Briefings in Bioinformatics, 22(2), 2021, 2085–2095

doi: 10.1093/bib/bbaa037
Advance Access Publication Date: 31 March 2020
Problem Solving Protocol

MeSHHeading2vec: a new method for representing MeSH headings as vectors based on graph embedding algorithm

Zhen-Hao Guo†, Zhu-Hong You†, De-Shuang Huang, Hai-Cheng Yi, Kai Zheng, Zhan-Heng Chen and Yan-Bin Wang

Corresponding author: Zhu-Hong You, The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; University of Chinese Academy of Sciences, Beijing 100049, China. Tel: +86-991-367-2967; E-mail: zhuhongyou@ms.xjb.ac.cn
†These authors contributed equally to this work.

Abstract

Effectively representing Medical Subject Headings (MeSH) headings (terms) such as disease and drug as discriminative vectors could greatly improve the performance of downstream computational prediction models. However, these terms are often abstract and difficult to quantify. In this paper, we converted the MeSH tree structure into a relationship network and applied several graph embedding algorithms on it to represent these terms. Specifically, the relationship network consisting of nodes (MeSH headings) and edges (relationships), which can be constructed by the tree num. Then, five graph embedding algorithms including DeepWalk, LINE, SDNE, LAP and HOPE were implemented on the relationship network to represent MeSH headings as vectors. In order to evaluate the performance of the proposed methods, we carried out the node classification and relationship prediction tasks. The results show that the MeSH headings characterized by graph embedding algorithms can not only be treated as an independent carrier for representation, but also can be utilized as additional information to enhance the representation ability of vectors. Thus, it can serve as an input and continue to play a significant role in any computational models related to disease, drug, microbe, etc. Besides, our method holds great hope to inspire relevant researchers to study the representation of terms in this network perspective.

Key words: MeSHHeading2vec; MeSH relationship network ; graph embedding; computational prediction model

Zhen-Hao Guo is a master student at the University of Chinese Academy of Sciences. His research interests include text data mining, network analysis and their applications in bioinformatics.
Dr. Zhu-Hong You is a professor at the University of Chinese Academy of Sciences and at the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences. His research interests include big data analysis, data mining, pattern recognition and their applications in bioinformatics.
Dr. De-Shuang Huang is a chaired professor in Tongji University. At present, he is the Director of Institute of Machines Learning and Systems Biology, Tongji University. Dr. Huang is currently IAPR Fellow and a senior member of the IEEE. His current research interest includes Bioinformatics, pattern recognition and machine learning.
Hai-Cheng Yi is a master student at the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences. He works on machine learning, network analysis and their applications in bioinformatics.
Kai Zheng is a master from China University of Mining and Technology. His research interests include pattern recognition, machine learning, intelligent information processing and their applications in bioinformatics.
Zhan-Heng Chen is currently a PhD candidate from the University of Chinese Academy of Sciences. His research interests include machine learning and its application in proteomics.
Yan-Bin Wang is currently a PhD candidate from Zhejiang University. He works on machine learning and their applications in bioinformatics and network security.
Submitted: 5 December 2019; **Received (in revised form):** 13 February 2020

Introduction

Technological advances over the past few decades, from high-throughput sequencing technologies to omics, have dramatically changed the paradigm of medicine and biology [1, 2]. In particular, since the official launch of the Human Genome Project in the 1990s, the large-scale genomic, chemical and pathological data has brought novel insights for humans to re-recognize life processes [3]. However, the information overload caused by tremendous growth of data makes it difficult to take full use of existing knowledge and literature. For instance, a premier database called MEDLINE contains about 26 million records from more than 5600 selected publications covering biomedical and life sciences to the present. So how to efficiently organize and manage the literature and explore the implicit value becomes a formidable challenge.

In response to this situation, the literature-based discovery (LBD) method was firstly proposed by Don R. Swanson which logically combines independent pieces of information to infer new interesting discoveries [4]. Many models were continuously developed to provide efficient and stable support for researchers such as co-occurrence-based approaches [5], semantic relation-based approaches [6], graph-based approaches [7] and hybrid approaches [8].

For the traditional LBD method, such as MeSH, Unified Medical Language System (UMLS), emMedDB and etc. are often treated as auxiliary knowledge sources to improve the performance of the model. Although significant progress has been made in this domain, most of them ignored the potential value behind the MeSH headings that itself is carefully designed. In addition, terms such as disease, drug and microbe are abstract entities that are difficult to be represented as concrete vectors as input for computational methods. In fact, models that predict potential relationships based on known experimental data are ubiquitous [9–11] and can be experimentally validated [12]. In this paper, we focus on analyzing MeSH to mine the hidden information. It is believed that this expert knowledge can be utilized to precisely quantify these terms.

MeSH is a kind of controlled and comprehensive vocabulary for subject indexing and searching books or journals in life sciences [13]. It was produced by National Library of Medicine (NLM) since 1960 and widely used around the world. More than half a century of heavy application has made MeSH increasingly perfect and made significant contributions to various fields. The MeSH consists of three parts including Main Headings, Qualifiers and Supplementary Concepts. Main Headings as the trunk of MeSH are used to describe the content or theme of the article. Qualifiers is the refinement of MeSH headings, i.e. how to be processed when it is in a specific area. Supplementary Concept is a complementary addition that is mostly related to drugs and chemistry. Some new substances have not yet become the main subject and will be included in Supplementary Concept to promote the integrity of MeSH. Here, we focus on discussing Main Headings which consists of MeSH headings (descriptors), corresponding entry term and tree num.

MeSH headings can be divided into 16 categories such as category A for anatomy, category B for organisms, category C for diseases, category D for Chemicals and Drugs, etc. Entry term is a kind of synonym or similar vocabulary for MeSH headings. Tree num is the tag of MeSH heading in tree structure. In MeSH tree structure, MeSH headings are organized as a 'tree' with 16 top categories in which the higher hierarchy has the broader meaning and the lower hierarchy has the specific meaning, considering that MeSH headings usually have many tree nums or can be defined from different perspectives. Compared with the tree structure, network (graph) is a more flexible data type which widely spreads in the real world and has been deeply researched [14]. In fact, many biological and medical research signal pathways exist in the form of unstructured networks [15,16]. Effective analysis of the Network not only can deeply understand the original data, but also facilitate downstream tasks such as node classification and relationship prediction. Hence, we construct the MeSH heading relationship network from tree structure through hierarchical tree num rules.

Graph embedding (network representation) is a kind of method to process the network problem which aims at transforming the node into low-dimensional vectors. In this process, it maximumly preserves both the local and global structure of the network. The mainstream graph embedding algorithms can be roughly divided into three categories: factorization-based methods, random walk-based methods and deep learning-based methods [17]. The random walk-based graph embedding method is to use the random walk on the network to obtain a series of node paths to mimic the sentences or text. Then, the Word2vec model can be applied to transform the node into vectors. The method of factorization takes the adjacency matrix as the structure of the graph and obtains the node representation vectors by the method of matrix decomposition. Explosive research on deep learning has rapidly expanded its field to the network. The deep learning-based method is to carry out the feature capture and dimensional reduction tasks on node original representation to get the new low-dimensional vectors.

In this paper, the mainstream idea of using MeSH as a dictionary for indexing is abandoned; we transform the MeSH tree structure into a relationship network and implement five common graph embedding algorithms on it to represent the MeSH headings as vectors. In general, the whole process can be divided into three steps. Firstly, MeSH headings, tree num and entry terms were downloaded from National Library of Medicine (NLM) in 22 September 2019. Then we connected different Mesh headings through the rules of tree num to convert the tree structure to the relationship network. The label (category) of each node (Mesh heading) in the relationship network can be defined by the mode of its corresponding tree num. Secondly, the network has been briefly analyzed, including the number of nodes and edges, the distribution of node degrees and labels. Thirdly, we applied five network representation (graph embedding) algorithms including DeepWalk [18], LINE [19], SDNE [20], LAP [21] and HOPE [22] to map the nodes into low-dimensional dense vectors which maximumly preserves the original network structure and the node relationship information. Then, we performed two types of tasks including node classification and relationship prediction. The node classification and relationship prediction tasks are used to assess the distinguishability of vectors between and within categories. In relationship prediction task, we performed drug–target interaction and miRNA-disease association prediction tasks to display that the term representation vectors can be as input for machine learning model. All competitive results achieved by our method implied that the representation vectors generated by MeSH relationship network are efficient and reliable. High quality MeSH heading representation will definitely improve the prediction performance of existing computational models. At the same time, we hope that this work can provide novel insight to inspire relevant medical and life science researchers to mine the semantic information in MeSH through the network method. The flowchart is shown in Figure 1.

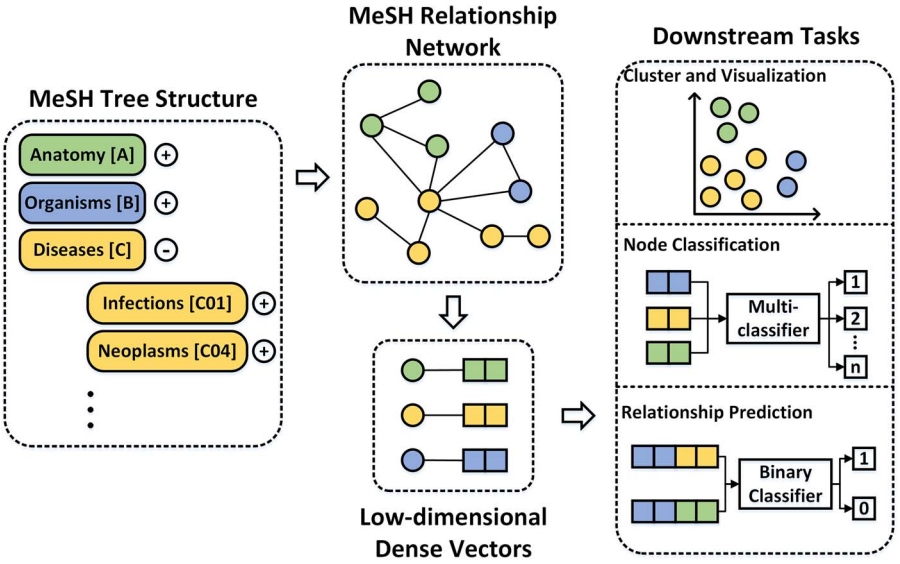

Figure 1. The flowchart of the proposed method includes three steps: construction, analysis and applications.

Materials & methods

MeSH headings, tree numbers and entry terms

The Medical Subject Headings (MeSH) is a controlled and hierarchically organized vocabulary directed by the National Library of Medicine (NLM), which is utilized for indexing, searching, and etc. in medical and life sciences. We downloaded MeSH headings, tree num and entry terms from NLM in 22 September 2019 and arranged them by routine standardized pretreatments including identifier unification and redundancy removal. After above operations, 29 349 MeSH headings including their corresponding tree num and entry terms are congregated together for network construction.

Each MeSH heading can be described by one or more tree nums to reflect its hierarchy in the tree structure and relationships with other MeSH headings. Tree num consists of letters and numbers, the first of which is uppercase letter representing category and the rest are made up of numbers. The first two digits are fixed design following the first capital letter and can be seen the top category except capital letter. Each three digits represent a hierarchy in the tree structure. There are some MeSH headings such as Lung Neoplasms (C04.588.894.797.520, C08.381.540, and C08.785.520) that are described by a single type of tree num, while others such as Reflex (E01.370.376.550.650, E01.370.600.550.650, F02.830.702 and G11.561.731) can be represented by different kinds of tree num.

Whenever the last hierarchy of tree num is removed, a new tree num and corresponding MeSH heading can be generated and contacted. The details can be seen in Figure 2. Through the formation of this kind of relationship, a MeSH heading network consisting of 29 349 nodes and 39 784 edges can be constructed. For the sake of simplicity, we treat the mode of the tree num category of MeSH heading as its label.

In order to unify identifiers and eliminate ambiguity, we create a MeSH Heading Term Correspondence Table to convert the entry terms to standard MeSH headings. All available data are uploaded in github: https://github.com/CocoGzh/MeSHHeading2vec.

Known drug-target interactions

A total of 28 211 known drug-target interactions were downloaded from DrugBank in 8 May 2019 [23]. After standardizing the identifiers via the Correspondence Table and STRING database, we got 7739 different drugs and 4975 different proteins. In order to avoid sparsity of associations, we selected drugs and proteins that are associated with more than five corresponding objects similar to the article described by Zhang *et al.* [24]. Finally, we obtained 7318 experimental valid drug-target interactions containing 641 different drugs and 317 different proteins.

The experimental-validated drug-target interaction pairs are regarded as the positive samples and the randomly selected equal unlabeled pairs are treated as negative samples. This is a typical strategy that equalizes training samples and is widely used in bioinformatics [25]. Each positive and negative sample is given a label 1 and 0, respectively.

Known miRNA-disease associations

A total of 35 547 known human miRNA-disease associations, which consist of 1206 different miRNAs and 894 different diseases, were downloaded from HMDD in 8 May 2019 [26]. Considering that the name of disease and miRNA in the original database are nonstandard such as 'breast neoplasms' and 'carcinoma, breast' are the same type of disease. After standardizing the identifiers via miRBase and the Correspondence Table described above, we obtained 11 109 experimental valid miRNA-disease associations containing 843 different miRNAs and 531 different diseases.

Negative sample selection is the same as the strategy mentioned in the section "Known drug-target interactions".

k-mer method

For a long time, how to transform sequences efficiently and reliably into numerical representations is a formidable challenge. In this article, a widely used bassline method called k-mer is applied, and the details of the algorithm are shown as follows.

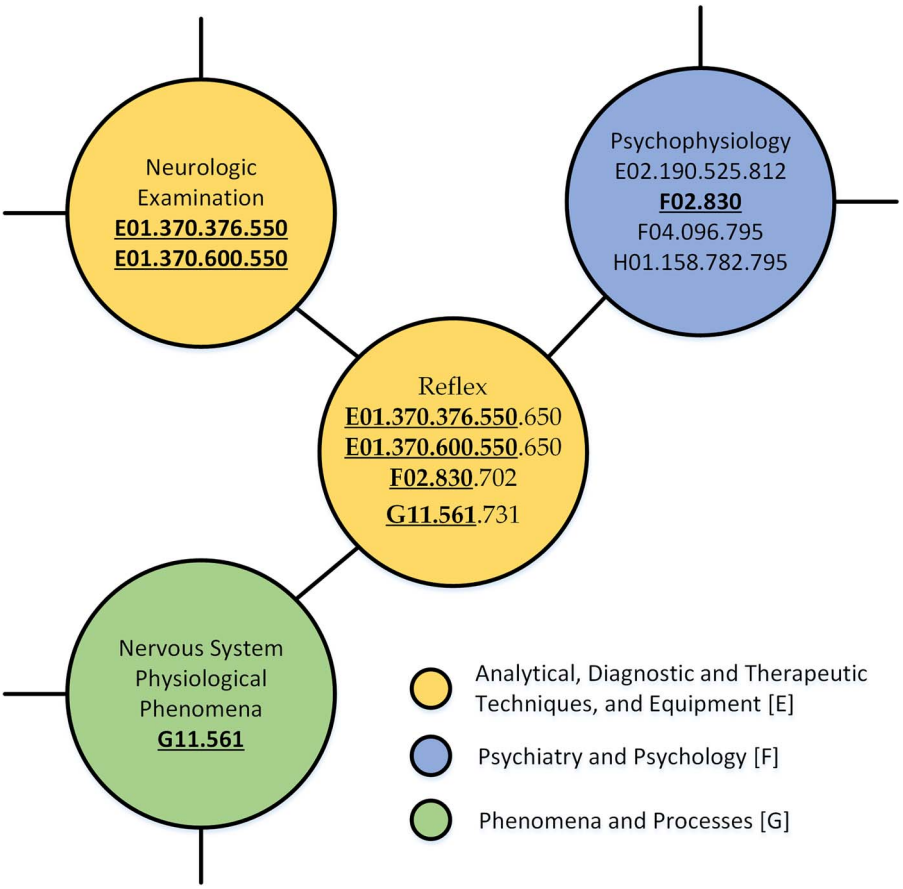

Figure 2. The construction of the MeSH relationship network. Reflex has four tree num including E01.370.376.550.650, E01.370.600.550.650, F02.830.702 and G11.561.731. The Neurologic Examination (E01.370.376.550, E01.370.600.550) can be obtained when the last three digits (.650 and .650) of Ref lex (E01.370.376 .550.650, E01.370.600.550.650) are removed. . The category (label) of each MeSH heading is the mode of its corresponding tree num.

For protein and miRNA, the sequences of them were downloaded from STRING [27] and miRBase [28], respectively. Inspired by Shen *et al.* [29], we represent proteins and miRNAs as vectors by analyzing and normalizing their components. For proteins, we classified 20 amino acids into four groups according to the polarity of the side chain, including (Ala, Val, Leu, Ile, Met, Phe, Trp and Pro), (Gly, Ser, Thr, Cys, Asn, Gln and Tyr), (Arg, Lys and His) and (Asp and Glu). For miRNA, there naturally exist four types of nucleotides including adenine (A), cytosine (C), guanine (G) and uracil (U) in the sequence. Then, each miRNA or protein can be abstracted into a vector by the method k-mer, in which all dimensions represent the full permutation of k nucleotide combinations and the value of each dimension is the normalized frequency of the corresponding k-mer appearing in the sequence. Here, we set k to 3, and the dimension of the representation vector is 64 (4^3).

Drug molecular fingerprint method

Molecular Fingerprint is one of the most popular methods to represent drugs by describing the structure of compounds. The basic idea is to segment the drug molecule and obtain structure fragments one after another. Then, these substructures are encoded into numbers according to certain rules, which can correspond to each of the binary strings. The whole binary string is used as the characterization of drug molecular structure. In this paper, the fingerprint method is chosen as the baseline to represent the drug.

The drug SMILES was downloaded from DrugBank and transformed into fingerprints by python package called RDKit [30].

Disease similarity-based method

Disease is an abnormal life activity process that occurs when a living organism is destructively affected by a certain cause. The semantic similarity of disease is a common method of abstracting disease into vectors [31]. For each disease, a Directed Acyclic Graph (DAG) can be constructed by the MeSH heading relationship in Section "MeSH headings, tree numbers and entry terms". Specifically, disease D's ancestor nodes can be obtained by continuously removing the last hierarchy of its tree num. D and its ancestor nodes together constitute a DAG. Then the similarity between two diseases can be calculated according to the generalized Jaccard formula, i.e. the larger the intersection, the more similar it is. According to the previous literature [32], the specific calculation process is as follows:

For disease D, $DAG(D) = (D, N(D), E(D))$, $N(D)$ is the point set that includes all $DAG(D)$'s diseases. $E(D)$ is the edge set that includes all $DAG(D)$'s relationships. The semantic value contribution of disease d in the set $N(D)$ to disease D can be defined as:

$$\begin{cases} D_D(d) = 1 & \text{if } d = D \\ D_D(d) = \max\left\{\Delta * D_D\left(d'\right) | d' \in \text{children of } d\right\} & \text{if } d \neq D \end{cases} \quad (1)$$

where Δ denotes a decline factor. In the $DAG(D)$, D can be seen as the disease that contributes the most to its own semantic value

and equals to 1, and the remaining diseases will contribute less and less to disease D as the distance increases. Then, the sum of the contributions of diseases which are in the set $N(D)$ to D can be calculated as follows:

$$DV(D) = \Sigma_{d \in N(D)} D_D(d) \tag{2}$$

Finally, the similarity between diseases m and n can be calculated by the following formula:

$$Similarity\,(m,n) = \frac{\sum_{d \in N(m) \cap N(n)} \left(D_m(d) + D_n(d)\right)}{DV(m) + DV(n)} \tag{3}$$

The disease similarity matrix of k rows and k columns containing k different diseases can be constructed, and the i-th row can be regarded as a representation vector of the i-th disease.

Autoencoder

In order to unify the dimensions of the vector and obtain a higher quality representation, autoencoder is applied to map the drug fingerprint and disease similarity from original space to the low-dimensional space. Hidden layer representation h and output layer representation y can be calculated by the following formula:

$$h = f\left(\mathrm{W}x + b\right) \tag{4}$$

$$y = g\left(W'h + b'\right) \tag{5}$$

where x is input, W and b are weights and thresholds, respectively, and f and g are the activation functions. Loss function can be obtained by minimizing the error between input and output:

$$L = \Sigma \left\| y - x \right\|^2 \tag{6}$$

Finally, all drug fingerprint and disease similarity can be normalized to 64-dimensional vectors.

Graph embedding methods

Mesh heading relationship network is a complex heterogeneous attribute network. Analysis of network can better help us understand this kind of unstructured data and benefit the exploration of the underlying knowledge. Graph embedding is an effective method to provide new insights on how to make good use the hidden information behind the graph. In this chapter, we first give a graph embedding formal definition, and then briefly introduce several algorithms used in this paper.

A graph $G(V, E)$ is a collection of vertices (node) set $V = \{v_1, \ldots, v_n\}$ and edge set $E = \{e_{i,j}\}_{i,j=1}^{n}$. The aim of graph embedding is to find a mapping function $f : v_i \to x_i \in R^d$, where $d \ll |V|$, and $X_i = \{x_1, x_2, \cdots, x_d\}$ is the embedded vector that captures the structural of vertex v_i.

In this paper, we apply five kinds of graph embedded methods on the network to perform downstream tasks including node classification and relationship prediction.

Deepwalk obtains a series of node sequences through random walks of vertexes in the network and inspired by the Skip-Gram model to analogize these paths to sentences for representation learning. The goal is to learn a latent representation and the mapping function is:

$$\Phi : v \in V \mapsto R^{|V| \times d} \tag{7}$$

The problem then, is to estimate the likelihood:

$$P_r\left(v_i \middle| \left(\Phi\left(v_1\right), \Phi\left(v_2\right), \ldots, \Phi\left(v_{i-1}\right)\right)\right) \tag{8}$$

The recent relaxation in language modeling turns the prediction problem and this yields the optimization problem:

$$\underset{\Phi}{\text{minimize}} = -\log P_r\left(\left\{v_{i-w}, \ldots, v_{i+w}\right\} \setminus v_i \middle| \Phi\left(v_i\right)\right) \tag{9}$$

Large-scale Information Network Embedding (LINE) is an efficient network representation learning algorithm that is quite different from random walk-based method. Low-dimensional dense vectors can be obtained by LINE by preserving first-order and second-order proximity. For first-order, the objective function can be defined as follows:

$$O_1 = d\left(\hat{p}_1\left(\cdot, \cdot\right), p_1\left(\cdot, \cdot\right)\right) \tag{10}$$

For the edge $e_{i,j}$ which from vertex v_i to vertex v_j, $\hat{p}_1(\cdot, \cdot)$and $p_1(\cdot, \cdot)$ are the empirical and joint distribution, respectively, between the latent embeddings r_{v_i} and r_{v_j}. $d(\cdot, \cdot)$ is the distance between the above two distributions.

For second-order, the objective function can be defined as follows:

$$O_2 = \sum_{v_i \in V} \lambda_i d\left(\hat{p}_2\left(\cdot \middle| v_i\right), p_1\left(\cdot \middle| v_i\right)\right) \tag{11}$$

where $\hat{p}_2(\cdot | v_i)$ and $p_1(\cdot | v_i)$ are empirical and context conditional distribution for each $v_i \in V$ under the model by vertex embeddings. For the sake of simplicity, λ_i is set to the degree of the vertex i.

Structural Deep Network Embedding (SDNE) is a semi-supervised deep autoencoder consisting of supervised and unsupervised component that can capture the nonlinear structure from the network. For the supervised part, the objective function can be defined as follows:

$$L_1 = \sum_{i,j=1}^{|V|} S_{ij} \left\| r_{v_i}^{(K)} - r_{v_j}^{(K)} \right\|_2^2 \tag{12}$$

where $r_{v_i}^{(K)}$ is the K-th layer representation of v_i.

For the unsupervised part, the objective function can be defined as follows:

$$L_2 = \sum_{i=1}^{|V|} S_{ij} \left\| r_{v_i}^{(0)} - r_{v_i}^{(0)} \odot b_i \right\|_2^2 \tag{13}$$

where $r_{v_i}^{(0)}$ is the representation of v_i and b_i is a weight vector.

Finally, the joint objective function can be defined as follows:

$$L = L_1 + L_2 + L_{reg} \tag{14}$$

where L_{reg} is a regularization term to prevent overfitting.

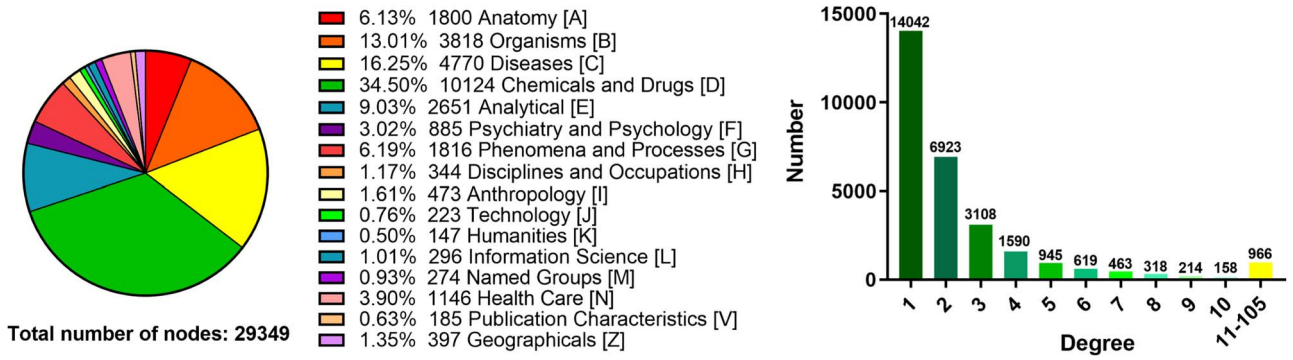

Figure 3. Distribution of node type and node degree in the relationship network.

High-order Proximity Preserved Embedding (HOPE) captures high order proximity of asymmetric transitivity in direct graph and symmetric transitivity in undirect graph. To achieve this goal, HOPE can obtain two vertex representation vectors $U^s, U^t \in R^{|V|\times d}$, where U^s and U^t are called source and target vectors. The objective function can be defined as follows:

$$\min_{U^s,U^t} \left\| S - U^s \cdot U^t \right\|_F^2 \tag{15}$$

The structure of the reserved graph can be considered as the similarity of the reserved nodes. Laplacian Eigenmaps is an embedding algorithm that obtains the representation vector when the similarity parameter W_{ij} is high. The objective function can be defined as follows:

$$\phi(Y) = \frac{1}{2} \sum_{i,j} \left(Y_i - Y_j \right) W_{ij} = Y^T L Y \tag{16}$$

Results

Evaluation criteria

The MeSH relationship network consisting of nodes and the edges contains a wealth of medical and biological knowledge. After mining the potential content by embedding algorithms, low-dimensional dense representation vectors can be used for downstream tasks such as visualization, node classification and relationship prediction. How to evaluate the merits and demerits of the proposed method in a fair and comprehensive way becomes a formidable challenge.

Firstly, we briefly analyzed the MeSH relationship network. Secondly, we not only perform the node classification in the whole network, but also extract drug and disease representation vectors to carry out the relationship prediction tasks. Both of them aim at evaluating the distinguishability of vectors. High-quality representation vectors make it easier to construct the classifier to make prediction results more accurate. The results can be seen in the following section.

Meanwhile, we applied a wide range of evaluation criteria to effectively assess the performance of our method [33]. Cross validation is a widely used method to measure model ability [34, 35]. For 5-fold cross-validation, the whole dataset is divided into 5 mutually exclusive subsets of roughly size, each subset is treated as the test set for evaluation in turn and the others are treated as the training sets for the model construction. At the same time, we draw ROC (receiver operating characteristic curve) and PR (precision-recall) to calculate AUC (area under ROC) and AUPR (area under PR), respectively, in order to visualize experimental results and facilitate comparison with other methods. In addition, a wide range of evaluation criteria including accuracy (Acc.), sensitivity (Sen.), specificity (Spec.), precision (Prec.) and MCC have been adopted to evaluate our approach more generally.

Network analysis

The MeSH heading relationship network is a heterogeneous network consisting of 29 349 nodes and 39 784 edges, where the nodes are included by 16 different kinds of descriptors. Node degree refers to the number of edges associated with the node, also known as correlation degree. The occurrence number of the node and degree can be statistics and visualized as the Figure 3.

Application 1: MeSH headings classification

As mentioned above, each node (MeSH heading) can be represented as a low-dimensional dense vector by graph embedding algorithm and can be labeled by the mode of its tree num. We want to verify the pros and cons of different graph embedding algorithms through the node classification experiment.

Specifically, five graph embedding algorithms including DeepWalk, LINE, SDAE, LAP and HOPE are applied on the relationship network to represent the nodes as 64-dimenson vectors and the labels can be labeled by the mode of the node's tree num. For example, reflex characterized by E01.370.376.550.650, E01.370.600.550.650, F02.830.702 and G11.561.731 will be given a label E. Then, 80% of the nodes and the corresponding labels are utilized to construct the multi-classifier, and the remaining 20% of the nodes and the corresponding labels are used for testing. Although there exist some noises and errors in labels, the accuracy of the classifier can reflect the quality of the representation vectors to some extent.

The keras library was applied to construct this multi-classifier. An artificial neural network with two layers was built where each layer consists of 512 neurons. The parameters including loss, optimizer, batch_size and epochs are set to categorical_crossentropy, RMSprop, 1024 and 100, respectively. The results including ACC and LOSS are shown in Figure 4 and Table 1.

The node classification task reflects the distinguishability between different type of term representation, such as anatomy and organisms. Compared with other method, DeepWalk obviously achieved the most competitive performance, which demonstrates that DeepWalk can indeed capture global structure and differences between various labels in the whole network.

Table 1. The test performance of different graph embedding methods on the node classification task

Test Performance	SDNE	HOPE	LINE	LAP	DeepWalk
Acc.	0.5056	0.7003	0.9068	0.9284	0.9824
Loss	1.7108	0.9164	0.4130	0.2105	0.0722

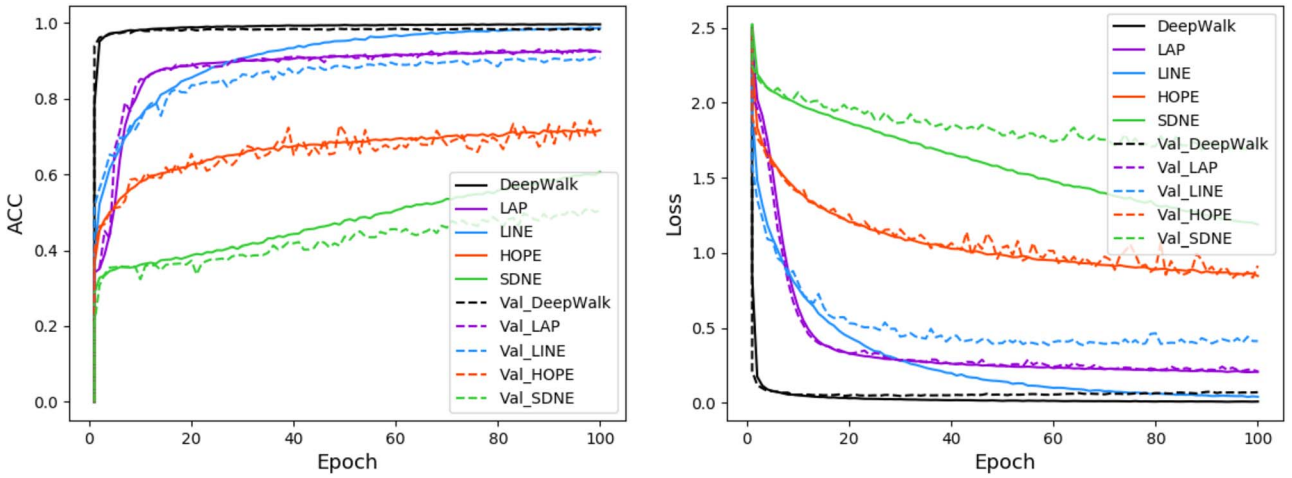

Figure 4. The performance of node classification task achieved by different graph embedding methods.

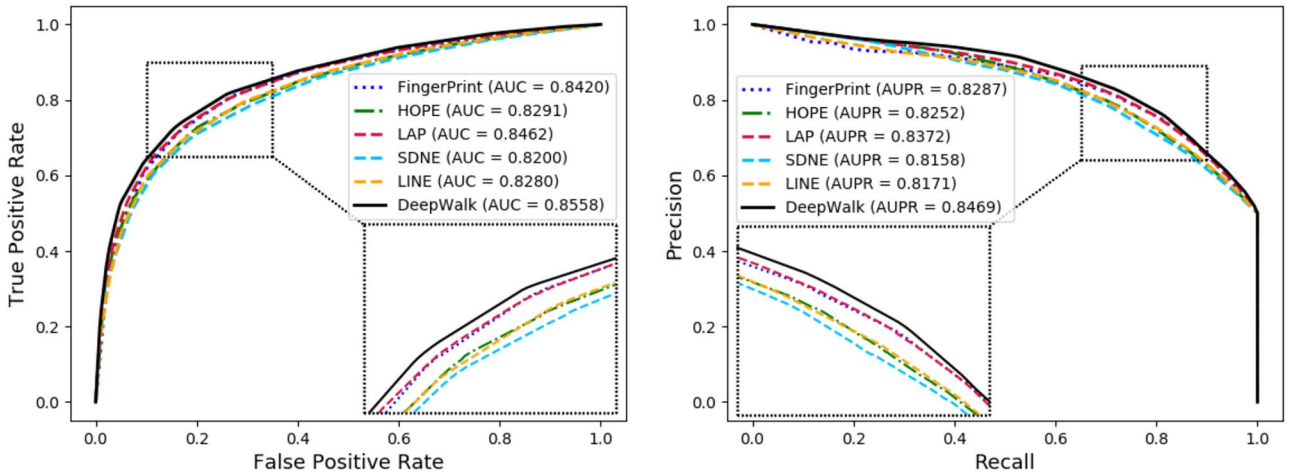

Figure 5. ROCs, AUCs, PRs and AUPRs of drug-target interaction prediction achieved by different graph embedding and drug Morgan molecular fingerprint methods.

Application 2: drug representation for drug-target interaction prediction

In this section, we choose drug-target interaction prediction as a specific research subject to evaluate the quality of the drug representation vector. Specifically, each drug and protein can be represented as a 64-dimenonal vector by graph embedding and k-mer method. We also treated drug Morgan molecular fingerprint method as a baseline for comparison. Then, each drug-target interaction pair is a 128-dimensional vector by concatenating drug and target. 5-fold cross validation was applied to evaluate the performance of each method. Random forest is chosen as the classifier to carry out the interaction prediction task. To evaluate the proposed method, we draw ROC and PR to calculate the AUC and AUPR, respectively. In addition, extensive evaluation criteria including Acc., Sen., Spec., Prec. and MCC are adopted. The results can be seen in Figure 5 and Table 2.

Compared with the node classification, the association prediction task reflects the distinguishability between the same type of term representation, such as drug.

In general, DeepWalk and LAP achieved pretty prediction effects. Considering the traditional method of analyzing the chemical structure of drugs, the satisfactory results prove that the proposed representation is novel and can adequate characterize the drug by semantic. We believe it will open up a new paradigm for semantic representation of drugs.

Application 3: disease representation for miRNA-disease association prediction

In this section, we choose miRNA-disease association prediction as a specific research subject to evaluate the quality of the disease representation vector. Specifically, each miRNA and disease can be represented as a 64-dimenonal vector by

Table 2. The performance of different graph embedding methods under 5-fold cross validation on the drug-target interaction prediction

Method	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC (%)
FingerPrint	77.49 ± 0.32	72.45 ± 0.85	82.53 ± 0.59	80.58 ± 0.43	55.27 ± 0.60	84.20 ± 0.37
HOPE	76.33 ± 0.84	72.63 ± 1.29	80.04 ± 1.02	78.44 ± 0.94	52.82 ± 1.69	82.91 ± 0.85
LAP	77.78 ± 1.00	73.07 ± 1.18	82.48 ± 1.10	80.66 ± 1.13	55.80 ± 2.00	84.62 ± 1.19
LINE	76.08 ± 0.46	69.88 ± 1.05	82.29 ± 0.80	79.79 ± 0.63	52.59 ± 0.90	82.80 ± 0.39
SDNE	75.63 ± 0.41	70.37 ± 0.57	80.88 ± 0.84	78.64 ± 0.70	51.54 ± 0.86	82.00 ± 0.27
DeepWalk	78.61 ± 0.55	73.24 ± 0.87	83.97 ± 1.02	82.06 ± 0.89	57.55 ± 1.13	85.58 ± 0.63

Table 3. The performance of different graph embedding methods under 5-fold cross validation on the miRNA-disease association prediction

Method	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC (%)
DAG	82.42 ± 0.28	79.01 ± 0.75	85.84 ± 0.92	84.81 ± 0.74	65.01 ± 0.59	89.95 ± 0.25
HOPE	78.12 ± 0.82	76.13 ± 0.96	80.11 ± 1.34	79.29 ± 1.13	56.28 ± 1.66	85.36 ± 0.80
LAP	81.53 ± 0.46	78.07 ± 0.49	84.99 ± 0.61	83.88 ± 0.59	63.22 ± 0.93	89.06 ± 0.52
SDNE	81.89 ± 0.46	78.69 ± 0.99	85.07 ± 0.96	84.07 ± 0.78	63.91 ± 0.93	89.53 ± 0.33
DeepWalk	82.68 + 0.62	79.17 + 1.61	86.18 + 1.42	85.16 + 1.16	65.53 + 1.22	90.15 + 0.52
LINE	83.02 ± 0.53	79.95 ± 0.24	86.09 ± 0.97	85.19 ± 0.88	66.17 ± 1.09	90.28 ± 0.27

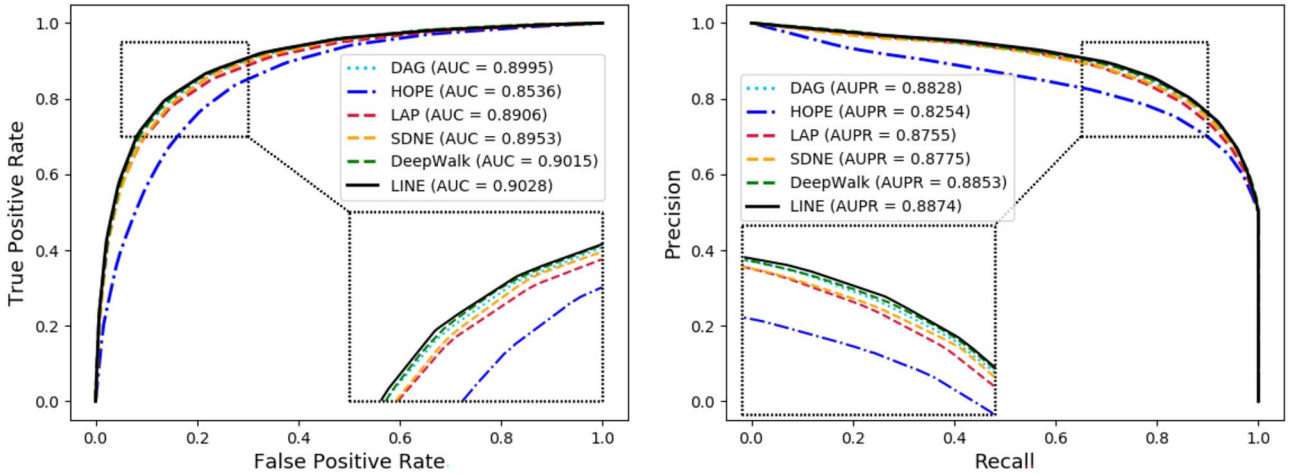

Figure 6. ROCs, AUCs, PRs and AUPRs of miRNA-disease association prediction achieved by different graph embedding and DAG methods.

k-mer and graph embedding method. We also performed disease semantics similarity method as a baseline for comparison. Then, each miRNA-disease association pair is a 128-dimensional vector by concatenating miRNA and disease. 5-fold cross validation was chosen to evaluate the performance. Random forest is applied as the classifier to carry out the association prediction task. To visualize the proposed method, we draw ROC and PR to calculate the AUC and AUPR, respectively. The results of all methods can be seen in Figure 6 and Table 3.

Briefly, the LINE method achieved the most remarkable with average Acc., Sen., Spec., Prec. and MCC of 83.02, 79.95, 86.09, 85.19, 66.17 and 90.28. The corresponding standard deviations of above evaluation criteria are 0.53, 0.24, 0.97, 0.88 and 1.09. The brilliant performance of the proposed method indicates that the representation vector generated by MeSH relationship network can be used as an independent carrier to characterize disease. Meanwhile, the lower standard deviation implied that the novel model was robust and stable.

Although the performance improvement relative to the disease semantics similarity is weak, the disease graph embedding representation has three obvious advantages. Firstly, compared with the similarity-based method, the graph-based method has faster calculation speed and less resource occupation. Secondly, the similarity-based method only calculates the similarity between the diseases in the current sample set. For example, the number of diseases in the miRNA-disease association benchmark data set proposed in this paper is 531. Based on graph representation algorithm, 29 349 MeSH heading vectors in network can be obtained at one time. Similarity-based method needs to be recalculated when facing a new sample, but the graph-based method can be generated once for permanent use.

Application 4: as addition information to enhance the ability of disease representation

In this section, we choose miRNA-disease association prediction as a specific research subject to prove that our representation method of disease can be utilized as the additional information. Specifically, inspired by Guo *et al.* [36], each miRNA and disease can be represented by two kinds of information including the behavior and the attribute feature. The behavior feature is the main information that is proposed by the idea of collaborative

Table 4. The performance of different features under 5-fold cross validation on the miRNA-disease association prediction

Method	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC (%)
Attribute	83.19 ± 0.48	79.75 ± 1.05	86.63 ± 0.36	85.65 ± 0.30	66.55 ± 0.91	90.33 ± 0.33
Behavior	83.56 ± 0.82	77.23 ± 1.30	89.89 ± 0.92	88.43 ± 0.97	67.67 ± 1.62	90.31 ± 0.39
Both	83.98 ± 0.59	78.57 ± 1.63	89.39 ± 0.73	88.11 ± 0.59	68.38 ± 1.05	91.40 ± 0.42

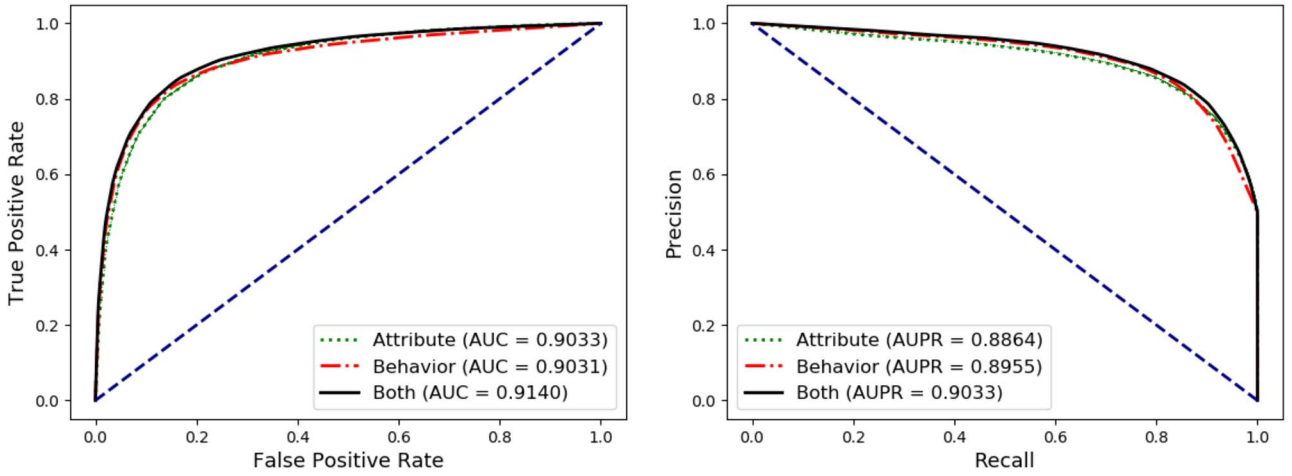

Figure 7. ROCs, AUCs, PRs and AUPRs achieved by different features.

filtering or similarity theory. It is known that miRNAs with similar functions are often associated with similar diseases and vice versa. Then, each miRNA and disease can be represented as a 64-dimensonal vector by known miRNA-disease associations through the LINE method. The attribute feature is the additional information including the RNA sequence, disease semantics, drug chemical structure and etc. The attribute feature of each node can be represented as a 64-dimensonal vector by miRNA sequence learned by k-mer and disease semantics learned by LINE. Then, each miRNA and disease can be viewed as a 128-dimensonal vector by connecting behavior and attribute feature. Finally, each miRNA-disease association pair is a 256-dimensional vector by concatenating miRNA and disease. 5-fold cross validation was applied to evaluate the proposed method. Random Forest classifier is chosen to carry out the association prediction task. The details of the results can be seen in the following Figure 7 and Table 4.

The results demonstrated that the attribute feature (disease semantics graph embedding representation) can play an auxiliary role. The representation vector combining the two feature is easier to distinguish, which can improve the prediction performance of the computational model.

Conclusion

Obtaining distinguishable vectors as the input of the computational prediction model has always been a hot topic of concern. Existing methods which manually define and measure similarity are limited considering the time and space complexity. In this paper, we constructed a MeSH heading relationship network and implemented five kinds of graph embedding algorithms on it. Then, the qualities of the vectors were evaluated based on the relationship network itself and the two benchmark datasets including drug-target interaction and miRNA-disease association. Obviously, the results of relationship prediction prove that the semantic representation of terms such as disease can not only be used as independent carrier for input, but also as additional information to enhance the distinguishability of vectors. Despite the limited performance of the upgrade, compared with the previous feature generation approach such as similarity-based or chemical structure method, the proposed method is a fully automatic and pure semantic way, which will bring new enlightenment to relevant researchers. Predictably, MeSHHeading2vec can be viewed as a foundation to establish interesting connections between network and semantic in both computer and life sciences. Briefly, our method will establish valuable insights in MeSH heading representation and disease-, drug- and etc.-related computational prediction model, bring beneficial inspiration to relevant scholars and expand the computational omics research paradigm.

Key Points

- Considering wet experiments are labor-intensive and time-consuming, computational prediction models are widely used to accelerate the process of biological experiments, boost diagnosis and treatment of diseases as well as new drug development.
- However, Medical Subject Headings (MeSH) terms such as diseases and drugs are abstract entities that are difficult to be quantified as input for machine learning model.
- In this paper, we converted the MeSH tree structure to a relationship network and proposed a new pure semantic approach to represent arbitrary terms as vectors.
- Compared with traditional methods such as drug chemical structure and disease similarity, experiment results have shown that the pure semantic method

still has definitely advantages. In addition, we constructed two benchmark data sets including drug-target interaction and miRNA-disease association for subsequent test and evaluation.
- Briefly, it can act as input and continue to play a significant role in any disease-, drug-, microbe- and etc.-related computational models in bioinformatics. Besides, it can inspire relevant researchers to study the representation of terms in this network perspective.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib. All available data are uploaded in github: https://github.com/CocoGzh/MeSHHeading2vec.

Author Contributions

Z-H.G., Z-H. Y considered the algorithm, arranged the datasets and performed the analyses. All authors wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the grant of National Key R&D Program of China (2018YFA0902600) and the grants of the National Science Foundation of China, Nos. 61722212, 61861146002, 61732012 & 61902342.

Competing interests

The authors declare that they have no competing interests.

References

1. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell* 2015;**58**:586–97.
2. Tyanova S, Temu T, Sinitcyn P, *et al*. The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nat Methods* 2016;**13**:731.
3. Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. *Science* 2003;**300**:286–90.
4. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986;**30**:7–18.
5. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* 1997;**91**:183–203.
6. Hu X, Zhang X, Yoo I, *et al*. Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule. *International Journal of Intelligent Systems* 2010;**25**:207–23.
7. Cameron D, Kavuluru R, Rindflesch TC, *et al*. Context-driven automatic subgraph creation for literature-based discovery. *J Biomed Inform* 2015;**54**:141–57.
8. Torvik VI, Smalheiser NR. A quantitative model for linking two disparate literatures in MEDLINE. *Bioinformatics* 2007;**23**:1658–65.
9. Milanese J-S, Tibiche C, Zou J, *et al*. Germline variants associated with leukocyte genes predict tumor recurrence in breast cancer patients. *NPJ precision oncology* 2019;**3**:1–9.
10. Zou J, Wang E. eTumorType, an algorithm of discriminating cancer types for circulating tumor cells or cell-free DNAs in blood. *Genomics Proteomics Bioinformatics* 2017;**15**:130–40.
11. Li J, Lenferink AE, Deng Y, *et al*. Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun* 2010;**1**:1–9.
12. Zaman N, Li L, Jaramillo ML, *et al*. Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. *Cell Rep* 2013;**5**:216–23.
13. Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc* 2000;**88**:265.
14. Cai H, Zheng VW, Chang KC-C. A comprehensive survey of graph embedding: problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* 2018;**30**:1616–37.
15. Li L, Tibiche C, Fu C, *et al*. The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. *Genome Res* 2012;**22**:1222–30.
16. Cui Q, Yu Z, Purisima EO, *et al*. Principles of microRNA regulation of a human cellular signaling network. *Mol Syst Biol* 2006;**2**:46.
17. Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. *Knowledge-Based Systems* 2018;**151**:78–94.
18. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014, p. 701-710. ACM.
19. Tang J, Qu M, Wang M *et al*. Line: large-scale information network embedding. In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, p. 1067-1077. International World Wide Web Conferences Steering Committee.
20. Wang D, Cui P, Zhu W. Structural deep network embedding. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, p. 1225-1234. ACM.
21. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 2003;**15**:1373–96.
22. Ou M, Cui P, Pei J *et al*. Asymmetric transitivity preserving graph embedding. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, p. 1105-1114. ACM.
23. Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2017;**46**:D1074–82.
24. Zhang W, Yue X, Lin W, *et al*. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics* 2018;**19**:233.
25. Ben-Hur A, Noble WS. Kernel methods for predicting protein–protein interactions. *Bioinformatics* 2005;**21**:i38–46.
26. Huang Z, Shi J, Gao Y, *et al*. HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res* 2018;**47**:D1013–7.
27. Szklarczyk D, Gable AL, Lyon D, *et al*. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2018;**47**:D607–13.

28. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2018;**47**:D155–62.
29. Shen J, Zhang J, Luo X, *et al*. Predicting protein–protein interactions based only on sequences information. *Proc Natl Acad Sci* 2007;**104**:4337–41.
30. Landrum G. Rdkit documentation. *Release* 2013;**1**:1–79.
31. Guo Z-H, You Z-H, Wang Y-B, *et al*. A learning-based method for LncRNA-disease association identification combing similarity information and rotation Forest. *iScience* 2019;**19**:786–95.
32. Wang D, Wang J, Lu M, *et al*. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**:1644–50.
33. Wang Y, You Z-H, Yang S, *et al*. A high efficient biological language model for predicting protein–protein interactions. *Cell* 2019;**8**:122.
34. Guo Z-H, You Z-H, Li L-P *et al*. Combining high speed ELM with a CNN feature encoding to predict LncRNA-disease associations. In: *International Conference on Intelligent Computing*. 2019, p. 406-417. Springer.
35. You Z-H, Huang Z-A, Zhu Z, *et al*. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput Biol* 2017;**13**: e1005455.
36. Guo Z-H, Yi H-C, You Z-H. Construction and comprehensive analysis of a molecular association network via lncRNA–miRNA–disease–drug–protein graph. *Cell* 2019;**8**: 866.