frontiers in Microbiology

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fmicb.2020.00579

Predicting Microbe-Disease Association by Learning Graph Representations and Rule-Based Inference on the Heterogeneous Network

Xiujuan Lei and Yueyue Wang*

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

More and more clinical observations have implied that microbes have great effects on human diseases. Understanding the relations between microbes and diseases are of profound significance for disease prevention and therapy. In this paper, we propose a predictive model based on the known microbe-disease associations to discover potential microbe-disease associations through integrating Learning Graph Representations and a modified Scoring mechanism on the Heterogeneous network (called LGRSH). Firstly, the similarity networks for microbe and disease are obtained based on the similarity of Gaussian interaction profile kernel. Then, we construct a heterogeneous network including these two similarity networks and microbe-disease associations' network. After that, the embedding algorithm Node2vec is implemented to learn representations of nodes in the heterogeneous network. Finally, according to these low-dimensional vector representations, we calculate the relevance between each microbe and disease by utilizing a modified rule-based inference method. By comparison with three other methods including LRLSHMDA, KATZHMDA and BiRWHMDA, LGRSH performs better than others. Moreover, in case studies of asthma, Chronic Obstructive Pulmonary Disease and Inflammatory Bowel Disease, there are 8, 8, and 10 out of the top-10 discovered disease-related microbes were validated respectively, demonstrating that LGRSH performs well in predicting potential microbe-disease associations.

Keywords: microbe-disease association, heterogeneous network, network embedding algorithm, Node2vec, skip-gram

OPEN ACCESS

Edited by:
Hyun-Seob Song, University of Nebraska–Lincoln, United States

Reviewed by:
Wen Zhang, Huazhong Agricultural University, China
Sridevi Maharaj, University of California, Irvine, United States

****Correspondence:***
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *18 November 2019*
Accepted: *17 March 2020*
Published: *15 April 2020*

Citation:
Lei X and Wang Y (2020) Predicting Microbe-Disease Association by Learning Graph Representations and Rule-Based Inference on the Heterogeneous Network. Front. Microbiol. 11:579. doi: 10.3389/fmicb.2020.00579

INTRODUCTION

Varieties of microbial communities are dominant throughout the human different body niches including skin, mouth, respiratory tract, throat, stomach, gut and colon, which mainly compose of bacteria, protozoa, archaeon, viruses, and fungi (Methe et al., 2012; Althani et al., 2016). It is generally that a wide range of them play fundamental roles in human health and diseases such as maintaining homeostasis (Bouskra et al., 2008), developing the immune system (Round and Mazmanian, 2010; Gollwitzer et al., 2014) and resisting pathogens (Methe et al., 2012).

For example, the majority of microbes reside in the gut, regulating human physiology and nutrition by modulating host metabolism and immunity. They can digest and convert dietary constituents into active forms (Qin et al., 2010; Ahn et al., 2013).

Microbial communities are considered as an essential "organ" governing health and disease, which can be influenced by host genetics and host environment such as feeding habits, life styles, seasons and antibiotics (Huttenhower et al., 2012; Althani et al., 2016). If the microbial communities become imbalanced, there may interfere with the symbiotic relationships and cause diseases. For instance, researchers found that the number of phylum Actinobacteria among diabetics was significantly lower than the healthy person (Long et al., 2017). In addition, some studies found a decrease in the relative percentage of Bacteroidetes in obese people compared to the general population (Ley et al., 2006). Moreover, low microbial diversity can lead to inflammatory bowel disease (IBD) (Qin et al., 2010). Thus, understanding the microbe-disease associations can help us know disease pathogenesis to boost disease diagnosis and therapy.

With the advances in sequencing technologies and bioinformatics, more and more microbes living in oceans, soil, human bodies and elsewhere began to be investigated by the scientific community (Gilbert and Dupont, 2011; Methe et al., 2012; Cenit et al., 2014). The Human Microbiome Project Consortium (HMP) was funded to explore the relationships between microbes and human diseases. It generates a wide range of quality-controlled resources and data to develop metagenomic protocols, which is available for scientific research (Methe et al., 2012). Ma et al. (2016) constructed The Human Microbe-Disease Association Database (HMDAD) through collecting correlations between microbes and diseases from 61 published literatures. These achievements provided the foundation for further research on using computational methods to predict potential associations.

In recent years, some computational methods have been conceived for predicting microbe-disease associations based on the assumption that similarly functioning microorganisms incline to share similar associations or non-associations with diseases. By using the Gaussian interaction profile (GIP) kernel similarity, Chen et al. (2017) developed a prediction method called KATZHMDA that infers potential associations based on the number and length of walks in a heterogeneous network. Li et al. (2019) constructed a bidirectional weighted network by combining a normalized Gaussian interaction scheme with a bidirectional recommendation model. Zou et al. (2017) used a bi-random walk and logistic function transformation on a heterogeneous network constructed based on the GIP kernel similarity. Through a combination of the GIP kernel similarity and LapRLS classification, Wang et al. (2017) designed a computing model LRLSHMDA, which is semi-supervised . Meanwhile, through integrating the GIP kernel similarity with disease symptom similarity, Qu et al. (2019) implemented the matrix decomposition and label propagation algorithm on the similarity network for associations' prediction . Huang et al. (2017) predicted potential associations based on known microbe-disease bipartite graph and neighbor collaborative filtering. Moreover, Fan et al. (2019) proposed a method called MDPH_HMDA for prediction by executing standardized HeteSim measurements to weight the relations in a heterogeneous network combined by the GIP kernel similarity, the microbe–microbe functional similarity and the symptom-based human disease similarity. Niu et al. (2019) identified the potential associations by introducing the concept of hypergraph, which put all disease-related microbes on a single hyperedge. In order to take the unequal contributions of microbe and disease information into consider, Zhang et al. (2018) developed a bidirectional similarity integral label propagation method with calculating the microbe functional similarity and the disease semantic similarity.

At the same time, many network embedded methods have been proposed, such as DeepWalk (Perozzi et al., 2014), SDNE (Wang et al., 2016), Node2vec (Grover and Leskovec, 2016), etc. In this study, inspired by the performance of graph representations for many real-world problems such as protein network research, text and visual processing (Cao et al., 2016). We utilize Node2vec (Grover and Leskovec, 2016) to predict potentially unknown associations (LGRSH) on a heterogeneous network. First, similarity networks for microbes and diseases are calculated by the GIP kernel similarity. Then, we construct a heterogeneous network integrating the two similarity networks and known microbe-disease associations' network. After that, the embedding algorithm Node2vec has been utilized to assign a low-dimensional vector representation to nodes in the heterogeneous network. Finally, according to the vector representation of each node, we calculate the degrees of correlation between microbes and diseases to discover potential associations with a modified rule-based inference method. In order to assess the prediction performance of LGRSH, we implemented Leave-one-out cross validation (LOOCV) and fivefold cross validation. The area under the receiver operating characteristic curve (AUC) obtained by LGRSH are 0.9260 and 0.9254, which is better than the compared methods. Moreover, case studies of asthma, Chronic Obstructive Pulmonary Disease (COPD) and IBD demonstrate that LGRSH can be considered as an effective method for association prediction.

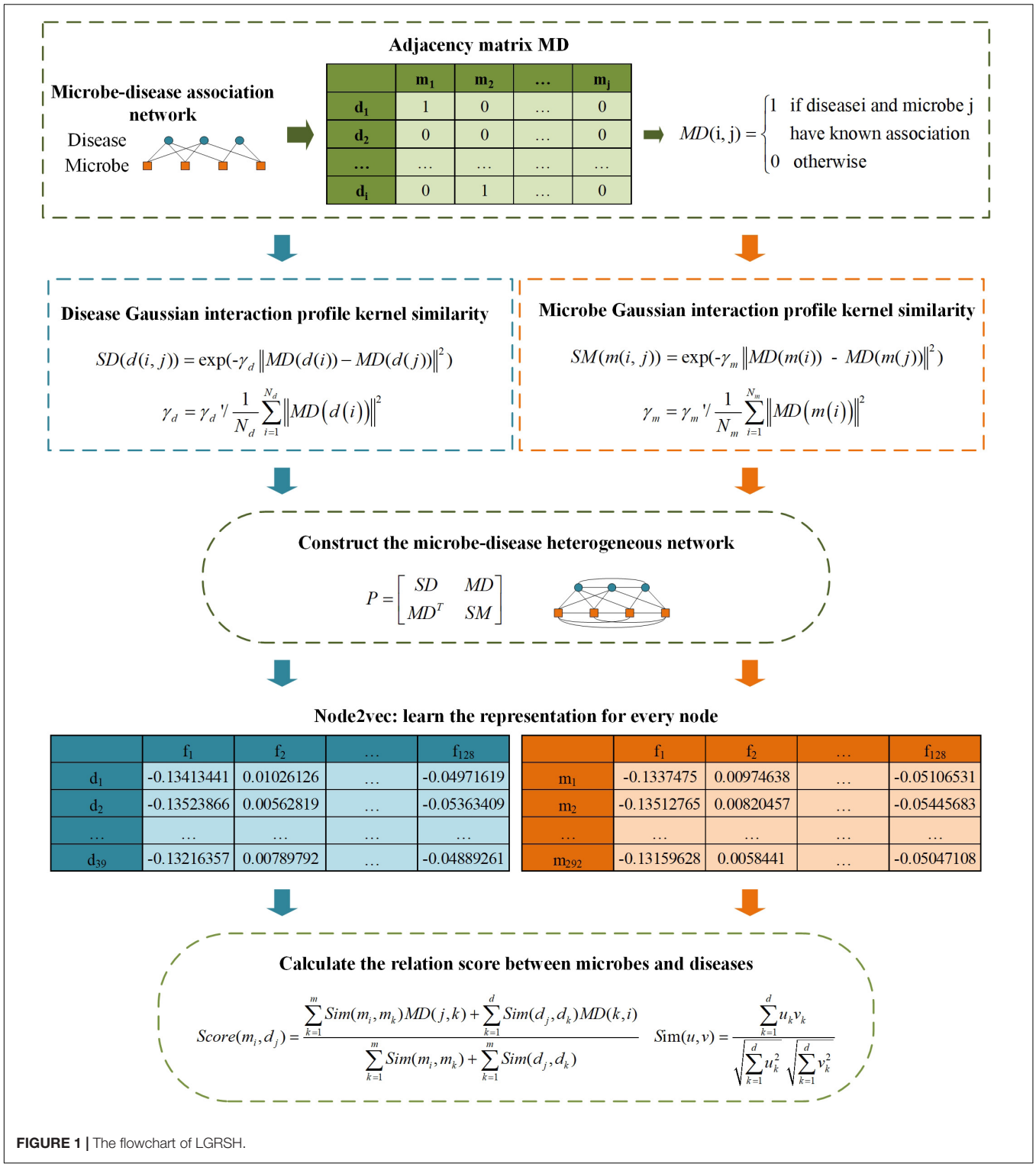
MATERIALS AND METHODS

Material

We download microbe-disease associations from HMDAD (Ma et al., 2016), which contains 483 verified associations' records between 292 microbes and 39 diseases. After removing the repetitive relationships, 450 distinct associations' records are obtained. Then we construct a 39×292 dimensional adjacency matrix MD of the associations' network. $MD(i, j)$ is 1 indicating that there is a known association between disease $d(i)$ and microbe $m(j)$, otherwise, $MD(i, j)$ is 0.

Methods

As illustrated in **Figure 1**, firstly, the similarity networks for microbe and disease have been constructed. And then, a heterogeneous network integrating two similarity networks and



	m_1	m_2	…	m_j
d_1	1	0	…	0
d_2	0	0	…	0
…	…	…	…	…
d_i	0	1	…	0

$$SD(d(i,j)) = \exp(-\gamma_d \|MD(d(i)) - MD(d(j))\|^2)$$

$$\gamma_d = \gamma_d' / \frac{1}{N_d}\sum_{i=1}^{N_d}\|MD(d(i))\|^2$$

$$SM(m(i,j)) = \exp(-\gamma_m \|MD(m(i)) - MD(m(j))\|^2)$$

$$\gamma_m = \gamma_m' / \frac{1}{N_m}\sum_{i=1}^{N_m}\|MD(m(i))\|^2$$

$$P = \begin{bmatrix} SD & MD \\ MD^T & SM \end{bmatrix}$$

	f_1	f_2	…	f_{128}
d_1	-0.13413441	0.01026126	…	-0.04971619
d_2	-0.13523866	0.00562819	…	-0.05363409
…	…	…	…	…
d_{39}	-0.13216357	0.00789792	…	-0.04889261

	f_1	f_2	…	f_{128}
m_1	-0.1337475	0.00974638	…	-0.05106531
m_2	-0.13512765	0.00820457	…	-0.05445683
…	…	…	…	…
m_{292}	-0.13159628	0.0058441	…	-0.05047108

$$Score(m_i, d_j) = \frac{\sum_{k=1}^{m} Sim(m_i, m_k)MD(j,k) + \sum_{k=1}^{d} Sim(d_j, d_k)MD(k,i)}{\sum_{k=1}^{m} Sim(m_i, m_k) + \sum_{k=1}^{m} Sim(d_j, d_k)}$$

$$Sim(u,v) = \frac{\sum_{k=1}^{d} u_k v_k}{\sqrt{\sum_{k=1}^{d} u_k^2}\sqrt{\sum_{k=1}^{d} v_k^2}}$$

FIGURE 1 | The flowchart of LGRSH.

microbe-disease associations' network can be obtained. After that, the embedding algorithm Node2vec is utilized to learn the representation for every node. Finally, according to the topology information based on Node2vec method, we calculate the relation score between every microbe vector and disease vector.

Calculation of Microbe Similarities Based on the GIP Kernel Similarity

Based on the assumption that two microbes are more likely to share functional similarities potentially if they are related to more common diseases. We calculate the GIP kernel similarity for

microbes based on known microbe-disease associations' network. For microbes *m(i)* and *m(j)*, the similarity score is obtained according to Eq. (1) (Wang et al., 2017):

$$SM(m(i,\ j)) = \exp(-\gamma_m \| MD(m(i)) - MD(m(j)) \|^2) \quad (1)$$

where *m(i, j)* represents two arbitrary microbes in matrix *MD*. Parameter γ_m is used to control the bandwidth and is affected by a new bandwidth parameter γ_m'(Wang et al., 2017), which can be obtained as Eq. (2):

$$\gamma_m = \gamma'_m / \frac{1}{N_m} \sum_{i=1}^{N_m} \| MD\,(m\,(i))\,\|^2 \quad (2)$$

here, N_m is equal to 292, which indicates the total number of microbes. The parameter γ_m' is set to 1 for simplicity (Wang et al., 2017).

Calculation of Disease Similarities Based on the GIP Kernel Similarity

In the similar way, we construct a disease similarity network by using the GIP kernel similarity for each disease pair. The similarity between disease *d(i)* and *d(j)* is obtained according to Eq. (3) (Wang et al., 2017):

$$SD(d(i,j)) = \exp(-\gamma_d \| MD(d(i)) - MD(d(j)) \|^2) \quad (3)$$

where *d(i, j)* represents two arbitrary diseases in matrix *MD*. The parameter γ_d can be obtained as Eq. (4):

$$\gamma_d = \gamma'_d / \frac{1}{N_d} \sum_{i=1}^{N_d} \| MD\left(d\,(i)\right)\|^2 \quad (4)$$

here, N_d is equal to 39, which indicates the total number of diseases. The parameter γ_d' is set to 1 for simplicity (Wang et al., 2017).

Constructing a Heterogeneous Network for Microbes and Diseases

According to the Eqs (1) and (3), we have constructed two similarity matrices SM and SD. Then we construct a heterogeneous network including the edges of microbe–microbe, microbe-disease and disease–disease associations, and it can be expressed as Eq. (5):

$$P = \begin{bmatrix} SD & MD \\ MD^T & SM \end{bmatrix} \quad (5)$$

where *P* represents the matrix of heterogeneous network. MD^T is the transpose of *MD*.

Using Node2vec to Learning Representations

Node2vec is a flexible neighborhood sampling strategy which can explore neighborhoods in the form of Breadth-First Sampling (*BFS*) and Depth-First Sampling (*DFS*) fashion by introducing two parameters (Grover and Leskovec, 2016). It maximizes the network neighborhood of nodes by mapping nodes to vector feature spaces. Therefore, we apply Node2vec to learn vector representations for nodes in the heterogeneous network.

Firstly, we utilize a bias random walk strategy to calculate the transition probabilities for every node. For a current node *u*, the probability of accessing the next node *x* can be calculated as follows:

$$P(c_i = x | c_{i-1} = u) = \begin{cases} \frac{\pi_{ux}}{Z} & if\ (u, x) \in E \\ 0 & otherwise \end{cases} \quad (6)$$

here, *Z* is a regularization constant. π_{ux} is denormalized transition probabilities on edges (*u*, *x*) leading from *u*, which is influenced by a weight adjustment parameter α. We suppose the walk just went from *t* to *u* and set$\pi_{ux} = \alpha_{pq}\,(t,\ x) \cdot w_{ux}$, where

$$a_{pq}(t,\ x) = \begin{cases} 1/p & if\ d_{tx} = 0 \\ 1 & if\ d_{tx} = 1 \\ 1/q & if\ d_{tx} = 2 \end{cases} \quad (7)$$

here, d_{tx} is in the range of {0, 1, 2}, representing the shortest distance from nodes *t* to *x*. Parameters *p* and *q* are used to strike a balance between *DFS* and *BFS*. As shown in **Figure 2**, parameter *p* is a return parameter that affects the possibility of re-traversing a node immediately during a walk. If *p* is set to be larger, it is less likely to revisit the node that was just accessed. This strategy can lead to moderate exploration and avoid repetitive sampling. If the value is set to be smaller, the walk is more likely to backtrack, and tends to reach nodes near the node. There is more concerned for the local information. Parameter *q* is an in-out parameter, which allows searches to distinguish "inward" and "outward" nodes (Zeng et al., 2019). If $q > 1$, the walk tends to be closer to node *u*. In contrast, if $q < 1$, it tends to traverse nodes far from node *u* (Zeng et al., 2019).

We first select one node *u* and mark it as the current node, and then select one node *v* from all the neighbors of the current node *u* based on the transition probabilities calculated above. Following, we mark this newly selected node *v* as the current node and repetitive such as a node sampling process. The algorithm terminates when the number of nodes in a sequence reaches a preset walking length *l*. By referring to the previous paper, we set *l* as10 (Munui et al., 2018).

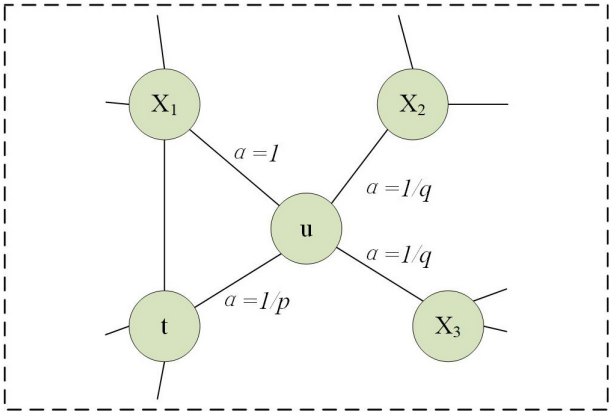


FIGURE 2 | Description of walking strategy in Node2vec when the traversal has just gone from *t* to *u*.

Node2vec uses Skip-gram model to generate eigenvectors of nodes (Jang et al., 2019). Skip-gram model is a word embedding algorithms for learning distributed vector representations from a large number of textual corpora which tries to categorize a word according to other words in the same sentence as much as possible (Mikolov et al., 2013). In fact, the sequence of nodes obtained by bias random walk algorithm, each node actually corresponds to a word. The input of this model is the sequence encoding of a node, and the output is the nodes before and after the sequence. In this paper, we set the context size to 10 and the dimension of these eigenvectors to 128 according to the original parameter selection for the best performance (Grover and Leskovec, 2016). The algorithm is detailed in **Figure 3**.

Association Discovering

According to the popular rule-based inference method for predicting novel drug-target associations based on indirect relationships in 2017 (Zong et al., 2017), we utilize a modified Scoring mechanism to grade microbe-disease relations based on the low-dimensional vector representation. Considering that indirect relationships do not fully predict the relationship if there are few known relations between some microbes and diseases, especially if there is only single relationship, we have used both direct and indirect connections to calculate correlations between microbes and diseases.

We use $Score(m_i, d_j)$ to represent the correlation score between the *ith* microbe and *jth* disease in the heterogeneous network. It can be calculated according to Eq. (8):

$$Score(m_i, d_j) = \frac{\sum_{k=1}^{m} Sim(m_i, m_k)MD(j, k) + \sum_{k=1}^{d} Sim(d_j, d_k)MD(k, i)}{\sum_{k=1}^{m} Sim(m_i, m_k) + \sum_{k=1}^{m} Sim(d_j, d_k)} \quad (8)$$

In this Equation, m and d indicate the numbers of microbe and disease, $MD(i, j)$ is the association between disease i and microbe j. The $Sim(u, v)$ is calculated as Eq. (9):

$$Sim(u, v) = \frac{\sum_{k=1}^{d} u_k v_k}{\sqrt{\sum_{k=1}^{d} u_k^2} \sqrt{\sum_{k=1}^{d} v_k^2}} \quad (9)$$

here, d represents the dimension for each vector, u_k, v_k represent the components of vectors u and v.

Learning representations
Input: Graph $P = (V, E, W)$, dimension d, walks per node r, walk length l, Context size k, Return parameter p, In-out parameter q
Output: Eigenvectors of each node

```
Node2vec(P, d, r, l, k, p, q)
    π = TP preprocessing (P, p, q)          // Calculate the transition probabilities
    P' = (V, E, π)                          // Normalize transition probabilities
    Initialize walks to Empty
    For iter = 1 to r:
        for all nodes u ∈ V:                // Simulate a random walk starting from start node
            Initialize walk to [u]
            for walk_length = 1 to l:
                curr = walk[- 1]
                Vcurr = Get Neighbors(curr, P')   // Find neighbors of the current node
                s = probability select(Vcurr, π)  // Walk based on transition probabilities
                Append s to walk
    Append walk to walks
    F = Skip-gram (k, d, walks)             // Generate eigenvectors of each node
    return F
```

FIGURE 3 | Description of algorithm Node2vec.

TABLE 1 | Effect of parameters p and q in fivefold cross validation.

	q = 0.25	q = 0.5	q = 1	q = 2	q = 4	q = 8	q = 16
p = 0.25	0.9251	0.9165	0.9178	0.9246	0.9229	0.9236	0.9244
p = 0.5	0.9253	0.9236	0.9251	0.9246	**0.9254**	0.9235	0.9229
p = 1	0.9240	0.9250	0.9190	0.9213	0.9234	0.9234	0.9242
p = 2	0.9214	0.9204	0.9239	0.9230	0.9251	0.9181	0.9208
p = 4	0.9215	0.9222	0.9206	0.9229	0.9241	0.9239	0.9235

Bold values: LGRSH achieves the best performance while p = 0.5, q = 4.

RESULTS

We implement LOOCV and fivefold cross validation on HMDAD to assess the prediction performance of LGRSH. In the LOOCV, we regard each known association as a test sample, with other known associations as training samples (Quan et al., 2014). All unverified microbe-disease associations are regarded as candidate samples. In the fivefold cross validation, we randomly divide all known microbe-disease associations into 5 average groups. Each of these five groups is regarded as testing sample, while other four groups are training samples. This process is conducted five times to mitigate the bias due to random sample partitioning (Niu et al., 2019). Based on the prediction score, we evaluate the predictive performance by ranking the test samples. The AUC can be calculated according to the receiver operating characteristic (ROC) curve. If there is a random prediction performance, the AUC value is 0.5.

Effect of Parameters

There are two important parameters in Node2vec. One is a return parameter p and another is an in-out parameter q. We set various values under the framework of fivefold cross validation in order to evaluate the impact of these parameters. According to the comparison results in **Table 1** and **Figure 4**, we can find that the performance of LGRSH is best with 0.9254 while $p = 0.5$, $q = 4$. Hence, we set $p = 0.5$, $q = 4$ in the subsequent experiments.

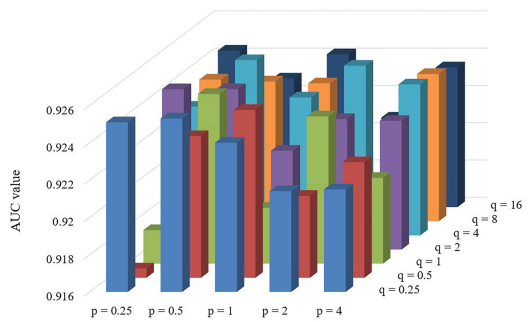


FIGURE 4 | Effect of parameters p and q in fivefold cross validation.

Comparison With Other Methods

We compare LGRSH with three methods including LRLSHMDA (Wang et al., 2017), KATZHMDA (Chen et al., 2017) and BiRWHMDA (Zou et al., 2017). These four methods are measured by Precision-recall curve. As illustrated in **Figures 5**, **6**, we can draw a conclusion that LGRSH performs better than other three methods.

Furthermore, we measure the top-level results of LGRSH and three other methods in LOOCV. As shown in **Figure 7**, LGRSH can find more known associations among the top 500 predicted microbes.

CASE STUDIES

To evaluate the ability of LGRSH for discovering unknown associations in HMDAD, we implement case studies in asthma, COPD and IBD. We conduct experiments for 10 times on each diseases to make the results more stable. After calculating the similarity of every microbe and disease, the scores are sorted in descending order to obtain the top-10 candidate microbes for every disease. The scores of top-10 disease-related microbes are provided in **Supplementary Tables S1–S3**, respectively.

Asthma

Asthma is a common inflammatory disease affecting more than 300 million people all over the world, which is more common in childhood with recurrent cough, wheezing and breathing difficulties. In recent years, asthma has been found to be closely linked with microbes (Caliskan et al., 2013). Hence, we consider Asthma for case studies. As shown in **Table 2**, 8 of top-10 discovered microbes were confirmed. For instance, Clostridium difficile colonization (ranked 1st in the list) in 1 month was associated with asthma between the ages of 6 and

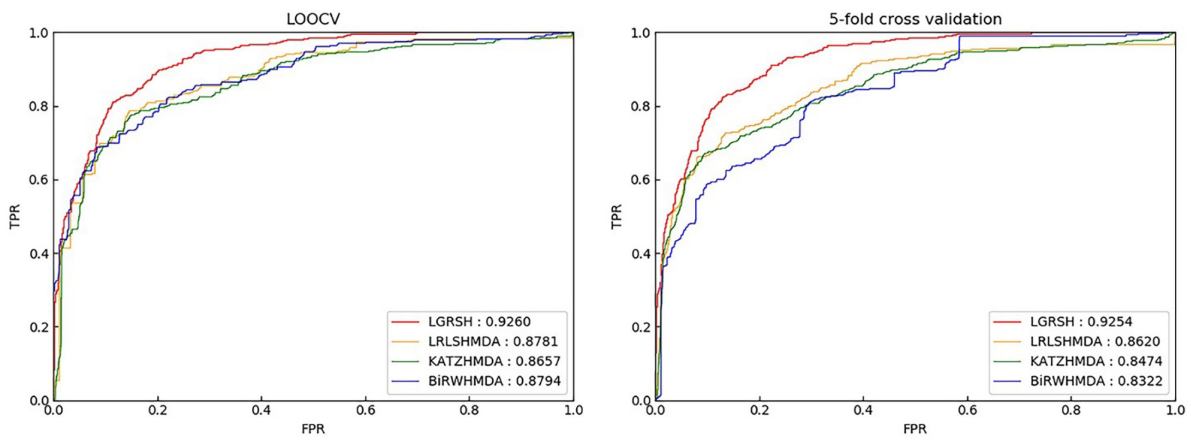


FIGURE 5 | Prediction comparison between LGRSH and other three methods in LOOCV and fivefold cross validation while $p = 0.5$, $q = 4$.

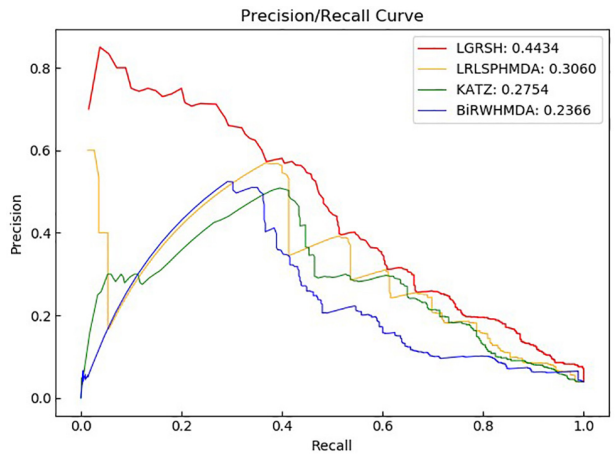


FIGURE 6 | Precision-recall curves for LGRSH and other three methods in fivefold cross validation.

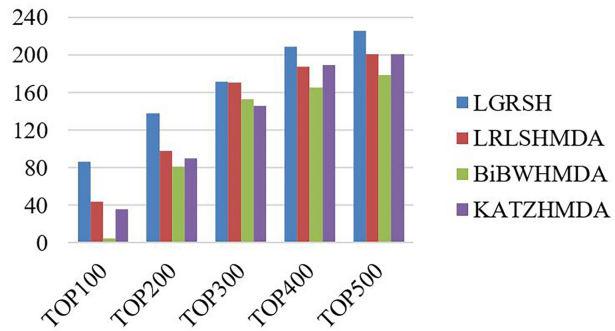


FIGURE 7 | The number of correctly predicted by LGRSH and other three methods on HMDAD.

TABLE 2 | Validation results for Top-10 predicted microbes related with asthma.

Rank	Microbe	Evidence
1	Clostridium difficile	PMID:21872915
2	Firmicutes	PMID:27078029
3	Clostridium coccoides	PMID:21477358
4	Actinobacteria	PMID:30286807
5	Enterobacteriaceae	PMID:28947029
6	Lactobacillus	PMID:30400588
7	Bacteroides	PMID:18822123
		PMID:29161087
8	Burkholderia	**Unconfirmed**
9	Lachnospiraceae	PMID:28912020
10	Enterococcus	**Unconfirmed**

7 (van Nimwegen et al., 2011). Researchers also proved that colonization with Clostridium coccoides (ranked 3rd in the list) and Bacteroides (ranked 7th in the list) at 3 weeks were associated with positive predictors of asthma at age 3 (Carl et al., 2008, 2011). In addition, the abundance of Firmicutes (ranked 2nd in the list) and Enterobacteriaceae (ranked 5th in the list) were higher in severe asthmatics compared with non-asthmatic people, while Actinobacteria (ranked 4th in the list) and Lachnospiraceae (ranked 9th in the list) with lower proportion (Marri et al., 2013; Ciaccio et al., 2015; Zhang et al., 2016; Li et al., 2017). Moreover, Huang et al. (2018) found that Lactobacillus (ranked 6th in the list) can reduce asthma severity and improve asthma control, which is beneficial to children with asthma.

TABLE 3 | Validation results for Top-10 predicted microbes related with COPD.

Rank	Microbe	Evidence
1	Proteobacteria	PMID:29579057
2	Prevotella	PMID:28542929
3	Helicobacter pylori	PMID:28558695
4	Actinobacteria	PMID:29709671
5	Bacteroidetes	PMID:29579057
6	Clostridium difficile	PMID:30430993
7	Clostridium coccoides	**Unconfirmed**
8	Lactobacillus	PMID:26630356
9	Lachnospiraceae	**Unconfirmed**
10	Staphylococcus aureus	PMID:30804927

TABLE 4 | Validation results for Top-10 predicted microbes related with IBD.

Rank	Microbe	Evidence
1	Prevotella	PMID:24013298
2	Bacteroidetes	PMID:29492876
3	Clostridium difficile	PMID:24838421
4	Helicobacter pylori	PMID:22221289
		PMID:28124160
5	Firmicutes	PMID:25307765
		PMID:29492876
6	Clostridium coccoides	PMID:19235886
7	Lactobacillus	PMID:26340825
8	Enterobacteriaceae	PMID:30319571
9	Veillonella	PMID:30573380
10	Haemophilus	PMID:24013298

Chronic obstructive pulmonary disease (COPD)

Chronic obstructive pulmonary disease is a progressive obstructive pulmonary disease with main symptoms of breathing difficultly and coughing (Rabe et al., 2007). It is more common among smokers, and is also influenced by factors like air pollution and genetics. Although the disease can be slowed down by treatment, there is still no clear treatment or pathogenesis for it. Recently, some findings indicate that changes in microbes may have significant effects in the development of COPD (Malhotra and Henric, 2015). Thus, we consider COPD for case studies. As shown in **Table 3**, 8 of top 10 discovered microbes were confirmed. For example, the main flora of Proteobacteria (ranked 1st in the list) and Bacteroidetes (ranked 5th in the list) increased with the deterioration of COPD (Rohde et al., 2004). Researchers also found that Helicobacter pylori (ranked 3rd in the list)

infection is associated with reduced lung function and systemic inflammation in COPD patients (Mammen and Sethi, 2016). In patients with COPD, the proportion of Prevotella (ranked 2nd in the list) is reduced compared with healthy people, but phyla Actinobacteria (ranked 4th in the list), Clostridium difficile (ranked 6th in the list) and Lactobacillus (ranked 8th in the list) are increased (Yadava et al., 2016; Larsen, 2017; de Miguel-Diez et al., 2018; Ghebre et al., 2018). For example, the Clostridium difficile is twice as high in COPD patients as in healthy person. Moreover, Staphylococcus aureus (ranked 10th in the list) has been found in the respiratory tract of patients with COPD (Uddin et al., 2019).

Inflammatory bowel disease (IBD)

Inflammatory bowel disease is a chronic, idiopathic gastrointestinal inflammatory disease that is thought to be influenced by environmental and host factors (D'Aoust et al., 2017). It is characterized by recurrent episodes, diverse clinical manifestations and severe complications such as bleeding, abscess formation and perforation (Cosnes et al., 2002). In this paper, we consider IBD for case studies. As shown in **Table 4**, 10 of top-10 discovered microbes were confirmed. For instance, researchers have found that IBD is related to gut microbiological disorders including expansion of Enterobacteriaceae facultative anaerobic bacteria (ranked 8th in the list) and decrease in some beneficial fecal bacteria such as Firmicutes (ranked 5th in the list) (Eom et al., 2018; Zuo and Ng, 2018). In patients with IBD, the dominant of Prevotella (ranked 1st in the list), Veillonella (ranked 9th in the list) and Haemophilus (ranked 10th in the list) were largely contribute to dysbiosis (Said et al., 2014). Bacteroidetes (ranked second in the list) and Lactobacillus (ranked 7th in the list) were significantly increased compared with healthy people, but the Clostridium coccoides (ranked 6th in the list) was less abundant (Sokol et al., 2009; Thomas et al., 2015; Eom et al., 2018). Researchers also found that Clostridium difficile (ranked 3rd in the list) infection has become a significant clinical challenge for patients suffering from IBD, which can worsen flares of IBD, inducing to emergent colectomies and mortality (Hashash and Binion, 2014). Moreover, recent experimental results found that chronic infection with Helicobacter pylori (ranked 4th in the list) is protective against IBD. And IBD patients are least likely to be infected with Helicobacter pylori compared to the normal population (Sonnenberg and Genta, 2012; Kyburz and Muller, 2017).

CONCLUSION

There are countless microbe communities inhabited in the human body, having important impacts on human health and disease by regulating the metabolism and immunity. With the establishment of relational databases for microbes and diseases, exploring their associations have become a hot topic for researchers. In this study, we propose a predictive approach called LGRSH by utilizing network embedding algorithm Node2vec to obtain the representation for every node in the heterogeneous network. According to the vector representation for every node, we rank the relevance of each microbe vector and disease vector to discover potential microbe-disease associations. In LOOCV and 5-fold cross validation, LGRSH performs better compared with three other methods with AUC reached 0.9260 and 0.9254. The case studies of asthma, COPD and IBD show that LGRSH can be used as a predictive tool for microbe-disease associations.

Certainly, there are still some deficiencies in LGRSH. For example, there are only 450 know micro-disease associations, which accounts for very small proportion of human microbial diseases. This may result in less comprehensive for prediction. We believe that the problem will be solved when more microbe-disease links are discovered. In addition, the embedding algorithm itself is a local method. In the future, we will learn more graph representation algorithms to improve the global capability. Moreover, we calculate the similarities for microbe and disease through the GIP kernel, which may biased toward microbes and diseases with more known associations. Hence, we will improve the efficiency of LGRSH by integrating some optimization strategies such as microbe functional similarity, disease semantic similarity and symptom-based disease similarity in the future work.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

XL and YW conceptualized the study and read and approved the final manuscript. YW conducted the experiments, analyzed the result, and wrote the manuscript. XL conceived the project, analyzed the result, and revised the manuscript.

FUNDING

This work was supported by the funding from National Natural Science Foundation of China (Nos. 61972451, 61672334, and 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00579/full#supplementary-material

REFERENCES

Ahn, J., Sinha, R., Pei, Z. H., Dominianni, C., Wu, J., Shi, J. X., et al. (2013). Human gut microbiome and risk for colorectal Cancer. *J. Natl. Cancer Inst.* 105, 1907–1911. doi: 10.1093/jnci/djt300

Althani, A. A., Marei, H. E., Hamdi, W. S., Nasrallah, G. K., El Zowalaty, M. E., Al Khodor, S., et al. (2016). Human microbiome and its association with health and diseases. *J. Cell. Physiol.* 231, 1688–1694. doi: 10.1002/jcp.25284

Bouskra, D., Brezillon, C., Berard, M., Werts, C., Varona, R., Boneca, I. G., et al. (2008). Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature* 456, 507–534. doi: 10.1038/nature07450

Caliskan, M., Bochkov, Y. A., Kreiner-Moller, E., Bonnelykke, K., Ober, C., et al. (2013). Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N. Engl. J. Med.* 368, 1398–1407. doi: 10.1056/NEJMoa1211592

Cao, S., Lu, W., and Xu, Q. (2016). *Deep Neural Networks for Learning Graph Representations. Paper presented at the AAAI*. Menlo Park, CA: AAAI Press.

Carl, V., Liesbeth, V., Kristine, N. D., and Herman, G. (2011). Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol.* 11:68. doi: 10.1186/1471-180-11-68

Carl, V., Vera, N., Verhulst, S. L., Herman, G., and Medicine, Desager, K. N. (2008). Early intestinal *Bacteroides* fragilis colonisation and development of asthma. *BMC Pulmon. Med.* 8:19. doi: 10.1186/1471-2466-8-19

Cenit, M. C., Matzaraki, V., Tigchelaar, E. F., and Zhernakova, A. (2014). Rapidly expanding knowledge on the role of the gut microbiome in health and disease. *Biochim. Biophy. Acta Mol. Basis Dis.* 1842, 1981–1992. doi: 10.1016/j.bbadis.2014.05.023

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2017). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739. doi: 10.1093/bioinformatics/btw715

Ciaccio, C. E., Barnes, C., Kennedy, K., Chan, M., Portnoy, J., and Rosenwasser, L. (2015). Home dust microbiota is disordered in homes of low-income asthmatic children. *J. Asthma* 52, 1–8. doi: 10.3109/02770903.2015.1028076

Cosnes, J., Cattan, S., Blain, A., Beaugerie, L., Carbonnel, F., Parc, R., et al. (2002). Long-term evolution of disease behavior of Crohn's disease. *Inflamm. Bowel Dis.* 8, 244–250. doi: 10.1097/00054725-200207000-00002

D'Aoust, J., Battat, R., and Bessissow, T. (2017). Management of inflammatory bowel disease with clostridium difficile infection. *World J. Gastroenterol.* 23, 4986–5003. doi: 10.3748/wjg.v23.i27.4986

de Miguel-Diez, J., Lopez-de-Andres, A., Esteban-Vasallo, M. D., Hernandez-Barrera, V., de Miguel-Yanes, J. M., Mendez-Bailon, M., et al. (2018). Clostridium difficile infection in hospitalized patients with COPD in Spain (2001-2015). *Eur. J. Intern. Med.* 57, 76–82. doi: 10.1016/j.ejim.2018.06.022

Eom, T., Kim, Y. S., Choi, C. H., Sadowsky, M. J., and Unno, T. (2018). Current understanding of microbiota- and dietary-therapies for treating inflammatory bowel disease. *J. Microbiol.* 56, 189–198. doi: 10.1007/s12275-018-8049-8

Fan, C. Y., Lei, X. J., Guo, L., and Zhang, A. D. (2019). Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. *Neurocomputing* 323, 76–85. doi: 10.1016/j.neucom.2018.09.054

Ghebre, M. A., Pang, P. H., Diver, S., Desai, D., Bafadhel, M., Haldar, K., et al. (2018). Biological exacerbation clusters demonstrate asthma and chronic obstructive pulmonary disease overlap with distinct mediator and microbiome profiles. *J. Allergy Clin. Immunol,* 141, 2027.e12–2036.e12. doi: 10.1016/j.jaci.2018.04.013

Gilbert, J. A., and Dupont, C. L. (2011). Microbial metagenomics: beyond the genome. *Annu. Rev. Mar. Sci.* 3, 347–371. doi: 10.1146/annurev-marine-120709-142811

Gollwitzer, E. S., Saglani, S., Trompette, A., Yadava, K., Sherburn, R., McCoy, K. D., et al. (2014). Lung microbiota promotes tolerance to allergens in neonates via PD-L1. *Nat. Med.* 20, 642–647. doi: 10.1038/nm.3568

Grover, A., and Leskovec, J. (2016). node2vec: scalable feature learning for networks. *KDD* 2016, 855–864. doi: 10.1145/2939672.2939754

Hashash, J. G., and Binion, D. G. (2014). Managing clostridium difficile in inflammatory bowel disease (IBD). *Curr. Gastroenterol. Rep.* 16:393. doi: 10.1007/s11894-014-0393-1

Huang, C.-F., Chie, W.-C., and Wang, I.-J. J. N. (2018). Efficacy of lactobacillus administration in school-age children with asthma: a randomized, placebo-controlled trial. *Nutrients* 10:1678. doi: 10.3390/nu10111678

Huang, Y. A., You, Z. H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017). Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* 15:209. doi: 10.1186/s12967-017-1304-7

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jang, B., Kim, I., and Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *Plos One* 14:e0220976. doi: 10.1371/journal.pone.0220976

Kyburz, A., and Muller, A. (2017). *Helicobacter* pylori and extragastric diseases. *Curr. Top. Microbiol. Immunol.* 400, 325–347. doi: 10.1007/978-3-319-50520-6_14

Larsen, J. M. (2017). The immune response to *Prevotella bacteria* in chronic inflammatory disease. *Immunology* 151, 363–374. doi: 10.1111/imm.12760

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023. doi: 10.1038/4441022a

Li, H., Wang, Y. Q., Jiang, J. W., Zhao, H. C., Feng, X., Wang, L., et al. (2019). A novel human microbe-disease association prediction method based on the bidirectional weighted network. *Front. Microbiol.* 10:676. doi: 10.3389/fmicb.2019.00676

Li, N., Qiu, R., Yang, Z., Li, J., Chung, K. F., Zhang, Q. J. R. M., et al. (2017). Sputum microbiota in severe asthma patients: relationship to eosinophilic inflammation. *Respiratory Med.* 131, 192–198. doi: 10.1016/j.rmed.2017.08.016

Long, J., Cai, Q., Steinwandel, M., Hargreaves, M. K., Bordenstein, S. R., Blot, W. J., et al. (2017). Association of oral microbiome with type 2 diabetes risk. *J. Periodontal Res.* 52, 636–643. doi: 10.1111/jre.12432

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., and Cui, Q. (2016). An analysis of human microbe–disease associations. *Briefings Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Malhotra, R., and Henric, O. (2015). Immunology, genetics and microbiota in the COPD pathophysiology: potential scope for patient stratification. *Expert Rev. Respirat. Med.* 9, 153–159. doi: 10.1586/17476348.2015.1000865

Mammen, M. J., and Sethi, S. (2016). COPD and the microbiome. *Respirology* 21, 590–599. doi: 10.1111/resp.12732

Marri, P. R., Stern, D. A., Wright, A. L., Billheimer, D., and Martinez, F. D. (2013). Asthma-associated differences in microbial composition of induced sputum. *J. Allergy Clin. Immunol.* 131, 346.–352.e1-13. doi: 10.1016/j.jaci.2012.11.013

Methe, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., et al. (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Science, J. D. J. C. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. Ithaca, NY: Cornell University.

Munui, K., Han, B. S., and Min, S. (2018). Relation extraction for biological pathway construction using node2vec. *Bmc Bioinform.* 19:206. doi: 10.1186/s12859-018-2200-8

Niu, Y. W., Qu, C. Q., Wang, G. H., and Yan, G. Y. (2019). RWHMDA: random walk on hypergraph for microbe-disease association prediction. *Front. Microbiol.* 10:1578. doi: 10.3389/fmicb.2019.01578

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: online learning of social representations," in *Paper presented at the Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and data Mining*, New York, NY: ACM.

Qin, J., Li, R., Raes, J., Arumugam, M., and Nature, M. K. J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qu, J., Zhao, Y., and Yin, J. (2019). Identification and analysis of human microbe-disease associations by matrix decomposition and label propagation. *Front. Microbiol.* 10:291. doi: 10.3389/fmicb.2019.00291

Quan, Z., Jinjin, L., and Chunyu, W. (2014). Approaches for Recognizing Disease Genes Based on Network. *Biomed. Res. Int.* 2014, 416323. doi: 10.1155/2014/416323

Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., and Respiratory, J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD Executive Summary. *Am. J. Respir. Crit. Care Med.* 176, 532–555. doi: 10.1164/rccm.200703-456SO

Rohde, G., Gevaert, P., Holtappels, G., Borg, I., Wiethege, A., Arinir, U., et al. (2004). Increased IgE-antibodies to Staphylococcus aureus enterotoxins in patients with COPD. *Respir. Med.* 98, 858–864. doi: 10.1016/j.rmed.2004.02.012

Round, J. L., and Mazmanian, S. K. (2010). Inducible Foxp3+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12204–12209. doi: 10.1073/pnas.0909122107

Said, H. S., Suda, W., Nakagome, S., Chinen, H., Oshima, K., Kim, S., et al. (2014). Dysbiosis of salivary microbiota in inflammatory bowel disease and its association with oral immunological biomarkers. *DNA Res.* 21, 15–25. doi: 10.1093/dnares/dst037

Sokol, H., Seksik, P., Furet, J. P., Firmesse, O., Nion-Larmurier, I., Beaugerie, L., et al. (2009). Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflamm. Bowel. Dis.* 15, 1183–1189. doi: 10.1002/ibd.20903

Sonnenberg, A., and Genta, R. M. (2012). Low prevalence of *Helicobacter* pylori infection among patients with inflammatory bowel disease. *Aliment. Pharmacol. Ther.* 35, 469–476. doi: 10.1111/j.1365-2036.2011.04969.x

Thomas, M., Langella, P., and Neyrolles, O. (2015). Lactobacillus acidophilus: a promising tool for the treatment of inflammatory bowel diseases?. *Med. Sci.* 31, 715–717. doi: 10.1051/medsci/20153108004

Uddin, M., Watz, H., Malmgren, A., and Pedersen, F. (2019). NETopathic inflammation in chronic obstructive pulmonary disease and severe asthma. *Front. Immunol.* 10:47. doi: 10.3389/fimmu.2019.00047

van Nimwegen, F. A., Penders, J., Stobberingh, E. E., Postma, D. S., Koppelman, G. H., Kerkhof, M., et al. (2011). Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. *J. Allergy Clin. Immunol.* 128, 948–955.e1-e3. doi: 10.1016/j.jaci.2011.07.027

Wang, D., Cui, P., and Zhu, W. (2016). "Structural deep network embedding," in *Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM.

Wang, F., Huang, Z. A., Chen, X., Zhu, Z. X., Wen, Z. K., Yan, G. Y., et al. (2017). LRLSHMDA: laplacian regularized least squares for human microbe-disease association prediction. *Sci. Rep.* 7:7601. doi: 10.1038/s41598-017-08127-2

Yadava, K., Pattaroni, C., Sichelstiel, A. K., Trompette, A., Gollwitzer, E. S., Salami, O., et al. (2016). Microbiota promotes chronic pulmonary inflammation by enhancing il-17a and autoantibodies. *Am. J. Respir. Crit. Care Med.* 193, 975–987. doi: 10.1164/rccm.201504-0779OC

Zeng, M., Li, M., Fei, Z., Wu, F., Li, Y., Pan, Y., et al. (2019). A deep learning framework for identifying essential proteins by integrating multiple types of biological information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* PP:1. doi: 10.1109/TCBB.2019.2897679

Zhang, Q., Cox, M., Liang, Z., Brinkmann, F., Cardenas, P. A., Duff, R., et al. (2016). Airway microbiota in severe asthma and relationship to asthma severity and phenotypes. *PloS One* 11:e0152724. doi: 10.1371/journal.pone.0152724

Zhang, W., Yang, W. T., Lu, X. T., Huang, F., and Luo, F. (2018). the bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access.* 6, 38052–38061. doi: 10.1109/ACCESS.2018.2851751

Zong, N., Kim, H., Ngo, V., and Harismendy, O. J. B. (2017). Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics* 33, 2337–2344. doi: 10.1093/bioinformatics/btx160

Zou, S., Zhang, J. P., and Zhang, Z. P. (2017). A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. *PloS One* 12:e0184394. doi: 10.1371/journal.pone.0184394

Zuo, T., and Ng, S. C. (2018). The gut microbiota in the pathogenesis and therapeutics of inflammatory bowel disease. *Front. Microbiol.* 9:2247. doi: 10.3389/fmicb.2018.02247

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.