Hindawi
Complexity
Volume 2020, Article ID 9342640, 11 pages
https://doi.org/10.1155/2020/9342640

Research Article

Predicting Metabolite-Disease Associations Based on Linear Neighborhood Similarity with Improved Bipartite Network Projection Algorithm

Xiujuan Lei and Cheng Zhang

School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710119, China

Correspondence should be addressed to Xiujuan Lei; xjlei@snnu.edu.cn

Received 11 February 2020; Revised 18 April 2020; Accepted 4 May 2020; Published 23 May 2020

Academic Editor: Dimitri Volchenkov

A large number of clinical observations have showed that metabolites are involved in a variety of important human diseases in the recent years. Nonetheless, the inherent noise and incompleteness in the existing biological datasets are tough factors which limit the prediction accuracy of current computational methods. To solve this problem, in this paper, a prediction method, IBNPLNSMDA, is proposed which uses the improved bipartite network projection method to predict latent metabolite-disease associations based on linear neighborhood similarity. Specifically, liner neighborhood similarity matrix about metabolites (diseases) is reconstructed according to the new feature which is gained by the known metabolite-disease associations and relevant integrated similarities. The improved bipartite network projection method is adopted to infer the potential associations between metabolites and diseases. At last, IBNPLNSMDA achieves a reliable performance in LOOCV (AUC of 0.9634) outperforming the compared methods. In addition, in case studies of four common human diseases, simulation results confirm the utility of our method in discovering latent metabolite-disease pairs. Thus, we believe that IBNPLNSMDA could serve as a reliable computational tool for metabolite-disease associations prediction.

1. Introduction

Metabolites, the final products of cellular regulatory process, whose levels can be considered as the ultimate response of biological systems to genetic or environmental changes have significant effects in human body [1]. Meanwhile, it is a trend for disease researches to find the effect in molecular level with the rapidly developing biomedical instruments, and analytical platforms [2, 3] and metabolisms disrupted by disease state are widely identified as disease signatures [4].

Although many metabolite signatures of diseases have been gradually identified by high-throughput metabolomics technologies in metabolomics [4, 5], the unconfirmed metabolite-disease associations still exist in large numbers. Furthermore, the efficiency of obtaining useful results by conventional biology experiments is not high due to the factories of time, fund, and accuracy. Thus, developing computational methods to efficiently and reliably excavate the potential metabolite-disease associations is significant for human health and medical advance, which also can solve time-consuming and labor-intensive problems. RWRMDA [6] is the first method to explore the latent associations between metabolites and diseases, which pushes the development of computational method in metabolomics. However, they do not consider the diseases similarity when calculating the last predicted results. Although KATZMDA [7] considers two similarities, less information about similarity integration is a disadvantage. Additionally, some similarity measurement methods with biological characteristic of diseases or metabolites have been widely taken advantage of other fields in bioinformatics such as functional similarity or semantic similarity. Howbeit, some biological characteristics between disease or metabolite pairs are insufficient. Other methods about measuring similarity such as

Gaussian interaction profile kernel similarity [8] or cosine similarity [9] based on pairwise topological similarities between diseases or metabolites are not robust enough as noted by [10].

In this paper, we put forward an improved bipartite network projection method based on linear neighborhood similarity for unconfirmed metabolite-disease association predictions (IBNPLNSMDA) (see Figure 1). Firstly, the new feature matrix is obtained by WKNKN and integrating metabolite (disease) similarities in order to make full use of existing data. Secondly, the relevant linear neighborhood similarity is constructed based on utilizing new feature matrix and reconstructing data points from neighbors. Thirdly, the improved bipartite network projection algorithm is utilized to predict the potential by combining the linear neighborhood similarity about metabolites (diseases) with the known metabolite-disease associations, which guarantees the accuracy of predictions. Finally, the IBNPLNSMDA obtains an AUC value of 0.9634 which outperforms the other methods in LOOCV. In addition, four types of case studies demonstrated the reliability and feasibility of IBNPLNSMDA.

2. Materials and Methods

2.1. Human Metabolite-Disease Associations. Firstly, the data of the known human metabolite-disease associations are extracted from human metabolome database (HMDB). Due to the calculation of disease semantic similarity and disease functional similarity, the data about diseases ontology [11] and DisGeNET [12] (http://www.disgenet.org/web/DisGeNET/menu) need to be considered. Secondly, we select the disease with DOID according to diseases ontology. Then, the common disease between DisGeNET and the disease we have selected in the last step become the final diseases data. Finally, the known human metabolite-disease network (see Figure 2) is constructed according to the final diseases data [13], which contains 3589 distinct experimentally confirmed human metabolite-disease associations about 2121 metabolites and 130 diseases. Based on these associations, we construct a $nd \times nm$ dimensional adjacency matrix M, where nd and nm are denoted as the number of diseases and metabolites. If a disease $d(i)$ has been experimentally verified to be associated with a metabolite $m(j)$, then $M(i,j)$ equals to 1, otherwise 0.

2.2. Diseases Functional Similarity. Based on the assumption that the more common the related genes between two diseases are, the larger the similarity between two diseases is. According to DisGeNET, we extract the associations between diseases and their relevant genes. Then, we construct an adjacency matrix GD in which the row represents diseases and the column represents genes and utilize the cosine similarity measurement to calculate disease similarity by calculating the angle cosine values of two vectors [9]:

$$DFS\left(d_i, d_j\right) = \left(\overrightarrow{GD\left(d_q\right)}, \overrightarrow{GD\left(d_p\right)}\right) = \frac{\overrightarrow{GD\left(d_q\right)}\ \overrightarrow{GD\left(d_q\right)}}{\left|\overrightarrow{GD\left(d_q\right)}\right|\left|\overrightarrow{GD\left(d_q\right)}\right|}, \tag{1}$$

$$G_n\left(d_q\right) = \begin{cases} 1, & \text{if } G_n \text{ is associated with } d_q, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\overrightarrow{GD(d_q)} = [G_1(d_q), \ldots, G_n(d_q), ..G_{nd}(d_q)]$ and $GD(d_q)$ denotes the associations of disease q with all the genes.

2.3. Diseases Semantic Similarity. A disease can be described as a directed acyclic graph (DAG) according to the mesh database [14, 15]. Taking disease D as an example, we use DAG (D, $T(D)$, $E(D)$) to represent it, where $T(D)$ is the node set consisting of the disease D and its ancestor nodes and $E(D)$ is the edge set including the direct edges from parent nodes to child nodes. And then, the semantic value of the disease D is given by the following equation [16]:

$$D_V(D) = \sum_{d \in T(D)} D_D(d), \tag{3}$$

$$D_D(d) = \begin{cases} 1, & \text{if } d = D, \\ \max\{\Delta * D_D(d') | d' \in \text{children of } d\}, & \text{if } d \neq D, \end{cases} \tag{4}$$

where Δ is the layer contribution factor which is set 0.5 in this study as in previous literature [17]. The diseases located in the same layer contribute the same semantic value to disease D, but the contribution of other diseases decreases by a factor Δ when the layer between these diseases and D increases. Sharing the larger parts of DAGs between 2 diseases is considered to be more similar. Thus, we define semantic similarity between d_i and d_j as follows:

$$DSS\left(d_i, d_j\right) = \frac{\sum_{t \in T(D_i) \cap T(D_j)} \left(D_i(t) + D_j(t)\right)}{D_V\left(D_i\right) + D_V\left(D_j\right)}. \tag{5}$$

2.4. Metabolite Functional Similarity. The metabolite functional similarity depends on the basic idea that two functional similar metabolites have the similar diseases. Assume that DT_A and DT_B represent a group of diseases associated with the metabolite m_A and m_B, respectively. Firstly, we, respectively, select the maximum semantic similarity between two diseases in DT_A and DT_B, representing the similarity of a disease and a disease group which is defined as follows [6]:

$$\text{Sim}(dt, DT) = \max_{1 \leq i \leq k} \left(DSS\left(dt, dt_i\right)\right), \tag{6}$$

where dt and $DT = \{dt_1, dt_2, \ldots, dt_k\}$ represent a disease and a disease group, respectively.

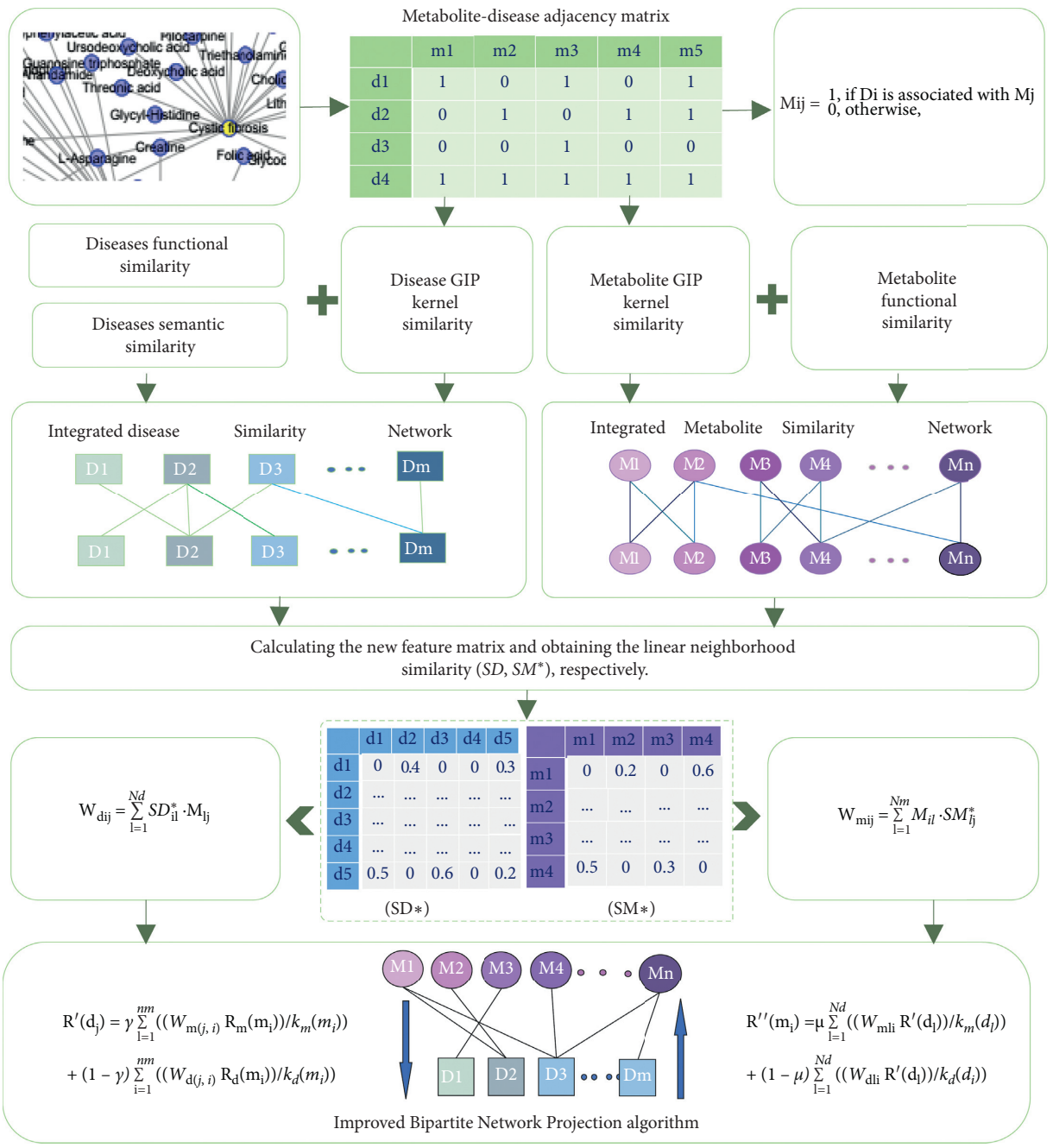

Figure 1: Flowchart of IBNPLNSMDA.

Then, the functional similarity between metabolite A and B is denoted as $MFS(m_A, m_B)$, which is calculated as

$$MFS(m_A, m_B) = \frac{\sum_{1 \le i \le |DT_A|} \mathrm{Sim}(dt_i, DT_A) + \sum_{1 \le j \le |DT_B|} \mathrm{Sim}(dt_j, DT_B)}{|DT_A| + |DT_B|}, \tag{7}$$

where two metabolites are connected if the similarity score is greater than 0 and the score is set as the weight in the metabolite functional similarity network.

2.5. Gaussian Interaction Profile Kernel Similarity. We use the vector IP to represent the interaction profiles in the known association network between metabolites and

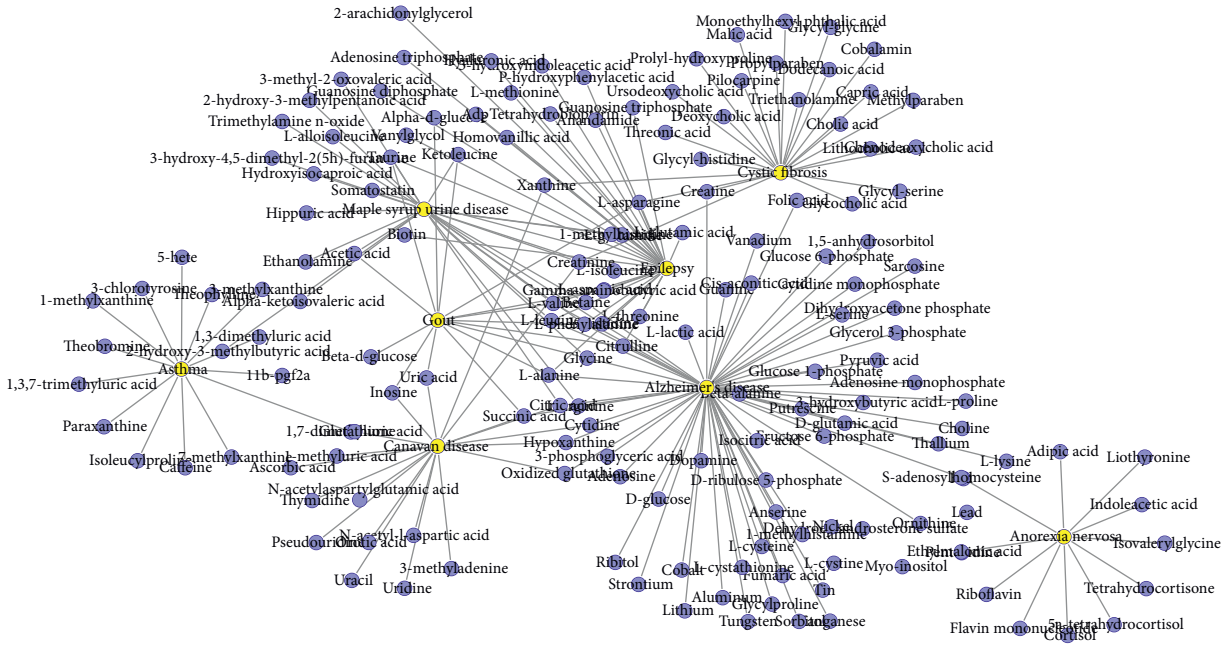

FIGURE 2: A part of the known metabolite-disease associations network. Yellow nodes represent diseases, and purple nodes represent metabolites.

diseases. Because the adjacency matrix M contains the information about the associations between diseases and metabolites, the $IP(d_i)$ and $IP(m_j)$ represent the ith row and jth column of the adjacency matrix M which is mentioned above. Therefore, the Gaussian interaction profile kernel similarity of diseases and metabolites can be given as follows [15]:

$$GD\left(d_i, d_j\right) = exp\left(-\omega_d \left\| IP\left(d_i\right) - IP\left(d_j\right)\right\|^2\right), \tag{8}$$

$$GM\left(m_i, m_j\right) = exp\left(-\omega_m \left\| IP\left(m_i\right) - IP\left(m_j\right)\right\|^2\right), \tag{9}$$

where parameters ω_d and ω_m are used to control the kernel bandwidth, which can be gained by normalizing the original ω_d' and ω_m' (both were set as 1 based on previous work), which can be calculated as follows:

$$\omega_d = \omega_d' \Big/ \left(\frac{1}{nd}\sum_{i=1}^{nd} IP\left(d_i\right)^2\right), \tag{10}$$

$$\omega_m = \omega_m' \Big/ \left(\frac{1}{nm}\sum_{k=1}^{nm} IP\left(m_i\right)^2\right). \tag{11}$$

Yet two matrix GD and GM are obtained, in which the entity $GD(d_i, d_j)$ represents the Gaussian interaction profile kernel similarity between disease d_i and diseases d_j and $GM(m_i, m_j)$represents the Gaussian interaction profile kernel similarity between metabolites m_i and m_j.

2.6. Integrated Similarity for Metabolites and Diseases. We denote the final similarity matric of metabolites as $SM \in (nm * nm)$ which is constructed by matrices MFS and GM. Meanwhile, the final similarity matric of diseases is defined as $SD \in (nd * nd)$ which consists of matrices DFS, DSS, and GD. By combining relevant matrices, the integrated similarity for metabolites and diseases is defined as follows:

$$SM\left(m_i, m_j\right) = \begin{cases} MFS\left(m_i, m_j\right), & \text{if } MFS\left(m_i, m_j\right) \neq 0, \\ \dfrac{MFS\left(m_i, m_j\right) + GM\left(m_i, m_j\right)}{2}, & \text{otherwise,} \end{cases} \tag{12}$$

$$SD\left(d_i, d_j\right) = \begin{cases} DSS\left(d_i, d_j\right), & \text{if } DFS\left(d_i, d_j\right) = 0, \\ GM\left(d_i, d_j\right), & \text{if } DSS\left(d_i, d_j\right) = 0,\ DFS\left(d_i, d_j\right) = 0, \\ \dfrac{DFS\left(d_i, d_j\right) + GD\left(d_i, d_j\right) + DSS\left(d_i, d_j\right)}{3}, & \text{otherwise.} \end{cases} \tag{13}$$

2.7. Features of Diseases and Metabolites. In this section, we use vector *IP* of diseases and metabolites from adjacency matrix *M* representing the initial feature vectors [18–20], respectively. However, most of associations between diseases and metabolites have not been verified which lead to feature very sparse. In order to solve this problem, WKNKN [21] as a preprocessing procedure is utilized to infer the interaction likelihood score for these latent pairs based on their known neighborhoods. Specifically, there are three steps when WKNKN replaces $M(i,j) = 0$ with an interaction likelihood value (Algorithm 1):

(1) Take metabolite i as an example, and we need to obtain the K known metabolites nearest to m_i according to integrated similarity matric (SM) and utilize their corresponding vector IP to estimate the interaction likelihood profile for m_i.

(2) Similar to the step (1), the interaction likelihood profile for d_j can be calculated.

(3) By taking the average of two interaction likelihood profile, the $M(i,j)$ is replaced if $M(i,j) = 0$. Finally, we get a new feature matrix MF about diseases and metabolites. The Algorithm 1 for WKNKN is demonstrated as follows.

2.8. Linear Neighborhood Similarity. Roweis et al. [22] reveal that it is close to a locally linear patch of the manifold between a data point and its neighbors, and Wang et al. [10] discover that each point can be optimally reconstructed by its neighbors. Besides, Zhang et al. [23, 24] apply the linear neighborhood similarity in bioinformatics which achieves better prediction performance. Based on these studies [6, 23, 25], we reconstruct the metabolite (disease) pairwise similarities. Take disease as an instance, let X_i denote the feature vector of the ith diseases in MF, and we use the following objective function, which minimizes the reconstruction error:

$$\varepsilon_{\mathrm{i}} = \left\| X_i - \sum_{i_{j:}X_{ij} \in N\left(X_i\right)} w_{ii_j} X_{i_j} \right\|^2, \quad \text{s.t.} \sum_{i_j: Xi_{j\in}(X_i)} w_{ii_j} = 1,\ w_{ii_j} \geq 0, \tag{14}$$

where $N(X_i)$ is defined the set of k (a free parameter) nearest neighbors which is calculated by Euclidean distance of X_i. X_{i_j} is the jth neighbor of X_i, and w_{ii_j} denotes the contribution of X_{i_j} to the reconstruction of X_i and could be regarded as their similarities. Let $G_{i_j i_k} = (X_i - X_{i_j})^T (X_i - X_{i_k})$. Then, ε_{i} can be rewritten as

$$\varepsilon_{\mathrm{i}} = \sum_{i_j, i_k: X_{i_j}, X_{i_k,} \in N\left(X_i\right)} w_{ii_j} G_{i_j i_k} w_{ii_k(2)}. \tag{15}$$

The Tikhonov regularization term that minimizes the norm of reconstructive weight w_i is adopted to avoid overfitting, and the objective function can be modified as

$$\varepsilon_i = \sum_{i_j, i_k: X_{i_j}, X_{i_k,} \in N\left(X_i\right)} w_{ii_j} G_{i_j i_k} w_{ii_k} + \alpha w_i^2 = w_i^T \left(G^i + \alpha I\right) w_i, \quad \text{s.t.} \sum_{i_j: Xi_{j\in}(X_i)} w_{ii_j} = 1,\ w_{ii_j} \geq 0, \tag{16}$$

where $w_i = \left\{w_{i_1}, w_{i_2}, \ldots, w_{i_k}\right\}$ and α is the penalty parameter for the regularization term which is set 1 for simplicity. Standard quadratic programming is used to solve (16), and the results are the reconstruction weights of X_i. After every feature vector about diseases in MF are calculated, we finally get a weight matrix W whose dimension is $nd * nd$ that could be treated as the disease linear neighborhood similarity (SD^*). Similarly, when we input feature vector about metabolites in MF, we also get a weight matrix W whose dimension is $nm * nm$ that could be treated as the metabolite linear neighborhood similarity (SM^*).

2.9. Improved Bipartite Network Projection Recommendation Algorithm. The baseline bipartite network projection recommendation algorithm [26] is a two-round resource transfer process which does not consider the weights of relevant similarities by just using the information of the known metabolite-disease matric (M). However, the bias for allocation of resources about each metabolite (disease) prefers to a specific disease (metabolite) together with their similar metabolites (or diseases). Simultaneously, enlightened by the idea that a potential metabolite (disease) could be predicted according to the related similar metabolites (diseases) [27–29], the similarity weights about metabolites and diseases are, respectively, considered when the resources are allocated which can be written as

$$\mathrm{W_m}(i,j) = \sum_{l=1}^{nm} M_{il} \cdot SM^*_{lj}, \tag{17}$$

$$\mathrm{W_d}(i,j) = \sum_{n=1}^{nd} SD^*_{in} \cdot M_{nj}, \tag{18}$$

where $W_m \in (nd * nm)$ and $W_d \in (nd * nm)$ represent the different weighted matrices when allocating resources according to different similarities. Nm represents the

Input: matrixes $M \in R^{nd*nm}$, $SD \in R^{nd*nd}$, $SM \in R^{nm*nm}$, neighborhood sizes K, and decay term T.
Output: new feature matrix MF.
for $p \longleftarrow 1$ to nm **do**
  mnn = KNN (p, SM, K)
or $i \longleftarrow 1$ to K do
    $w_i = T^{i-1} SM(p, mnn_i)$
  end for
  $Qm = \sum_{i=1}^{K} SM(p, mnn_i)$
  $M_m(p) = (1/Qm) \sum_{i=1}^{K} w_i M(mnn_i)$
end for
for $q \longleftarrow 1$ to nd **do**
  dnn = KNN (q, SD, K)
for $j \longleftarrow 1$ to K **do**
  $w_j = T^{i-1} SD(q, dnn_j)$
end for
$Qd = \sum_{j=1}^{K} SD(q, dnn_j)$
$M_d(q) = (1/Qd) \sum_{j=1}^{K} w_j M(dnn_j)$
end for
$M_{dm} = (M_m + M_d)/2$
MF = max(M, M_{dm})
return MF
end function

Algorithm 1: WKNKN.

number of metabolites, and nd represents the number of diseases. M is an adjacency matrix which is mentioned before. Next, each metabolite would be allocated with a score after two-round resource distribution. In the first round, the initial resource in $MM = \{m_1, m_2, ..., m_{nm}\}$ flows to $D = \{d_1, d_2, ..., d_{nd}\}$ [30] according to W_m and W_d, and the jth D node gains the resource as follows:

$$R'(d_j) = \gamma \sum_{i=1}^{nm} \frac{W_m(j,i) R_m(m_i)}{k_m(m_i)} + (1-\gamma) \sum_{i=1}^{nm} \frac{W_d(j,i) R_d(m_i)}{k_d(m_i)}. \tag{19}$$

Then, all the resources located on the D node returns back to MM by W_m and W_d [28], and the final resource located on the m_i node is

$$R''(m_i) = \mu \sum_{l=1}^{nd} \frac{W_m(l,i) R'(d_l)}{k_m(d_l)} + (1-\mu) \sum_{l=1}^{nd} \frac{W_d(l,i) R'(d_l)}{k_d(d_l)}, \tag{20}$$

where $k_m(m_i)$ is the sum of the ith row of the weighted matrix W_m and the $k_d(d_l)$ is the sum of the lth row of the weighted matrix W_d. $R_m(m_i)$ is the initial resource located on M_i based on the weighted matrix W_m, and $R_d(m_i)$ is the initial resource based on M_i using the weighted matrix W_d. For simplification, the damping factors γ and μ which are used to balance the scores between W_m and W_d are set 0.5.

3. Results

Leave-one-out across validation (LOOCV) is utilized to evaluate the prediction accuracy of IBNPLNSMDA. Each known metabolite-disease association is selected in turn as the test sample, and the rest of associations are regarded as training samples. Moreover, all metabolite-disease pairs whose associations are not confirmed would be considered negative samples, while the positive samples consist of the known associations. Thereafter, we rank each test sample with all metabolite-disease pairs without known associations based on the predicted scores. Additionally, test samples with rankings above the given threshold are regarded to be successful samples. According to the results of the LOOCV, AUC which is the area under the ROC (receiver operating characteristic) curve containing true-positive rate (TPR) and the false-positive rate (FPR) and AUPR which is the area under PR (precision-recall) curve containing precision and recall plays significant roles in evaluation performance of method. After LOOCV, IBNPLNSMDA obtains reliable AUC value of 0.9634 and AUPR value of 0.4971 which indicates that IBNPLNSMDA has satisfactory prediction performances.

3.1. Comparison. In this section, we explore the influence of main parts on the accuracy of our method and compare other methods such as RWRMDA and KATZMDA based on the same data about known metabolite-disease associations as follows: firstly, we compare baseline bipartite network projection method (BBNP) which only contains the information of known metabolite-disease pairs and RWRMDA [6] which uses the random walk model and only considers

metabolite similarity with our method (IBNPLNS). These methods get relevant results about AUC (BBNP = 0.6611 and RWRMDA = 0.73200) and AUPR (BBNP = 0.2879 and RWRMDA = 0.0916), while our method gets 0.9634 for AUC and 0.4917 for AUPR, which indicates that, respectively, considering related information of diseases and metabolites and making it for the foundation of constructing relevant similar network as the input of prediction method is beneficial to improve prediction performance. Secondly, here are three methods to compare such as the method which we make traditional similarity as input (IBNP with integrated similarity), the method which we do not use WKNKN before calculating linear neighborhood similarity (IBNP without WKNKN), the method named KATZMDA which is based on the path searching method and only uses simple data when constructing relevant similarities. The compared results shown in Figures 3 and 4 indicate that our method is better than compared methods. The reason for a higher predictive performance is that we use WKNKN to build new feature matrix before the construction of linear neighborhood similarity which cuts down the influence by lack and imbalance of relevant known information. According to above tests, we find main parts are crucial to our methods for improvement of accuracy about prediction, and our method is expected to be a reliable biomedical research tool for predicting latent metabolite-disease associations.

3.2. Parameter Analysis. According to the previous study [31], the K in WKNKN is set 5 and T, which is a decay term with $T \leq 1$, is selected 0.5 for convenience. Then, the number of linear neighborhood of diseases and metabolites represent kd and km, respectively, which are set as [10,50] and utilize the result of 10-fold cross validation to analyse relevant parameters. In this study, we set km = 50 according to Table 1 (deep multiview (Figure 5)).

3.3. Case Study. In this section, four kinds of diseases such as hepatitis, leukemia, obesity, and Alzheimer's disease are selected for case studies to explore their pathogenic mechanism from the perspective of metabolites. There are 10, 10, 9, and 7 out of the top 10 predicted metabolites that could be verified for the four diseases by literature. The predicting network with several diseases and their relevant top 10 metabolites which include their associations and their relevant neighborhood is shown in Figure 6.

Hepatitis, a general term for inflammation of the liver usually refers to the destruction of liver cells and the damage of liver function, which is caused by many pathogenic factors, such as virus, bacteria, parasite, chemical poison, drug, alcohol, and autoimmune factor. We conducted a case study on hepatitis using our calculation method. As illustrated in Table 3, the top 10 predicted metabolites interrelated with hepatitis are selected and verified to be correlative. For instance, high concentrations of homocysteine (2ed) could have favorable consequences in HCV (chronic hepatitis virus C) infection [32].

Leukemia is a group of life-threatening malignant disorder of the blood and bone marrow. Most of patients have

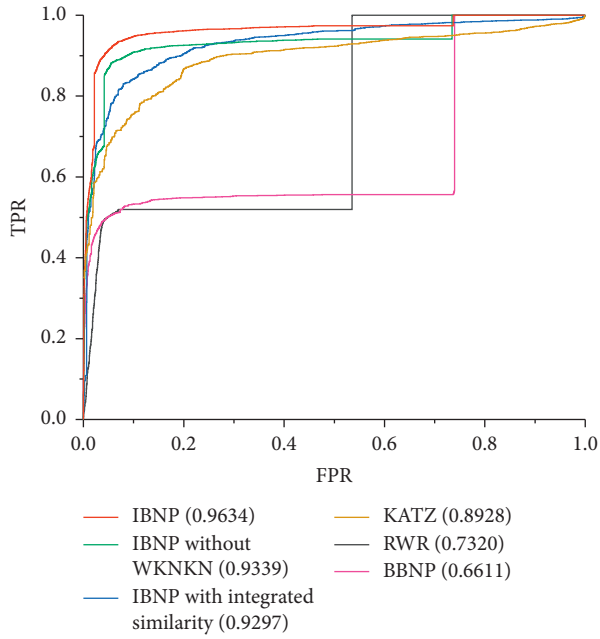

Figure 3: The ROC about LOOCV.

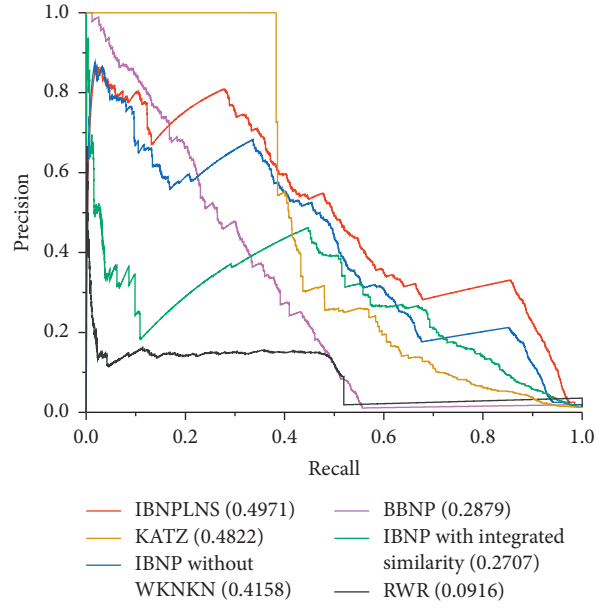

Figure 4: Performance comparison about the PR curve.

Table 1: Parameter analysis for km.

km	10	20	30	40	50
AUCs	0.7598	0.7901	0.8087	0.8370	0.8553

km is changed, and kd is set 10 under 10-fold cross validation.

Table 2: Parameter analysis for kd.

kd	10	20	30	40	50
AUCs	0.8553	0.8542	0.8589	0.8579	0.8563

kd is changed, and km is set 50 under 10-fold cross validation.

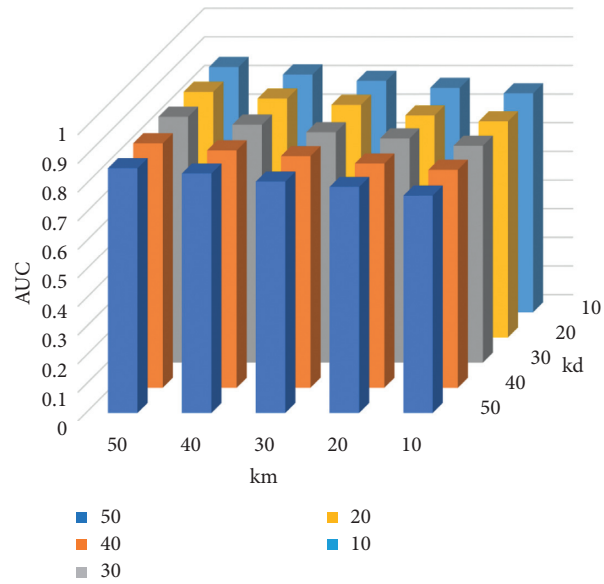

Figure 5: AUCs of different parameters under 10-fold cross validation. It shows the comparison of the parameters by integrating the data in Tables 1 and 2.

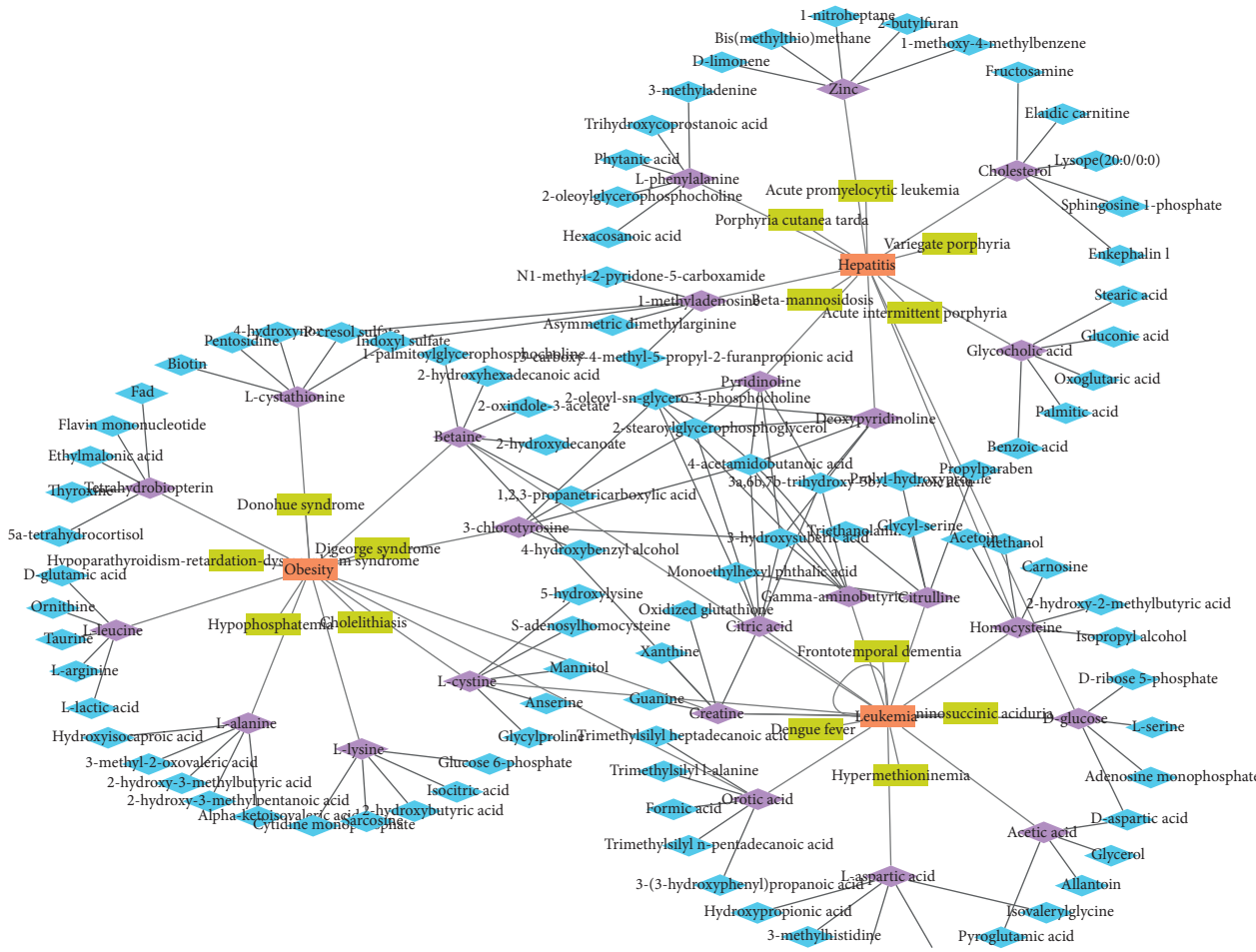

Figure 6: The predicting disease-metabolite association network. Three diseases and their relevant top 10 metabolites in cases studies are selected. Moreover, their top 5 relevant neighborhoods are shown. The rectangle represents diseases, and diamond represents metabolites. The orange represents target diseases, and the green represents their relevant neighborhoods. The purple represents the predicted metabolites, and the blue represents their relevant neighborhoods.

Table 3: Candidate metabolites of hepatitis.

Hepatitis		
Rank	Metabolite name	Evidences
1	Deoxypyridinoline	PMID: 8887033
2	Homocysteine	PMID: 17483780
3	Pyridinoline	PMID: 24073717
4	Cholesterol	PMID: 31517857
5	Zinc	PMID: 29897788
6	Glycocholic acid	PMID: 8630789
7	L-Phenylalanine	PMID: 22191466
8	1-Methyladenosine	PMID: 31648804
9	D-Glucose	PMID: 27760925
10	L-Cysteine	PMID: 30610573

Table 4: Candidate metabolites of leukemia.

Leukemia		
Rank	Metabolite name	Evidences
1	L-Cystine	PMID: 29773592
2	Orotic acid	PMID: 1958838
3	Citrulline	PMID: 19688831
4	Citric acid	PMID: 27465658
5	Betaine	PMID: 520651
6	Homocysteine	PMID: 27874212
7	L-Aspartic acid	PMID: 22356135
8	Creatine	PMID: 28191887
9	Acetic acid	Unconfirmed
10	Gamma-aminobutyric acid	PMID: 27080467

common first symptoms including fever, progressive anemia, significant bleeding tendency, or bone and joint pain, which is prevalent for the adolescent and young adult (AYA) population. We carried out a case study of leukemia disease with our method, and 9 out of the top 10 predicted metabolites interrelated with leukemia are verified to be correlative (see Table 4). Taking the following as instances, orotic acid is verified that its level is higher in milk of cows with leukemia [33]. It is confirmed that combining parthenolide with inhibitors of L-cystine uptake will achieve a greater toxicity to childhood T-cell acute lymphoblastic leukemia [34]. Furthermore, it is shown that several GABAAR (gamma-aminobutyric acid) subunits are significantly increased in ALL (acute lymphoblastic leukemia) children compared with the data of non-ALL children [35].

Obesity is a common disease caused by metabolic disorder. When the human body takes in more calories than required, the rest of calories are stored in the body in the form of fat, which exceeds the normal physiological requirements and becomes obese when it reaches a certain value. The detailed metabolic phenotype of the obese will play a valuable role in understanding the pathophysiology of metabolic disorders. In the obesity-related metabolite prediction results, 9 out of the top 10 predicted metabolites have been verified by published references (see Table 5). For example, L-alanine (Ala) has been reported to regulate pancreatic β-cell physiology and to prevent body fat accumulation in diet-induced obesity [36].

Alzheimer's disease which is a neurodegenerative disorder with insidious onset and slow progression is a growing global health concern with huge implications for individuals and society. It is reported that about 5.4 million Americans have Alzheimer's disease. Today, the number of people living with Alzheimer's disease in the United States is still growing. Someone in the country develops Alzheimer's disease every 66 seconds. The costs of Alzheimer's care may place a substantial financial burden on families. Thus, it is a new therapeutic strategy with the aim of moving from treatment to prevention. Studying disease-related metabolism and observing their concentration changes is also one of the preventive measures. In this study, we select the top 10 latent associations with Alzheimer's disease, and 7 out of the top 10 predicted metabolites have been verified by published references (see Table 6).

Table 5: Candidate metabolites of obesity.

Obesity		
Rank	Metabolite name	Evidences
1	L-Cystine	PMID: 30186675
2	L-Alanine	PMID: 27317126
3	L-Lysine	PMID:22083525
4	Orotic acid	PMID: 7996267
5	Betaine	PMID:29373534
6	Creatine	PMID:28844881
7	L-Cystathionine	PMID:30526049
8	Tetrahydrobiopterin	PMID:26830550
9	3-Chlorotyrosine	Unconfirmed
10	L-Leucine	PMID:27256112

Table 6: Candidate metabolites of Alzheimer's disease.

Alzheimer's disease		
Rank	Metabolite name	Evidences
1	Biotin	PMID: 29150274
2	Cholesterol	PMID: 26944571
3	Taurine	PMID: 31450076
4	Acetic acid	Unconfirmed
5	Phosphate	Unconfirmed
6	Glutamyllysine	Unconfirmed
7	Cobalamin	PMID: 25523421
8	Inosine	PMID: 29363833
9	Trimethylamine *N*-oxide	PMID: 30579367
10	L-Methionine	PMID: 26590557

4. Discussion

In this paper, we propose the improved bipartite network projection based on linear neighborhood similarity for metabolite-disease association prediction (IBNPLNSMDA). We take advantage of the integrated similarities for obtaining the new feature matric to construct linear neighborhood similarity at the beginning of the method. Furthermore, we improve the baseline Algorithm 1 of bipartite network recommendation by adding similarity weights when resources are allocated. Furthermore, LOOCV and several case studies on important human diseases have been implemented. As a result, IBNPLNSMDA performs well both in LOOCV and the case studies.

The excellent performance of IBNPLNSMDA mainly attributes to the following several important factors. Firstly,

different data such as the data of mesh database and DisGeNET are considered to construct integrated disease similarity and integrated metabolite similarity which could make full use of various similarity information to lay a foundation for obtaining new features. Secondly, the application of linear neighborhood similarity (LNS) with WKNKN alleviates the sparsity and incompleteness problems in the current dataset. Last but not least, similarities such as weights are taken into account in the baseline bipartite network recommendation algorithm which has a significant improvement for prediction results.

Despite the efficiency and practicability of the proposed method, it still has some limitations in identifying disease-related metabolites. First of all, more known confirmed human metabolite-disease associations would improve the development and performance of computational human metabolite-disease prediction methods. Furthermore, some reliable metabolite (disease) similarity matrices from other biological features could integrate with relevant linear neighborhood similarity.

Abbreviations

DAG: Directed acyclic graph
GIP: Gaussian interaction profile
LOOCV: Leave-one-out across validation
TPR: True-positive rate
FPR: False-positive rate
ROC: Receiver operating characteristics
AUC: Area under the curve.

Data Availability

The data about metabolite-disease associations used are from the website https://hmdb.ca/.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

CZ carried out the method IBNPLNSMDA to predict the latent associations of metabolites and diseases, participated in designing, and drafted the manuscript. XJL helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors thank the financial support from the National Natural Science Foundation of China (61672334, 61972451, and 61902230) and the Fundamental Research Funds for the Central Universities (no. GK201901010).

References

[1] W. B. Dunn and D. I. J. T. T. I. A. C. Ellis, "Metabolomics: current analytical platforms and methodologies," *TrAC Trends in Analytical Chemistry*, vol. 24, no. 4, pp. 285–294, 2005.

[2] L. Cheng, "MetSigDis: a manually curated resource for the metabolic signatures of diseases," *Briefings in Bioinformatics*, vol. 34, 2017.

[3] C. Gregorio, "Lipid peroxidation, nitric oxide metabolites, and their ratio in a group of subjects with metabolic syndrome," *Oxidative Medicine & Cellular Longevity*, vol. 45, pp. 1–8, 2014.

[4] H. Xianlin, "Metabolomics in early Alzheimer's disease: identification of altered plasma sphingolipidome using shotgun lipidomics," *PLoS One*, vol. 6, no. 7, Article ID e21643, 2011.

[5] B. G. Chenggang Yan, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 56, 2020.

[6] Y.-A. Huang, K. C. C. Chan, and Z.-H. You, "Constructing prediction models from expression profiles for large scale lncRNA-miRNA interaction profiling," *Bioinformatics*, vol. 34, no. 5, pp. 812–819, 2018.

[7] X. Lei and C. J. B. M. Zhang, "Predicting metabolite-disease associations based on KATZ model," *BioData Mining*, vol. 12, no. 1, p. 19, 2019.

[8] C. Fan, X. Lei, and F.-X. Wu, "Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks," *International Journal of Biological Sciences*, vol. 14, no. 14, pp. 1950–1959, 2018.

[9] G. J. I. J. O.B. S. Rui, "PRWHMDA: human microbe-disease association prediction by random walk on the heterogeneous network with PSO," *International Journal of Biological Sciences*, vol. 14, no. 8, pp. 849–857, 2018.

[10] F. Wang, C. J. I. T. O. K. Zhang, and D. Engineering, "Label propagation through linear neighborhoods," *IEEE Transactions on Knowledge & Data Engineering*, vol. 20, no. 1, pp. 55–67, 2007.

[11] W. A. Kibbe, "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic Acids Research*, vol. 43, pp. 1071–1078, 2015.

[12] J. Piñero, "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, no. 3, 2015.

[13] C. Liu, Y. Ma, J. Zhao et al., "Computational network biology: data, models, and applications," *Physics Reports*, vol. 846, pp. 1–66, 2020.

[14] X. Chen, C. C. Yan, and X. J. S. R. Zhang, "WBSMDA within and between score for MiRNA-disease association prediction," *Science Reports*, vol. 6, no. 1, p. 21106, 2016.

[15] T. V. Laarhoven, S. B. Nabuurs, and E. J. B. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *PLoS Computational Biology*, vol. 27, no. 21, pp. 3036–3043, 2011.

[16] C. Liang, S. Yu, and J. J. P. C. B. Luo, "Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs," *PLoS Computational Biology*, vol. 15, no. 4, Article ID e1006931, 2019.

[17] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[18] M. Wan, M. Li, G. Yang, S. Gai, and Z. Jin, "Feature extraction using two-dimensional maximum embedding difference," *Information Sciences*, vol. 274, pp. 55–69, 2014.

[19] M. L. Z. Wan, Z. Lai, G. Yang, Z. Yang, F. Zhang, and H. Zheng, "Local graph embedding based on maximum margin criterion via fuzzy set," *Fuzzy Sets and Systems*, vol. 318, pp. 120–131, 2017.

[20] B. S. Y. Chenggang, H. Zhao, R. Ning, Y. Zhang, and F. Xu, "3D room layout estimation from a single RGB image," *IEEE Transactions on Multimedia*, vol. 45, 2020.
[21] A. Ezzat, P. Zhao, M. Wu, X.-L. Li, and C.-K. Kwoh, "Drug-target interaction prediction with graph regularized matrix factorization," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 3, pp. 646–656, 2017.
[22] S. T. Roweis and L. K. J. S. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[23] W. Zhang, Y. Chen, and D. Li, "Drug-target interaction prediction through label propagation with linear neighborhood information," *Molecules*, vol. 22, no. 12, p. 2056, 2017.
[24] W. Zhang, "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 34, 2019.
[25] G. Li, "Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity," *Journal of Biomedical Informatics*, vol. 82, pp. 169–177, 2009.
[26] T. Zhou, "Bipartite network projection and personal recommendation," *Physical Review E*, vol. 76, no. 4, Article ID 046115, 2007.
[27] D. Sun, A. Li, H. Feng, and M. Wang, "NTSMDA: Prediction of miRNA-disease associations by integrating network topological similarity," *Molecular Biosystems*, vol. 12, no. 7, pp. 2224–2232, 2016.
[28] F. Cheng, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Computational Biology*, vol. 8, no. 5, 2012.
[29] C. L. Zi-Ke Zhang and Z. Yi-Cheng, "Solving the cold-start Problem in recommender Systems with social tags," *EPL (Europhysics Letters)*, vol. 34, 2010.
[30] F. Zhi-An, "Novel link prediction for large-scale miRNA-lncRNA interaction network in a bipartite graph," *BMC Medical Genomics*, vol. 34, 2018.
[31] M. M. Gao, "Dual-network L 2,1 -graph regularized matrix factorization for predicting miRNA-disease associations," *Molecular Biosystems*, vol. 15, 2019.
[32] X. Roblin, J. Pofelski, and J.-P. Zarski, "Rôle de l'homocystéine au cours de la stéatose hépatique et de l'hépatite chronique C," *Gastroentérologie Clinique et Biologique*, vol. 31, no. 4, pp. 415–420, 2007.
[33] T. Motyl, J Krzemiński, M Podgurniak, C Witeszczak, and P Zochowski, "Variability of orotic acid concentration in cow's milk," *Endocrine Regulations*, vol. 25, no. 25, pp. 79–82, 1991.
[34] B. C. Ede, R. R. Asmaro, J. P. Moppett, P. Diamanti, and A. Blair, "Investigating chemoresistance to improve sensitivity of childhood T-cell acute lymphoblastic leukemia to parthenolide," *Haematologica*, vol. 103, no. 9, pp. 1493–1501, 2018.
[35] H. Wang, M. Feng, Y. Liu et al., "Up-regulation of GABAergic signal events in bone marrow lymphocytes in childhood acute lymphoblastic leukemia," *The Chinese Journal of Physiology*, vol. 59, no. 59, pp. 119–125, 2016.
[36] T. R. Araujo, I. N. Freitas, J. F. Vettorazzi et al., "Benefits of L-alanine or L-arginine supplementation against adiposity and glucose intolerance in monosodium glutamate-induced obesity," *European Journal of Nutrition*, vol. 56, no. 6, pp. 2069–2080, 2017.