doi: 10.1093/bib/bbaa212
Problem Solving Protocol

Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches

Tianyi Zhao, Yang Hu and Liang Cheng

Corresponding author: Liang Cheng. E-mail: liangcheng@hrbmu.edu.cn

Abstract

Motivation: The functional changes of the genes, RNAs and proteins will eventually be reflected in the metabolic level. Increasing number of researchers have researched mechanism, biomarkers and targeted drugs by metabolites. However, compared with our knowledge about genes, RNAs, and proteins, we still know few about diseases-related metabolites. All the few existed methods for identifying diseases-related metabolites ignore the chemical structure of metabolites, fail to recognize the association pattern between metabolites and diseases, and fail to apply to isolated diseases and metabolites. Results: In this study, we present a graph deep learning based method, named Deep-DRM, for identifying diseases-related metabolites. First, chemical structures of metabolites were used to calculate similarities of metabolites. The similarities of diseases were obtained based on their functional gene network and semantic associations. Therefore, both metabolites and diseases network could be built. Next, Graph Convolutional Network (GCN) was applied to encode the features of metabolites and diseases, respectively. Then, the dimension of these features was reduced by Principal components analysis (PCA) with retainment 99% information. Finally, Deep neural network was built for identifying true metabolite-disease pairs (MDPs) based on these features. The 10-cross validations on three testing setups showed outstanding AUC (0.952) and AUPR (0.939) of Deep-DRM compared with previous methods and similar approaches. Ten of top 15 predicted associations between diseases and metabolites got support by other studies, which suggests that Deep-DRM is an efficient method to identify MDPs. Contact: . Availability and implementation: https://github.com/zty2009/GPDNN-for-Identify-ing-Disease-related-Metabolites.

Key words: disease-related metabolites; deep learning; graph convolutional network

Introduction

Metabolism is an important biochemical reaction that accompanies the entire life cycle, and it is extremely susceptible to the occurrence and development of diseases, which in turn causes abnormalities in metabolites in the blood and urine [1]. The functional changes of the upstream (nucleic acid, protein, etc.) macromolecules will eventually be reflected in the metabolic level [2], such as changes in neurotransmitters, hormone regulation, receptor effect, cell signal release, energy transmission and intercellular communication, etc., so metabolism is located downstream of the gene regulation network and protein interaction network, and provides the terminal information of biology. Therefore, genomics and proteomics tell us what might happen, and metabolomics tells us what has happened [3].

Tianyi Zhao is a PhD student in the Department of Computer Science at the Harbin Institute of Technology. He currently works as a Bioinformatician in the Beth Israel Deaconess Medical Center.
Yang Hu is an Associate Professor in the Department of Life Science at the Harbin Institute of Technology. His expertise is bioinformatics.
Liang Cheng is a Professor in NHC and CAMS Key Laboratory of Molecular Probe and Targeted Theranostics, College of Bioinformatics Science and Technology at Harbin Medical University. He is also the Chief Editor of *Current Gene Therapy*.

Submitted: 21 June 2020; **Received (in revised form):** 27 July 2020

Using metabolomics data to study diseases also has the following advantages [4]. First, small changes in gene and protein expression at the functional level will be amplified on the metabolite, which makes detection easier. Second, many nonfunctional changes in genes and proteins will not be reflected on the metabolite, which makes metabolism play the role of 'noise filtering' in the process of upstream information transmission to downstream; Third, the types of metabolites are much smaller than the number of genes and proteins, and the molecular structure of the substance is much simpler, so studying diseases through metabolites is easier. In addition, common metabolites are similar in various biological systems (such as the primary metabolism of plants, microorganisms, animals), so the platform technology used in metabolomics research can be applied in different biological systems.

Recently, increasing number of researchers have devoted to discover disease mechanism, biomarkers and targeted drugs by metabolites. Mathewson *et al.* [5] found butyrate restoration can decrease intestinal epithelial cells apoptosis and mitigate graft-versus-host disease (GVHD). They also pointed out local and specific alteration of microbial metabolites has direct salutary effects on GVHD target tissues and can mitigate disease severity. Martínez-Reyes and Chandel [6] claimed that tricarboxylic acid cycle (TCA cycle), which is a ubiquitous metabolic pathway in aerobic organisms, is associated with cancers, immune and stem cells functions. Chang *et al.* [7] found 15 metabolites with CD4+ T-cell bioactivity and the biological activity of related T cells is quite high among Crohn's disease–associated metabolites. In addition, researchers also found multiple diseases such as cardiovascular disease [8], neurodegenerative disease [9], liver disease [10], neuroimmune disease [11], chronic kidney disease [12], etc. are associated with metabolites.

Overall, metabolites have shown their strong power of helping understand and against disease. However, compared with genes, RNAs and proteins, people knew few about disease-related metabolites. This is mainly caused by two reasons. (i) The proportion of human metabolites that can be identified via untargeted mass spectrometry (MS)-based metabolomics techniques is typically <2% of identified MS peaks [13]. (ii) Identifying related metabolites for each disease is time and money consuming. For the first problem, a number of very accurate open access and commercial tools for *in silico* metabolite and spectral prediction have recently become available [14]. However, for the second one, few tools have been developed for identifying disease-related metabolites.

In 2018, we constructed metabolites similarity network based on the similarity of diseases and predicted diseases-related metabolites by Random Walk (RW) [15]. Following our research, Wang *et al.* [16] added text mining scores into the metabolites similarity network and applied RW to traverse this network to obtain potentially relevant metabolites of diseases. Similarly, Lei and Tie [17] proposed 'MDBIRW' that used Gaussian Interaction Profile (GIP) to calculate disease similarity and then inferred metabolites similarity network. In addition, bi-random walk was implemented to traverse both diseases and metabolites network.

However, these methods all have three drawbacks. (i) The similarity of metabolites is calculated only based on similarity of their corresponding diseases, which ignored the chemical properties of metabolites. (ii) The principle of these methods is all based on the hypothesis of 'similarity metabolites are associated with similar diseases', which did not recognize the pattern of associations between metabolites and diseases. (iii) These methods are not available for those isolated diseases/metabolites in the networks. Specifically, the similarity between metabolites without corresponding disease and other metabolites cannot be calculated, and the potential metabolites of diseases without known related metabolites also cannot be predicted.

Therefore, we propose a novel method 'Deep-DRM' to overcome these drawbacks. First, the similarities of metabolites are obtained based on their chemical structure. Then, Graph Convolutional Network (GCN) is used to encode both metabolites and diseases network. Since metabolites are substrates or products of proteins, the features of metabolites are encoded by their corresponding proteins by one-hot encoding. Principal components analysis (PCA) is applied to reduce the dimension of features. Finally, the features are input into Deep Neural Network (DNN) to recognize the pattern of associations between metabolites and diseases. In addition, our method can also identify corresponding diseases and metabolites for isolated diseases/metabolites.

Methods

We propose a novel method called Deep-DRM, which is the fusion of GCN, PCA and DNN to identify potential diseases-related metabolites. Deep-DRM includes three steps (Figure 1).

Construction of networks

Metabolites network

Chemical property is the most important characteristic for metabolites getting involved in biochemical reactions. The development of software 'PaDEL-Descriptor' [18] gives us an opportunity to calculate molecular descriptors and fingerprints based on chemical structure. It can provide us the chemical property of metabolites such as atom-type electrotopological state descriptors, Crippen's logP and MR, and extended topochemical atom (ETA) descriptors.

The 1D&2D descriptors and fingerprints were calculated for each metabolite by PaDEL-Descriptor. A 2325 dimensional vector (1441 for 1D&2D descriptors and 881 for fingerprints) was used to describe the chemical property of each metabolite. Since the scales of dimensions are different, normalization is needed for each dimension. Z-score normalization was applied as following:

$$\hat{m}_i^k = \frac{m_i^k - \text{mean}\left(m^k\right)}{\text{std}\left(m^k\right)}, \tag{1}$$

where $\hat{m}_i^k$ denotes the kth dimension of ith metabolite after normalization, mean(m^k) denotes the average of original kth dimension of all metabolites and std(m^k) denotes the standard deviation of original kth dimension of all metabolites.

Then, the similarity of metabolites could be obtained by these vectors as following:

$$\text{sim}\left(\hat{m}_i, \hat{m}_j\right) = \frac{\sum_k^{2325} \hat{m}_i^k \times \hat{m}_j^k}{\sqrt{\sum_k^{2325} \left(\hat{m}_i^k\right)^2} \times \sqrt{\sum_k^{2325} \left(\hat{m}_j^k\right)^2}}, \tag{2}$$

where sim($\hat{m}_i, \hat{m}_j$) represents the similarity between ith metabolite and jth metabolite.

Finally, the metabolites network could be built in which metabolites are nodes and the similarities are the edges of the network.

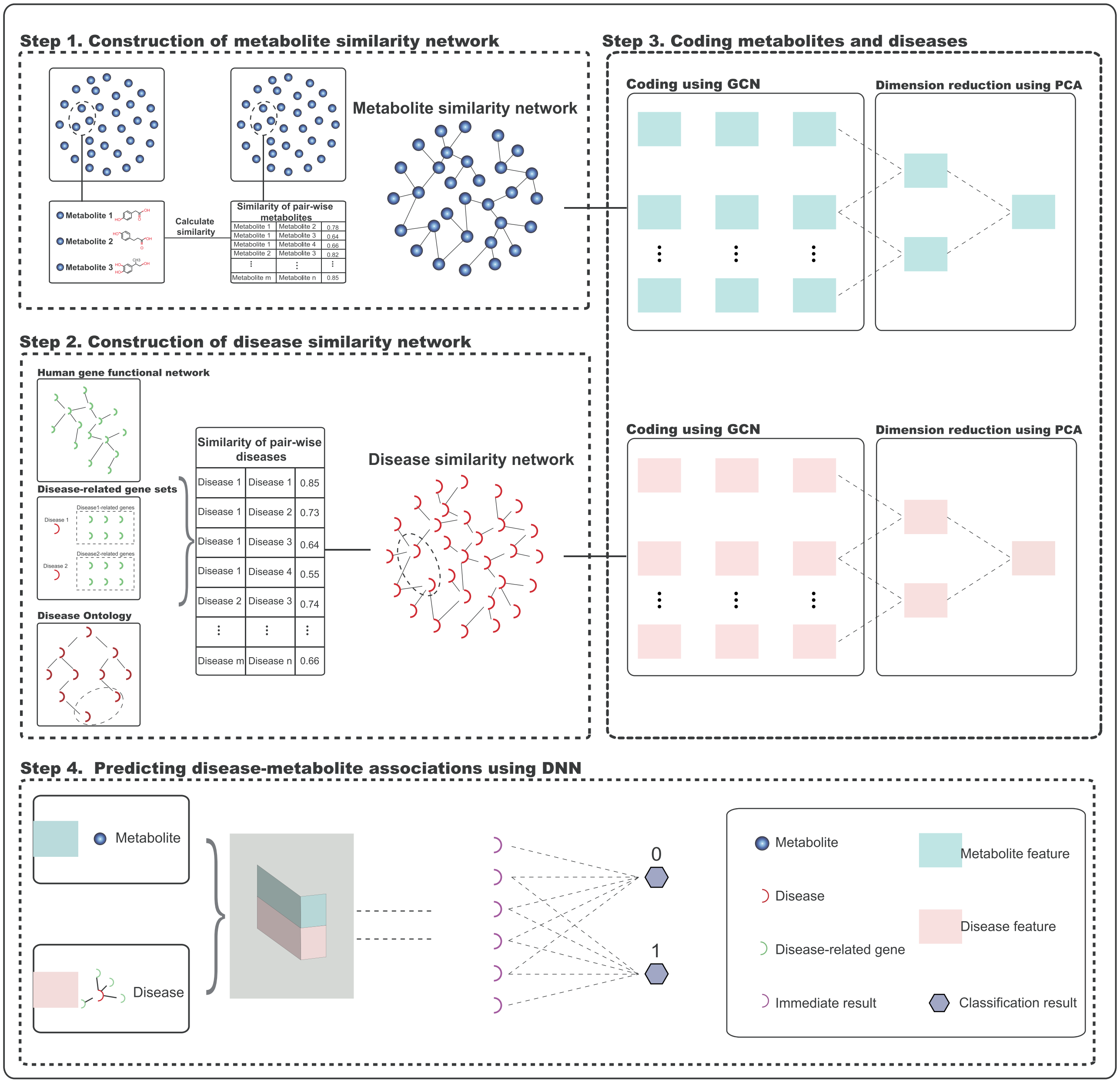
Step 1. Construction of metabolite similarity network
Metabolite 1
Metabolite 2
Metabolite 3
HO
OH
O
CH3
Calculate similarity
Similarity of pair-wise metabolites
Metabolite 1 | Metabolite 2 | 0.78
Metabolite 1 | Metabolite 3 | 0.64
Metabolite 1 | Metabolite 4 | 0.66
Metabolite 2 | Metabolite 3 | 0.82
⋮ | ⋮ | ⋮
Metabolite m | Metabolite n | 0.85
Metabolite similarity network
Step 2. Construction of disease similarity network
Human gene functional network
Disease-related gene sets
Disease1-related genes
Disease 1
Disease2-related genes
Disease 2
Disease Ontology
Similarity of pair-wise diseases
Disease 1 | Disease 1 | 0.85
Disease 1 | Disease 2 | 0.73
Disease 1 | Disease 3 | 0.64
Disease 1 | Disease 4 | 0.55
Disease 2 | Disease 3 | 0.74
⋮ | ⋮ | ⋮
Disease m | Disease n | 0.66
Disease similarity network
Step 3. Coding metabolites and diseases
Coding using GCN
Dimension reduction using PCA
Coding using GCN
Dimension reduction using PCA
Step 4. Predicting disease-metabolite associations using DNN
Metabolite
Disease
0
1
Metabolite
Metabolite feature
Disease
Disease feature
Disease-related gene
Immediate result
Classification result

Figure 1. Workflow of Deep-DRM. The first step is to construct metabolites network by calculating chemical similarity using structure-data file (SDF) of metabolites. Then, disease similarity network could be built based on semantic association in Disease Ontology (DO) and related genes. In third step, GCN were implemented to encode features of metabolites and diseases network. Afterwards, PCA was applied to reduce the dimension of metabolites and diseases features. Finally, combining features of diseases and metabolites, the features of metabolite-disease pairs (MDP) could be constructed. Finally, DNN was applied to learn the pattern of associations between diseases and metabolites by MDPs.

Diseases network

The similarity between diseases stems from the semantic association between diseases and the functional associations between disease-related genes.

SemFunSim [19] which is a method we proposed before not only uses disease-related gene sets to calculate disease similarity in a weighted network of human gene functions but also uses the relationship between two diseases in Disease Ontology (DO) to calculate disease similarity. Finally, the two kinds of similarities are combined together to obtain the final similarity of diseases.

The functional similarity score between a pair of genes is defined as $\mathrm{FunSim}(g_i, g_j)$

$$\mathrm{FunSim}\left(g_i, g_j\right) = \begin{cases} 1 & i = j \\ \mathrm{LLS_N}\left(g_i, g_j\right) & i \neq j \ \& \ \in \mathbb{E}\,(\mathrm{HumanNet})\,, \\ 0 & i \neq j \ \& \ \notin \mathbb{E}\,(\mathrm{HumanNet}) \end{cases} \qquad (3)$$

where $\mathrm{LLS_N}\left(g_i, g_j\right)$ measures the probability of a functional linkage between genes by log likelihood score (LLS) in HumanNet.

Then, we define the functional association between a gene g and a gene set $G = \{g_1, g_2, \cdots, g_k\}$ as $F_G(g)$.

$$F_G(g) = \max_{1 \le i \le k} \left(\text{FunSim}\,(g, g_i)\right), \quad g_i \in G, \tag{4}$$

where k indicates the number of genes in G.

If disease d_1 is related to gene set,$G_1 = \{g_{11}, g_{12}, \cdots, g_{1m}\}$ and if disease d_2 is related to gene set, $G_2 = \{g_{21}, g_{22}, \cdots, g_{2n}\}$. The similarity between d_1 and d_2 could be calculated as following:

$$\text{FunSim}\,(d_1, d_2) = \frac{\sum\limits_{1 \le i \le m} F_{G_2}\,(g_{1i}) + \sum\limits_{1 \le j \le n} F_{G_1}\,(g_{2j})}{m + n} \tag{5}$$

where $g_{1i} \in G_1, g_{2j} \in G_2$, m is the number of genes in G_1 and n is the number of genes in G_2.

Then, semantic similarity between disease pair d_1 and d_2 could be calculated as equation (6).

$$\text{SemSim}\,(d_1, d_2) = \frac{|\,G_1\,|}{|\,G_{\text{MICA}}\,|} \cdot \frac{|\,G_2\,|}{|\,G_{\text{MICA}}\,|} \tag{6}$$

where G_{MICA}represents the most informative common ancestor (MICA) of d_1 and d_2 in the directed acyclic graph (DAG) of DO.

Finally, the similarity of diseases is the product of FunSim() and SemSim().

$$\text{Sim}\,(d_1, d_2) = \text{FunSim}\,(G_1, G_2) \cdot \text{SemSim}\,(d_1, d_2) \tag{7}$$

In this way, the diseases network could be built in which diseases are nodes and the similarities are the edges of the network.

Network encoding

After building metabolites and diseases networks, each node of the two networks should contain its intrinsic feature. Then, GCN could encode the networks based on the structure of networks and feature of nodes. Finally, reduction of dimension could be achieved by PCA.

Feature extraction

Metabolites are substrates or products of proteins so their relationship with proteins could represent the feature of them.

One-hot code was used to encode the feature of metabolites. Assuming that there are k proteins, then the feature of metabolites could be denoted as following:

$$F_m = [P_1, P_2, \ldots, P_k], \tag{8}$$

where F_m denotes the feature of metabolites, P_i denotes the relationship between the metabolite and ith protein ($1 < i < k$). If the metabolite is the substrate or product of ith protein, $P_i = 1$, otherwise, $P_i = 0$.

Since enough information has been contained in the similarity of diseases, we only used one-hot code to encode the index of diseases.

Assuming that there are n diseases, then the feature of ith disease could be denoted as following:

$$F_{d_i} = [D_1, D_2, \ldots, D_n] \;, \tag{9}$$

where F_{d_i} denotes the feature of ith disease. Among D_1 to D_n, only $D_i = 1$ and all other elements are 0.

Graph encoding

GCN is a neural network algorithm that can directly extract the structural information and node information of networks. It has been widely used in many different fields of bioinformatics [20, 21].

The adjacency matrix A could be obtained from diseases and metabolites networks, respectively. A describes the connection between nodes.

Since the features of metabolites and diseases should contain their own information, the adjacency matrix should be further processed.

$$A' = A + I, \tag{10}$$

where I is the identity matrix.

The next step is to obtain the inverse degree matrix D′.

$$D'_{ii} = \sum\nolimits_j A'_{ij} \tag{11}$$

Finally, the feature of networks could be extracted as following:

$$X' = \text{ReLu}\left(D'^{-\frac{1}{2}} A' D'^{-\frac{1}{2}} X\right) \tag{12}$$

where X is the information of each node. For metabolites network, X could be obtained by formula (8). For diseases network, X could be obtained by formula (9). In addition, ReLu is the rectified linear unit. The formula of which is

$$f(x) = \max\,(x, 0) \tag{13}$$

Finally, we could obtain the feature of metabolites and diseases after GCN encoding, respectively.

Reduction of dimension

Since the number of proteins and diseases are large, the dimensions of metabolites and diseases are large too. Therefore, PCA was introduced to reduce the dimension of features.

As a well-established method, we won't explain the process of PCA in detail. We retained 99% of the feature information for both metabolites and diseases.

Classification of metabolites-diseases pairs

Since the features of metabolites and diseases could all be represented as a vector through the above processing, we constructed the feature of metabolites-diseases pairs (MDP) by combining their features.

The feature of MDP could be denoted as following:

$$F_{MDP} = \left[F_{m_1}, F_{m_2}, \ldots, F_{m_i}, F_{d_1}, F_{d_2}, \ldots F_{d_j}\right], \tag{14}$$

where $\left[F_{m_1}, F_{m_2}, \ldots, F_{m_i}\right]$ and $\left[F_{d_1}, F_{d_2}, \ldots F_{d_j}\right]$ are the feature of metabolites and diseases, respectively, after GCN and PCA encoding, and F_{MDP} is the feature of MDP. As shown in formula (14), the feature of MDP is the combination of the corresponding metabolite and disease.

Table 1. The parameters of DNN model

Structure	Parameters
Layer 1	Units: 512 Activation function: Sigmoid Dropout rate: 0.4
Layer 2	Units: 256 Activation function: Sigmoid Dropout rate: 0.3
Layer 3	Units: 128 Activation function: Sigmoid Dropout rate: 0.2
Layer 4	Units: 2 Activation function: Sigmoid
Loss function	Binary cross entropy
Optimizer	Rmsprop

Table 2. Data description in the three aims

	K_i	U_d	U_m
Metabolites	1436	1436	2293 (857)
Diseases	242	402 (160)	242
Known associations	3124	3124	3124
Unknown associations	344 388	229 760	13 794

Then, we can input the feature of MDP and their labels into DNN to identify true MDP. We built a DNN model with four layers. The parameters were set as Table 1.

The DNN model have two outputs. One is the probability of the test MDP being true and the other one is the probability of being false.

Dataset

The Human Metabolome Database (HMDB) [13] is the most comprehensive web resource about the human metabolome. We obtained the associations between metabolites and diseases from HMDB. In addition, the chemical structure and related proteins of metabolites were also obtained from HMDB.

We totally obtained 3524 diseases from DO to calculate similarity of diseases.

We divided our task into 3 aims. The first aim is to identify novel associations between known metabolites and known diseases. 'Known' means the metabolite/disease has related diseases/metabolites. We call this aim K_i. The second aim is to identify associations between known metabolites and unknown diseases. 'Unknown diseases' denotes those diseases without known related metabolites. We call this aim U_d. The last aim is to identify associations between unknown metabolites and known diseases. 'Unknown metabolites' represents those metabolites without known related diseases. We call this aim U_m. Table 2 shows the number of data used in these three aims.

As shown in Table 2, we obtained 3124 known associations (MDPs) between 1436 metabolites and 242 diseases from HMDB. Therefore, the unknown associations between these metabolites and diseases should be $1436 \times 242 - 3124 = 344\,388$. For aim U_d, we only selected diseases that have similarities with at least one of the 242 known diseases higher than 0.3 because the more closely related diseases are to 242 known diseases, the more potential our approach has to accurately find metabolites associated with them. Finally, we found 160 unknown diseases with high potential from 3524 diseases. Therefore, the number of unknown associations should be $160 \times 1436 = 229\,760$. For aim U_m, we selected metabolites associated with at least 120 proteins as unknown metabolites because the more associations between the metabolites and proteins, the more likely the metabolites are to participate in disease-related biochemical reactions. Finally, 857 unknown metabolites were obtained from 114 100 metabolites in HMDB. Therefore, the number of unknown associations should be $857 \times 242 = 13\,794$.

Training and testing

To verify the validity of Deep-DRM, we conducted 10-cross validation on the three aims separately.

Since the number of unknown MDPs are far more than the number of known MDPs, we randomly selected negative samples from unknown MDPs with the same number of positive sets (3124 samples). Therefore, for each aim, 3124 positive samples and 3124 negative samples were constructed as a new dataset. Then, we conducted 10-cross validation on this new dataset.

To test the stability of Deep-DRM, all the processes mentioned above (construct a new dataset by random selection of negative samples and 10-cross validation) were repeated 5 times for each aim.

Results

Verifying the validity of Deep-DRM

Since our method is the fusion of GCN, PCA and DNN, we compared Deep-DRM with two similar methods. PCA was selected due to the dimension reduction function. However, Deep Belief Network (DBN) that is constructed by Restricted Boltzmann machine (RBM) also has the ability of reducing dimension. Therefore, we fused GCN, DBN and DNN to construct a novel method named 'GRDNN'. We compared Deep-DRM with GRDNN to show the difference between PCA and DBN. Then, to show the power of GCN, we only used PCA and DNN to identify MDPs. We called this method 'PDNN'.

To show the performance of these three methods, we draw Figure 2.

As shown in Figure 2, all these three methods were tested on the three aims. The red bar denotes the performance of Deep-DRM, and the blue and green bars represent GRDNN and PDNN, respectively. Since we repeated the 10-cross validation 5 times by constructing different datasets, the error bars show the standard deviations of Area Under Curve (AUC) and Area Under Precision Recall Curve (AUPR). Deep-DRM performed best in all the three aims with high stability, and GRDNN was the worst. Analyzing the comparison experiments between Deep-DRM and GRDNN, we could know that the performance of DBN was worse than PCA because DBN extracts highly abstract features, which makes it is more suitable for image processing compared Deep-DRM and PDNN, the similarities of metabolites and diseases play an important role in recognizing the pattern of associations between metabolites and diseases.

Overall, Deep-DRM showed high AUC and AUPR in this comparison experiment, which showed its effectiveness of identifying potential MDPs.

Comparison with previous methods

To show the improvement of our method, we compared Deep-DRM with RW [15] and MDBIRW [17]. The parameters of

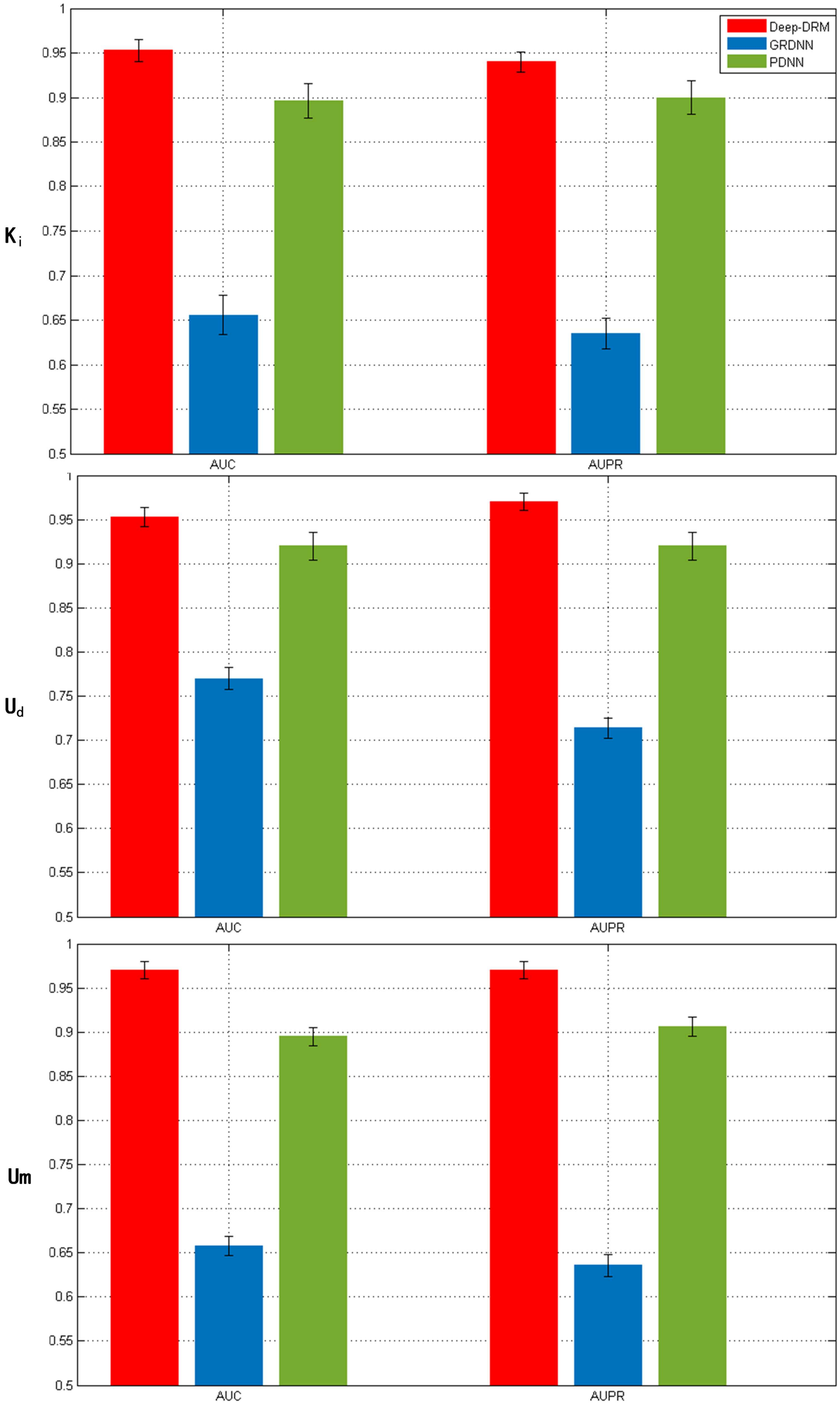

Figure 2. The comparison results between Deep-DRM and other two similar methods.

algorithms were set as same as their papers. As we mentioned before, these methods cannot deal with the isolated nodes in the network, so the two methods can only be tested on the K_i aim. In addition, RW-based methods are only suitable for leave-one-out verification because if a large number of connections are deleted in the network, the accuracy will be extremely reduced. Therefore, we did leave-one-out verification for Deep-DRM, RW and MDBIRW, respectively.

Table 3. The comparison results between Deep-DRM and previous methods

Method	AUC	AUPR
Deep-DRM	0.952	0.939
RW	0.959	0.065
MDBIRW	0.967	0.743
RW_{NN}	0.912	0.781
$BIRW_{NN}$	0.924	0.811

In addition, one of the most advantages of Deep-DRM is the way of building networks. We applied RW and Bi-random Walks (BIRW) in our networks. We called these two methods 'RW_{NN}' and '$BIRW_{NN}$', respectively.

Table 3 shows the comparison results which include Deep-DRM, two previous methods and previous methods on our networks. Although the AUC of Deep-DRM is less than RW and MDBIRW, there was no big difference between these methods. However, Deep-DRM shows significantly high AUPR than other methods'. In other words, although RW and MDBIRW can identify almost all the true MDPs, the false-positive rate is very high. Another thing worth noting is that the AUPR of RWNN and BIRWNN is significant higher than RW and MDBIRW, which means our networks are much better than the networks in previous studies, and our network can significantly reduce the false positive rate.

Overall, although the AUC of Deep-DRM is slightly lower than the previous methods, it significantly reduces the false positive rate and improves the AUPR. In addition, it can also identify corresponding diseases and metabolites for isolated diseases/metabolites, which cannot be achieved by RW-based methods.

Verifying novel MDPs

Because we demonstrated the effectiveness of our method in sections 'Verifying the validity of Deep-DRM' and 'Comparison with previous methods', we used it to identify associations between metabolites and diseases in the entire set of unknown MDPs.

We set 0.5 as the threshold and identified that 22 587 of 229 760, 2704 of 13 794 and 50 670 of 344 388 unknown MDPs are true for U_d, U_m and K_i, respectively.

Similar diseases have a greater number of overlapping metabolites

The first test was based on the hypothesis that similar diseases should be related to similar metabolites.

First, the similarity between any pair of diseases can be obtained by SemFunSim. Then, we selected metabolites related to these two diseases from the predicted results. The overlapping rate (OR) of metabolites can be calculated as follows:

$$OR_{i,j} = \frac{M_i \cap M_j}{M_i \cup M_j}, \tag{15}$$

where $OR_{i,j}$ denotes the overlapping rate between ith disease and jth disease. M_i represents the metabolites set of ith disease. Therefore, the ratio of the number of intersections to the number of unions is the overlapping rate for any pair of diseases.

Finally, we could test the Pearson correlation coefficient (PCC) between OR and similarity of diseases.

Table 4. PCC between OR and similarity of diseases in different aims

Aims	PCC	P-value
K_i	0.37	<2.2e−16
U_d	0.59	<2.2e−16
U_m	0.1	2.59e−58

As shown in Table 4, the predicted results show high correlation between OR and similarity of diseases, which indicates similar diseases do share similar metabolites.

From this experiment, we not only verified the accuracy of our predicted novel MDPs but also provided strong evidence for the hypothesis.

Case study

To further verify our predicted MDPs, we selected top 5 MDPs of three aims to do case studies. All these 15 MDPs are not recorded in the HMDB and predicted by Deep-DRM. Table 5 shows the evidences that support these MDPs in the literatures.

As shown in Table 5, 10 of 15 predicted MDPs have been reported by other researchers and only 5 MDPs do not have the support of literatures.

To highlight the function of metabolites in complex diseases, we discussed functions of two novel MDPs. Anders and Dekant [22] found that the HMDB0000512 (*N*-acetyl-L-phenylalanine) plays a role in xenobiotic detoxification and bioactivation and are important in the interorgan processing of xenobiotic-derived amino acid conjugates, which are related to Cerebral degeneration. Zanatta et al. [23] explored the function of HMDB0000459 (3-methylcrotonylglycine (3MCG)) in cerebral cortex of young rats. They found 3MCG would cause alterations of the cellular redox homeostasis, which could be involved in the pathophysiology of the neurological dysfunction and structural brain alterations.

These case studies show the accuracy of our predicted MDPs.

Discussion

As the final product of biological processes, metabolites are very promising biomarkers and an important part of understanding the pathogenesis. However, although people have known a lot about disease-related genes, RNAs and proteins, few knowledge about disease-related metabolites were discovered till now.

Although some computational methods based on RW have been developed to identify disease-related metabolites, they all had problems such as failure to effectively utilize the chemical characteristics of metabolites, unable to identify isolated metabolites/diseases-related diseases or metabolites.

In this paper, we proposed a novel method 'Deep-DRM' to overcome the drawbacks of previous methods. First, we calculated similarities of metabolites based on the chemical structures of them and used their related proteins as the features. Then, the similarities of diseases were obtained based on their functional gene network and semantic associations. After constructing metabolites and network networks, GCN was implemented to encode the features of metabolites and diseases, respectively. Therefore, the features of nodes contain not only their own information but also their associations with other nodes. Then, the dimension of features was reduced by PCA with retainment 99% original information. Afterward, the features of metabolites and diseases were combined together to construct

Table 5. Top ranked 15 novel MDPs predicted by Deep-DRM

Aim	Disease	Metabolite	Evidence
U_m	Eosinophilic esophagitis	HMDB0011147	PMID: 6267133
	Eosinophilic esophagitis	HMDB0006497	None
	Alzheimer's disease	HMDB0000010	PMID: 10873554
	Alzheimer's disease	HMDB0002142	PMID: 20164570
	Eosinophilic esophagitis	HMDB0001338	None
U_d	Sleep disorder	HMDB0000052	doi.org/10.1016/B978-0-323-03354-1.50111-5
	Cerebral degeneration	HMDB0000512	PMID: 8068563
	Sleep disorder	HMDB0000265	PMID: 32120028
	Adrenal gland disease	HMDB0000939	PMID: 597268
	Sleep disorder	HMDB0000192	PMID: 20163603
K_i	Alzheimer's disease	HMDB0000459	PMID: 23053545
	Autistic disorder	HMDB0000459	None
	Alzheimer's disease	HMDB0000824	PMID: 29370177
	Crohn's disease	HMDB0000459	None
	Alzheimer's disease	HMDB0000791	None

MDPs. Finally, DNN model with four layers was constructed to identify true MDPs.

The performance Deep-DRM was verified by comparing it with similar methods and previous methods. It showed outstanding precision in identifying novel-related metabolites for known diseases (U_m), known metabolites for novel diseases (U_d) and novel associations between known metabolites and diseases (K_i). To overcome the overfitting problem, we randomly selected negative samples from unknown MDPs and did 10-cross validation. The sampling and cross validation process has been repeated 5 times for each aim, we calculated both AUC and AUPR based on the average of these experiments. In addition, the standard deviation has been given to show the stability of Deep-DRM in different sample sets. Finally, we verified our predicted MDPs by the hypothesis of 'similar diseases should be related to similar metabolites' and case studies. The overlapping rate of predicted metabolites between any two diseases showed high correlation with similarity of the two diseases. In addition, 10 of top 15 predicted MDPs got literatures to support their accuracy.

Overall, we demonstrated that Deep-DRM achieves significantly more accurate results than the other state-of-the-art methods under different prediction tasks settings and different methods of performance evaluation. It is a powerful tool for identifying potential diseases-related metabolites.

Key Points

- Chemical structures of metabolites were used to calculate similarities of metabolites, which constructs metabolites network. The similarities of diseases were obtained based on their functional gene network and semantic associations, which constructs diseases network.
- We employed a Graph Convolutional Network (GCN)–based model to encode both metabolites and diseases networks.
- The results of our evaluation of Deep-DRM show that our method outperforms state-of-the-art approaches for disease-related metabolites prediction.

Funding

Tou-Yan Innovation Team Program of the Heilongjiang Province (2019–15); National Natural Science Foundation of China (grant no. 61871160); Heilongjiang Province Postdoctoral Fund (grant no. LBH-TZ20); Young Innovative Talents in Colleges and Universities of Heilongjiang Province (grant no. 2018–69).

References

1. Cani PD. Microbiota and metabolites in metabolic diseases. *Nat Rev Endocrinol* 2019;**15**:69–70.
2. Cedernaes J, Schönke M, Westholm JO, *et al.* Acute sleep loss results in tissue-specific alterations in genome-wide DNA methylation state and metabolic fuel utilization in humans. *Sci Adv* 2018;**4**:eaar8590.
3. Wishart D. Development of an assay for dietary and exposome measurements for precision medicine. *Scr Sci Pharm* 2017;**4**:1–38.
4. Gonçalves S, Romano A. Production of plant secondary metabolites by using biotechnological tools. In: Vijayakumar R, Raja SSS (eds). *Secondary Metabolites-Sources and Applications*. Rijeka – Croatia: IntechOPen, 2018, 81–99.
5. Mathewson ND, Jenq R, Mathew AV, *et al.* Gut microbiome–derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease. *Nat Immunol* 2016;**17**:505.
6. Martínez-Reyes I, Chandel NS. Mitochondrial TCA cycle metabolites control physiology and disease. *Nat Commun* 2020;**11**:1–11.
7. Chang Y-L, Rossetti M, Vlamakis H, *et al.* A screen of Crohn's disease-associated microbial metabolites identifies ascorbate as a novel metabolic inhibitor of activated human T cells. *Mucosal Immunol* 2019;**12**:457–67.
8. Roe AJ, Zhang S, Bhadelia RA, *et al.* Choline and its metabolites are differently associated with cardiometabolic risk factors, history of cardiovascular disease, and MRI-documented cerebrovascular disease in older adults. *Am J Clin Nutr* 2017;**105**:1283–90.
9. Camandola S, Plick N, Mattson MP. Impact of coffee and cacao purine metabolites on neuroplasticity and neurodegenerative disease. *Neurochem Res* 2019;**44**:214–27.

10. Chu H, Duan Y, Yang L, *et al*. Small metabolites, possible big changes: a microbiota-centered view of non-alcoholic fatty liver disease. *Gut* 2019;**68**:359–70.
11. Morris G, Berk M, Carvalho A, *et al*. The role of the microbial metabolites including tryptophan catabolites and short chain fatty acids in the pathophysiology of immune-inflammatory and neuroimmune disease. *Mol Neurobiol* 2017;**54**:4432–51.
12. McMahon GM, Hwang S-J, Clish CB, *et al*. Urinary metabolites along with common and rare genetic variations are associated with incident chronic kidney disease. *Kidney Int* 2017;**91**:1426–35.
13. Wishart DS, Feunang YD, Marcu A, *et al*. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 2018;**46**:D608–17.
14. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, *et al*. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Chem* 2015;**7**:44.
15. Hu Y, Zhao T, Zhang N, *et al*. Identifying diseases-related metabolites using random walk. *BMC Bioinformatics* 2018;**19**:116.
16. Wang Y, Juan L, Peng J, *et al*. Prioritizing candidate diseases-related metabolites based on literature and functional similarity. *BMC Bioinformatics* 2019;**20**: 574.
17. Lei X, Tie J. Prediction of disease-related metabolites using bi-random walks. *PLoS One* 2019;**14**(11):e0225380.
18. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 2011;**32**:1466–74.
19. Cheng L, Li J, Ju P, *et al*. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One* 2014;**9**(6):e99415.
20. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 2018;**34**:i457–66.
21. Parisot S, Ktena SI, Ferrante E, *et al*. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med Image Anal* 2018;**48**:117–30.
22. Anders MW, Dekant W. Aminoacylases. *Adv Pharmacol* 1994;**27**:431–48.
23. Ângela Z, Alana PM, Anelise MT, *et al*. Neurochemical evidence that the metabolites accumulating in 3-Methylcrotonyl-CoA carboxylase deficiency induce oxidative damage in cerebral cortex of young rats. *Cell Mol Neurobiol* 2013;**34**:137–46.