frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 March 2022
doi: 10.3389/fmicb.2022.834982

Novel Collaborative Weighted Non-negative Matrix Factorization Improves Prediction of Disease-Associated Human Microbes

Da Xu[1], *Hanxiao Xu*[1], *Yusen Zhang*[1]* *and Rui Gao*[2]*

[1] *School of Mathematics and Statistics, Shandong University, Weihai, China,* [2] *School of Control Science and Engineering, Shandong University, Jinan, China*

Extensive clinical and biomedical studies have shown that microbiome plays a prominent role in human health. Identifying potential microbe–disease associations (MDAs) can help reveal the pathological mechanism of human diseases and be useful for the prevention, diagnosis, and treatment of human diseases. Therefore, it is necessary to develop effective computational models and reduce the cost and time of biological experiments. Here, we developed a novel machine learning-based joint framework called CWNMF-GLapRLS for human MDA prediction using the proposed collaborative weighted non-negative matrix factorization (CWNMF) technique and graph Laplacian regularized least squares. Especially, to fuse more similarity information, we calculated the functional similarity of microbes. To deal with missing values and effectively overcome the data sparsity problem, we proposed a collaborative weighted NMF technique to reconstruct the original association matrix. In addition, we developed a graph Laplacian regularized least-squares method for prediction. The experimental results of fivefold and leave-one-out cross-validation demonstrated that our method achieved the best performance by comparing it with 5 state-of-the-art methods on the benchmark dataset. Case studies further showed that the proposed method is an effective tool to predict potential MDAs and can provide more help for biomedical researchers.

Keywords: microbe, disease, association prediction, collaborative weighted non-negative matrix factorization, graph Laplacian regularized least squares

OPEN ACCESS

Edited by:
Qi Zhao,
University of Science and Technology Liaoning, China

Reviewed by:
Wen Zhang,
Huazhong Agricultural University, China
Lihong Peng,
Hunan University of Technology, China

Correspondence:
Yusen Zhang
zhangys@sdu.edu.cn
Rui Gao
gaorui@sdu.edu.cn

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

Received: *14 December 2021*
Accepted: *19 January 2022*
Published: *10 March 2022*

Citation:
Xu D, Xu H, Zhang Y and Gao R (2022) Novel Collaborative Weighted Non-negative Matrix Factorization Improves Prediction of Disease-Associated Human Microbes. Front. Microbiol. 13:834982. doi: 10.3389/fmicb.2022.834982

INTRODUCTION

Extensive clinical and biomedical studies have shown that microbiome has a prominent role in human health and disease. More than 100 trillion (10^{14}) microbes inhabit the human gut and constitute a nutrient-rich environment where symbiotic relationships are of benefit to the host (Ley et al., 2006; Lozupone et al., 2012). Therefore, gut flora is often referred to as the "forgotten organ" (O'Hara and Shanahan, 2006). Once the balance is broken or the symbiotic relationship is disturbed, this close relationship will carry risks for the development of the disease, including cardiovascular disease (Wang et al., 2011), neurological disease (Tremlett et al., 2017), cancer (Schwabe and Jobin, 2013), inflammatory bowel disease (IBD) (Hossen et al., 2020), and so on. To better understand the medical and biological significance of the human microbiome, some large projects have been launched and made substantial progress, such as the project of metagenomics of the human intestinal tract (Ehrlich, 2011; Cho and Blaser, 2012) and the Human Microbiome Project (HMP) (Turnbaugh et al., 2007).

Studies investigating microbiomes demonstrated a critical role for microbes in the disease and health of humans. Considering the complexity and diversity of the microbial community, it is still a challenge to fully understand the interaction mechanism between microorganisms and human diseases, healthy composition, and functional states of the human microbiome. Because of the known disease-related microbes being insufficient, developing effective computational methods is necessary for reducing the cost and time of biological experiments. Recently, with the deepening of studies on computational biology, many computation-based methods have been proposed and achieved successful applications in the bioinformatics field, such as miRNA–disease (Peng et al., 2018a; Chen et al., 2019) or drug–target (Chen et al., 2016) association prediction, and lncRNA–miRNA (Zhang et al., 2021), protein–protein (Xu et al., 2020a), or lncRNA–protein (Peng et al., 2021; Zhou et al., 2021) interaction prediction.

Fortunately, in 2016, a human microbe–disease association database was constructed by Ma et al. (2017). It provided a foundation for identifying potential MDAs through computational methods. A basic assumption is mainly used in the developed methods that microbes will share similar interaction patterns with phenotype diseases if they have similar functions (Zhao et al., 2020). Chen et al. (2017) proposed the first computational model called KATZHMAD for MDA prediction using the KATZ measure. With the rapid development of artificial intelligence and machine learning (Camacho et al., 2018; Xu et al., 2020b), some machine learning-based models were proposed. For instance, Wang et al. (2017) developed the LRLSHMDA method using the Laplacian regularized least squares. In 2021, Xu et al. (2021b) developed a novel prediction model named MDAKRLS using multisimilarity and Kronecker regularized least squares for prediction and achieved better performance. Shi et al. (2018) designed a prediction model by binary matrix completion.

In addition, there are some network-based computational methods. For example, Zou et al. (2017) and Luo and Long (2020) developed BiRWHMDA and NTSHMDA by random walk for prediction only using the Gaussian interaction profile (GIP) kernel similarity, respectively. Recently, several integrated model methods have also been proposed. For example, Huang et al. (2017) built a computational model by combining two single computational methods (graph-based and neighbor-based models). Qu et al. (2019) constructed an integrated model based on label propagation and matrix decomposition. Peng et al. (2020) developed a reliable negative sample selection method based on the random walk with restart and positive unlabeled learning, then used the logistic matrix factorization with neighborhood regularization for prediction. Yin et al. (2020) also designed an integrated method using label propagation and network consistency projection. Some matrix factorization-based computational methods have been proposed to solve microbe–disease association prediction tasks or similar questions. For example, He et al. (2018) designed a graph regularized non-negative matrix factorization (NMF) framework for prediction. In 2020, Gao et al. (2021) developed multilabel fusion collaborative matrix factorization to solve lncRNA–disease association prediction task. In 2021, Xu et al. (2021a) developed regularized NMF and obtained better prediction results in the lncRNA–protein interaction prediction. However, these models may not achieve better prediction results if the dataset is very sparse.

Some existing methods inevitably have certain limitations. For example, some methods used a single similarity that may cause these methods to be biased toward the fully studied diseases or microbes. Besides, constructions of some algorithms contain many artificial parameters, and it is not easy to select the best parameters for a new dataset, which may reduce the robustness of the model. The imbalance problem of the contribution of microbes and diseases needs to be considered since their numbers are different. The benchmark microbe–disease dataset is very sparse; it is essential to weaken the effect caused by the sparse dataset and let known observed data provide more effective information. Effective methods are still scarce since most MDAs remain unknown (Fan et al., 2019; Long et al., 2021). It is necessary to overcome or weaken these limitations and develop new computational methods to improve prediction performance.

In general, from the algebraic view, biological problems of association prediction could be transformed into matrix completion problems. With the rapid development of machine learning, matrix factorization is a useful tool that has been widely used for matrix completion and solving recommendation system problems. In addition, graph regularization-based methods have been successfully applied to semisupervised learning. Considering some limitations of the previous computation-based methods, to improve the prediction performance, we designed a novel method called CWNMF-GLapRLS for MDA prediction. It used the proposed collaborative weighted NMF technique to recover the sparse association matrix and used the developed graph Laplacian regularized least squares for prediction. The experimental results showed our method achieved superior performance. It is an effective tool to predict potential MDAs and can provide more help for biomedical researchers.

MATERIALS AND METHODS

Dataset

In this study, a widely used benchmark dataset (HMDAD) was used in our experiments. It can be downloaded from http://www.cuilab.cn/hmdad, which was collected by Ma et al. (2017). It contains 292 human microbes, 39 diseases, and 483 experimentally confirmed associations. After filtering out repetitive associations, we obtained 450 associations for prediction. The summary of the microbe–disease association dataset is tabulated in **Table 1**.

Overview of the Proposed Method

To predict potential MDAs, we proposed a novel machine learning-based joint framework named CWNMF-GLapRLS based on the collaborative weighted non-negative matrix factorization (CWNMF) and graph Laplacian regularized least squares (GlapRLS). **Figure 1** illustrates the flowchart of the prediction method. It can be decomposed into the following main steps. First, we calculate the functional similarity of microbes through the microbe–disease association network and symptom-based disease similarity. Second, we obtain the GIP kernel similarity based on the topological structure information of the known association matrix, respectively. Third, we calculate the integrated similarities by similarity fusion. Fourth, the proposed CWNMF technique is implemented to reconstruct the association matrix. Finally, we use the designed GlapRLS to score the microbe–disease pairs.

Similarity Measures

For convenience, we set two sets $D = \left\{d_1, d_2, \ldots, d_i, \ldots d_{nd}\right\}$ and $M = \left\{m_1, m_2, \ldots, m_j, \ldots, m_{nm}\right\}$, which represent all diseases and microbes, where nd represents the number of diseases and nm denotes the number of microbes. We constructed a binary matrix $XR^{nd\times nm}$ to represent the microbe–disease association network:

$$X\left(i,j\right) = \begin{cases} 1, & \textit{if disease } d_i \textit{ is associated with microbe } m_j \\ 0, & \textit{otherwise} \end{cases} \tag{1}$$

For disease d_i, its interaction profile is represented by $IP\left(d_i\right)\{0,\ 1\}^{1*nm}$, which denotes the ith row of the binary matrix X. For microbe m_p, its interaction profile is denoted by $IP\left(m_p\right)\{0,\ 1\}^{nd*1}$, which represents the pth column of the binary matrix X.

Symptom-Based Disease Similarity

Some similarity calculation methods of diseases have been proposed using different kinds of disease information. Symptom-based disease similarity has been increasingly demonstrated that it can provide effective information for MDA prediction (Peng et al., 2018b; Zou et al., 2018). In this work, we also introduced symptom-based disease similarity and utilized $S_d^S R^{nd\times nd}$ to represent the similarity matrix. $S_d^S\left(d_i,\ d_j\right)$ represents the similarity between diseases d_i and d_j. More details of the calculation method could be found in a previous study (Zhou et al., 2014). They used a vector of symptoms to represent every disease and used the cosine similarity and term frequency-inverse document frequency (TF-IDF) technique to calculate the similarity of diseases.

TABLE 1 | Summary of microbe–disease association dataset.

Types	Statistical information
Microbes	292
Diseases	39
Associations	450
Sparsity (%)	96.05

Microbe Functional Similarity

In this section, inspired by previous work (Zhang et al., 2018; Li et al., 2019) and the basic assumption that microbes will have similar interaction patterns with phenotype diseases that have similar symptoms, we proposed a method to calculate the functional similarity of microbes through the symptom-based disease similarity and association network.

Firstly, we suppose microbes m_i and m_j are associated with M and N diseases, respectively. Then, set $D_i = \left\{d_{i1}, d_{i2}, \ldots, d_{ip}, \ldots, d_{iM}\right\}$ and $D_j = \left\{d_{j1}, d_{j2}, \ldots, d_{jq}, \ldots, d_{jN}\right\}$ represent two subsets of diseases in the database, in which all diseases are related to the microbe m_i and microbe m_j, respectively. Subsequently, we define the microbe functional similarity as follows:

$$S_m^F\left(m_i,\ m_j\right) = \frac{\sum_{p=1}^{M}\left(\max\limits_{1\le q\le N} S_d^S\left(d_{ip}, d_{jq}\right)\right) + \sum_{q=1}^{N}\left(\max\limits_{1\le p\le M} S_d^S\left(d_{jq}, d_{ip}\right)\right)}{M+N} \tag{2}$$

where S_d^S denotes the symptom-based disease similarity matrix; $\max\limits_{1\le q\le N} S_d^S\left(d_{ip}, d_{jq}\right)$ represents the maximum similarity score between disease d_{ip} and all diseases of subset D_j; S_m^F is defined as the microbe functional similarity matrix.

Gaussian Interaction Profile Kernel Similarity

In this work, symptom-based disease similarity matrix S_d^S and microbe functional similarity matrix S_m^F are both sparse. To integrate more effective information and mine the topology information of known association networks as much as possible, we further introduced popular GIP kernel similarity to calculate the similarity of diseases and microbes (van Laarhoven et al., 2011; Xu et al., 2021b). First, $IP\left(d_i\right)$ of disease d_i and $IP\left(d_j\right)$ of disease d_j were extracted from the training microbe–disease association matrix. Then, we measure the GIP kernel similarity between disease pairs as follows:

$$S_d^G\left(d_i,\ d_j\right) = exp\left(-\sigma_d||IP\left(d_i\right) - IP\left(d_j\right)||^2\right) \tag{3}$$

$$\sigma_d \ = \sigma_d^{'}/\left(\frac{1}{nd}\sum_{k=1}^{nd}||IP\left(d_k\right)||^2\right) \tag{4}$$

where σ_d is a normalized kernel bandwidth and updated through Eq. (4); $\sigma_d^{'}$ is an adjustment coefficient and was set to 1; S_d^G denotes the GIP kernel similarity matrix of diseases.

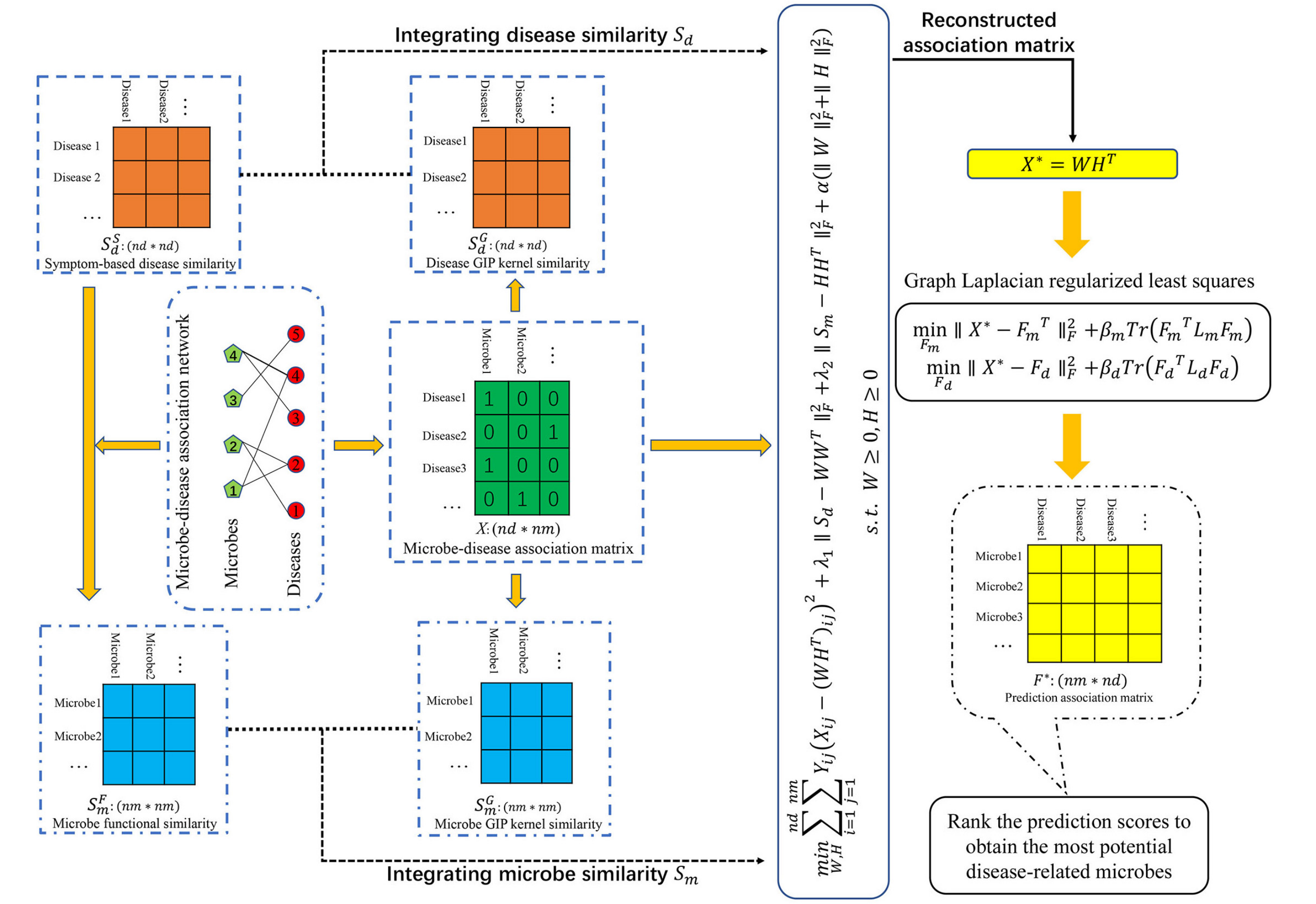
Integrating disease similarity S_d
Disease1
Disease2
Disease 1
Disease 2
$S_d^S: (nd * nd)$
Symptom-based disease similarity
$S_d^G: (nd * nd)$
Disease GIP kernel similarity
Microbe-disease association network
Microbes
Diseases
Microbe1
Microbe2
Disease3
1 0 0
0 0 1
1 0 0
0 1 0
$X: (nd * nm)$
Microbe-disease association matrix
$S_m^F: (nm * nm)$
Microbe functional similarity
$S_m^G: (nm * nm)$
Microbe GIP kernel similarity
Integrating microbe similarity S_m
$\min_{W,H} \sum_{i=1}^{nd}\sum_{j=1}^{nm} Y_{ij}(X_{ij} - (WH^T)_{ij})^2 + \lambda_1 \| S_d - WW^T \|_F^2 + \lambda_2 \| S_m - HH^T \|_F^2 + \alpha(\| W \|_F^2 + \| H \|_F^2)$
$s.t.\ W \geq 0, H \geq 0$
Reconstructed association matrix
$X^* = WH^T$
Graph Laplacian regularized least squares
$\min_{F_m} \| X^* - F_m{}^T \|_F^2 + \beta_m Tr(F_m{}^T L_m F_m)$
$\min_{F_d} \| X^* - F_d \|_F^2 + \beta_d Tr(F_d{}^T L_d F_d)$
Microbe3
$F^*: (nm * nd)$
Prediction association matrix
Rank the prediction scores to obtain the most potential disease-related microbes

FIGURE 1 | The flowchart of CWNMF-GLapRLS framework for prediction.

Similarly, we can calculate the GIP kernel similarity of microbes:

$$S_m^G(m_p,\ m_q) = exp\left(-\sigma_m||IP(m_p) - IP(m_q)||^2\right) \quad (5)$$

$$\sigma_m = \sigma'_m / \left(\frac{1}{nm}\sum_{k=1}^{nm}||IP(m_k)||^2\right) \quad (6)$$

where σ'_m is an adjustment coefficient and was set to 1; σ_m is a normalized kernel bandwidth and updated through Eq. (6); S_m^G represents the microbe GIP kernel similarity matrix.

Integrated Similarities

Multisimilarity fusion is an effective technique that can fuse different feature information and improve performance. However, the microbe functional similarity matrix is sparse; not every microbe has a functional similarity. It may be unreasonable if the integrated similarity is calculated as a mean of functional similarity and GIP kernel similarity. This approach will dilute the GIP kernel similarity of the integrated similarity. To supplement and integrate more effective biological information for microbes, we defined an integrated similarity for microbes. The calculation of similarity between microbes m_p and m_q is defined as follows:

$$S_m(m_p, m_q) = \begin{cases} \frac{S_m^F(m_p,m_q)+S_m^G(m_p,m_q)}{2}, & if\ S_m^F(m_p, m_q) \neq 0 \\ S_m^G(m_p, m_q), & otherwise \end{cases} \quad (7)$$

where $S_m R^{nm\times nm}$ denotes the integrated microbe similarity matrix. Specifically, the final similarity will be calculated as a mean if the microbe pair has a functional similarity. Otherwise, the GIP kernel similarity will be assigned to the integrated similarity.

Similarly, the integrated similarity calculation method of diseases d_i and d_j is defined as follows:

$$S_d(d_i,\ d_j) = \begin{cases} \frac{S_d^S(d_i,\ d_j)+S_d^G(d_i,\ d_j)}{2}, & if\ S_d^S(d_i,\ d_j) \neq 0 \\ S_d^G(d_i,\ d_j), & otherwise \end{cases} \quad (8)$$

where $S_d R^{nd\times nd}$ denotes the integrated disease similarity matrix.

Collaborative Weighted Non-negative Matrix Factorization

In general, to recover the association matrix, we could transform this biological problem into a recommendation task. NMF enforced non-negativity constraints on factor matrixes for a low-rank approximation of the non-negative matrix (Lee and Seung, 1999), which could ensure that every element can be represented as an additive linear combination of canonical coordinates. Microbe–disease binary association data X is a non-negative matrix. We could use the NMF for matrix completion or association prediction.

In this work, microbe–disease association data X is incomplete and sparse. To deal with missing values and effectively overcome the data sparsity problem, we introduced weighted non-negative matrix factorization (WNMF), which slightly changed classical NMF by introducing a weighting term. WNMF was first proposed to cope with missing values in large-scale networks for predicting and representing distances (Mao and Saul, 2004) and has been used for recommendation systems (Gu et al., 2010) to solve the incomplete data problem. The biological problem can be translated into minimizing the following objective:

$$J=\sum_{i=1}^{nd}\sum_{j=1}^{nm}Y_{ij}\left(X_{ij}-\left(WH^{\mathrm{T}}\right)_{ij}\right)^{2} \tag{9}$$

$$s.t.\quad W\geq 0,\ H\geq 0$$

where $XR^{nd\times nm}$ are the training association data; the product of non-negative matrices $WR^{nd\times k}$ and $HR^{nm\times k}$ is the best approximation of X, $k\ll min\left\{nd,\ nm\right\}$. Microbes and diseases are mapped into a shared latent space with a low-dimensionality k. Y is a non-negative weight matrix used to reduce the influence of missing values on matrix factorization, where $Y_{ij}=0$ indicates X_{ij} is a missing value and $Y_{ij}=1$ indicates X_{ij} is an observed value. The objective function will degenerate into the standard NMF when all weights of matrix Y are equal to one.

In 2000, Lee and Seung (2001) have shown that the iterative update algorithm can ensure NMF objective function convergence and is very easy to use and code. At the same time, an iterative multiplicative updating algorithm was also used to solve WNMF (Zhang et al., 2006). The objective function leads to the following updated formulas:

$$w_{ik}=w_{ik}\frac{(Y\odot XH)_{ik}}{\left(Y\odot\left(WH^{\mathrm{T}}\right)H\right)_{ik}} \tag{10}$$

$$h_{jk}=h_{jk}\frac{\left((Y\odot X)^{\mathrm{T}}W\right)_{jk}}{\left(\left(Y\odot\left(WH^{\mathrm{T}}\right)\right)^{\mathrm{T}}W\right)_{jk}} \tag{11}$$

where $\odot$ is the Hadamard product. These updated rules are computationally efficient.

In 2021, Xu et al. (2021a) developed regularized NMF and obtained better prediction results in the lncRNA–protein interaction prediction. This study proved that collaborative factorization of the similarity matrix can effectively guide matrix factorization and improve prediction performance. To introduce more effective similarity information to guide the matrix factorization, two collaborative regularization terms were incorporated into the WNMF framework to fuse similarity information and constrain two low-dimensional representations. It can be turned into a constrained optimization problem and formulated a joint matrix factorization framework of association data and similarity data. Then, we can obtain a novel objective function as follows:

$$\begin{aligned}J=&\sum_{i=1}^{nd}\sum_{j=1}^{nm}Y_{ij}\left(X_{ij}-\left(WH^{\mathrm{T}}\right)_{ij}\right)^{2}\\&+\lambda_1\parallel S_d-WW^{\mathrm{T}}\parallel_F^2+\lambda_2\parallel S_m-HH^{\mathrm{T}}\parallel_F^2\\&s.t.\quad W\geq 0,\ H\geq 0\end{aligned} \tag{12}$$

where $||\cdot||_F$ is the Frobenius norm; λ_1 and λ_2 are non-negative regularization parameters balancing two collaborative regularization terms and the reconstruction error. The objective function will degenerate into WNMF if λ_1 and λ_2 are equal to zero.

To prevent overfitting and adjust the smoothness of W and H, we introduced the Tikhonov (L_2) regularization terms (Xiao et al., 2018) into the objective function and obtained the final collaborative weighted non-negative matrix factorization (CWNMF) objective function as follows:

$$\begin{aligned}J=&\sum_{i=1}^{nd}\sum_{j=1}^{nm}Y_{ij}\left(X_{ij}-\left(WH^{\mathrm{T}}\right)_{ij}\right)^{2}\\&+\lambda_1\parallel S_d-WW^{\mathrm{T}}\parallel_F^2+\lambda_2\parallel S_m-HH^{\mathrm{T}}\parallel_F^2\\&+\alpha\left(\parallel W\parallel_F^2+\parallel H\parallel_F^2\right)\\&s.t.\quad W\geq 0,\ H\geq 0\end{aligned} \tag{13}$$

where α is used to adjust the Tikhonov regularization terms, which is a regularization coefficient. To improve the robustness of the model, we set the same value for the same Tikhonov regularization terms, and α was set to 1 for the dataset.

Since the objective function is not convex in both variables W and H, the iterative update algorithm was used to search the local minimum. Here, we used the Lagrange multipliers method and Karush–Kuhn–Tucker (KKT) conditions to optimize the objective function. Eventually, we obtained the following multiplicative updates:

$$w_{ik}=w_{ik}\frac{(Y\odot XH+2\lambda_1 S_d W)_{ik}}{\left(Y\odot\left(WH^{\mathrm{T}}\right)H+\alpha W+2\lambda_1 WW^{\mathrm{T}}W\right)_{ik}} \tag{14}$$

$$h_{jk}=h_{jk}\frac{\left((Y\odot X)^{\mathrm{T}}W+2\lambda_2 S_m H\right)_{jk}}{\left(\left(Y\odot\left(WH^{\mathrm{T}}\right)\right)^{\mathrm{T}}W+\alpha H+2\lambda_2 HH^{\mathrm{T}}H\right)_{jk}} \tag{15}$$

Then, we can obtain the reconstructed association matrix $X^{*}=WH^{\mathrm{T}}$. The low-dimensionality representation k was set as 35 in the process of prediction.

Graph Laplacian Regularized Least Squares

In this section, to improve the prediction performance, we developed a semisupervised learning method named graph Laplacian regularized least squares based on the reconstructed association matrix X^*. Graph regularization is used to fully exploit data geometric structure for semisupervised learning. Specifically, in the prediction space of microbes, with the above defined integrated microbe similarity matrix S_m, the graph Laplacian regularization term was incorporated into the least-squares framework to enhance the learning performance. The optimization problem can be formularized as follows:

$$\min_{F_m} \parallel X^* - F_m^{\mathrm{T}} \parallel_F^2 + \beta_m \frac{1}{2}\left(\sum_{i,j=1}^{nm} \parallel F_{mi} - F_{mj} \parallel^2 S_{m_{ij}}\right) \tag{16}$$

where $X^* R^{nd\times nm}$ is a reconstructed association matrix obtained by the CWNMF method; β_m is the regularization coefficient; F_m is the prediction score matrix based on the microbes; F_{mi} denotes the ith row of $F_m \in R^{nm\times nd}$; and F_{mj} denotes the jth row of F_m. The graph Laplacian regularization term (Xiao et al., 2018; Cai et al., 2020) can be transformed into a matrix form by some algebraic manipulations:

$$\frac{1}{2}\left(\sum_{i,j=1}^{nm} \parallel F_{mi} - F_{mj} \parallel^2 S_{m_{ij}}\right) = Tr\left(F_m^{\mathrm{T}} L_m F_m\right) \tag{17}$$

where $Tr\,(?)$ denotes the trace of a matrix; $L_m = D_m - S_m$ is the graph Laplacian matrix for S_m. D_m is the diagonal matrix whose entries are calculated as the column sums of S_m. Therefore, Eq. (16) can be transformed into the following equation:

$$\min_{F_m} \parallel X^* - F_m^{\mathrm{T}} \parallel_F^2 + \beta_m Tr\left(F_m^{\mathrm{T}} L_m F_m\right) \tag{18}$$

where $F_m = S_m \alpha_m$, $\alpha_m \in R^{nm\times nd}$ is a matrix (Xia et al., 2010). To improve the robustness of the model and according to the choice of previous similar work (van Laarhoven et al., 2011), β_m was set to 1. We can obtain the solution of the optimization problem by some manipulations, $\alpha_m^* = (S_m + L_m S_m)^{-1} X^{*\mathrm{T}}$. Then, in the microbe prediction space, the prediction score matrix can be calculated as follows:

$$F_m = S_m\,(S_m + L_m S_m)^{-1}\, X^{*\mathrm{T}} \tag{19}$$

Similarly, for disease prediction space, the optimization problem can be formularized as the following equation:

$$\min_{F_d} \parallel X^* - F_d \parallel_F^2 + \beta_d Tr\left(F_d^{\mathrm{T}} L_d F_d\right) \tag{20}$$

where β_d was also set to 1. We can obtain the prediction score matrix in the disease prediction space.

$$F_d = S_d\,(S_d + L_d S_d)^{-1}\, X^* \tag{21}$$

Finally, the predicted microbe–disease association matrix is calculated as $F^* = \eta F_m^{\mathrm{T}} + (1-\eta)\,F_d$, where η is a tradeoff parameter describing the importance of microbe and disease space. The microbe-related diseases can be prioritized by the size of the prediction scores in matrix F^*. The detailed steps of the CWNMF-GlapRLS procedure are detailed in **Algorithm 1**.

Algorithm 1 | CWNMF-GlapRLS Algorithm.

```
Input: Matrices $XR^{nd\times nm}$, $S_dR^{nd\times nd}$ and $S_mR^{nm\times nm}$; non-negative weight matrix $YR^{nd\times nm}$; regularization coefficients $\lambda_1$ and $\lambda_2$; tradeoff parameter $\eta$.
Output: Predicted score matrix $F^*$.
Randomly initialize two non-negative matrices $WR^{nd\times k}$ and $HR^{nm\times k}$.
Repeat
Update $W$ and $H$ by the following rules:
$w_{ik} = w_{ik}\frac{(Y\odot XH+2\lambda_1 S_d W)_{ik}}{(Y\odot(WH^{\mathrm{T}})H+\alpha W+2\lambda_1 WW^{\mathrm{T}}W)_{ik}}$
  $h_{jk} = h_{jk}\frac{\left((Y\odot X)^{\mathrm{T}}W+2\lambda_2 S_m H\right)_{jk}}{\left((Y\odot(WH^{\mathrm{T}}))^{\mathrm{T}}W+\alpha H+2\lambda_2 HH^{\mathrm{T}}H\right)_{jk}}$
Until convergence
Reconstruct association matrix $X^* = WH^{\mathrm{T}}$.
Calculate diagonal matrix $D_m$;
$L_m = D_m - S_m$;
$F_m = S_m\,(S_m + L_m S_m)^{-1}\,X^{\mathrm{T}}$//calculate the score matrix $F_m$ based on the microbe prediction space.
Calculate diagonal matrix $D_d$;
$L_d = D_d - S_d$;
$F_d = S_d\,(S_d + L_d S_d)^{-1}\,X$//calculate the score matrix $F_d$ based on the disease prediction space.
Return $F^* = \eta F_m^{\mathrm{T}} + (1-\eta)\,F_d$.
```

RESULTS

Evaluation Metrics

To ensure the reliability of experimental results, we implemented the global leave-one-out cross-validation (LOOCV) framework to validate the performance of models (Bao et al., 2017). In each round cross-validation of the LOOCV framework, the integrated similarity of diseases and microbes should be recalculated, which can guarantee independence between the validation set and the training set. Specifically, under this framework, every known microbe–disease pair will be regarded as a test set, the rest of the known pairs are treated as the training set in the dataset, and all pairs without observed association are used as candidate samples. We calculated the predicted microbe–disease score matrix by running the model. Then, the prediction score is compared with all candidate samples to get the ranking of each test sample. This testing sample will be regarded as a successful prediction if the rank is higher than the threshold. We used the receiver operating characteristic (ROC) curve to vividly describe the performance of the model by calculating sensitivity (true positive rates) and 1-specificity (false positive rates) with different thresholds. In addition, we calculated the area under curve (AUC) to intuitively describe the performance. Similarly, fivefold cross-validation (CV) was also applied to evaluate the effectiveness of the models. The experiment was repeatedly performed 10 times to reduce potential bias caused by random segmentation of the dataset. At the same time, the

ROC curves and average AUC values were also obtained under the fivefold CV framework.

Parameter Sensitivity and Model Setting

It is necessary to evaluate the influence of model parameters on the prediction performance of CWNMF-GLapRLS. We studied the influence of two regularization parameters λ_1 and λ_2. The grid search method was adopted to find better model parameters. In the experiments, we first tuned the range of two parameters from 0 to 0.5, and each step is 0.01. Then, the proposed method was run to find the optimal model parameter values based on the AUC values on the 50×50 grid. **Figure 2A** shows the relationship between the AUC value and the parameter pair (λ_1, λ_2) under the fivefold framework. Finally, we selected the parameter pair of (0.02, 0.04) as the optimal value of (λ_1, λ_2) based on the grid search results under the two evaluation frameworks. Then, we fixed the parameter pair and adjusted the parameter η. The effects between parameter η and the AUC value are shown in **Figure 2B**. Finally, η was set at 0.15 as the optimal value for the following analysis.

Iterative update algorithm can ensure objective function convergence and guarantee to converge to a locally optimal. **Figure 3** shows the objective function convergence curve of CWNMF. From the figure, we can see that the convergence is fast, and the objective function value decreases as the iterations. The number of iterations is usually very small (fewer than 100) before practical convergence. Thus, the proposed method can scale to larger datasets. Finally, the number of iterations was set at 300 in the process of prediction.

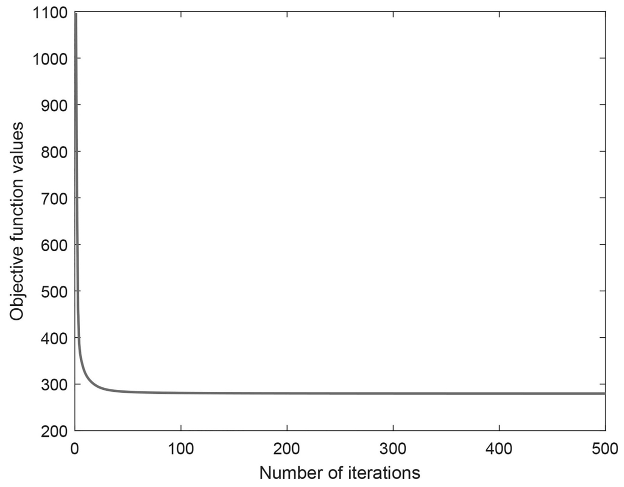

FIGURE 3 | Convergence behavior of CWNMF objection function.

Performance Analysis

Here, we compared five different forms (proposed method, proposed without microbe functional similarity, proposed method without weight, proposed method without GLapRLS, and proposed method without CWNMF) of the introduced method to analyze the proposed method. Especially, to improve the prediction performance and fuse more similarity information, we calculated microbe functional similarity. To deal with missing values and effectively overcome the data sparsity problem, we introduced WNMF, which slightly changed classical NMF by introducing a weighting term, and proposed the technique CWNMF for recovering the association matrix. The proposed method is a joint framework. The CWNMF technique was first used to recover the original matrix; then, the GLapRLS method was used for prediction. **Figure 4** shows the performance comparison of methods with different forms on the HMDAD dataset. The proposed method performs better than the other four methods. From the figure, we can obtain that the combination of CWNMF and GLapRLS can significantly improve the prediction performance. The comparison results indicate that microbe functional similarity

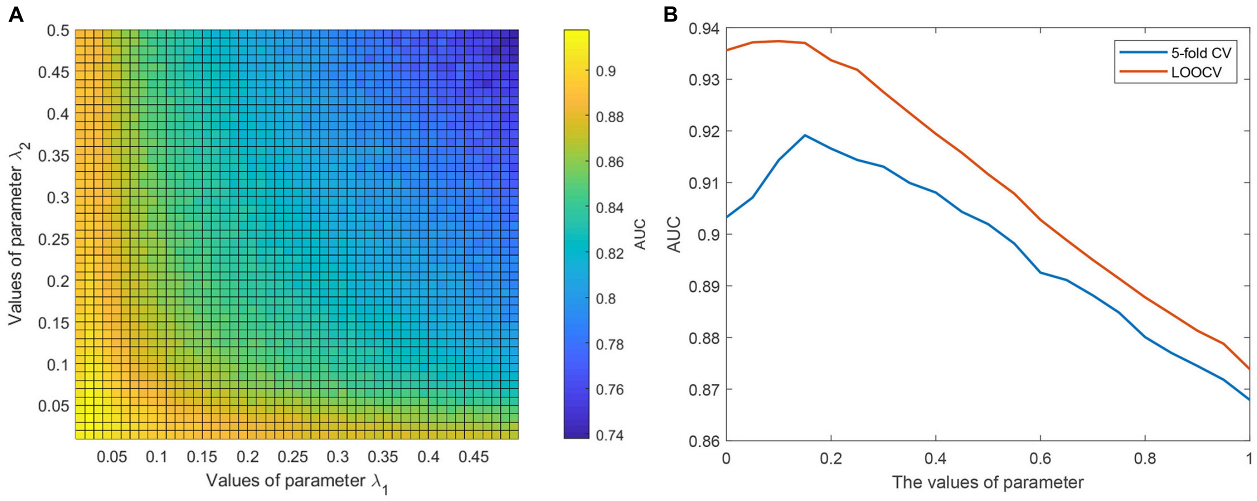

FIGURE 2 | (A) The illustration of determining the optimal values of parameter pair (λ_1, λ_2) under grid search. **(B)** The effects between parameter η and AUC value.

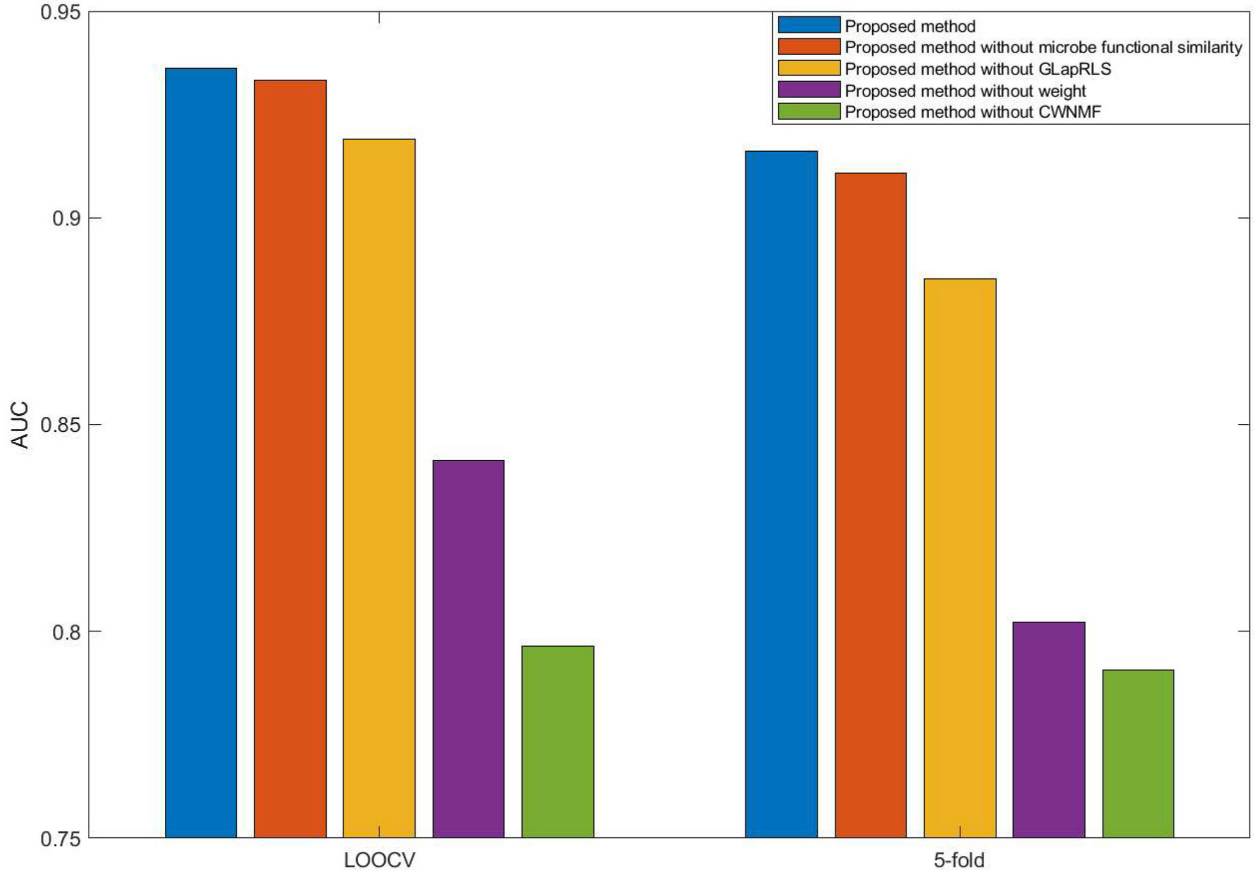

FIGURE 4 | The performance comparison of different methods.

and weighting term are also effective in improving the performance of prediction.

Comparison With State-of-the-Art Prediction Methods

In this section, to evaluate the effectiveness of the proposed method, we compared it with 5 state-of-the-art methods, including graph regularized non-negative matrix factorization (GRNMFHMDA) (He et al., 2018), KATZ measure (KATZHMDA) (Chen et al., 2017), bi-random walk (BiRWHMDA) (Zou et al., 2017), Laplacian regularized least squares (LRLSHMDA) (Wang et al., 2017), and network topological similarity (NTSHMDA) (Luo and Long, 2020) for human MDA prediction methods. Optimal parameter combinations for 5 comparison methods are listed in **Supplementary Table 1**.

First, under the LOOCV framework, the ROC curves and AUC values of six methods have been shown in **Figure 5**. From the figure, we can see that the proposed method outperforms other methods with an AUC of 0.9362 under the LOOCV framework, while GRNMFHMDA, KATZHMDA, LRLSHMDA, BiRWHMDA, and NTSHMDA obtained AUC values of 0.8719, 0.8382, 0.8916, 0.8964, and 0.9040, respectively. In addition, the ROC curves and average AUC values of six methods under the fivefold CV framework have been shown in **Figure 6**. We can see that the proposed method is more outstanding than other methods with an AUC of 0.9161 under the fivefold CV framework, while GRNMFHMDA, KATZHMDA, LRLSHMDA, BiRWHMDA, and NTSHMDA obtained AUC values of 0.8555, 0.8324, 0.8809, 0.8839, and 0.8918, respectively. These experimental results proved that our method is effective and reliable, and may be an effective tool for seeking potential disease-related microbes.

Case Studies

Accumulating evidence has shown that the development and occurrence of human disease are closely related to the imbalance of the microbial community. To infer potential association, in this section, case studies were implemented on two different common human diseases (asthma and IBD). In this way, we used the number of validated predicted microbes of the top 15 prediction results to further measure the predictive capability, respectively. If the genus of a microbe is related to the disease, this microbe will be related to the disease. This assumption has been widely used in related studies (Niu et al., 2019; Wang et al., 2019). Specifically, for a given disease, all pairs without

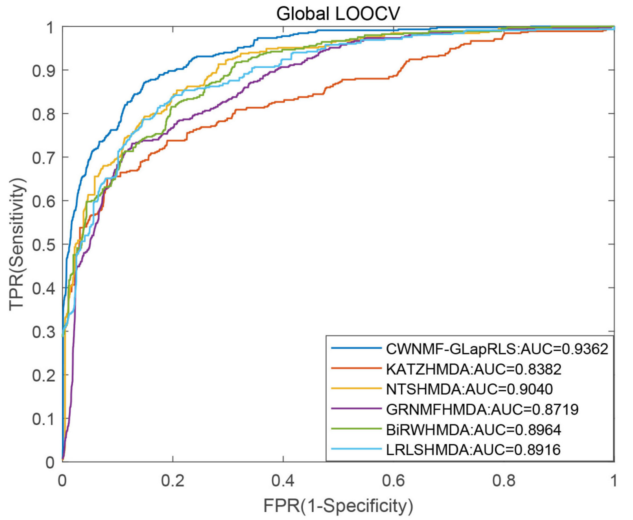

FIGURE 5 | The ROC curves and AUC values of six methods under LOOCV framework.

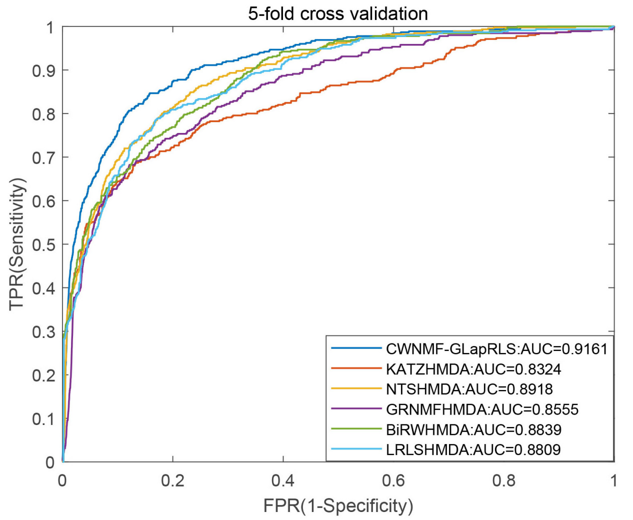

FIGURE 6 | The ROC curves and average AUC values of six methods under fivefold CV framework.

observed association were regarded as candidate samples. We calculated the association scores for all microbes based on the joint framework. All candidate microbe samples were prioritized based on their scores.

Asthma is a common chronic inflammatory disease, which affects the daily lives of 300 million people worldwide (Lambrecht and Hammad, 2015). To investigate asthma-causing microbes, the prediction results have been tabulated in **Table 2**. There are 13 out of the top 15 candidate microbes that have been successfully supported to be associated with asthma based on previously published medical or biological literature. According to the table, our method has an excellent effect. Increasing evidence has shown that the development and occurrence of human asthma are closely related to the imbalance of the microbial community. For example, some clinical evidence has shown that asthmatic patients have lower Actinobacteria, Firmicutes, and Bacteroides proportions (Björkstén et al., 1999; Marri et al., 2013). The colonization by *Clostridium coccoides* subcluster XIVa species at age 3 weeks may serve as an early indicator of possible asthma (Vael et al., 2011). In addition, colonization by *Clostridium difficile* at age 1 month was closely associated with asthma at 6–7 years old (Van Nimwegen et al., 2011). One study showed that *Streptococcus* increases the risk of asthma by early asymptomatic colonization (Teo et al., 2015). *Lactobacillus* has been shown to be beneficial to asthmatic children (Huang et al., 2018).

TABLE 2 | Prediction results of the top 15 asthma-associated microbes.

Rank	Microbe	Evidence
1	*Firmicutes*	PMID:23265859
2	*Clostridium coccoides*	PMID:21477358
3	*Actinobacteria*	PMID:26220531
4	*Clostridia*	Unconfirmed
5	*Bacteroides*	PMID:10202341
6	*Clostridium difficile*	PMID:21872915
7	*Lactobacillus*	PMID:30400588
8	*Bifidobacterium*	PMID:24735374
9	*Lachnospiraceae*	PMID:31958431
10	*Veillonella*	PMID:26424567
11	*Streptococcus*	PMID:25865368
12	*Staphylococcus aureus*	PMID:25533526
13	*Fusobacterium nucleatum*	Unconfirmed
14	*Faecalibacterium prausnitzii*	PMID:30208875
15	*Fusobacterium*	PMID:27838347

IBD starts with inflammation and is a collective term for a wide range of intestinal diseases, which is a worldwide healthcare problem (Hossen et al., 2020). IBD has become one of the most studied human diseases linked to gut microbiota (Kostic et al., 2014). We listed the top 15 IBD-associated microbes in **Table 3**. As a result, 14 out of the top 15 candidate microbes

TABLE 3 | Prediction results of the top 15 IBD-associated microbes.

Rank	Microbe	Evidence
1	*Bacteroidetes*	PMID:25307765
2	*Prevotella*	PMID:25307765
3	*Firmicutes*	PMID:25307765
4	*Clostridium coccoides*	PMID:19235886
5	*Helicobacter pylori*	PMID:22221289
6	*Bacteroides*	PMID:25307765
7	*Clostridia*	PMID:31142855
8	*Haemophilus*	PMID:24013298
9	*Clostridium difficile*	PMID:24838421
10	*Lactobacillus*	PMID:24478468
11	*Bifidobacterium*	PMID:24478468
12	*Veillonella*	PMID:24013298
13	*Staphylococcus aureus*	PMID:19809406
14	*Staphylococcus*	Unconfirmed
15	*Faecalibacterium prausnitzii*	PMID:32815163

have been successfully validated to be associated with the IBD based on published literature. Emerging evidence showed that many microbes are closely related to IBD. For example, the infection of *Clostridium difficile* is a significant clinical challenge for IBD patients, which can result in morbidity and mortality (Hashash and Binion, 2014). Some studies showed *Bacteroidetes, Bacteroides, Firmicutes*, and *Prevotella* are associated with the development of IBD (Juste et al., 2014; Walters et al., 2014). In IBD patients, *Prevotella*, *Veillonella*, and *Haemophilus* were found, which can contribute largely to dysbiosis, which is associated with inflammatory responses (Said et al., 2014). The study confirmed that *Helicobacter pylori* was inversely associated with IBD (Sonnenberg and Genta, 2012). In addition, *Veillonella* and *Bifidobacterium* decreased, while the proportion of *Lactobacillus* increased in the feces of IBD patients (Takaishi et al., 2008). Case studies indicated that our method has a practical effect on potential association prediction.

CONCLUSION AND DISCUSSION

Studies investigating microbiomes demonstrated a critical role for microbes in human health and disease. Identifying potential disease-related microbes is essential for understanding the mechanisms of host–microbe interactions and revealing the pathological mechanism of human diseases. Here, we designed a joint framework for association prediction based on the proposed CWNMF and graph Laplacian regularized least squares. The experimental results showed that our method achieved the best performance by comparing it with 5 state-of-the-art models. Case studies of asthma and IBD also further demonstrated that the proposed method is a useful tool to infer potential associations. All experimental results adequately demonstrated that the proposed method has reliable and effective prediction performance.

There are several key factors that make the proposed method have effective performance. Firstly, compared with graph regularized NMF and collaborative matrix factorization, we introduced a weighting term and changed the NMF for prediction to deal with missing values and weaken the effect caused by a sparse dataset. Secondly, we calculated the functional similarity of microbes and introduced symptom-based disease similarity for fusing more similarity information. Thirdly, to restructure the sparse association matrix, two collaborative regularization terms were incorporated into the framework to fuse similarity information and constrain two low-dimensional representations, guiding the matrix factorization process. We used the iterative update algorithm to solve the matrix factorization objective function, which is easy to use and code. Semisupervised learning provides more effective information in the process of prediction. We hope that the proposed method can help biomedical researchers conduct follow-up research, and a growing number of potential disease-related microbes could be verified through biological or clinical experiments.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

DX: methodology, software, formal analysis, and writing—original draft, writing—review and editing. HX: data curation, methodology, software, and writing—original draft. YZ: supervision, funding acquisition, funding acquisition, and writing—review and editing. RG: formal analysis, supervision, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been supported by the National Natural Science Foundation of China under (Grant Nos. 61877064, U1806202, and 61533011).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.834982/full#supplementary-material

REFERENCES

Bao, W., Jiang, Z., and Huang, D. S. (2017). Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinform.* 18:543. doi: 10.1186/s12859-017-1968-2

Björkstén, B., Naaber, P., Sepp, E., and Mikelsaar, M. (1999). The intestinal microflora in allergic Estonian and Swedish 2-year-old children. *Clin. Exp. Allergy* 29, 342–346. doi: 10.1046/j.1365-2222.1999.00560.x

Cai, D., He, X., Han, J., and Huang, T. S. (2020). Graph Regularized Nonnegative Matrix Factorization for Data Representation. *Appl. Intell.* 50, 438–447. doi: 10.1007/s10489-019-01539-9

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell* 173, 1581–1592. doi: 10.1016/j.cell.2018.05.015

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

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2019). MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* 20, 515–539. doi: 10.1093/bib/bbx130

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016). Drug-target interaction prediction: Databases, web servers and computational models. *Brief. Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Cho, I., and Blaser, M. J. (2012). The human microbiome: At the interface of health and disease. *Nat. Rev. Genet.* 13, 260–270. doi: 10.1038/nrg3182

Ehrlich, S. D. (2011). “MetaHIT: The European Union Project on Metagenomics of the Human Intestinal”. In: Nelson K. (eds) Metagenomics of the Human Body. (New York, NY: Springer)

Fan, C., Lei, X., Guo, L., and Zhang, A. (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

Gao, M. M., Cui, Z., Gao, Y. L., Wang, J., and Liu, J. X. (2021). Multi-Label Fusion Collaborative Matrix Factorization for Predicting LncRNA-Disease Associations. *IEEE J. Biomed. Heal. Informatics* 25, 881–890. doi: 10.1109/JBHI.2020.2988720

Gu, Q., Zhou, J., and Ding, C. (2010). Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. *Proc. 10th SIAM Int. Conf. Data Mining SDM* 2010, 199–210. doi: 10.1137/1.9781611972801.18

Hashash, J. G., and Binion, D. G. (2014). Managing Clostridium difficile in Inflammatory Bowel Disease (IBD). *Curr. Gastroenterol. Rep.* 16, 14–19. doi: 10.1007/s11894-014-0393-1

He, B. S., Peng, L. H., and Li, Z. (2018). Human microbe-disease association prediction with graph regularized non-negative matrix factorization. *Front. Microbiol.* 9:2560. doi: 10.3389/fmicb.2018.02560

Hossen, I., Hua, W., Ting, L., Mehmood, A., Jingyi, S., Duoxia, X., et al. (2020). Phytochemicals and inflammatory bowel disease: a review. *Crit. Rev. Food Sci. Nutr.* 60, 1321–1345. doi: 10.1080/10408398.2019.1570913

Huang, C. F., Chie, W. C., and Wang, I. J. (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

Juste, C., Kreil, D. P., Beauvallet, C., Guillot, A., Vaca, S., Carapito, C., et al. (2014). Bacterial protein signals are associated with Crohn's disease. *Gut* 63, 1566–1577. doi: 10.1136/gutjnl-2012-303786

Kostic, A. D., Xavier, R. J., and Gevers, D. (2014). The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology* 146, 1489–1499. doi: 10.1053/j.gastro.2014.02.009

Lambrecht, B. N., and Hammad, H. (2015). The immunology of asthma. *Nat. Immunol.* 16, 45–56. doi: 10.1038/ni.3049

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Lee, D. D., and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* 13, 1–7.

Ley, R. E., Peterson, D. A., and Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124, 837–848. doi: 10.1016/j.cell.2006.02.017

Li, G., Luo, J., Liang, C., Xiao, Q., Ding, P., and Zhang, Y. (2019). Prediction of LncRNA-Disease Associations Based on Network Consistency Projection. *IEEE Access* 7, 58849–58856. doi: 10.1109/ACCESS.2019.2914533

Long, Y., Luo, J., Zhang, Y., and Xia, Y. (2021). Predicting human microbe-disease associations via graph attention networks with inductive matrix completion. *Brief. Bioinform.* 22, 1–13. doi: 10.1093/bib/bbaa146

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489, 220–230. doi: 10.1038/nature11550

Luo, J., and Long, Y. (2020). NTSHMDA: Prediction of Human Microbe-Disease Association Based on Random Walk by Integrating Network Topological Similarity. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 17, 1341–1351. doi: 10.1109/TCBB.2018.2883041

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Mao, Y., and Saul, L. K. (2004). Modeling distances in large-scale networks by matrix factorization. *Proc. 2004 ACM SIGCOMM Internet Meas. Conf. IMC* 2004, 278–287. doi: 10.1145/1028788.1028827

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.e–352.e. doi: 10.1016/j.jaci.2012.11.013

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, 102–300. doi: 10.3389/fmicb.2019.01578

O'Hara, A. M., and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep.* 7, 688–693. doi: 10.1038/sj.embor.7400731

Peng, L., Shen, L., Liao, L., Liu, G., and Zhou, L. (2020). RNMFMDA: A Microbe-Disease Association Identification Method Based on Reliable Negative Sample Selection and Logistic Matrix Factorization With Neighborhood Regularization. *Front. Microbiol.* 11:592430. doi: 10.3389/fmicb.2020.592430

Peng, L., Wang, C., Tian, X., Zhou, L., and Li, K. (2021). Finding lncRNA-protein Interactions Based on Deep Learning with Dual-net Neural Architecture. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* doi: 10.1109/TCBB.2021.3116232 [Epub ahead of print].

Peng, L. H., Sun, C. N., Guan, N. N., Li, J. Q., and Chen, X. (2018a). HNMDA: heterogeneous network-based miRNA–disease association prediction. *Mol. Genet. Genomics* 293, 983–995. doi: 10.1007/s00438-018-1438-1

Peng, L. H., Yin, J., Zhou, L., Liu, M. X., and Zhao, Y. (2018b). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

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

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

Schwabe, R. F., and Jobin, C. (2013). The microbiome and cancer. *Nat. Rev. Cancer* 13, 800–812. doi: 10.1038/nrc3610

Shi, J. Y., Huang, H., Zhang, Y. N., Cao, J. B., and Yiu, S. M. (2018). BMCMDA: A novel model for predicting human microbe-disease associations via binary matrix completion. *BMC Bioinform.* 19:281. doi: 10.1186/s12859-018-2274-3

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

Takaishi, H., Matsuki, T., Nakazawa, A., Takada, T., Kado, S., Asahara, T., et al. (2008). Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. *Int. J. Med. Microbiol.* 298, 463–472. doi: 10.1016/j.ijmm.2007.07.016

Teo, S. M., Mok, D., Pham, K., Kusel, M., Serralha, M., Troy, N., et al. (2015). The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 17, 704–715. doi: 10.1016/j.chom.2015.03.008

Tremlett, H., Bauer, K. C., Appel-Cresswell, S., Finlay, B. B., and Waubant, E. (2017). The gut microbiome in human neurological disease: A review. *Ann. Neurol.* 81, 369–382. doi: 10.1002/ana.24901

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Vael, C., Vanheirstraeten, L., Desager, K. N., and Goossens, H. (2011). Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol.* 11:68. doi: 10.1186/1471-2180-11-68

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

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.e–955.e. doi: 10.1016/j.jaci.2011.07.027

Walters, W. A., Xu, Z., and Knight, R. (2014). Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett.* 588, 4223–4233. doi: 10.1016/j.febslet.2014.09.039

Wang, F., Huang, Z. A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., 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

Wang, L., Wang, Y., Li, H., Feng, X., Yuan, D., and Yang, J. (2019). A bidirectional label propagation based computational model for potential microbe-disease association prediction. *Front. Microbiol* 10:684. doi: 10.3389/fmicb.2019.00684

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., Dugar, B., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 472, 57–65. doi: 10.1038/nature09922

Xia, Z., Wu, L. Y., Zhou, X., and Wong, S. T. C. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* 4:S6. doi: 10.1186/1752-0509-4-6

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Xu, D., Xu, H., Zhang, Y., Chen, W., and Gao, R. (2020a). Protein-Protein Interactions Prediction Based on Graph Energy and Protein Sequence Information. *Molecules* 25:1841.

Xu, D., Zhang, J., Xu, H., Zhang, Y., Chen, W., Gao, R., et al. (2020b). Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data. *BMC Genomics* 21:650. doi: 10.1186/s12864-020-07038-3

Xu, D., Xu, H., Zhang, Y., Chen, W., and Gao, R. (2021a). LncRNA-protein interaction prediction based on regularized nonnegative matrix factorization and sequence information. *Match* 85, 555–574.

Xu, D., Xu, H., Zhang, Y., Wang, M., Chen, W., and Gao, R. (2021b). MDAKRLS: Predicting human microbe-disease association based on Kronecker regularized least squares and similarities. *J. Transl. Med.* 19:66. doi: 10.1186/s12967-021-02732-6

Yin, M.-M., Liu, J.-X., Gao, Y.-L., Kong, X.-Z., and Zheng, C.-H. (2020). NCPLP: A Novel Approach for Predicting Microbe-Associated Diseases With Network Consistency Projection and Label Propagation. *IEEE Trans. Cybern.* doi: 10.1109/tcyb.2020.3026652 [Epub ahead of print].

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using Network Distance Analysis to Predict lncRNA–miRNA Interactions. *Interdiscip. Sci. Comput. Life Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhang, S., Wang, W., Ford, J., and Makedon, F. (2006). Learning from incomplete ratings using non-negative matrix factorization. *Proc. Sixth SIAM Int. Conf. Data Min.* 2006, 549–553. doi: 10.1137/1.9781611972764.58

Zhang, W., Yang, W., Lu, X., 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

Zhao, Y., Wang, C.-C., and Chen, X. (2020). Microbes and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 22:bbaa158. doi: 10.1093/bib/bbaa158

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA–protein interaction identification. *BMC Bioinform.* 22:479. doi: 10.1186/s12859-021-04399-8

Zhou, X., Menche, J., Barabási, A. L., and Sharma, A. (2014). Human symptoms-disease network. *Nat. Commun* 5:4212. doi: 10.1038/ncomms5212

Zou, S., Zhang, J., and Zhang, Z. (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

Zou, S., Zhang, J., and Zhang, Z. (2018). Novel human microbe-disease associations inference based on network consistency projection. *Sci. Rep.* 8:8034. doi: 10.1038/s41598-018-26448-8

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.