Xu et al. J Transl Med (2021) 19:66
https://doi.org/10.1186/s12967-021-02732-6

Journal of Translational Medicine

RESEARCH Open Access

MDAKRLS: Predicting human microbe-disease association based on Kronecker regularized least squares and similarities

Da Xu[1], Hanxiao Xu[1], Yusen Zhang[1*], Mingyi Wang[2*], Wei Chen[1] and Rui Gao[3]

Abstract

Background: Microbes are closely related to human health and diseases. Identification of disease-related microbes is of great significance for revealing the pathological mechanism of human diseases and understanding the interaction mechanisms between microbes and humans, which is also useful for the prevention, diagnosis and treatment of human diseases. Considering the known disease-related microbes are still insufficient, it is necessary to develop effective computational methods and reduce the time and cost of biological experiments.

Methods: In this work, we developed a novel computational method called MDAKRLS to discover potential microbe-disease associations (MDAs) based on the Kronecker regularized least squares. Specifically, we introduced the Hamming interaction profile similarity to measure the similarities of microbes and diseases besides Gaussian interaction profile kernel similarity. In addition, we introduced the Kronecker product to construct two kinds of Kronecker similarities between microbe-disease pairs. Then, we designed the Kronecker regularized least squares with different Kronecker similarities to obtain prediction scores, respectively, and calculated the final prediction scores by integrating the contributions of different similarities.

Results: The AUCs value of global leave-one-out cross-validation and 5-fold cross-validation achieved by MDAKRLS were 0.9327 and 0.9023 ± 0.0015, which were significantly higher than five state-of-the-art methods used for comparison. Comparison results demonstrate that MDAKRLS has faster computing speed under two kinds of frameworks. In addition, case studies of inflammatory bowel disease (IBD) and asthma further showed 19 (IBD), 19 (asthma) of the top 20 prediction disease-related microbes could be verified by previously published biological or medical literature.

Conclusions: All the evaluation results adequately demonstrated that MDAKRLS has an effective and reliable prediction performance. It may be a useful tool to seek disease-related new microbes and help biomedical researchers to carry out follow-up studies.

Keywords: Association prediction, Microbe, Disease, Machine learning, Kronecker regularized least squares

Background

With the fast development of advanced analytical techniques and high-throughput methods for exploring complex microbial communities, in human disease and health, the role of the microbiome has gained widespread attention over the past decade [1, 2]. The microbial community is complex and immensely diverse, research

*Correspondence: zhangys@sdu.edu.cn; wangmingyi1973@outlook.com
[1] School of Mathematics and Statistics, Shandong University, Weihai 264209, China
[2] Department of Central Lab, Weihai Municipal Hospital, Cheeloo College of Medicine, Shandong University, Weihai, Shandong, China
Full list of author information is available at the end of the article

showed that about 100 trillion archaeal and bacterial cells in the human gut which belong to more than 1000 species, tenfold the number of human cells [3, 4]. Microbes are closely related to human health and disease. Generally, most of the gut microbes are either harmless or even beneficial to the human body, such as which can contribute to normal immune function, improve metabolic capability and protect against enteric pathogens [5, 6]. Therefore, microbes are also considered as "forgotten organs" in host [7]. But, if the normal balance between the host and microbiota is broken, which may possibly induce many diseases, including asthma [8], inflammatory bowel disease (IBD) [9], brain disorders or neurodevelopmental deficits [10] and even cancer [11], and so on.

Unquestionably, it has a great significance to identify microbes related to diseases for revealing the pathological mechanism of human diseases and understanding the mechanisms of microbe-host interactions. Some large-scale projects have been initiated, such as the Human Microbiome Project (HMP) [12] and Metagenomics of the Human Intestinal Tract of European Union Project (MetaHIT) [4, 13]. It can help us to initially understand the significance of medicine and biology, functional states and healthy composition of the human microbiome [5]. It is still a challenge to understand how the microbiome influences human diseases, since the microbial community is complex and diverse. Effective computational methods could significantly reduce the time and cost of traditional culture-based microbial experiments. Researchers could select potential MDAs for experimental verification. In 2016, Ma et al. [14] manually collected and developed a human microbe–disease association database (HMDAD), which provided the foundation for identifying the MDAs through computational methods.

In general, we could transform this biological problem of predicting disease-related microbes into a link prediction task. In fact, some computational methods have been widely developed to solve the association or interaction problem such as miRNA-disease [15], drug-target [16], lncRNA-protein [17] and protein–protein interaction [18] prediction problems, and so on. However, to the best of our knowledge, until 2016, there are almost no related MDAs prediction researches from a computational point of view. Thereafter, in 2016, Chen et al. [19] designed the first computational method called KATZHMDA for the prediction of MDAs. It is a KATZ measure-based network prediction method to solve the problem of MDAs prediction by calculating the Gaussian interaction profile (GIP) kernel similarity. Beyond that, in recent years, some network-based methods were also proposed only using the GIP kernel similarity for prediction, which are primarily based on the fusion of known associations and heterogeneous data to construct the network, including random walking-based methods [20, 21], label propagation-based method [22], path-based method [23]. In 2017, Huang et al. [24] presented the NGRHMDA method by integrating two single recommendation methods (graph-based scoring and neighbor-based collaborative filtering prediction model), and achieved a good prediction result. With the fast development of machine learning technology [25, 26], some machine learning-based methods were also presented for MDAs prediction. For example, in 2017, Wang et al. [27] proposed a semi-supervised method called LRLSHMDA based on the Laplacian regularized least squares method. In addition, in 2018, He et al. [28] and Shi et al. [29] developed machine learning-based method named GRNMFHMDA and BMCMDA for MDAs prediction, respectively, based on the graph regularized non-negative matrix factorization and binary matrix completion.

In recent years, the above computational methods mainly utilized a basic assumption that microbes with similar functions will share similar non-interaction or interaction patterns with phenotype diseases [30, 31]. With the fast development of machine learning technology, the regularized least squares algorithm is a useful tool and has been widely used in the recommended system [32–34]. Although some computational methods have been developed, most disease-related microbes remain unknown and effective methods are still scarce [5, 35]. We could address or reduce some limitations to improve the prediction performance of the computational method. For example, some existing methods only used the GIP kernel similarity for extracting the efficacious information, which may lead to the algorithm inevitably biased against well-researched microbes and diseases, multivariate information fusion will be more helpful for prediction. Beyond that, some existing methods did not consider that the effective contribution of diseases and microbes is uneven due to the number of diseases and microbes is different in the database [36]. It is necessary to improve calculation speed since some methods integrate multiple calculation methods which may be complex and time-consuming. Some methods used many model parameters which may reduce robustness and do not apply to new data.

In this paper, considering some of the above limitations, we developed a novel computational method called MDAKRLS based on the Kronecker regularized least squares method to identify potential MDAs. It is a machine learning-based method and uses fewer model parameters, which can save time and obtain robust performance. First, we calculated Kronecker Gaussian similarity and Kronecker Hamming similarity of microbe-disease pairs based on the known microbe-disease association network. Then, the Kronecker regularized least

squares algorithm used two different Kronecker similarities to obtain prediction scores, respectively. Finally, we obtained the final prediction results by integrating the contributions of different similarities. The experimental results of 5-fold cross-validation (5-CV) and global leave-one-out cross-validation (LOOCV) indicated that MDAKRLS can achieve superior performance by comparing it with five state-of-the-art methods. In addition, case studies further demonstrated that MDAKRLS is a useful tool that can effectively identify potential MDAs.

Materials and methods

In this work, we proposed a novel method called MDAKRLS for inferring latent MDAs. Figure 1 describes the overall flow chart of MDAKRLS for prediction. The framework of prediction method consists of three steps. First, we constructed Kronecker Gaussian similarity K_G and Kronecker Hamming similarity K_H of microbe-disease pairs by fully exploiting Gaussian interaction profile (GIP) kernel similarity and Hamming interaction profile (HIP) similarity from known microbe–disease association matrix, respectively. Second, we introduced the Kronecker regularized least squares algorithm based on two Kronecker similarity to construct loss function for prediction. Third, we used an integration strategy to get the final predicted association matrix. Finally, the final possibility score of each microbe-disease pair can be calculated.

Human microbe–disease association data set

In this study, we used a widely-used benchmark data set (HMDAD) to evaluate the reliability and effectiveness of MDAKRLS. It was manually collected by Ma et al. [14] and can be available at http://www.cuilab.cn/hmdad. The database contains a total of 483 verified associations, 292 human microbes and 39 diseases. The microbe-disease association data set adopted by us was downloaded from HMDAD in June, 2020. We finally obtained 450 verified associations after we removed repetitive associations. In fact, we represented the advantages of MDAKRLS through the overall HMDAD data set. For a better description, we constructed an adjacency matrix $A \epsilon R^{39\times 292}$ to express the associations network.

Similarity measures

For a better description, in this study, set $D = \{d_1, d_2, \ldots, d_i, \ldots d_{nd}\}$ and $M = \{m_1, m_2, \ldots, m_j, \ldots, m_{nm}\}$ denote the sets of diseases and microbes, respectively. We introduced an adjacency matrix $A \epsilon R^{nd\times nm}$ to express the associations network, where variable nd denotes the numbers of diseases; nm represents the numbers of microbes. Besides, the adjacency matrix $A \epsilon R^{nd\times nm}$ is defined as follows:

$$A(i,j) = \begin{cases} 1, & \text{if disease } d_i \text{ is related to microbe } m_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Set $A(d_i) \epsilon \{0,1\}^{1*nm}$ represents the i th row of A, which is a binary vector and denotes the interaction profile of the disease d_i. Similarly, $A(m_j) \epsilon \{0,1\}^{nd*1}$ denotes the j th column of A, which represents the interaction profile of the microbe m_j. According to the basic assumption, microbes with similar functions will share similar non-interaction or interaction patterns with phenotype diseases, which is widely used in the related studies. To integrate more effective information and uncover potential associations, we calculated the GIP kernel similarity and HIP similarity of human microbes and diseases, respectively.

GIP kernel similarity for microbes and diseases

To mine conveniently the topological structure information of association matrix A, we used the GIP kernel similarity [19, 37] for measuring similarity of human microbes. Specifically, for two given microbes m_i and m_j, we first extracted their interaction profiles $A(m_i)$ and $A(m_j)$ from the training adjacency matrix A, respectively. Subsequently, the GIP kernel similarity of microbes can be calculated as follows:

$$S_G^m(m_i, m_j) = exp\left(-\sigma_m ||A(m_i) - A(m_j)||^2\right) \tag{2}$$

$$\sigma_m = \sigma_m^{'} / \left(\frac{1}{nm}\sum_{k=1}^{nm} ||A(m_k)||^2\right) \tag{3}$$

where S_G^m is defined as the microbe GIP kernel similarity matrix; $\sigma_m^{'}$ is a trade-off parameter and we set $\sigma_m^{'} = 1$ in the experiments; parameter σ_m is applied to tune-up bandwidth of GIP kernel, which can be updated by the Eq. (3).

Similarly, we also obtained the disease GIP kernel similarity as follows:

$$S_G^d(d_p, d_q) = exp\left(-\sigma_d ||A(d_p) - A(d_q)||^2\right) \tag{4}$$

$$\sigma_d = \sigma_d^{'} / \left(\frac{1}{nd}\sum_{k=1}^{nd} ||A(d_k)||^2\right) \tag{5}$$

where S_G^d is defined as the disease GIP kernel similarity matrix; $\sigma_d^{'}$ is a trade-off parameter and we set $\sigma_d^{'} = 1$ in the experiments; parameter σ_d is applied to tune-up bandwidth of GIP kernel, which can be updated by the Eq. (5).

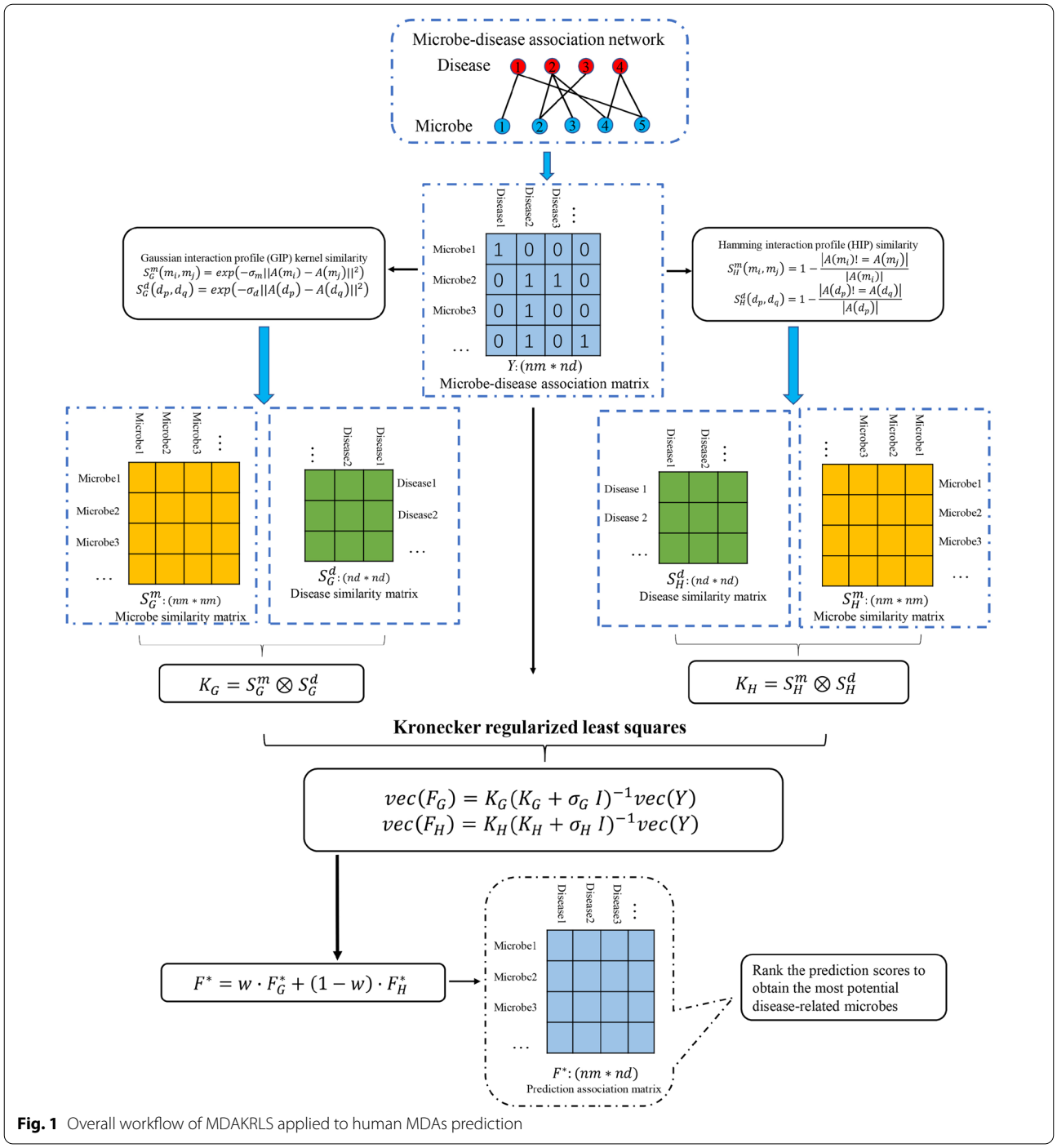

Fig. 1 Overall workflow of MDAKRLS applied to human MDAs prediction

HIP similarity for microbes and diseases

In this work, inspired by the Jiang et al.' s work [38], we introduced HIP similarity to measure the interaction profile similarity between microbe pairs from the training adjacency matrix A. For HIP similarity, two microbes will have a lower similarity if they have more different corresponding values in the interaction profiles. Further, the HIP similarity of microbes m_i and m_j is defined as follows:

$$S_H^m(m_i, m_j) = 1 - \frac{|A(m_i)! = A(m_j)|}{|A(m_i)|} \tag{6}$$

where S_H^m denotes the microbe HIP similarity matrix; $|\cdot|$ denotes the number of elements in the interaction profile.

Similarly, based on the interaction profiles of diseases, the HIP similarity of diseases can be calculated as follows:

$$S_H^d(d_p, d_q) = 1 - \frac{|A(d_p)! = A(d_q)|}{|A(d_p)|} \tag{7}$$

where S_H^d denotes the disease HIP similarity matrix.

MDAKRLS for microbe-disease association prediction

Regularized least squares (RLS) and its extended versions are popular machine learning methods. In this work, to boost the predictable performance, a novel predict model called MDAKRLS is proposed to calculate the relevance scores between microbes and diseases by integrating the Kronecker product and RLS method.

We first obtained the Kronecker Gaussian similarity between microbe-disease pairs by the GIP similarity matrix of microbes and diseases. Specifically, we use the following equation to define the similarity between the microbe-disease pairs $(m(i), d(p))$ and $(m(j), d(q))$:

$$K_G\big((m(i), d(p)), \big(m(j), d(q)\big)\big) = S_G^m\big(m(i), m(j)\big) * S_G^d(d(p), d(q)) \tag{8}$$

where S_G^m and S_G^d represent the GIP similarity matrix of microbes and diseases defined above, respectively. Let $N = nd \times nm$ which represents the number of microbe-disease pairs. The above equation can be represented by the Kronecker product as follows:

$$K_G = S_G^m \otimes S_G^d \tag{9}$$

where $K_G \epsilon R^{N \times N}$ is defined as the Kronecker Gaussian similarity of microbe-disease pairs. In the same manner, the Kronecker Hamming similarity matrix $K_H \epsilon R^{N \times N}$ of microbe-disease pairs can be measured:

$$K_H = S_H^m \otimes S_H^d \tag{10}$$

For a better description, in this work, we set $X = (x_1, x_2, \ldots, x_i, \ldots, x_N)$, where x_i denotes the i th microbe-disease pair. $vec(Y) = (y_1, y_2, \ldots, y_i, \ldots, y_N)$, where $Y \epsilon R^{nd \times nm}$ denotes the training microbe-disease adjacency matrix in the process of forecasting; $vec(\cdot)$ is a vector operator that stacks the elements of all columns into a vector; $y_i \epsilon \{0, 1\}$ denotes the corresponding label of microbe-disease pair x_i. The biological problem of predicting disease-related microbes can be transformed to learn a mapping function f_G and calculate a corresponding association score. $vec(F_G) = (f_G(x_1), f_G(x_2), \ldots, f_G(x_i), \ldots, f_G(x_N))$, where F_G denotes the prediction score matrix based on the Kronecker Gaussian similarity; $f_G(x_i)$ represents the prediction score of microbe-disease pair x_i obtained by prediction function f_G.

In further work, first, we constructed the Kronecker regularized least squares [39] based on the Kronecker Gaussian similarity to solve the microbe-disease prediction problem. The objective function based on the Tikhonov minimization problem is formulated as follows:

$$J(f_G) = \frac{1}{2}\sum_{i=1}^{N}\big(y_i - f_G(x_i)\big)^2 + \frac{\sigma_G}{2}\| f_G \|_k^2 \tag{11}$$

where $\sigma_G > 0$ is a regularization coefficient used to adjust the regularization term and loss function of the objective function; $\| f_G \|_k$ is the norm of mapping function f_G in Reproducing Kernel Hilbert Space (RKHS) [40] associated to the kernel k. Based on the classical Representer Theorem [41], the solution of the Tikhonov regularization problem exists in the RKHS and can be calculated as follows:

$$f_G(x_i) = \sum_{j=1}^{N}\alpha_j K_G(x_i, x_j) \tag{12}$$

According to the previous studies [37, 42], the optimal solution of the objective function can be further calculated as follows:

$$vec(F_G) = K_G(K_G + \sigma_G I)^{-1} vec(Y) \tag{13}$$

where I denotes the identity matrix.

Eigen decompositions were implemented on the GIP similarity matrix S_G^m of microbes and GIP similarity matrix S_G^d of diseases. We can get $S_G^m = V_G^m \Lambda_G^m V_G^{m\mathrm{T}}$ and $S_G^d = V_G^d \Lambda_G^d V_G^{d\mathrm{T}}$, respectively. According to the property of the Kronecker product, we can obtain the $K_G = S_G^m \otimes S_G^d = V_G \Lambda_G V_G^{\mathrm{T}}$, where $V_G = V_G^m \otimes V_G^d$ and $\Lambda_G = \Lambda_G^m \otimes \Lambda_G^d$. Then, we can transform the Eq. (13) as follows:

$$\begin{aligned} vec(F_G) &= V_G \Lambda_G V_G^{\mathrm{T}} \left(V_G \Lambda_G V_G^{\mathrm{T}} + \sigma_G I\right)^{-1} vec(Y) \\ &= V_G \Lambda_G (\Lambda_G + \sigma_G I)^{-1} V_G^{\mathrm{T}} vec(Y) \end{aligned} \tag{14}$$

According to another property of the Kronecker product [43], $(N^{\mathrm{T}} \otimes M) vec(C) = vec(MCN)$, Eq. (14) can be rewritten as follows:

$$
\begin{aligned}
vec(F_G) =& (V_G^m \otimes V_G^d)\left(\Lambda_G^m \otimes \Lambda_G^d\right)\left(\Lambda_G^m \otimes \Lambda_G^d + \sigma_G I\right)^{-1} \\
& (V_G^{m\mathrm{T}} \otimes V_G^{d^{\mathrm{T}}}) vec(Y) \\
= & (V_G^m \otimes V_G^d)\left(\Lambda_G^m \otimes \Lambda_G^d\right)\left(\Lambda_G^m \otimes \Lambda_G^d + \sigma_G I\right)^{-1} vec(V_G^{d^{\mathrm{T}}} Y V_G^m) \\
= & (V_G^m \otimes V_G^d) vec(X_G) \\
= & vec\left(V_G^d X_G V_G^{m\mathrm{T}}\right)
\end{aligned} \tag{15}
$$

Finally, we will obtain the score matrix based on the Kronecker Gaussian similarity by the following equation:

$$F_G^* = V_G^d X_G V_G^{m\mathrm{T}} \tag{16}$$

where $vec(X_G) = \left(\Lambda_G^m \otimes \Lambda_G^d\right)\left(\Lambda_G^m \otimes \Lambda_G^d + \sigma_G I\right)^{-1} vec\left(V_G^{d^{\mathrm{T}}} Y V_G^m\right)$.

In addition, we also can get another objective function and optimal solution based on the Kronecker Hamming similarity in a similar manner:

$$J(f_H) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - f_H(x_i)\right)^2 + \frac{\sigma_H}{2}\| f_H \|_k^2 \tag{17}$$

$$vec(F_H) = K_H(K_H + \sigma_H I)^{-1} vec(Y) \tag{18}$$

We implemented eigen decompositions on the HIP similarity matrix S_H^m of microbes and HIP similarity matrix S_H^d of diseases, and obtained the second score matrix based on the Kronecker Hamming similarity:

$$F_H^* = V_H^d X_H V_H^{m\mathrm{T}} \tag{19}$$

where $vec(X_H) = \left(\Lambda_H^m \otimes \Lambda_H^d\right)\left(\Lambda_H^m \otimes \Lambda_H^d + \sigma_H I\right)^{-1} vec\left(V_H^{d^{\mathrm{T}}} Y V_H^m\right)$.

After obtaining the prediction matrix F_G^* and F_H^* based on the two different Kronecker similarities, respectively, we obtain the final prediction matrix by integrating their contributions as follows:

$$F^* = w \cdot F_G^* + (1 - w) \cdot F_H^* \tag{20}$$

where w is a trade-off parameter. Eventually, we will obtain the score matrix F^*. In the future research, the association with the high score will have a priority to be verified by biological experiment.

Results and discussion

Evaluation metrics

To measure the reliability and effectiveness of the proposed method, in the same experimental conditions, we implemented our method and reran the other five state-of-the-art computational methods for comparison, under 5-CV and global LOOCV framework. Notably, the GIP kernel similarity and HIP similarity of microbes and diseases should be recalculated in every round of the global LOOCV and 5-CV framework.

Specifically, in the global LOOCV framework, all of the microbe-disease pairs without associations were used as candidate samples, each of the known MDA was treated as a testing sample and the rest of the known MDAs were treated as a training set to conduct experiments. We can obtain the rank of every testing sample by comparing it with candidate samples. To visualize the prediction performance, 1-specificity (false positive rates) and sensitivity (true positive rates) were calculated to plot the receiver operating characteristic (ROC) curves by setting different thresholds. For convenient observation, we calculated the area under the ROC curve (AUC) values to measure the ability of prediction method.

In the validation framework of 5-CV, all observed microbe-disease associations are randomly split into 5 subsets. Each of the 5 subsets is specified as an independent testing set and the rest of the 4 subsets are regarded as training sets. To weaken potential experimental bias caused by random sample division, the process of the experiment of every method was performed 100 times. Furthermore, the corresponding 1-specificity and sensitivity were obtained for plotting the ROC curves. The corresponding AUC values were also calculated for evaluation. The AUC value of 1 means best prediction, while the AUC value of 0.5 indicates random prediction.

Parameter sensitivity analysis

There are three parameters (σ_G, σ_H and w) in our model. In general, the prediction performance of the model depends on some parameters, and different scale values of the parameter will produce different prediction results. Here, to explore the properties of the proposed method and the influences of parameter and find the optimal parameter, we calculated the AUCs and made some comparison experiments with different initial parameters under the 5-CV and LOOCV frameworks.

σ_G and σ_H are self-tuned parameters of MDAKRLS. To promote robust performance and simplify the complex problem, we set the same variable value for parameters σ_G and σ_H. The experimental results of the parameters have been shown in Fig. 2a. From the figure, the average AUC of MDAKRLS is greatly enhanced when the parameter increases from 0 to 5, and the performance remains almost unchanged as the value of the parameter increases from 5 to 35 under two kinds of frameworks. Finally, the values of parameters σ_G and σ_H were set as 30 to obtain a stable and optimal prediction result. Then, we fixed σ_G and σ_H, and adjusted the trade-off parameter w. The relationship between the AUC value and the parameter

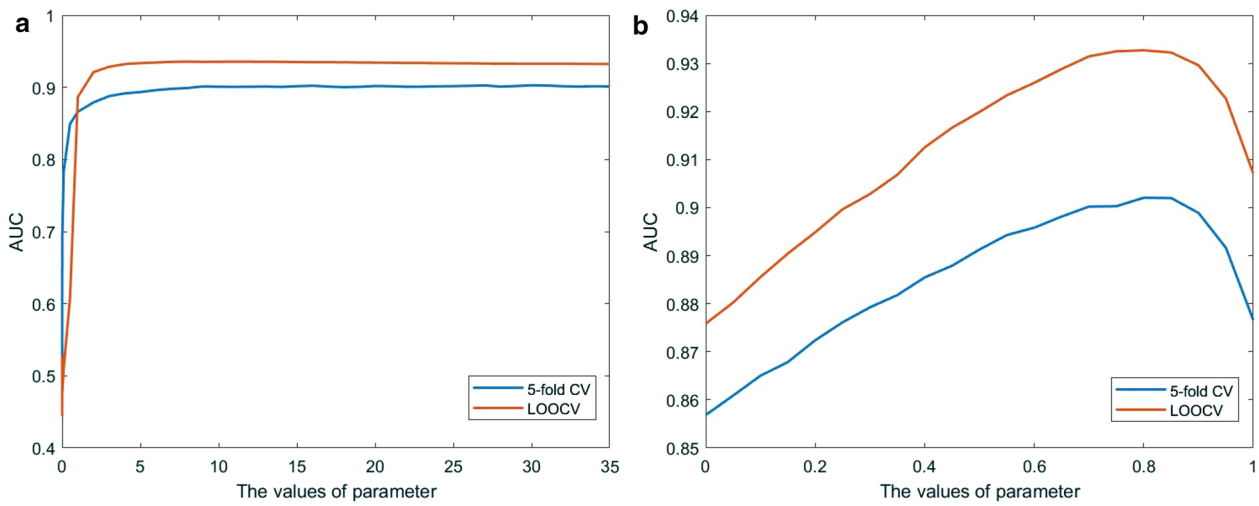

Fig. 2 **a** The relationship between parameters σ_G and σ_H and AUC value under the 5-CV and LOOCV frameworks. **b** The relationship between parameter w and AUC value under the 5-CV and LOOCV frameworks

w is shown in Fig. 2b. It can be seen that MDAKRLS will obtain the highest AUC whn $w = 0.8$, indicating that the Kronecker Gaussian similarity can provide more effective information for prediction. Finally, we obtained the best parameters for the following analysis, which can achieve better performance. The average AUC value of 5-CV achieved by our proposed method based on the optimal parameters was 0.9023 ± 0.0015, and the AUC value of global LOOCV was 0.9327. The standard deviation and evaluation results demonstrate that used parameter values are reliable and robust for the proposed model.

Comparison with other methods

To validate the effectiveness of MDAKRLS, we compared it with five state-of-the-art computational methods under the same experimental conditions, including KATZ measure (KATZHMDA) [19], Laplacian Regularized Least Squares (LRLSHMDA) [27], Bi-Random Walk (BiRWHMDA) [20], Network Topological Similarity (NTSHMDA) [21] and Graph Regularized Non-negative Matrix Factorization (GRNMFHMDA) [28] for human microbe–disease association prediction. Previous studies showed that these methods achieved effective prediction results. Here, we implemented the above 5 prediction methods for comparison under the global LOOCV and 5-CV frameworks on the same benchmark data set. The comparison results are shown in Figs. 3 and 4, respectiv ely.

Specifically, Fig. 3 shows AUC values and ROC curves of different methods under the global LOOCV framework. It can be observed from the figure that the

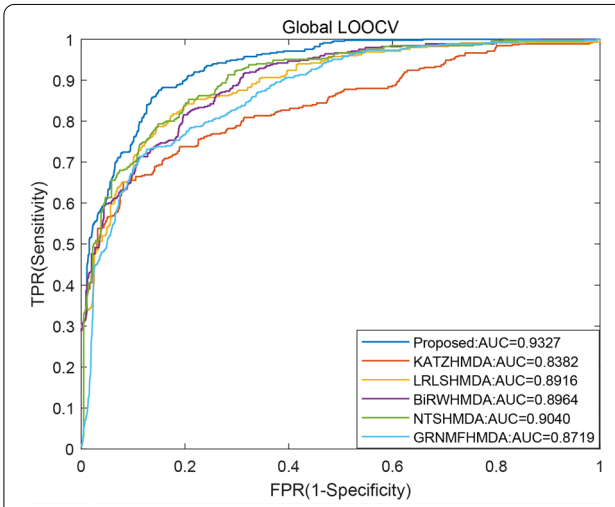

Fig. 3 The AUC values and ROC curves of different methods under global LOOCV

AUC values of five comparative methods are the following: KATZHMDA (0.8382), LRLSHMDA (0.8916), BiRWHMDA (0.8964), NTSHMDA (0.9040) and GRNMFHMDA (0.8719). Our method obtained the highest AUC value (0.9327), which is superior to the other five methods. Similarly, we compared all methods in the framework of 5-CV. The corresponding average AUC values and ROC curves of different methods have been shown in Fig. 4. As a result, the average AUC value of the proposed method is 0.9023, which performs better than KATZHMDA (0.8324), LRLSHMDA

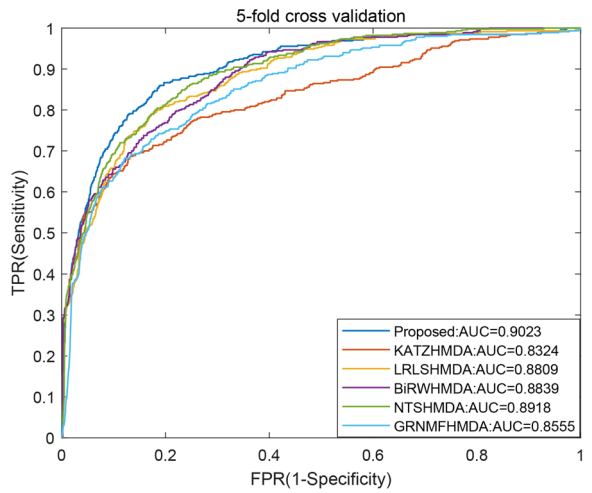

Fig. 4 The average AUC values and ROC curves of different methods under 5-CV

Table 1 Comparison of calculation speed between proposed method and five state-of-the-art prediction methods

Method	5-fold CV Average running time (s)	Global LOOCV Running time (s)
Proposed method	1.5604 ± 0.1290	7.4807
KATZHMDA	1.7240 ± 0.1190	8.6369
LRLSHMDA	1.8275 ± 0.2236	29.1313
BiRWHMDA	1.6293 ± 0.1057	11.7000
GRNMFHMDA	1.9593 ± 0.1669	30.1469
NTSHMDA	1.8302 + -0.8365	19.4239

(0.8809), BiRWHMDA (0.8839), NTSHMDA (0.8918) and GRNMFHMDA (0.8555). The experimental results demonstrate that the proposed method is an effective and reliable prediction tool in inferring possible associations.

Practically, the traditional experiment-based methods are time-consuming, some computational methods were proposed to save time. It is necessary to improve calculation speed and prediction accuracy for developing new reliable computational methods. Thus, calculation speed is an important metric for performance evaluation of different computational methods. Therefore, it is a fresh perspective and we implemented a runtime analysis. Specifically, we compared the calculation speed between the proposed method and five state-of-the-art prediction methods under the global LOOCV and 5-CV frameworks. The comparisons of calculation speed analysis are shown in Table 1. Our proposed method obtained a faster average running time under the 5-CV framework. Moreover, the proposed method can use the shortest time for prediction under the global LOOCV framework. In brief, these results indicate the proposed method is reliable, effective and time-saving. It may be a useful tool for seeking disease-related new microbes.

Case studies

To further access the practical effect of the proposed method in inferring associated microbes with a disease without any known associated microbes, case study analysis [28, 44] was implemented on the MDAKRLS. For a given disease, we removed all known microbe associations in the HMDAD. Then, the proposed method was trained on the rest of the known associations and tested on the candidate microbe samples to seek the disease-related microbes, it can guarantee the independence between training data sets and validation data sets. In other words, the prediction model only depends on the rest of the known association information and the similarity measures of the training data sets. Specifically, in the microbe-disease adjacency matrix A, we converted all 1 to 0 for a given disease and ranked all microbe samples based on the prediction scores. The top ranked microbes will be further verified by the relevant literature and the method will be effective if the top prediction results have more verified microbes. To reveal the pathological relationship of diseases and microbes, in the framework of the MDAKRLS, we implemented independent case studies on two kinds of important human diseases: asthma and inflammatory bowel disease (IBD). It should be noted that we assume that the genus of this microbe will be associated with the disease if the microbe is associated with the disease when we validate microbes [21, 45].

Asthma is a common chronic inflammatory disease, which has substantial morbidity. According to statistics, more than 300 million patients were affected by asthma worldwide [8]. In this study, the top 20 prediction results of asthma-related microbes are tabulated in Table 2. In the prediction list, there are some predictions have been validated by the HMDAD, the rest could be validated by the previously published biological and medical literature for asthma-related microbes. Finally, 19 of the top 20 prediction microbes could be manually verified that they are related to asthma patients. For example, *Actinobacteria*, *Firmicutes* and *Bacteroides* have lower proportions in all sputum samples of asthmatic patients, while *Proteobacteria* and *Staphylococcus aureus* were higher [46, 47]. Moreover, *Clostridium difficile* colonized at 1 month of age, which was closely related to asthma at 6 to 7 years of age [48]. The clustering results of bacterial composition showed *Enterobacteriaceae* family were more abundant in healthy people, while *Lachnospiraceae* and *Bifidobacterium* were more abundant in the asthma patients [49].

Table 2 Prediction results of the top 20 associated microbes with asthma

Rank	Microbe	Evidence	Score
1	Proteobacteria	Confirmed by HMDAD	0.0840
2	Firmicutes	PMID:23265859	0.0698
3	Clostridium difficile	PMID:21872915	0.0687
4	Bacteroidetes	Confirmed by HMDAD	0.0683
5	Prevotella	Confirmed by HMDAD	0.0623
6	Helicobacter pylori	Confirmed by HMDAD	0.0571
7	Clostridium coccoides	PMID:21477358	0.0506
8	Actinobacteria	PMID:23265859	0.0503
9	Staphylococcus aureus	PMID:18822123	0.0450
10	Lachnospiraceae	PMID:28912020	0.0411
11	Lactobacillus	PMID:20592920	0.0388
12	Clostridia	Unconfirmed	0.0367
13	Enterobacteriaceae	PMID:28947029	0.0349
14	Bacteroides	PMID:18822123	0.0336
15	Veillonella	PMID:25329665	0.0301
16	Haemophilus	Confirmed by HMDAD	0.0297
17	Fusobacterium	PMID:27838347	0.0285
18	Stenotrophomonas maltophilia	PMID:16351036	0.0269
19	Bifidobacterium	PMID:24735374	0.0260
20	Bacteroides vulgatus	PMID:28966614	0.0250

Table 3 Prediction results of the top 20 related microbes with IBD

Rank	Microbe	Evidence	Score
1	Proteobacteria	Confirmed by HMDAD	0.0820
2	Bacteroidetes	PMID:25307765	0.0798
3	Prevotella	PMID:25307765	0.0732
4	Firmicutes	PMID:25307765	0.0703
5	Clostridium difficile	PMID:24838421	0.0692
6	Helicobacter pylori	PMID:22221289	0.0684
7	Clostridium coccoides	PMID:19235886	0.0508
8	Staphylococcus aureus	PNID:19809406	0.0454
9	Haemophilus	PMID:24013298	0.0401
10	Lactobacillus	PMID:26340825	0.0389
11	Clostridia	PMID:25307765	0.0370
12	Actinobacteria	Confirmed by HMDAD	0.0370
13	Enterobacteriaceae	PMID:24629344	0.0351
14	Bacteroides	PMID:25307765	0.0336
15	Staphylococcus	PMID:28174737	0.0306
16	Veillonella	PMID:28842640	0.0301
17	Lachnospiraceae	Confirmed by HMDAD	0.0291
18	Fusobacterium	PMID:25307765	0.0282
19	Stenotrophomonas maltophilia	Unconfirmed	0.0271
20	Bifidobacterium	PMID:24478468	0.0260

In addition, in a study about children and infants, the fecal colonization of *Clostridium coccoides* subcluster XIVa species and *Bacteroides fragilis* subgroup can be served as early indicators, which will be good for the prevention of asthma [50]. *Lactobacillus* has been shown to be beneficial for children with asthma [51].

IBD is a chronic disabling gastrointestinal disease with a continually increasing incidence, which is a worldwide health-care problem [9]. Similar to asthma, the top 20 prediction results of inflammatory bowel disease (IBD)-related microbes are tabulated in Table 3. In the prediction list, based on the HMDAD and recently published biological and medical literature for IBD-related microbes, 19 of the top 20 prediction microbes could be manually verified that they are related to the IBD patients. For example, previous studies showed *Bacteroidetes*, *Prevotella* and *Firmicutes* were associated with the formation of IBD [52, 53]. *Clostridium difficile* can aggravate flares of IBD, resulting in mortality and morbidity [54]. There is a negative relevant relation between IBD and *Helicobacter pylori* [55]. Compared with healthy people, *Clostridium coccoides* was less represented in active IBD patients [56]. In the salivary microbiota of IBD patients, *Haemophilus*, *Veillonella* and *Prevotella* were found that can largely contribute to dysbiosis [57]. In addition, in the faeces of IBD patients, the proportion of *Lactobacillus* increased, while *Bifidobacterium* decreased [58].

In addition, we also implemented case studies for three metabolic diseases including Obesity, Type 1 diabetes and Type 2 diabetes (see Additional file 1). Case studies indicate if one of the 39 human diseases does not have any known related microbes in the HMDAD, MDAKRLS can calculate the possibility of association between the disease and 292 microbes. The proposed method may be an effective tool for seeking disease-related possible new microbes. Then we further used MDAKRLS to rank all candidate microbes for all the diseases involved in HMDAD (see Additional file 2). We hope that the prediction list can provide aid, and more and more potential microbe-disease pairs could be verified by clinical or biological experiment observation.

Conclusion

Identifying of MDAs could help us better understand the pathogenesis of human diseases, which is also useful for the prevention, diagnosis and treatment of human diseases. In this study, we developed a novel computational method called MDAKRLS based on the Kronecker regularized least squares. Firstly, we not only calculated the Kronecker Gaussian similarity of microbe-disease

pairs through the GIP kernel similarity of microbes and diseases, but also obtained the Kronecker Hamming similarity by the HIP similarity. Then, we developed the Kronecker regularized least squares based on the Kronecker product and RLS method to calculate the correlation scores of MDAs. A comparison of calculation speed showed our method has the advantage of fast calculating speed. The evaluation results of the 5-CV and the global LOOCV framework demonstrated that MDAKRLS improved calculation accuracy and had a reliable prediction performance. In addition, case studies of IBD and asthma further indicated that MDAKRLS can effectively discover potential associations.

Several critical factors that make MDAKRLS has a reliable prediction performance. Firstly, different from some methods only using the GIP kernel similarity for prediction, we also introduced the HIP similarity to measure the similarities of microbes and diseases. Secondly, we used the Kronecker product to construct two kinds of Kronecker similarities of microbe-disease pairs, which is complementary and can effectively mine the topological structure information of the network. Thirdly, In the process of solving the Tikhonov minimization problem, we introduced eigen decompositions to reduce the computational complexity. Kronecker regularized least squares is a machine learning-based method and uses fewer model parameters, thus saving time and improving robust performance. Of course, MDAKRLS needs to be improved in future work, such as some prior information of microbes or diseases could be introduced to improve the prediction performance; the insufficient number of experimentally verified MDAs limits the performance and development of the computational model.

The development of a reasonable and effective calculation model is conducive to the study of the microbial community. MDAKRLS has a good transplantation character, which is easily implemented to solve similar biological problems. The insufficient number of experimentally verified MDAs limits the performance and development of the computational model. At present, most disease-related microbes remain unknown in HMDAD. Therefore, it will be feasible and be of great practical significance to develop prediction algorithms that can effectively overcome the data sparsity problem. In addition, it is necessary to add more experimentally verified MDAs to improve the database, which can provide a foundation for improving the performance of computational method. We hope that our method could help biomedical researchers to carry out follow-up studies, and more and more potential microbe-disease associations could be verified by clinical or biological experimental observation.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-021-02732-6.

Additional file 1. The prediction results of the top 20 associated microbes with Obesity, Type 1 diabetes and Type 2 diabetes.

Additional file 2. We further used MDAKRLS to rank all candidate microbes for all the diseases involved in HMDAD. The prediction results may help biomedical researchers conduct experimental validation and follow-up research.

Abbreviations

MDAs: Microbe-disease associations; IBD: Inflammatory bowel disease; HMDAD: Human microbe–disease association database; GIP: Gaussian interaction profile; HIP: Hamming interaction profile; 5-CV: 5-fold cross-validation; LOOCV: Leave-one-out cross-validation; ROC: Receiver operating characteristic; AUC: Area under the ROC curve.

Acknowledgements

MYW thanked the Weihai Engineering Technology Research Center for financial support. The authors thank the editors and anonymous reviewers for their reading time and constructive comments.

Authors' contributions

DX conceived and designed the study, wrote the manuscript. HXX, YSZ obtained and processed the data. DX, HXX and MYW performed the experiments and analyzed the results. MYW, RG and WC provided suggestions and supervised the research. All authors read and approved the final manuscript.

Funding

YSZ was supported by the National Natural Science Foundation of China under Grant (Nos. 61877064, U1806202). MYW thanked the Weihai Engineering Technology Research Center for financial support. RG was supported by the National Natural Science Foundation of China under Grant (Nos. U1806202 and 61533011).

Availability of data and materials

The data set analyzed during the current study can be available at: http://www.cuilab.cn/hmdad.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] School of Mathematics and Statistics, Shandong University, Weihai 264209, China. [2] Department of Central Lab, Weihai Municipal Hospital, Cheeloo College of Medicine, Shandong University, Weihai, Shandong, China. [3] School of Control Science and Engineering, Shandong University, Jinan 250061, China.

Received: 9 October 2020 Accepted: 1 February 2021
Published online: 12 February 2021

References

1. Cho I, Blaser MJ. The human microbiome: At the interface of health and disease. Nat Rev Genet. 2012;13:260–70.
2. Consortium THMP. A framework for human microbiome research. Nature. 2012;486:215–21.
3. Tremaroli V, Bäckhed F. Functional interactions between the gut microbiota and host metabolism. Nature. 2012;489:242–9.

4. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
5. Zhao Y, Wang C-C, Chen X. Microbes and complex diseases: from experimental results to computational models. Brief Bioinform. 2020;1:21.
6. Niu YW, Qu CQ, Wang GH, Yan GY. RWHMDA: Random Walk on Hypergraph for Microbe-Disease Association Prediction. Front Microbiol. 2019;10:1–10.
7. Clemente JC, Ursell LK, Parfrey LW, Knight R. The Impact of the Gut Microbiota on Human Health: An Integrative View. Cell. 2012;148:1258–70.
8. Lambrecht BN, Hammad H. The immunology of asthma. Nat Immunol. 2015;16:45–56.
9. Zhang YZ, Li YY. Inflammatory bowel disease: Pathogenesis. World J Gastroenterol. 2014;20:91–9.
10. Borre YE, O'Keeffe GW, Clarke G, Stanton C, Dinan TG, Cryan JF. Microbiota and neurodevelopmental windows: implications for brain disorders. Trends Mol Med. 2014;20:509–18.
11. Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer Nature Publishing Group. 2013;13:800–12.
12. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007;449:804–10.
13. Ehrlich SD, Consortium TM. MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract. Metagenomics Hum Body. 2011;2:307–16.
14. Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, et al. An analysis of human microbe-disease associations. Brief Bioinform. 2017;18:85–97.
15. Chen X, Xie D, Zhao Q, You ZH. MicroRNAs and complex diseases: From experimental results to computational models. Brief Bioinform. 2019;20:515–39.
16. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, et al. Drug-target interaction prediction: Databases, web servers and computational models. Brief Bioinform. 2016;17:696–712.
17. Zhang H, Liang Y, Han S, Peng C, Li Y. Long noncoding RNA and protein interactions: From experimental results to computational models based on network methods. Int J Mol Sci. 2019;20:9.
18. Xu D, Xu H, Zhang Y, Chen W, Gao R. Protein-Protein Interactions Prediction Based on Graph Energy and Protein Sequence Information. Molecules. 2020;25:1–15.
19. Chen X, Huang YA, You ZH, Yan GY, Wang XS. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2017;33:733–9.
20. Zou S, Zhang J, Zhang Z. A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. PLoS ONE. 2017;12:1–16.
21. Luo J, Long Y. NTSHMDA: Prediction of Human Microbe-Disease Association based on Random Walk by Integrating Network Topological Similarity. IEEE/ACM Trans Comput Biol Bioinforma. 2020;17:1341–51.
22. Qu J, Zhao Y, Yin J. Identification and Analysis of Human Microbe-Disease Associations by Matrix Decomposition and Label Propagation. Front Microbiol. 2019;10:1–10.
23. Huang ZA, Chen X, Zhu Z, Liu H, Yan GY, You ZH, et al. PBHMDA: Path-Based Human Microbe-Disease Association Prediction. Front Microbiol. 2017;8:1–10.
24. Huang YA, You ZH, Chen X, Huang ZA, Zhang S, Yan GY. Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. J Transl Med. 2017;15:1–11.
25. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. Cell. 2018;173:1581–92.
26. Xu D, Zhang J, Xu H, Zhang Y, Chen W, Gao R, et al. Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data. BMC Genomics. 2020;21:1–17.
27. Wang F, Huang ZA, Chen X, Zhu Z, Wen Z, Zhao J, et al. LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe-Disease Association prediction. Sci Rep. 2017;7:1–11.
28. He BS, Peng LH, Li Z. Human Microbe-Disease Association Prediction With Graph Regularized Non-Negative Matrix Factorization. Front Microbiol. 2018;9:1–11.
29. Shi JY, Huang H, Zhang YN, Cao JB, Yiu SM. BMCMDA: A novel model for predicting human microbe-disease associations via binary matrix completion. BMC Bioinformatics. 2018;19:77.
30. Bao W, Jiang Z, Huang DS. Novel human microbe-disease association prediction using network consistency projection. BMC Bioinformatics. 2017;18:173–81.
31. Wang L, Wang Y, Li H, Feng X, Yuan D, Yang J. A bidirectional label propagation based computational model for potential microbe-disease association prediction. Front Microbiol. 2019;10:44.
32. Chen X, Ren B, Chen M, Wang Q, Zhang L, Yan G. NLLSS: predicting synergistic drug combinations based on semi-supervised learning. PLoS Comput Biol. 2016;12:1–23.
33. Xie G, Meng T, Luo Y, Liu Z. SKF-LDA: Similarity Kernel Fusion for Predicting lncRNA-Disease Association. Mol Ther - Nucleic Acids. 2019;18:45–55.
34. Zhao Y, Chen X, Yin J, Qu J. SNMFSMMA: using symmetric nonnegative matrix factorization and Kronecker regularized least squares to predict potential small molecule-microRNA association. RNA Biol. 2020;17:281–91.
35. Fan C, Lei X, Guo L, Zhang A. Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. Neurocomputing. 2019;323:76–85.
36. Zhang W, Yang W, Lu X, Huang F, Luo F. The Bi-Direction Similarity Integration Method for Predicting Microbe-Disease Associations. IEEE Access. 2018;6:38052–61.
37. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics. 2011;27:3036–43.
38. Jiang L, Ding Y, Tang J, Guo F. MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association. Front Genet. 2018;9:7.
39. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, et al. Toward more realistic drug-target interaction predictions. Brief Bioinform. 2015;16:325–37.
40. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res. 2006;7:2399–434.
41. Rifkin R, Yeo G, Poggio T. Regularized Least-Squares Classification. Nato-Science Ser Sub Ser III Comput Syst Sci. 2003;190:131–54.
42. Luo J, Xiao Q, Liang C, Ding P. Predicting MicroRNA-Disease Associations Using Kronecker Regularized Least Squares Based on Heterogeneous Omics Data. IEEE Access. 2017;5:2503–13.
43. Pahikkala T, Airola A, Stock M, De Baets B, Waegeman W. Efficient regularized least-squares algorithms for conditional ranking on relational data. Mach Learn. 2013;93:321–56.
44. Peng LH, Yin J, Zhou L, Liu MX, Zhao Y. Human microbe-disease association prediction based on adaptive boosting. Front Microbiol. 2018;9:1–9.
45. Long Y, Luo J. WMGHMDA: A novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network. BMC Bioinform. 2019;20:1–18.
46. Marri PR, Stern DA, Wright AL, Billheimer D, Martinez FD. Asthma-associated differences in microbial composition of induced sputum. J Allergy Clin Immunol. 2013;131:346–52.
47. Vael C, Nelen V, Verhulst SL, Goossens H, Desager KN. Early intestinal Bacteroides fragilis colonisation and development of asthma. BMC Pulm Med. 2008;8:1–6.
48. Van Nimwegen FA, Penders J, Stobberingh EE, Postma DS, Koppelman GH, Kerkhof M, et al. Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. J Allergy Clin Immunol. 2011;128:948–55.
49. Abdel-Aziz MI, Vijverberg SJH, Neerincx AH, Kraneveld AD, Maitland AH. The crosstalk between microbiome and asthma: Exploring associations and challenges. Clin Exp Allergy. 2019;49:1067–86.
50. Vael C, Vanheirstraeten L, Desager KN, Goossens H. Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. BMC Microbiol. 2011;11:e23.
51. Huang CF, Chie WC, Wang IJ. Efficacy of lactobacillus administration in school-age children with asthma: a randomized. Placebo-Controlled Trial Nutr. 2018;10:1–11.
52. Juste C, Kreil DP, Beauvallet C, Guillot A, Vaca S, Carapito C, et al. Bacterial protein signals are associated with Crohn's disease. Gut. 2014;63:1566–77.
53. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett. 2014;588:4223–33.

54. Hashash JG, Binion DG. Managing Clostridium difficile in Inflammatory Bowel Disease (IBD). Curr Gastroenterol Rep. 2014;16:14–9.
55. Sonnenberg A, Genta RM. Low prevalence of Helicobacter pylori infection among patients with inflammatory bowel disease. Aliment Pharmacol Ther. 2012;35:469–76.
56. Sokol H, Seksik P, Furet JP, Firmesse O, Nion-Larmurier I, Beaugerie L, et al. Low Counts of Faecalibacterium prausnitzii in Colitis Microbiota. Inflamm Bowel Dis. 2009;15:1183–9.
57. Said HS, Suda W, Nakagome S, Chinen H, Oshima K, Kim S, et al. Dysbiosis of salivary microbiota in inflammatory bowel disease and its association with oral immunological biomarkers. DNA Res. 2014;21:15–25.
58. Takaishi H, Matsuki T, Nakazawa A, Takada T, Kado S, Asahara T, et al. Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. Int J Med Microbiol. 2008;298:463–72.

Publisher's Note