IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 15, 2018, accepted June 18, 2018, date of publication June 29, 2018, date of current version July 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2851751

The Bi-Direction Similarity Integration Method for Predicting Microbe-Disease Associations

WEN ZHANG, WEITAI YANG, XIAOTING LU, FENG HUANG, AND FEI LUO
School of Computer Science, Wuhan University, Wuhan 430072, China

Corresponding author: Wen Zhang (zhangwen@whu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772381 and Grant 61572368 and in part by the Fundamental Research Funds for the Central Universities under Grant 2042017kf0219.

ABSTRACT Identification of microbe-disease associations provides insight into the mechanism that microbes cause diseases at the molecular level. Existing microbe–disease association prediction methods mainly utilize microbe–disease association profiles to calculate microbe–microbe similarities and disease–disease similarities, and then build similarity-based prediction models. However, they ignore important biological knowledge, e.g., disease Medical Subject Headings (MeSH), and do not consider unequal contributions of microbe information and disease information. In this paper, we propose the bi-direction similarity integration label propagation (BDSILP) method for predicting microbe–disease associations. First, BDSILP introduces disease MeSH to calculate the disease–disease semantic similarity and the microbe–microbe functional similarity. Although MeSH is not available for all diseases, BDSILP presents a strategy for integrating multiple similarities for microbes and diseases. Second, two graphs are constructed by using integrated disease similarity and integrated microbe similarity, and BDSILP implements the label propagation on the graphs to score microbe–disease pairs. Third, BDSILP adopts the weighted averages of their scores as final predictions. BDSILP produces better performances than existing state-of-the-art methods, achieving the AUC of 0.9131 and the AUPR of 0.5343 in leave-one-out cross validation, and achieving the AUC of 0.9051 and the AUPR of 0.3037 in five-fold cross validation. Moreover, case studies and discussion demonstrate that BDSILP is promising for predicting novel microbe–disease associations.

INDEX TERMS Microbe-disease association, MeSH, label propagation.

I. INTRODUCTION

A microbe is a microscopic organism, which may exist in its single-celled form, or in a colony of cells. The human microbiota is the aggregate of microorganisms that consists of bacteria, viruses, eukaryotes and archaea [1], and microbes reside in and on different body niches such as oral cavity, throat, esophagus, stomach, colon, urogenital tract, respiratory tract and skin [2]. There is a great number of work and tools [3]–[6] about the dynamic behaviors of microbes. These studies show that the human ecosystem has more than 10000 microbial species, which produce nearly 8 million proteins [7], and thus control metabolic functions, such as obesity control, brain development, resistance to pathogens, immune response against infections and injuries. Therefore, microbes can greatly influence human health, and variances of microbiome may disturb the microbiota-human symbiotic relationship and cause diseases. For example, obesity is associated with phylum-level changes in the microbiota, reduced bacterial diversity and altered representation of bacterial genes [8]. National Institutes of Health (NIH) launched Human Microbiome Project (HMP) in 2008, which facilitates characterization of the human microbiota and understanding of how microbes cause diseases and influence human health. Identifying microbe-disease associations can find the disease-causing microbes and help the diagnosis and therapy of diseases. The culture-based wet methods for identifying microbe-disease associations are time-consuming and costly. In contrast, computational methods can accelerate microbe-disease association predictions and reduce costs.

With the development of artificial intelligence and machine learning technology [9]–[11], computational methods are widely applied in the field of bioinformatics [12]–[20]. To the best of our knowledge, several computational methods have been proposed to predict microbe-disease associations.

Most methods utilized microbe information or disease information to calculate microbe-microbe similarities or disease-disease similarities, and then construct similarity-based network to predict microbe-disease associations. Shen *et al.* utilized a symptom-based disease network, a Spearman correlation-based microbe network and a known microbe-disease network to construct a heterogeneous network, and used a random walk with restart algorithm on the heterogeneous network to predict microbes for a specific disease [21]. Zou *et al.* constructed a heterogeneous network from the microbe-disease association network and Gaussian interaction profile kernel similarity networks for microbes and diseases, and developed a bi-random walk method on the heterogeneous network [22]. Chen *et al.* constructed a heterogeneous network from the known microbe-disease associations and Gaussian interaction profile kernel similarities for microbes and diseases, and developed a novel KATZ measure with variable-length walks [23] to predict novel microbe-disease associations. Huang *et al.* put forward a path-based human disease-microbe association prediction model PBHMDA [24], which scores a candidate microbe-disease pair by traversing all possible paths between the microbe and disease in a heterogeneous network based on the known microbe-disease associations and Gaussian interaction profile kernel similarities for diseases and microbes. Huang *et al.* developed NGRHMDA [25], which combines collaborative filtering and a graph-based scoring method based on Gaussian kernel-based microbe similarity and symptom-based disease similarity. Besides, Wang *et al.* presented a semi-supervised learning method based on Gaussian interaction profile kernel similarity and Laplacian regularized least squares classifier LRLSHMDA for human microbe-disease association prediction [26]; Shen *et al.* proposed the computational model of collaborative matrix factorization method CMFHMDA [27] based on known microbe-disease associations.

Although many computational methods have been proposed, we can address several issues to improve the performances of prediction models. Existing methods utilize known microbe-disease associations to calculate Gaussian interaction profile kernel similarities for microbes and diseases, but ignore the biological knowledge about microbes and diseases, e.g. Medical Subject Headings (MeSH) information. Many previous experiments [28]–[32] show that multiple information is more helpful to make a prediction than individual information. Moreover, existing methods use microbe information or disease information to build prediction models, but do not take into their unequal contributions to the microbe-disease association prediction.

In this paper, we propose the bi-direction similarity integration label propagation method "BDSILP" for microbe-disease association prediction. In addition to the Gaussian interaction profile kernel similarities, BDSILP introduces disease Medical Subject Headings(MeSH) to calculate the disease-disease semantic similarity and the microbe-microbe functional similarity. Although MeSH is not available for all diseases, BDSILP presents a strategy of integrating multiple similarities for microbes and diseases [33]. Then, two graphs are constructed by using integrated disease similarity and integrated microbe-similarity, and BDSILP implements the label propagation [34]–[36] on the graphs to score microbe-disease pairs. At last, BDSILP adopts the weighted averages of their scores as final predictions. BDSILP produces better performances than existing state-of-the-art methods, achieving the AUC of 0.9131 and AUPR of 0.5343 in leave-one-out cross validation, and achieving the AUC of 0.9051 and AUPR of 0.3037 in 5-fold cross validation. Moreover, case studies and discussion demonstrate that BDSILP is promising for predicting microbe-disease associations.

II. MATERIALS AND METHODS

A. DATASETS

Recently, researchers collected microbe-disease associations data, and constructed datasets to facilitate related studies. The Human Microbe-Disease Association Database (HMDAD) [37] is a resource, which collected and curated the human microbe-disease associations. Currently, HMDAD includes 483 experimentally confirmed human microbe-disease associations between 292 microbes and 39 diseases, which were curated from 61 publications. We downloaded the data from HMDAD, and then removed redundant records. Thus, we obtained a dataset, which includes 292 microbes, 39 diseases and 450 microbe-disease associations.

Besides, we collected Medical Subject Headings (MeSH) descriptors of diseases from U.S. National Library of medicine. MeSH descriptors are a comprehensive controlled vocabulary, and these descriptors or subject headings are arranged in a hierarchy. However, the MeSH is only available for 28 out of 39 diseases.

B. OVERVIEW OF THE BI-DIRECTION SIMILAIRTY INTEGRATED METHOD

Given r microbes $m_1, m_2, \cdots, m_r$ and s diseases $d_1, d_2, \cdots, d_s$, the associations between microbes and diseases are denoted by a $r \times s$ adjacency matrix A. $A(i,j) = 1$, if there is an association between microbe m_i and disease d_j; otherwise, $A(i,j) = 0$.

Fig.1 illustrates the flowchart of the bi-direction similarity integration label propagation method "BDSILP", which predicts microbe-disease associations. First, we calculate the semantic similarity and association profile similarity for diseases as well as the functional similarity and association profile similarity for microbes. Second, we integrate multiple disease-disease similarities and multiple microbe-microbe similarities respectively [38], and obtain the integrated similarity for diseases and integrated similarity for microbes. Third, we respectively construct the

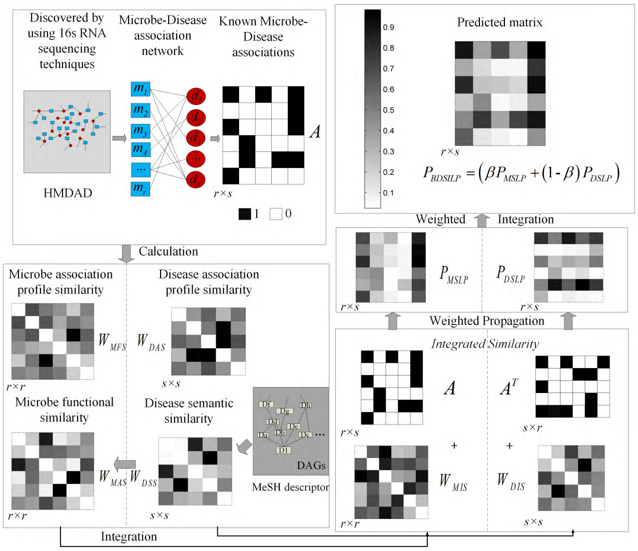

Discovered by using 16s RNA sequencing techniques
Microbe-Disease association network
Known Microbe-Disease associations
HMDAD
m_1 m_2 m_3 m_4 ... m_r
d_1 d_2 d_3 d_4 d_s
A
r×s
■1 □0
Predicted matrix
0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
r×s
$P_{BDSILP}=(\beta P_{MSLP}+(1-\beta)P_{DSLP})$
Weighted Integration
P_{MSLP} P_{DSLP}
r×s r×s
Calculation
Weighted Propagation
Microbe association profile similarity
Disease association profile similarity
W_{MFS} W_{DAS}
r×r s×s
Integrated Similarity
A A^T
r×s s×r
+ +
W_{MIS} W_{DIS}
r×r s×s
Microbe functional similarity
Disease semantic similarity
W_{MAS} W_{DSS}
r×r s×s
DAGs
MeSH descriptor
Integration

FIGURE 1. Flowchart of the bi-direction similarity integration label propagation method (BDSILP).

disease integrated similarity-based graph and microbe integrated similarity-based graph. Finally, we implement the label propagation process on two graphs to score microbe-disease pairs and adopt the weighted averages as final predictions.

C. DISEASE SIMILARITIES

In this section, we define two similarities: the semantic similarity and the association profile similarity for diseases, and then integrate them to obtain the integrated similarity of diseases.

1) DISEASE SEMANTIC SIMILARITY

According to MeSH descriptors, which are the representation of the objects in MeSH database, e.g. "Asthma" is described as "C08.127.108; C08.381.495.108; C08.674.095; C20.543.480.680.095". A disease d can be represented as a directed acyclic graph $DAG_d = (V_d, E_d)$, where V_d contains the nodes of this disease d itself and its ancestor diseases, and E_d consists of all the directed edges from parent nodes to child nodes. Fig.2 is an example of the DAG of the disease "Asthma". The semantic contribution of disease t in V_d to d is calculated by:

$$SC_d(t) = \begin{cases} 1 \quad if\ t = d \\ max\left\{ \Delta \times SC_d(t') \middle| t' \in children\ of\ t \right\} \\ \quad if\ t \neq d \end{cases} \tag{1}$$

where Δ is the semantic contribution factor, and we set $\Delta = 0.5$ according to previous work [33].

The semantic value of disease d is calculated by summing up the weighted contributions of parent nodes to disease d and its contribution to itself as follows:

$$SV_d = \sum\nolimits_{t \in V_d} SC_d(t) \tag{2}$$

The semantic similarity between disease d_i and disease d_j is calculated by:

$$W_{DSS}(d_i, d_j) = \frac{\sum_{t \in V_{d_i} \cap V_{d_j}} \left(SC_{d_i}(t) + SC_{d_j}(t)\right)}{SV_{d_i} + SV_{d_j}} \tag{3}$$

where t is a common ancestor disease of d_i and d_j. $SC_{d_i}(t)$ is the semantic contribution of t to disease d_i, and SV_{d_i} is the semantic value of d_i; $SC_{d_j}(t)$ is the semantic contribution of t to disease d_j, and SV_{d_j} is the semantic value of d_j.

2) DISEASE ASSOCIATION PROFILE SIMILARITY

The association profile of disease d_i is a binary vector, which represents the presence or absence of observed associations between the disease and each microbe. The association profile of disease d_i is actually the ith column of the microbe-disease association matrix A, i.e. $A(:,i)$. Then, the similarity between disease d_i and d_j is calculated by using the Gaussian kernel function:

$$W_{DAS}(d_i, d_j) = exp\left(-\gamma_d \|A(:,i) - A(:,j)\|^2\right) \tag{4}$$

where γ_d is responsible for controlling the kernel bandwidth, and $\gamma_d = \gamma / \left(\frac{1}{s}\sum_{i=1}^{s} \|A(:,i)\|^2\right)$. s is the total number of diseases, and γ is set to 1.

3) INTEGRATED SIMILARITY

We integrate the semantic similarity and the association similarity for diseases, and the integrated similarity between disease d_i and disease d_j is calculated by:

$$W_{DIS}(d_i, d_j) = \begin{cases} \frac{W_{DSS}(d_i,d_j) + W_{DAS}(d_i,d_j)}{2} \\ d_i\ and\ d_j\ have\ MeSH\ descriptors \\ W_{DAS}(d_i, d_j) \quad otherwise \end{cases} \tag{5}$$

Since not all diseases have MeSH descriptors, we cannot obtain semantic similarity for any two diseases. If d_i and d_j have MeSH descriptors, the integrated similarity is the average of the semantic similarity and the association profile similarity; otherwise, the integrated similarity is the association profile similarity. Then, the similarity matrix W_{DIS} for s diseases can be normalized as [39].

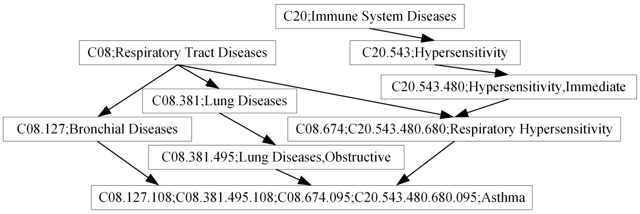

C20;Immune System Diseases
C08;Respiratory Tract Diseases
C20.543;Hypersensitivity
C20.543.480;Hypersensitivity,Immediate
C08.381;Lung Diseases
C08.127;Bronchial Diseases
C08.674;C20.543.480.680;Respiratory Hypersensitivity
C08.381.495;Lung Diseases,Obstructive
C08.127.108;C08.381.495.108;C08.674.095;C20.543.480.680.095;Asthma

FIGURE 2. DAG of the disease "Asthma".

D. MICROBE SIMILARITIES

In this section, we introduce the functional similarity [33] and the association profile similarity for microbes, and then integrate two similarities.

1) MICROBE FUNCTIONAL SIMILARITY

First, the similarity between a disease d' and a set of diseases D is defined as:

$$SIM\left(d', D\right) = \max_{d \in D}\left(W_{DSS}(d', d)\right) \tag{6}$$

where $W_{DSS}(d, d_i)$ is the semantic similarity between disease d and disease d_i. Then, the functional similarity between microbes m_i and m_j is calculated by:

$$W_{MFS}\left(m_i, m_j\right) = \frac{\sum_{d \in D_j} SIM\left(d, D_i\right) + \sum_{d \in D_i} SIM\left(d, D_j\right)}{|D_i| + \left|D_j\right|} \tag{7}$$

where D_i is a set of diseases which are associated with the microbe m_i; D_j is a set of diseases which are associated with the microbe m_j.

2) MICROBE ASSOCIATION PROFILE SIMILARITY

Similar to disease association profile, the association profile of microbe m_i is a binary vector, which represents the presence or absence of observed associations between the microbe and each disease. The association profile of microbe m_i is actually the ith row of the microbe-disease association matrix A, i.e. $A(i, :)$. Then, the similarity between microbe m_i and microbe m_j is calculated by using the Gaussian kernel function:

$$W_{MAS}\left(m_i, m_j\right) = exp\left(-\gamma_m \left\|A\left(i, :\right) - A\left(j, :\right)\right\|^2\right) \tag{8}$$

where γ_m is responsible for controlling the kernel bandwidth, and $\gamma_m = \gamma / \left(\frac{1}{r}\sum_{i=1}^{r}\left\|A\left(i, :\right)\right\|^2\right)$. r is the total number of microbes, and γ is set to 1.

3) INTEGRATED SIMILARITY

We integrate the functional similarity and the association similarity for microbes, and the integrated similarity between microbes m_i and m_j is calculated by:

$$W_{MIS}\left(m_i, m_j\right) = \begin{cases} \frac{W_{MFS}\left(m_i, m_j\right) + W_{MAS}\left(m_i, m_j\right)}{2} \\ all\ related\ diseases\ have\ descriptors \\ W_{MAS}\left(m_i, m_j\right) \quad otherwise \end{cases} \tag{9}$$

Calculation of the functional similarity relies on the semantic similarity. If all diseases related with microbes m_i and m_j have MeSH descriptors, the integrated similarity is average of the functional similarity and the association profile similarity; otherwise, the integrated similarity is the association profile similarity. Then, the integrated similarity matrix is normalized as [39].

E. BI-DIRECTION SIMILARITY INTEGRATED METHOD

In this study, we propose a novel computational method "BDSILP" to predict human microbe-disease associations by using label propagation [40].

We construct an undirected graph based on the microbe-microbe integrated similarity matrix W_{MIS}, in which r microbes are regarded as nodes and the similarity between microbes m_i and m_j is recognized as the weight of edges. For the disease d_j, the initial labels of nodes are the jth column of microbe-disease association matrix A, i.e. $A\left(:, j\right)$. These labels information is propagated from one node to the nodes adjacent to it. Then the labels of nodes are updated by labels of their neighbor nodes with probability α and retaining the initial labels with probability $1 - \alpha$. Let P_j^0 represent the initial labels of nodes for the jth disease, and the labels of the kth iteration are denoted as p_j^k, the update from step k to step $k + 1$ is,

$$P_j^{k+1} = \alpha W_{MIS} P_j^k + (1 - \alpha) A\left(:, j\right) \tag{10}$$

Taking labels for all diseases $d_1, d_2, \cdots, d_s$ into consideration, we can refine the formulas (10):

$$P^{k+1} = \alpha W_{MIS} P^k + (1 - \alpha) A \tag{11}$$

Eq. (11) can be written as,

$$\begin{aligned} P^{k+1} &= \alpha W_{MIS} P^k + (1 - \alpha) A \\ &= (\alpha W_{MIS})^2 P^{k-1} + (1 - \alpha)\left(I + \alpha W_{MIS}\right) A \\ &= \cdots = (\alpha W_{MIS})^{k+1} A + (1 - \alpha) \sum_{i=0}^{k} (\alpha W_{MIS})^i A \end{aligned} \tag{12}$$

Since the spectral radius $\rho\left(W_{MIS}\right) \leq 1$ and $0 < \alpha < 1$, then $\lim_{k \to \infty} (\alpha W_{MIS})^k = 0$, $\lim_{k \to \infty} \sum_{i=0}^{k} (\alpha W_{MIS})^i = (I - \alpha W_{MIS})^{-1}$. The iteration will converge,

$$P = \lim_{k \to \infty} P^{k+1} = (1 - \alpha)\left(I - \alpha W_{MIS}\right)^{-1} A \tag{13}$$

Thus, we can develop microbe similarity-based label propagation method (MSLP), and the predicted association matrix is:

$$P_{MSLP} = (1 - \alpha)\left(I - \alpha W_{MIS}\right)^{-1} A \tag{14}$$

Similarly, we construct an undirected graph based on the disease-disease integrated similarity matrix W_{DIS}. Then, we can develop disease similarity-based label propagation method (DSLP), and the predicted association matrix is:

$$P_{DSLP} = \left((1 - \alpha)\left(I - \alpha W_{DIS}\right)^{-1} A^T\right)^T \tag{15}$$

where A^T is the transpose of the microbe-disease association matrix A.

By using MSLP and DSLP as two components, we develop the bi-direction similarity integration method (BDSILP), and the predicted association matrix is:

$$P_{BDSILP} = \beta P_{MSLP} + (1 - \beta) P_{DSLP} \tag{16}$$

where β is decay factor, which controls the weight of P_{MSLP} and P_{DSLP}.

TABLE 1. **LOOCV performances of BDSILP models using integrated similarities and association profile similarities.**

Metrics Similarity	AUPR	AUC	F1	ACC	REC	SPEC	PRE
Association profile similarities	0.5138	0.9105	0.4880	0.9644	0.4289	0.9865	0.5660
Integrated Similarities	0.5343	0.9131	0.5160	0.9628	0.5022	0.9817	0.5305

III. RESULT AND DISCUSSION

A. PERFORMANCE AND EVALUATION

We adopt leave-one-out cross validation (LOOCV) and 5-fold cross validation (5-fold CV) to evaluate the performance of prediction models. In LOOCV, each microbe-disease pair is left out in turn as the testing sample, and other microbe-disease pairs are used as the training set. In each fold, we construct prediction models based on the training set, and then score the testing sample. We repeat the training process and testing process until we have prediction scores for all pairs. Finally, we take prediction scores and real labels (associations or non-associations) for all microbe-disease pairs to calculate evaluation metrics. In each fold, we recalculate similarities by using associations in the training set. In addition to LOOCV, 5-fold CV randomly splits known microbe-disease associations into five subsets. In each fold, one subset is used as the testing set, and others are used as the training set in turns.

We adopt several evaluation metrics to evaluate performances of prediction models, i.e. the area under receiver-operating characteristic curve (AUC), the area under precise-recall curve (AUPR), sensitivity (SEN), specificity (SPEC), precision (PRE), accuracy (ACC) and F-measure (F). The area under receiver-operating characteristic curve (AUC) is evaluating the prediction performance of a model by considering the true positive rate and the false positive rate over different thresholds. The area under precise-recall curve (AUPR) takes into account the recall and precision over different thresholds. Sensitivity (SEN), specificity (SPEC), precision (PRE), accuracy (ACC) and F-measure (F) are also popular metrics. These experiments are conducted in python under 64-bit Windows system.

B. PERFORMANCES OF BDSILP

BDSILP has two parameters: the propagation probability α and the weighting factor β. α is the probability that labels of nodes are updated by neighbor nodes' labels in the label propagation. α influences the process of label propagation, and thus has impact on prediction performance. β controls the contributions from the microbe-based component MSLP and the disease-based component DSLP. Here, we consider parameters $\alpha \in \{0.05, 0.1, \cdots, 0.95\}$ and $\beta \in \{0.05, 0.1, \cdots, 0.95\}$, and then build BDSILP models to test the influence of parameters. First, we tentatively set $\beta = 0.5$, and build BDSILP models based on different α values. Fig.3 (a) shows the influence of α on AUC scores of BDSILP models in LOOCV. We observe that BDSILP produces the best AUC score of 0.9019 when $\alpha = 0.25$, indicating that it is likely to retain the known label information with high probability. Then, we fix $\alpha = 0.25$, and demonstrate the LOOCV AUC scores of BDSILP models using different β values in Fig.3 (b). BDSILP can achieve the best AUC score of 0.9131 when $\beta = 0.7$. The results demonstrate that the microbe-based component P_{MSLP} has the greater weight than the disease-based component P_{DSLP}, indicating that they make unequal contributions to BDSILP. For comparison, we evaluate the performances of two components P_{MSLP} and P_{DSLP}. P_{MSLP} produces the LOOCV AUC score of 0.8599; P_{DSLP} can produce the LOOCV AUC score of 0.3472. That is the reason why the component P_{MSLP} has the greater weight β in BDSILP. Based on above discussion, we fix $\alpha = 0.25$ and $\beta = 0.7$ for BDSILP in the following studies.

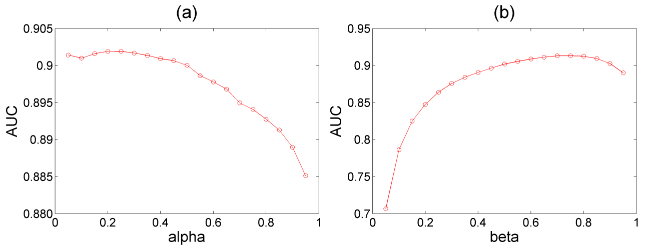


FIGURE 3. **AUC scores of BDSILP models using different parameter values evaluated by LOOCV.**

BDSILP utilizes the integrated similarities for microbes and diseases. The microbe integrated similarity combines the microbe association profile similarity and microbe functional similarity; the disease integrated similarity combines the disease association profile similarity and disease semantic similarity. The microbe functional similarity and disease semantic similarity rely on the MeSH descriptors for diseases. MeSH descriptors provide the category information of diseases, and thus lead to the good performances of BDSILP. Since MeSH descriptors are not available for all diseases, they are supplementary information to the association profile similarity in the integrated similarity. For comparison, we only use the association profile similarity for microbes and diseases to build BDSILP models. As shown in Table 1, BDSILP models using integrated similarities significantly improve the performances of BDSILP models only using association profile similarities in LOOCV, revealing the usefulness of MeSH information.

C. COMPARISON WITH OTHER STATE-OF-THE-ART METHODS

In this section, we consider several state-of-the-art microbe-disease associations prediction methods and make

TABLE 2. Performances of different methods evaluated by LOOCV.

Methods \ Metrics	AUPR	AUC	F1	ACC	REC	SPEC	PRE
BDSILP	0.5343	0.9131	0.5160	0.9628	0.5022	0.9817	0.5305
KATZHMDA	0.3321	0.8380	0.4865	0.9616	0.4600	0.9823	0.5162
BiRWHMDA	0.4304	0.8790	0.4521	0.9608	0.4089	0.9835	0.5055
NGRHMDA	0.3102	0.8337	0.4875	0.9622	0.4556	0.9830	0.5243
LRLSHMDA	0.4877	0.8936	0.4543	0.9534	0.4911	0.9724	0.4226

comparisons to demonstrate superior performances of our proposed method BDSILP. KATZHMDA [23] predicts microbe-disease associations by measuring Katz distances in a heterogeneous network. BiRWHMDA [22] predicts microbe-disease associations by capturing circular bigraph patterns on a global heterogeneous network. NGRHMDA [25] combines two single recommendation system-based models to make predictions. LRLSHMDA [26] prioritizes candidate microbe-disease association pairs by optimizing a cost function. These representative methods have good results in experiments. Therefore, we adopt KATZHMDA, BiRWHMDA, NGRHMDA and LRLSHMDA as benchmark methods for comparison. We replicate the benchmark methods according to publications and evaluate all models on the benchmark dataset by using LOOCV and 5-fold CV. As shown in Table 2, BDSILP produces best performances in LOOCV, achieving the AUC score of 0.9131 and the AUPR score of 0.5343, while KATZHMDA, BiRWHMDA, NGRHMDA and LRLSHMDA yield AUC scores of 0.8380, 0.8790, 0.8337 and 0.8936, AUPR scores of 0.3321, 0.4304, 0.3102 and 0.4877. The results in Table 3 show that BDSILP also produces best performances in 5-fold CV. Clearly, BDSILP outperforms benchmark methods in terms of different evaluation metrics.

The main aim of computational methods is screening microbe-disease associations, and then guiding the wet experimental determination of real associations. A prediction method by yielding a score for each microbe-disease pair, which represents the probability of having an association. For a perfect model, real associations should have high rank in prediction scores of all microbe-disease pairs. We check up on top predictions of prediction methods, and count the number of real associations that prediction methods can discover.

Further, we make analysis based on the LOOCV results. We consider a wide range of top predictions from top 100 to top 10000 in a step size of 100, and compare the capability of different methods for discovering real associations in top predictions. We use numbers of top predictions as X-axis and numbers of discovered real associations as Y-axis, and visualize the results in Fig.4. Clearly, our method can find out more associations than benchmark methods in top predictions and has the great potential of detecting microbe-disease associations.

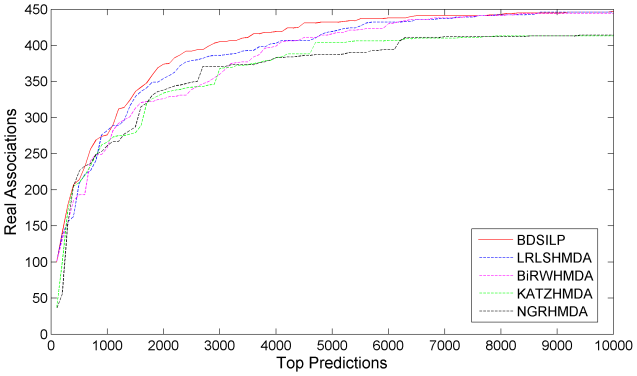


FIGURE 4. Top predictions and real associations for different methods based on LOOCV.

Further, we study the performances of prediction methods for predicting microbes associated with a specific disease and predicting diseases associated with a specific microbe. For this purpose, we adopt two different evaluation ways: LOOCV$_{\mathrm{D}}$ and LOOCV$_{\mathrm{M}}$ to evaluate the LOOCV results. For a specific disease, LOOCV$_{\mathrm{D}}$ uses the prediction scores for every microbe and the disease to calculate metric scores. For a specific microbe, LOOCV$_{\mathrm{M}}$ uses the prediction scores for every disease and the microbe to calculate metric scores. Since our dataset has 292 microbes and 39 diseases, we calculate AUC scores for every disease by using LOOCV$_{\mathrm{D}}$ and calculate AUC scores for every microbe by using LOOCV$_{\mathrm{M}}$. We conduct the statistical analysis on the results of different methods for microbes and diseases, and draw the boxplots of AUC scores for every microbe and every disease in Fig.5. The most important indicators in the boxplot are the median position and the interval between maximum and minimum values. For AUC scores of diseases, BDSILP and LRLSHMDA have larger median and smaller interval, indicating that the two approaches have better prediction performances than other methods. By contrast, in terms of AUC scores of microbes, BDSILP, BiRWHMDA and KATZHMDA have smaller interval and achieve better performances than LRLSHMDA and NGRHMDA. Clearly, our method can produce satisfying results for predicting microbe-associated diseases and predicting

TABLE 3. Performances of different methods evaluated by 5-fold CV.

Metrics Methods	AUPR	AUC	F1	ACC	REC	SPEC	PRE
BDSILP	0.3037	0.9051	0.3882	0.9921	0.3089	0.9977	0.5232
KATZHMDA	0.1128	0.8363	0.3248	0.9873	0.3067	0.9929	0.2102
BiRWHMDA	0.2669	0.8935	0.3656	0.9913	0.3044	0.9970	0.4721
NGRHMDA	0.1023	0.8284	0.3027	0.9862	0.2689	0.9921	0.1376
LRLSHMDA	0.2226	0.8821	0.3785	0.9922	0.2911	0.9980	0.5427

TABLE 4. Top 10 microbes associated with type 1 diabetes.

Rank	Microbe	Evidence
1	*Bacilli*	[41]
2	*Betaproteobacteria*	N.A.
3	*Verrucomicrobiaceae*	N.A.
4	*Desulfovibrio*	[42]
5	*Prevotella copri*	N.A.
6	*Corynebacterium*	[43]
7	*Acinetobacter*	[44]
8	*Faecalibacterium prausnitzii*	[45]
9	*Clostridium*	[46]
10	*Tropheryma whipplei*	N.A.

N.A.-not available

disease-associated microbes and outperform other methods. Moreover, comparing to diseases, microbes have more extreme outliers in boxplots, but AUC scores for most microbes are densely concentrated and distributed in higher intervals. It demonstrates that microbe-based prediction can produce better performances than disease-based prediction.

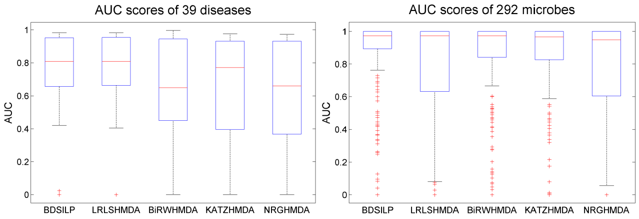


FIGURE 5. Boxplots of AUC scores for diseases and microbes.

D. CASE STUDIES

Microbes are closely related with human health, and microbe-disease associations are indicators how microbes cause diseases. Therefore, exploring disease-caused microbes is meaningful and quite urgent. In order to investigate into disease-causing microbes (pathogens), we take two diseases of wide interests: type 1 diabetes and bacterial vaginosis as examples. We construct the BDSILP model by using all microbe-disease associations in the benchmark dataset, and predict microbe-disease associations, which are not included in HMDAD. We list the top 10 microbes associated with type 1 diabetes in Table 4 and list the top 10 microbes associated with bacterial vaginosis in Table 5.

Type 1 diabetes is a form of diabetes mellitus that leads to high blood sugar levels. Although the causes of type 1 diabetes are still unclear, the disease is no doubt related to factors such as genes, microbes and the environment. The disease usually begins in children and young adults, and about 80,000 children develop the disease each year. Table 4 shows the top 10 predicted microbes associated with type 1 diabetes, and we can find evidences from public resources to confirm six type 1 diabetes-related microbes. Bacilli is a genus of gram-positive, rod-shaped bacteria and a member of the phylum Firmicutes. Bacilli was proved as one of the causes of the diabetes, and the abundance in children with type 1 diabetes was 8.5% [41]. Desulfovibrio is a genus of Gram-negative sulfate-reducing bacteria. The relative abundance of Desulfovibrio affects the glucose concentration in the human body, and then controls the incidence of type 1 diabetes [42]. Corynebacterium is a genus of bacteria that are Gram-positive and aerobic. As reported in [43], mice treated with Corynebacterium avoided the development of diabetes. Acinetobacter is a genus of Gram-negative bacteria belonging to the wider class of Gammaproteobacteria, and Acinetobacter levels significantly increased in patients with type 1 diabetes [44]. Faecalibacterium prausnitzii is one of the most

TABLE 5. Top 10 microbes associated with bacterial vaginosis.

Rank	Microbe	Evidence
1	*Prevotella copri*	[47]
2	*Desulfovibrio*	N.A.
3	*Corynebacterium*	[48]
4	*Acinetobacter*	[49]
5	*Tropheryma whipplei*	N.A.
6	*Verrucomicrobiaceae*	N.A.
7	*Bacteroidales*	N.A.
8	*Erysipelotrichales*	N.A.
9	*Enterococcus faecium*	www.allthingsvagina.com/enterococcus-faecalis/
10	*Clostridium leptum*	N.A.

N.A.-not available

abundant and important commensal bacteria of the human gut microbiota. In patients with type 1 diabetes, the level of Faecalibacterium prausnitzii may be increased within control [45]. Clostridium is a genus of Gram-positive bacteria which includes several significant human pathogens, a significant increase of Clostridium, may result in a disturbance in the ecological balance and then cause type 1 diabetes [46].

Bacterial vaginosis is a disease of the vagina caused by excessive growth of bacteria or imbalance of the naturally occurring bacteria in the vagina. BV is the most common vaginal infection in women of reproductive age, and the percentage of women affected at any given time varies between 5% and 70%. Meanwhile, the high rate of recurrence despite appropriate treatment hint at the complex nature of this condition. New insights about BV and BV-associated bacterial communities will widely flow from researches at the intersection of molecular microbiology, conventional microbiology, genomics, immunity, and the ecological determinants of the vaginal bacterial population. Table 5 shows the top 10 predicted microbes associated with BV, and we can find evidences from public resources to confirm four BV-related microbes. As reported in [47], Prevotella corporis is one of the prominent bacteria in the normal vaginal ecosystem. Pyrosequencing technology has found that the Prevotella group is the main member of BV bacterial community. Investigation in [48] revealed that an isolated unique coryneform bacterium from infection site of patients with BV is belong to Corynebacterium but represented as a new species, and their interactions are unclear. There is no single bacteria considered as the only special markers for diseases. Acinetobacter and Actinomycetes have been confirmed to have strong correlation with BV as main plant bacterial species by pyrosequencing of barcoded 16S rRNA genes technology from vaginal bacterial communities of 396 asymptomatic North American women, and its control has good effect on BV [49]. Enterococcus faecalis can be widely found in the vagina tract, being a cause of BV (linked to aerobic vaginitis) and with the increase of community number in human body, the development of the BV tends to be worsen (www.allthingsvagina.com/enterococcus-faecalis/).

IV. CONCLUSION

There is the mutualism relationship between human microorganisms and the human body. Microbes play a critical role in the metabolism activities of the human body and are closely related with human diseases. The identification of microbe-disease associations can reveal mechanism of microbe influencing diseases at the molecular level, and cure diseases. In this paper, we propose the bi-direction similarity integration label propagation method ''BDSILP'' to predict microbe-disease associations. BDSILP make uses of diverse information and also take unequal contributions of microbe information and disease information into account. The experiments demonstrate that BDSILP has the good performances for microbe-disease association prediction.

However, BDSILP still has several limitations, because of the data dilemma. On the one hand, only hundreds of microbe-disease associations are known or available; on the other hand, only known microbe-disease associations and semantic information can be used as features for modeling. Usually, researchers have to use biological features as additional information to make predictions when we do not know any disease information of a microbe. However, semantic information is only available for a portion of interested diseases. Therefore, our method focuses on the task that predicts unobserved or potential associations between microbes and diseases in the case that some associations has been observed, and can't be applied to a microbe without any disease information or a disease without any microbe information. In future, we will try to make de novo prediction for microbe-disease associations when more data is available.

REFERENCES

[1] B. A. Methé *et al.*, ''A framework for human microbiome research,'' *Nature*, vol. 486, pp. 215–221, Jun. 2012.

[2] M. C. Cénit, V. Matzaraki, E. F. Tigchelaar, and A. Zhernakova, ''Rapidly expanding knowledge on the role of the gut microbiome in health and disease,'' *Biochim. Biophys. Acta (BBA)-Mol. Basis Disease*, vol. 1842, no. 10, pp. 1981–1992, Oct. 2014.

[3] R. R. Stein *et al.*, ''Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota,'' *PLoS Comput. Biol.*, vol. 9, p. e1003388, Dec. 2013.

[4] B. Sánchez, S. Delgado, A. Blanco-Míguez, A. Lourenço, M. Gueimonde, and A. Margolles, "Probiotics, gut microbiota, and their influence on host health and disease," *Mol. Nutrition Food Res.*, vol. 61, p. 1600240, Jan. 2016.

[5] M. Imani and U. M. Braga-Neto, "Point-based methodology to monitor and control gene regulatory networks via noisy measurements," *IEEE Trans. Control Syst. Technol.*, to be published.

[6] M. Imani and U. M. Braga-Neto, "Control of gene regulatory networks with noisy measurements and uncertain inputs," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 2, pp. 760–769, Jun. 2018.

[7] L. Rup, "The human microbiome project," *Indian J. Microbiol.*, vol. 52, p. 315, Sep. 2012.

[8] P. J. Turnbaugh *et al.*, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, pp. 480–484, Jan. 2009.

[9] D. S. Huang, *Systematic Theory of Neural Networks for Pattern Recognition*. Beijing, China: Publishing House of Electronic Industry of China, May 1996.

[10] D. S. Huang, "Radial basis probabilistic neural networks: Model and application," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 13, pp. 1083–1101, Nov. 1999.

[11] D. S. Huang and J. X. Du, "A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks," *IEEE Trans. Neural Netw.*, vol. 19, no. 12, pp. 2099–2115, Dec. 2008.

[12] W. Zhang, X. Liu, Y. Chen, W. Wu, W. Wang, and X. Li, "Feature-derived graph regularized matrix factorization for predicting drug side effects," *Neurocomputing*, vol. 287, pp. 154–162, Apr. 2018.

[13] J. F. Xia, X. M. Zhao, J. N. Song, and D. S. Huang, "APIS: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," *BMC Bioinf.*, vol. 11, p. 174, Apr. 2010.

[14] D. S. Huang and C. H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, pp. 1855–1862, Aug. 2006.

[15] D. S. Huang, X. M. Zhao, G. B. Huang, and Y. M. Cheung, "Classifying protein sequences using hydropathy blocks," *Pattern Recognit.*, vol. 39, pp. 2293–2300, Dec. 2006.

[16] D. S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein Peptide Sci.*, vol. 15, pp. 553–560, Sep. 2014.

[17] D. S. Huang and H. J. Yu, "Normalized feature vectors: A novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 2, pp. 457–467, Mar. 2013.

[18] D. S. Huang and X. Huang, "Improved performance in protein secondary structure prediction by combining multiple predictions," *Protein Peptide Lett.*, vol. 13, pp. 985–991, Oct. 2006.

[19] S. P. Deng, L. Zhu, and D. S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *BMC Genomics*, vol. 16, p. S4, Jan. 2015.

[20] S. P. Deng and D. S. Huang, "SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, pp. 207–212, Oct. 2014.

[21] X. J. Shen, Y. Chen, X. Jiang, X. Hu, T. He, and J. Yang, "Predicting disease-microbe association by random walking on the heterogeneous network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 771–774.

[22] S. Zou, J. Zhang, and Z. Zhang, "A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network," *PLoS ONE*, vol. 12, p. e0184394, Sep. 2017.

[23] X. Chen, Y. A. Huang, Z. H. You, G. Y. Yan, and X. S. Wang, "A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 33, pp. 733–739, Mar. 2017.

[24] Z. A. Huang *et al.*, "PBHMDA: Path-based human microbe-disease association prediction," *Frontiers Microbiol.*, vol. 8, p. 233, Feb. 2017.

[25] Y. A. Huang, Z. H. You, X. Chen, Z. A. Huang, S. W. Zhang, and G. Y. Yan, "Prediction of microbe disease association from the integration of neighbor and graph with collaborative recommendation model," *J. Transl. Med.*, vol. 15, p. 209, Oct. 2017.

[26] F. Wang *et al.*, "LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction," *Sci. Rep.*, vol. 7, Aug. 2017, Art. no. 7601.

[27] Z. Shen, Z. Jiang, and W. Bao, "CMFHMDA: Collaborative matrix factorization for human microbe-disease association prediction," presented at the Int. Conf. Intell. Comput., Liverpool, U.K., 2017, pp. 261–269.

[28] W. Zhang, J. Shi, G. Tang, W. Wu, X. Yue, and D. Li, "Predicting small RNAs in bacteria via sequence learning ensemble method," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 643–647.

[29] W. Zhang, X. Zhu, Y. Fu, J. Tsuji, and Z. Weng, "Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods," *BMC Bioinf.*, vol. 18, p. 464, Dec. 2017.

[30] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," *BMC Bioinf.*, vol. 18, p. 18, Jan. 2017.

[31] W. Zhang, Y. Chen, S. Tu, F. Liu, and Q. Qu, "Drug side effect prediction through linear neighborhoods and multiple data source integration," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 427–434.

[32] L. Luo, D. Li, W. Zhang, S. Tu, X. Zhu, and G. Tian, "Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features," *PLoS ONE*, vol. 11, p. e0153268, Apr. 2016.

[33] 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, pp. 1644–1650, Jul. 2010.

[34] W. Zhang, Y. Chen, and D. Li, "Drug-target interaction prediction through label propagation with linear neighborhood information," *Molecules*, vol. 22, p. 2056, Dec. 2017.

[35] W. Zhang, Q. Qu, Y. Zhang, and W. Wang, "The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions," *Neurocomputing*, vol. 273, pp. 526–534, Jan. 2018.

[36] W. Zhang, X. Yue, F. Liu, Y. Chen, S. Tu, and X. Zhang, "A unified frame of predicting side effects of drugs by using linear neighborhood similarity," *BMC Syst. Biol.*, vol. 11, p. 101, Dec. 2017.

[37] W. Ma *et al.*, "An analysis of human microbe–disease associations," *Briefings Bioinf.*, vol. 18, pp. 85–97, Jan. 2016.

[38] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Sci. Rep.*, vol. 5, Nov. 2015, Art. no. 16840.

[39] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Comput. Biol.*, vol. 6, p. e1000641, Jan. 2010.

[40] W. Zhang *et al.*, "Predicting drug-disease associations based on the known association bipartite network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 503–509.

[41] M. C. de Goffau *et al.*, "Aberrant gut microbiota composition at the onset of type 1 diabetes in young children," *Diabetologia*, vol. 57, pp. 1569–1577, Aug. 2014.

[42] N. Nagata *et al.*, "Glucoraphanin ameliorates obesity and insulin resistance through adipose tissue Browning and reduction of metabolic endotoxemia in mice," *Diabetes*, vol. 66, pp. 1222–1236, May 2017.

[43] E. Kounoue *et al.*, "The significance of T cells, B cells, antibodies and macrophages against encephalomyocarditis (EMC)-D virus-induced diabetes in mice," *Arch. Virol.*, vol. 153, p. 1223, Jul. 2008.

[44] S. Ljubić, A. Balachandran, I. Pavlić-Renar, A. Barada, and Ž. Metelko, "Pulmonary infections in diabetes mellitus," *DiabetologiaCroatica*, vol. 33, no. 4, pp. 115–124, Mar. 2005.

[45] O. Vaarala, "Gut microbiota and type 1 diabetes," *Rev. Diabetic Stud.*, vol. 9, no. 4, pp. 251–259, 2012.

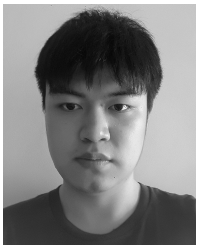
[46] M. Murri *et al.*, "Gut microbiota in children with type 1 diabetes differs from that in healthy children: A case-control study," *BMC Med.*, vol. 11, p. 46, Feb. 2013.

[47] E. Margolis and D. N. Fredricks, "Bacterial vaginosis-associated bacteria," in *Molecular Medical Microbiology*, M. Sussman, D. Liu, I. Poxton, and J. Schwartzman, Eds., 2nd ed. Boston, MA, USA: Academic, 2015, ch. 83, pp. 1487–1496.

[48] G. Funke, R. A. Hutson, M. Hilleringmann, W. R. Heizmann, and M. D. Collins, "*Corynebacterium lipophiloflavum* sp. nov. isolated from a patient with bacterial vaginosis," *FEMS Microbiol. Lett.*, vol. 150, pp. 219–224, May 1997.

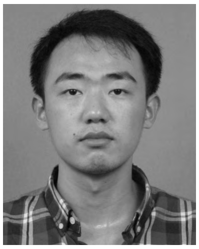
[49] S. Bengmark, "Gut microbiota, immune development and function," *Pharmacol. Res.*, vol. 69, pp. 87–113, Mar. 2013.

WEN ZHANG received the bachelor's and master's degree in computational mathematics and the Ph.D. degree in computer science from Wuhan University, China, in 2003, 2006, and 2009, respectively. He is currently an Associate Professor with the School of Computer Science, Wuhan University. His research interests include machine learning and bioinformatics.

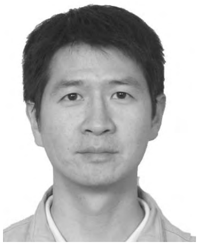


WEITAI YANG is currently pursuing the master's degree with the School of Computer Science, Wuhan University, China. His research interests include machine learning and data mining.

XIAOTING LU is currently pursuing the master's degree with the School of Computer Science, Wuhan University, China. Her research interests include machine learning and data mining.



FENG HUANG is currently pursuing the master's degree with the School of Computer Science, Wuhan University, China. His research interests include machine learning and data mining.



FEI LUO was born in Wuhan, Hubei, China, in 1980. He received the Ph.D. degree in computer science from Wuhan University. He currently serves as an Assistant Professor with Wuhan University. His research interests include bioinformatics and data mining.