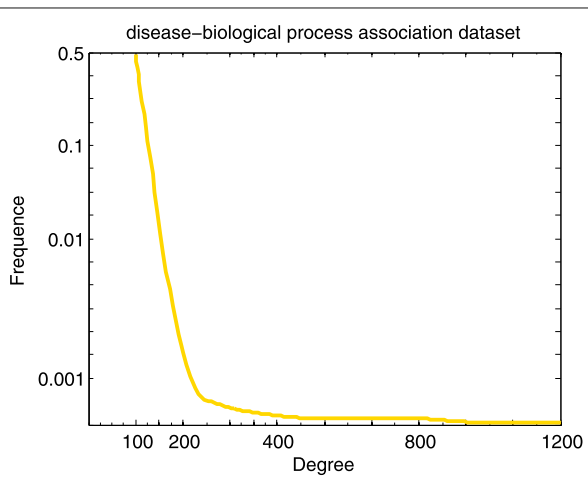

Fig. 2 Degree distribution of disease node in the small similarity network constructed based on GO biological process-disease association dataset

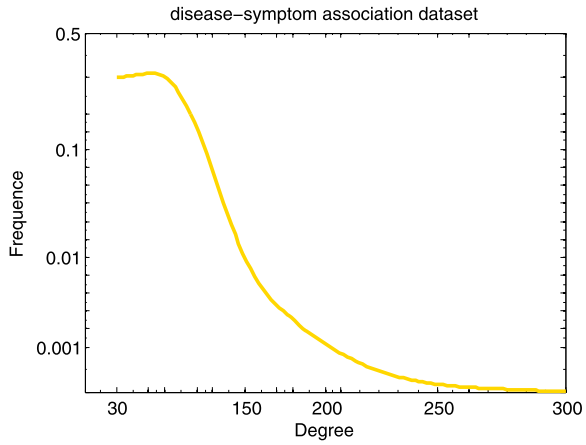

Fig. 3 Degree distribution of disease node in the small similarity network constructed based on disease-symptom association dataset

Oculocerebrorenal Syndrome (MeSH: D009800) associated with 256 diseases has the maximum degree (Fig. 3). From the above calculation we can draw a conclusion that the density of GDN is the largest compared to BPDN and SDN.

After obtaining the integrated disease similarity network (GPSN), the distribution of these similarity scores are also counted. The distribution is represented in Fig. 4, the similarity scores for most disease pairs across the network ranges from 0 to 0.6. The number of disease pairs in the 0.2-0.3 similarity bin is the highest, followed by the 0.3-0.4 bin.

Benchmark

The benchmark set which is adopted in this experiment contains 40 pairs of highly similar diseases. It is derived from the work of Suthram et al. [1] and Pakhomov et al. [32], and cancers are deleted. The benchmark set consists of pairs of diseases that are confirmed to be interrelated, such as Polycystic Ovary Syndrome(MeSH: D011085) and Obesity(MeSH: D009765), Chronic Obstructive Airway Disease(MeSH: D029424) and Asthma (MeSH: D001249). It also contains some diseases pairs which have no apparent correlations, but have proved to be correlated through various evidences, such as Obesity and Asthma, Malaria (MeSH: D008288) and Anemia (MeSH: D000740). Moreover, we randomly choose 500 disease pairs from the similarity network as a random set, where the disease pairs in the benchmark set are deleted.

Parameter selection

There are two parameters (α and d) to be tuned in MultiSourcDSim. The parameter α is the restart probability. According to previous practical experience [33], it is set to 0.5. The parameter d denotes the feature dimension of each node. We compare the performance for different numbers of dimensions based on the benchmark set. We calculate the values of AUC when d is increasing from 500 to 800 with step size 100. As shown in Fig. 5, the results show that the performance of MultiSourcDSim is stable over a wide range of values for the number of dimensions, implying that our method is robust to over-fitting. On the whole, the AUC comes to the max value when d equals 600. Hence, d is set to 600 in this paper.

Performances assessment

To evaluate the disease similarity results calculated by MultiSourcDSim, we make a comparison on the disease classification of MeSH. MeSH is an authoritative medical thesaurus and the basis for biomedical indexing. MeSH divides the disease (C) sections into 26 categories according to the tree code (excluding some ambiguous categories). To discuss whether GPSN is related to the MeSH disease category, we examine the difference between the

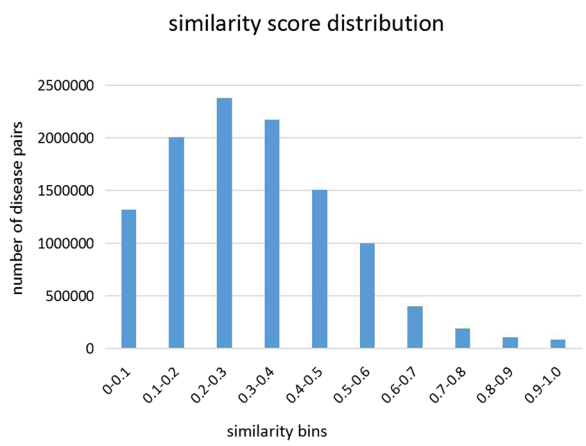

Fig. 4 Histogram of similarity scores between 8126 disease nodes. Most disease-disease pairs have a low similarity score

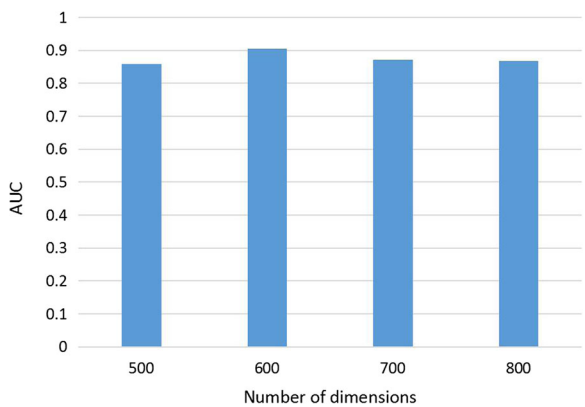

Fig. 5 Comparision for different numbers of dimensions

similarity scores of disease pairs belonging to the same MeSH category and the similarity scores of disease pairs of different MeSH categories. As demonstrated in Fig. 6, the average similarity scores for disease pairs from the same MeSH are significantly higher than those from different MeSH categories. In conclusion, the experiment demonstrates that the similarity scores of disease pairs are closely relevant to MeSH disease category.

Moreover, in order to verify that the performance of the network integrating the three data sets is better than that of the network formed by the single data set, we compare GDN, BPDN, SDN and GPSN based on the banchmark set and random set. AS shown in Fig. 7, MultiSourcDSim achieves the best AUC of 0.906, and the AUC values of GDN, BPDN and SDN are 0.771, 0.774 and 0.797, respectively. This result indicates that compared to individual networks without integration, MultiSourcDSim has a more stable and stronger power for discovering disease-disease associations. The performance improvement is partially attributed to the fact that synthetical analyzing the structure of the the multiple networks can uncover fine-grained topological patterns. Another important factor is the compactness of the feature representations, which help capture the relevant topological patterns apart from noise in the data.

The performance of MultiSourcDSim is further evaluated by comparing it with other three recent approaches: the text-based approach, namely MimMiner [34], an integrated semantic and functional approach, called MedNetSim [35], and the web-based approach, HSDN [27].

To fairly compare the performance of these methods, we select widely used metrics, such as accuracy (ACC), the area under the ROC curve (AUC), F1-score (F1), the Matthew's correlation coefficient (MCC), precision (PRE), sensitivity (SEN/Recall) and specificity (SPE). Based on the four approaches, we compute the the similarity scores of disease pairs in the benchmark set and the random set, and sort them in descending order, respectively. Moreover, we look on the disease pairs in the benchmark set and the random set as positive and negative samples,

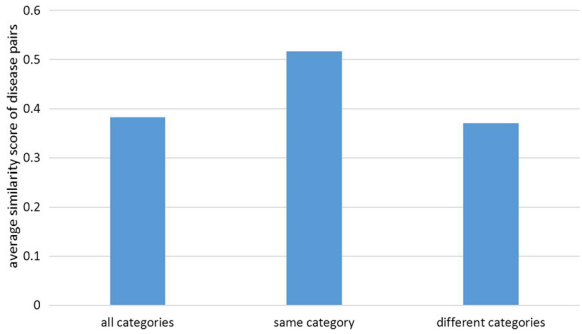

Fig. 6 Evaluation of MultiSourcDSim against MeSH classification

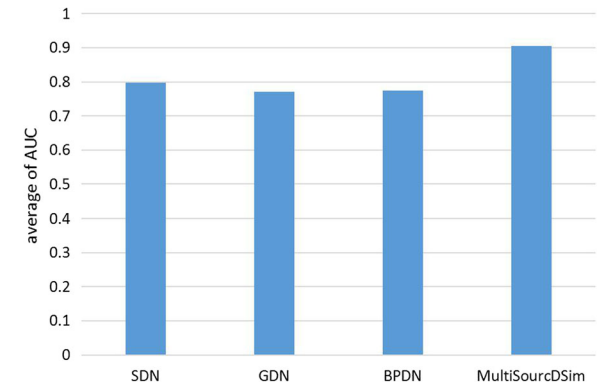

Fig. 7 Integrating Multiple Networks Outperforms Individual Networks

respectively. The disease pairs correctly predicted in the benchmark set are considered to be true positive samples, and the disease pairs in the random set which are predicted to be highly correlated are thought of as false positive samples. The results of the evaluation are shown in Table 1, where the AUC value of the HSDN method is the minimum, which is 0.818. The MimMiner method applies text mining to disease classification and improves performance, resulting in an AUC of 0.836. The MedNetSim method takes the entire protein interactions and the biomedical literature corpus into consideration, increasing its AUC to 0.854. Our approach integrates multiple disease-related data sets and further improves the performance with an AUC value of 0.905, which is the best in the four methods. In addition, our method also achieves the highest values for ACC, F1, MCC, PRE, and SEN, which are 0.815, 0.684, 0.273, 0.601, and 0.750, separately.

The results in Table 1 demonstrate that calculating disease similarity by integrating multiple disease-related data sources is an effective method. In order to test the stability of our method, we randomly select 100 disease pairs and compute their similarity scores. The calculations are repeated 100 times and the average AUC of the four methods are depicted in Fig. 8. The average values are respectively 0.819 (HSDN), 0.835 (MimMiner), 0.855 (MedNetSim) and 0.906 (MultiSourcDsim), which are consistent with the AUC column in Table 1. We further compare the ranking of disease pairs derived from the benchmark set. As shown in Fig. 9, The number of the solution disease pairs which are found by MultiSourcDsim always are the largest in the top 220 disease pairs.

In addition, by using the lowest ranked disease pairs in 540 disease pairs (500 random disease pairs and 40 benchmark pairs), MultiSourcDSim can find all 40 benchmark pairs, which represents quite good performance. For example, Obesity (MeSH: D009765) and Asthma (MeSH: D001249) are disease pairs belonging to the benchmark

Table 1 Prediction performance of MultiSourcDsim in comparison with other three methods on the benchmark set and random set

Methods	ACC	AUC	F1	MCC	PRE	SEN	SPE
MultiSourcDSim	**0.815**	**0.905**	**0.684**	**0.273**	**0.601**	**0.750**	0.656
HSDN	0.688	0.818	0.409	0.263	0.375	0.450	0.750
MimMiner	0.652	0.836	0.400	0.259	0.334	0.500	**0.875**
MedNetSim	0.630	0.854	0.391	0.224	0.361	0.425	0.874

The black bold fonts represent the optimal value

set, which ranks last in our approach. As shown in Table 2, the average ranking of Obesity and Asthma is very low among all the four methods. Nevertheless, compared to the other three methods, our approach has increased the ranking of Obesity and Asthma by 9%-14%.

Integrated disease similarity network

We construct a disease similarity network by using the top-ranking 0.3% of the similarity values in 8126 diseases. As shown in Fig. 10, there are 2604 diseases in the network and they are connected to each other by 121787 edges. The maximum connected component consists of 283 nodes. Martin-Probst Deafness-Mental Retardation Syndrome (MeSH: C564495), which is connected to 511 diseases, has the maximum degree. In Fig. 10, nodes in the network represent diseases, and the nodes are colored different colors. Each color is corresponding to a different MeSH category, such as Virus Diseases (MeSH: C02), Digestive System Diseases (MeSH: C06), Eye Diseases (MeSH: C11), Immune System Diseases (MeSH: C20) and so on. For each classification, diseases in the same MeSH category are usually similar to each other, such as disease of Musculoskeletal Diseases (MeSH: C05) category, disease of Nervous System Diseases (MeSH: C10) category, and so on. Figure 11 also shows the feature that diseases within one class are more probable to gather in the same neighbourhood with each other. For instance, 5 diseases belonging to the Otorhinolaryngologic Diseases classification constitute a small component. As shown in the Fig. 11a, all of these 5 diseases are deafness. Six diseases generate another connected component (Fig. 11b), five of which are Otorhinolaryngologic Diseases and the other is Stomatognathic Diseases. These demonstrations further indicate that the similarity scores of disease pairs belonging to the same category in the results computed by MultiSourcDSim are greater than those between belonging to different categories.

Besides identifying relationships between diseases belonging to the same disease classification, our approach can also find the associations beween diseases belonging to different classifications. For instance, as shown in Fig. 11c, three Musculoskeletal Diseases are linked to two Immune System Diseases by our method. Among the three Musculoskeletal Diseases, it has been reported that people with Lymphopenia might have immune system diseases.

Discussion and conclusion

Determining the correlation between diseases helps to deepen understanding of the potential mechanisms among diseases. There are many studies about the association between diseases, such as predicting disease-related genes [36–38] and new drug indications [2]. In addition, a huge challenge for researchers in modern biology [39, 40] is how to get more information about the disease. In the

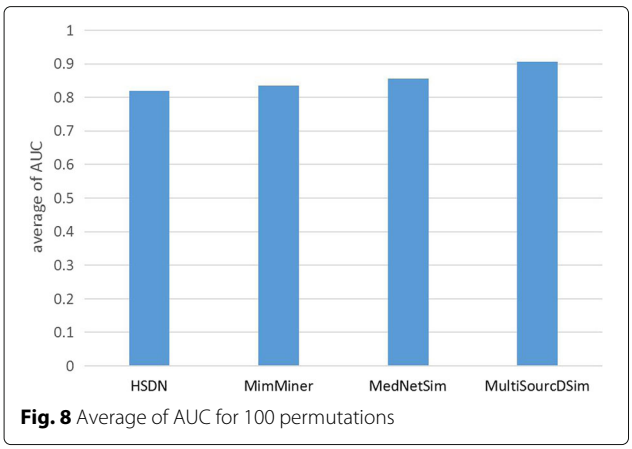

Fig. 8 Average of AUC for 100 permutations

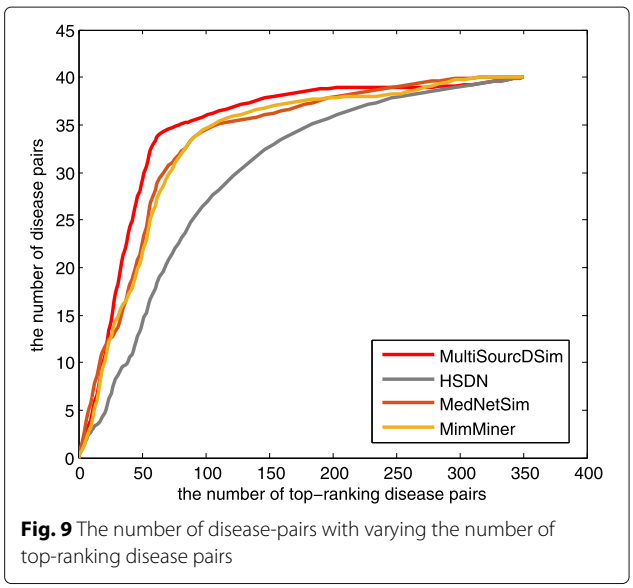

Fig. 9 The number of disease-pairs with varying the number of top-ranking disease pairs

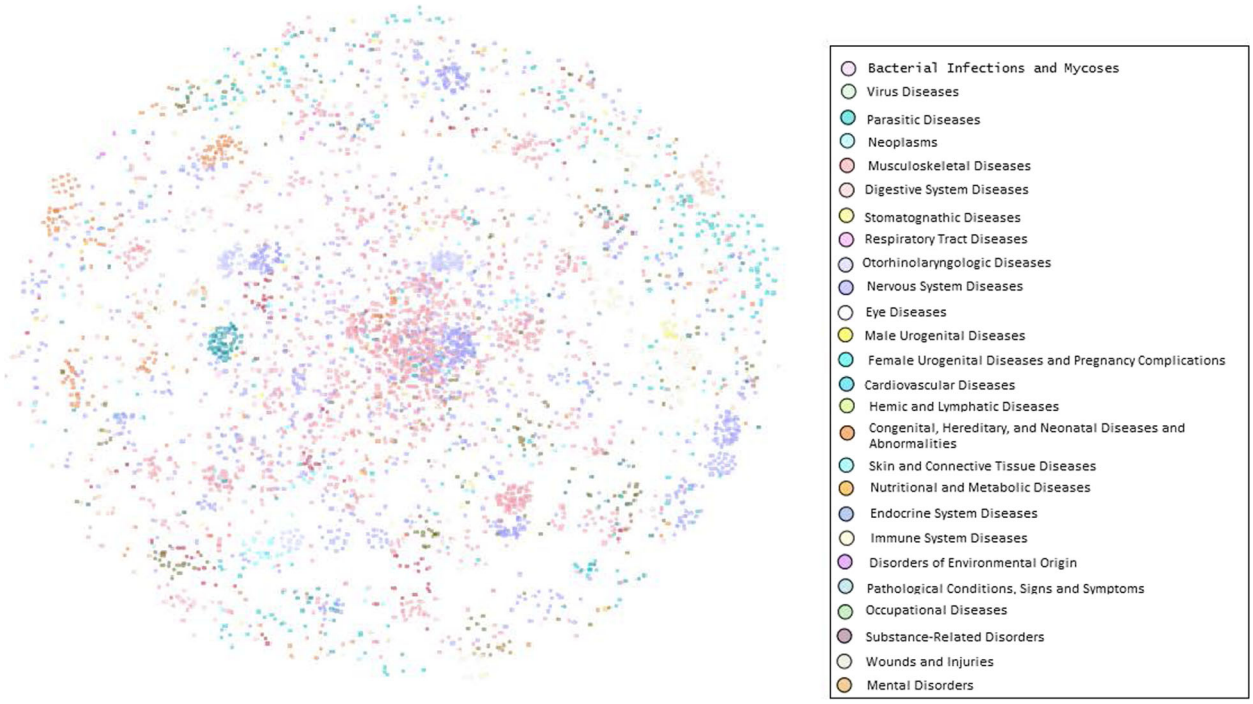

Fig. 10 An overview of disease similarity network (GPSN) based on our method results. Nodes were coloured according to the MeSH category to which they belong

past few decades, many researchers have proposed a number of methods to predict the similarity between diseases (for example, build a network of disease similarity) based on biological data and make a great progress. However, these methods use only a single biological data and do not consider combining multiple biological data as a basis for predicting disease similarity.

In this paper, we propose a novel method, MultiSourcDSim, to predict similarity between diseases, which builds a disease similarity network based on multi-faceted biological data related to disease. According to the similarity scores computed by our method, we can conclude that the similarity scores of disease pairs belonging to the same MeSH classification are significantly higher than those of disease pairs belonging to different MeSH classifications. And, comparing the performance of the MultiSourcDSim method with the other three methods (MimMiner [34], MedNetSim [35] and HSDN [27]) under the same benchmark set, we have found that our method is superior. Furthermore, the disease similarity network constructed by our method can also uncover latent relationships between diseases.

Although multiple disease-related data sources are integrated to compute similarities between diseases, there may be some bias due to incomplete data. In addition to considering the integration of multiple biological data, we

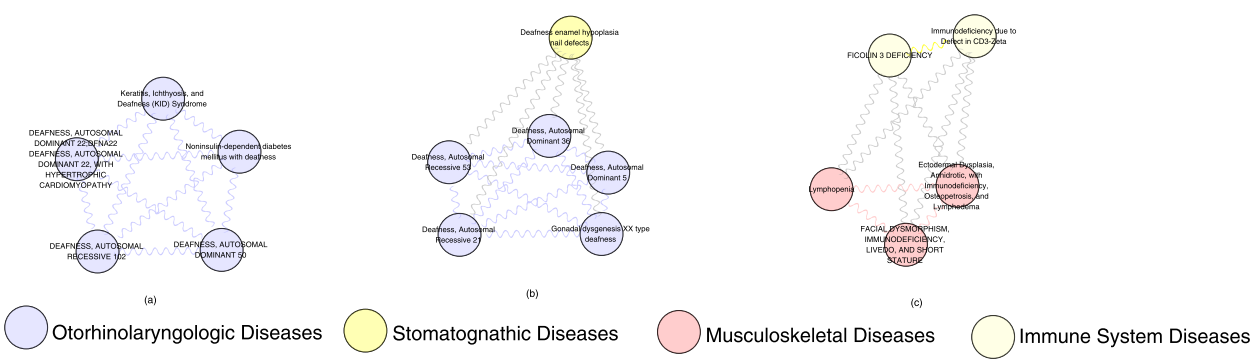

Fig. 11 Three connected components from the disease similarity network constructed by our method

Table 2 The average ranking of the disease pair (Obesity and Asthma) in 540 disease pairs

	HSDN	MimMiner	MedNetSim	MultiSourcDSim
average ranking	252.5	257.4	242.9	**220.6**

also need to take into account the modular nature of each disease in further study of the similarities between diseases, since the modularity of each disease module can give more information [41–43]. Moreover, disease networks have proven useful for predicting novel therapeutic applications of known compounds [44] and inferring novel disease genes [45].

Abbreviations

ACC: Accuracy; AUC: The area under the ROC curve; BPDN: GO biological process-disease association network; CTD: The comparative toxicogenomics database; DAG: Directed acyclic graph; DCA: Diffusion component analysis; F1: F1 score; GDN: Gene-disease association network; GO: Gene ontology; GPSN: The integrated disease similarity network; HPO: Human phenotype ontology; MCC: The Matthew's correlation coefficient; MeSH: Medical subject headings; NCBI: National center for biotechnology information; OMIM: Online mendelian inheritance in man; PRE: Precision; RWR: Random walk with restart; SDN: Symptom-disease association network; SPE: Specificity

Acknowledgements

This work was supported by National Natural Science Foundation of China under grants No. 61972422, No. 61672541 and No. 61672113.

Authors' contributions

LD, DY, JZ and JP designed the study and conducted experiments. LD and DY performed statistical analyses. LD, DY and JP drafted the manuscript. DY prepared the experimental materials and benchmarks. All authors have read and approved the final manuscript.

Funding

Publication costs are funded by National Natural Science Foundation of China under grant No. 61672541.

Availability of data and materials

The datasets used in this study is available at http://ctdbase.org.

About this supplement

This article has been published as part of *BMC Medical informatics and Decision Making Volume 19 Supplement 6, 2019: Selected articles from the IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2018: medical informatics and decision making*. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-19-supplement-6.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer Science and Engineering, Central South University, 410075, Changsha, China. [2]School of Computer and Data Science, Henan University of Urban Construction, 467000, Pingdingshan, China.

Published: 19 December 2019

References

1. Suthram S, Dudley JT, Chiang AP, Rong C, 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):1000662.
2. Gottlieb A, Stein GY, Ruppin E, Sharan R. Predict: a method for inferring novel drug indications with application to personalized medicine. Mole Syst Biol. 2011;7(1):496.
3. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Nat Acad Sci USA. 2007;104(21):8685–90.
4. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. Plos One. 2009;4(8):6536.
5. Zhang X, Zhang R, Jiang Y, Sun P, Tang G, Wang X, Lv H, Li X. The expanded human disease network combining protein-protein interaction information. Eur J Human Genet Ejhg. 2011;19(7):783–8.
6. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási AL. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci USA. 2008;105(29):9880–5.
7. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genet. 2003;33(33 Suppl):228–37.
8. Emmert-Streib F, Dehmer M. Analysis of Microarray Data: A Network-Based Approach: Wiley; 2008.
9. Emmertstreib F, Glazko GV. Network biology: a direct approach to study biological function. Wiley Interdiscipl Rev Syst Biol Med. 2011;3(4):379–91.
10. Jin L, Min L, Wei L, Wu FX, Yi P, Wang J. Classification of alzheimer's disease using whole brain hierarchical network. IEEE/ACM Trans Comput Biol Bioinforma. 2018;PP(99):624–32.
11. Chen B, Li M, Wang J, Shang X, Wu FX. A fast and high performance multiple data integration algorithm for identifying human disease genes. Bmc Med Genomics. 2015;8(S3):1–11.
12. Consortium TGO, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS. Gene ontology: tool for the unification of biology. Nature Genet. 2000;25(1):25–9.
13. Zeng C, Zhan W, Deng L. SDADB: A functional annotation database of protein structural domains. Database. 20181–8.
14. Zhang Z, Zhang J, Fan C, Tang Y, Deng L. Katzlgo: large-scale prediction of lncrna functions by using the katz measure based on multiple networks. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16(2):407–16.
15. Jimenezsanchez G, Childs B, Valle D. Human disease genes. Nature. 2001;409(6822):853–5.
16. López-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 2004;32(10):3108.
17. Pereziratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nature Genet. 2002;31(3):316–9.
18. Mathur S, Dinakarpandian D. Automated ontological gene annotation for computing disease similarity. Transl. Bioinforma. 2010;2010:12.
19. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Informa. 2012;45(2):363–71.
20. Li J. Dosim: An r package for similarity between diseases based on disease ontology. Bmc Bioinformatics. 2011;12(1):266.
21. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. 1995;1995:448Ű453.
22. Lin D. An information-theoretic definition of similarity. In: International Conference on Machine Learning(Citeseer); 1998. p. 296–304.
23. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. Proc. Int. Conf. Res. Comput. Linguist. 199719–33.
24. Deng Y, Gao L, Wang B, Guo X. Hposim: An r package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. Plos One. 2015;10(2):0115692.
25. Lipscomb CE. Medical subject headings (mesh). Bull Med Libr Assoc. 2000;88(3):265–6.
26. Tong H, Faloutsos C, Pan JY. Fast random walk with restart and its applications. In: International Conference on Data Mining(IEEE); 2006. p. 613–22.
27. Zhou XZ, Menche J, Barabási A, Sharma A. Human symptoms–disease network. Nature Commun. 2014;5:4212.
28. Cho H, Berger B, Peng J. Diffusion component analysis: Unraveling functional topology in biological networks. Comput Sci. 2016;9029(4):62–4.

29. Zhang J, Zhang Z, Wang Z, Liu Y, Deng L. Ontological function annotation of long non-coding rnas through hierarchical multi-label classification. Bioinformatics. 2018;34(10):1750–7.
30. Deng L, Wu H, Liu C, Zhan W, Zhang J. Probing the functions of long non-coding rnas by exploiting the topology of global association and interaction network. Comput Biol Chem. 2018;74:360–7.
31. Wang S, Cho H, Zhai C, Berger B, Peng J. Exploiting ontology graph for predicting sparsely annotated gene function. Bioinformatics. 2015;31(12): 357–64.
32. 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/ AMIA Symp. AMIA Symposium. 2010;2010:572.
33. Cho H, Berger B, Peng J. Compact integration of multi-network topology for functional analysis of genes. Cell Syst. 2016;3(6):540.
34. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Human Genet. 2006;14(5):535–42.
35. Li P, Nie Y, Yu J. Fusing literature and full network data improves disease similarity computation. Bmc Bioinformatics. 2016;17(1):326.
36. Lan W, Wang J, Li M, Peng W, Wu F. Computational approaches for prioritizing candidate disease genes based on ppi networks. Tsinghua Sci Technol. 2015;20(5):500–512.
37. Zhang J, Zhang Z, Chen Z, Deng L. Integrating multiple heterogeneous networks for novel lncrna-disease association inference. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16(2):396–406.
38. Deng L, Zhang W, Shi Y, Tang Y. Fusion of multiple heterogeneous networks for predicting circrna-disease associations. Sci Rep (Nat Publ Group). 2019;9:1–10.
39. Guo X, Zhang J, Cai Z, Du DZ, Pan Y. Searching genome-wide multi-locus associations for multiple diseases based on bayesian inference. IEEE/ACM Trans Comput Biol Bioinforma. 2017;PP(99):1–1.
40. Teng B, Yang C, Liu J, Cai Z, Wan X. Exploring the genetic patterns of complex diseases via the integrative genome-wide approach. IEEE/ACM Trans Comput Biol Bioinforma. 2016;13(3):557–64.
41. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microrna function and prioritizing disease-related microrna using biological interaction networks. Brief Bioinforma. 2016;17(2):193.
42. Zou Q, Li J, Hong Q, Lin Z, Wu Y, Shi H, Ying J. Prediction of microrna-disease associations based on social network analysis methods. Biomed Res Int. 2015;2015(10):810514.
43. Yan C, Wang J, Ni P, Lan W, Wu F, Pan Y. Dnrlmf-mda:predicting microrna-disease associations based on similarities of micrornas and diseases. IEEE/ACM Trans Comput Biol Bioinforma. 2017;PP(99):1–1.
44. Liang C, Li J, Peng J, Peng J, Wang Y. Semfunsim: A new method for measuring disease similarity by integrating semantic and gene functional association. Plos One. 2014;9(6):99415.
45. Ghiassian SD, Menche J, Barabási AL. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. Plos Comput Biol. 2015;11(4):1004120.

Publisher's Note