PRWHMDA: Human Microbe-Disease Association Prediction by Random Walk on the Heterogeneous Network with PSO

Chuanyan Wu¹, Rui Gao¹, Daoliang Zhang¹, Shiyun Han², Yusen Zhang³

1. School of Control Science and Engineering, Shandong University, Jinan, 250061, China.
2. General Clinic, The No. 2 People’s Hospital of Tianshui, Jinan, 250032, China.
3. School of Mathematics and Statistics, Shandong University, Weihai, 264209, China.

Corresponding author: Rui Gao
School of Control Science and Engineering, Shandong University, No.17923 of Jingshi Road, Jinan 250061, China. E-mail: gaorui@sdu.edu.cn.

© Ivyspring International Publisher. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY-NC) license (https://creativecommons.org/licenses/by-nc/4.0/). See http://ivyspring.com/terms for full terms and conditions.

Received: 2017.12.24; Accepted: 2018.02.28; Published: 2018.05.22

Abstract

Microorganisms resided in human body play a vital role in metabolism, immune defense, nutrition absorption, cancer control and protection against pathogen colonization. The changes of microbial communities can cause human diseases. Based on the known microbe-disease association, we presented a novel computational model employing Random Walking with Restart optimized by Particle Swarm Optimization (PSO) on the heterogeneous interlinked network of Human Microbe-Disease Associations (PRWHMDA) (see Figure 1). Based on the known human microbe-disease associations, we constructed the heterogeneous interlinked network with Cosine similarity. The extended random walk with restart (RWR) method was derived to get the potential microbe-disease associations. PSO was utilized to get the optimal parameters of RWR. To evaluate the prediction effectiveness, we performed leave one out cross validation (LOOCV) and 5-fold cross validation (CV), which got the AUC (The area under ROC curve) of 0.915 (LOOCV) and the average AUCs of 0.8875 ± 0.0046 (5-fold CV). Moreover, we carried out three case studies of asthma, inflammatory bowel disease (IBD) and type 1 diabetes (T1D) for the further evaluation. The result showed that 10, 10 and 9 of top-10 predicted microbes were verified by previously published experimental results, respectively. It is anticipated that PRWHMDA can be effective to identify the disease-related microbes and maybe helpful to disclose the relationship between microorganisms and their human host.

Key words: disease-microbe association prediction; random walk with restart; particle swarm optimization.

Introduction

Microorganisms include bacteria, archaea, protists, fungi and viruses [1]. In human body, there are numerous microorganisms colonizing the gut, skin, uterus, lung, saliva, and so on [2]. They play an important role in human health, metabolism, immune defense, nutrition absorption, cancer control and protection against pathogen colonization [3, 4].

Numerous lab experiments have revealed that many diseases are associated with the changes of microorganisms. For example, evidences from both rodent models and human researches suggested that the composition of the gut microorganisms had a significant influence on the immune system and might influence TID risks [5]. Fusobacterium was much more abundant in asthmatic patients comparing with healthy group [6]. Some evidences revealed that lecitinase-negative Clostridium and Lactobacillus were detected to be more abundant in colorectal carcinoma patients [7, 8]. More and more clinic studies revealed new associations between microbes and human diseases.
Computational biology can help identify essential gene [9, 10] and secretory protein [11] in disease, identify immunoglobulin [12], reveal the relationship between diseases and environmental factors, predict drug reactions, identify the early diagnostic markers of diseases, and carry out computer-aided drug design and production [13]. Recently, a lot of computational biologic methods provide novel and effective tools for identifying the key associations between microbes and diseases. Ma et al. constructed a human microbe-disease associations database, which facilitated the study of the relationships between microbes and diseases [8]. Based on the Gaussian kernel similarity, some prediction methods were studied as follows. Chen et al. presented a computational model (KATZHMDA) based on KATZ method, the number of walks, and their lengths for calculating the potential association probability between microbes and diseases [14]. For high prediction accuracy, Huang et al. proposed the prediction model (PBHMDA) with a depth-first search algorithm to get all possible paths between microbes and diseases [15]. Wang et al. adopted Laplacian regularized least squares classifier (LRLSHMDA) to construct a prediction model [16]. Shen et al. developed a model utilizing Collaborative Matrix Factorization for Human Microbe-Disease Association prediction (CMFHMDA) [17]. Various Random Walk Methods were utilized to rank the probabilities of predicted microbe-disease association. Based on traditional random walk with three parameters on the heterogeneous network constructed by Spearman correlation, Shen et al. derived a method (RWRHMDA) for prioritization of candidate microbes to predict disease-microbe association [18]. Zou et al. developed a computational model (BiRWHMDA) to predict potential microbe-disease associations by bi-random walk on the heterogeneous network [19]. It is anticipated that various computational prediction models could improve the identification of novel microbe-disease association.

Researching the interaction of the microorganism with disease will provide critical insight into the pathogenesis of disease and the development of strategies to prevent and treat disease. However, the knowledge in this field is far from satisfied. The aim of this article is to present a computational model to predict potential microbe-disease associations based on the known ones. In this study, we put forward a novel computational model of extended Random Walking with Restart optimized by PSO on the heterogeneous interlinked network of Human Microbe-Disease Associations (PRWHMDA) (see Figure 1). Firstly, under the hypothesis that microbes involved in similar disease tend to be functionally similar, we computed the similarities of diseases and microbes and diseases [15]. Wang et al. adopted Laplacian regularized least squares classifier (LRLSHMDA) to construct a prediction model [16]. Shen et al. developed a model utilizing Collaborative Matrix Factorization for Human Microbe-Disease Association prediction (CMFHMDA) [17]. Various Random Walk Methods were utilized to rank the probabilities of predicted microbe-disease association. Based on traditional random walk with three parameters on the heterogeneous network constructed by Spearman correlation, Shen et al. derived a method (RWRHMDA) for prioritization of candidate microbes to predict disease-microbe association [18]. Zou et al. developed a computational model (BiRWHMDA) to predict potential microbe-disease associations by bi-random walk on the heterogeneous network [19]. It is anticipated that various computational prediction models could improve the identification of novel microbe-disease association.
probability of reaching the node was, the more believable the association was. PSO was utilized to get
the four optimal parameters of RWR. With PSO, PRWHMDA model performed well with 325 of 450
associations ranked in top 10. To evaluate the prediction performance, we performed the LOOCV
and 5-fold CV, made the comparison with some state-of-the-art methods and analyzed the three case
studies on the proposed model. The CV verified the effectiveness of the proposed model with the AUC of
0.915 (LOOCV) and average AUCs of 0.8875 ± 0.0046 (5-fold CV), which are higher than some other existing
methods. In our case studies, 10, 10 and 9 of top-10 inferred microbes have been confirmed to have
associations with asthma, IBD and T1D according to the literature evidences. The results indicate that
PRWHMDA is effective to identify the disease-related microbes and could be helpful in future to disclose the
relationship between microorganisms and their human host.

The paper is organized as follows. In Materials and Methods section, we describe the similarity
calculation using Cosine measurement. The construction of networks of microbes, diseases and
microbe-disease associations are also presented. Furthermore, the random walk with restart method is
outlined in this section. In Results and Discussion section, we show the optimized parameters by PSO of
RWR. In addition, the results of CV, comparison results with state-of-the-art methods and the top 10
most potential microbes for asthma, IBD and T1D are presented, respectively. In Conclusion part, we
discuss the results succinctly and conclude the paper.

Materials and Methods

Materials

We adopted the microbe-disease association data set called HMDAD which was constructed by Ma et al.
[8]. HMDAD contains manually curated 483 microbe-disease associations, which involves 39
diseases and 292 microbes. As some associations have more than one evidences, we extracted 450 distinct
disease-microbe associations. Based on these known microbe-disease associations, we constructed the
network of diseases, the network of microbes, and the network of microbe-disease associations, respectively. In this paper, \( N_d = 39 \) denotes the number of diseases, and \( N_m = 292 \) denotes the number of microbes.

Construction of the disease network

In the disease-disease similarity network, there are 39 nodes representing the diseases. The value of
the edge between two nodes represents the similarity of the two diseases. We utilized the Cosine similarity
measurement to calculate disease similarity with the known microbe-disease associations.

The disease-disease similarity matrix was

\[
SD = (sd(d_i,d_j))_{N_d \times N_d}, \tag{1}
\]

where \( sd(d_i,d_j) \) calculated by (2) denoted the similarity of the two diseases \( d_i \) and \( d_j \).

The similarity of two diseases \( d_i \) and \( d_j \) was calculated by Cosine similarity as

\[
sd(d_i,d_j) = \cos(MS(d_i), MS(d_j)) \tag{2}
\]

\[
= \frac{MS(d_i) \cdot MS(d_j)}{|MS(d_i)||MS(d_j)|},
\]

where

\[
MS(d_i) = [M_1(d_i),\ldots,M_m(d_i),\ldots,M_{N_d}(d_i)], \tag{3}
\]

\( MS(d_i) \) denoted the associations of disease \( i \) with all the microbes,

\[
M_{m i}(d_i) = \begin{cases} 
1, & \text{if } m_i \text{ is associated with } d_i, \\
0, & \text{otherwise},
\end{cases}
\tag{4}
\]

\( M_{m i}(d_i) \) denoted whether microbe \( m_i \) was associated with disease \( i \).

Therefore, the disease-disease similarity network could be constructed by nodes representing the
diseases and edges representing the similarities defined by (1).

Construction of microbe network

Similarly, in the microbe-microbe similarity network, there are 292 nodes representing the
microbes. The value of the edge between two nodes is the similarity of the two microbes.

The microbe-microbe similarity matrix was

\[
SM = (sm(m_i,m_j))_{N_m \times N_m}, \tag{5}
\]

where \( sm(m_i,m_j) \) calculated by (6) denoted the similarity of the two microbes \( m_i \) and \( m_j \).

The similarity of two microbes \( m_i \) and \( m_j \) was defined by Cosine similarity as

\[
sm(m_i,m_j) = \cos(DS(m_i), DS(m_j)) \tag{6}
\]

\[
= \frac{DS(m_i) \cdot DS(m_j)}{|DS(m_i)||DS(m_j)|},
\]

where

\[
DS(m_i) = [D_1(m_i),\ldots,D_j(m_i),\ldots,D_{N_d}(m_i)], \tag{7}
\]

\( DS(m_i) \) denoted the associations of microbe \( i \) with all the diseases,

\[
D_j(m_i) = \begin{cases} 
1, & \text{if } m_i \text{ is associated with } d_j, \\
0, & \text{otherwise},
\end{cases}
\]

which denoted whether disease \( j \) was associated
with microbe i.

Therefore, the microbe-microbe similarity network could be constructed according to adjacency matrix defined by (5).

**Construction of heterogeneous interlinked network**

Based on the two networks, i.e., microbe-microbe network and disease-disease network, we used the known microbe-disease associations to construct the heterogeneous interlinked network. In the heterogeneous interlinked network, there are two types of node sets defined as \( D = \{d_1, d_2, \ldots, d_{Nd}\} \) and \( M = \{m_1, m_2, \ldots, m_{Nm}\} \) corresponding to diseases and microbes, respectively.

The associations between microbes and diseases was

\[
MD = (md(m_i, d_j))_{Nm \times Nd},
\]

where

\[
md(m_i, d_j) = \begin{cases} 1, & \text{if } m_i \text{ is associated with } d_j, \\ 0, & \text{otherwise}, \end{cases}
\]

\((i = 1, 2, \ldots, Nm, j = 1, 2, \ldots, Nd)\),

\(md(m_i, d_j)\) denoted the association or non-association pattern between microbe \( m_i \) and disease \( d_j \).

The adjacency matrix of the network was

\[
A = \begin{pmatrix}
SM & MD \\
DM & SD
\end{pmatrix}_{(Nm+Nm) \times (Nm+Nd)},
\]

where \( SM \) denoted the similarity of microbes inferred by (5), \( SD \) denoted the similarity of diseases inferred by (1), \( MD \) defined by (8) denoted the associations of microbes and diseases, \( DM \) denoted the transpose of the matrix \( MD \). In this way, the heterogeneous interlinked network was constructed.

**Random walk with restart on heterogeneous interlinked network**

RW was first put forward by Karl Pearson in 1905 [20]. In Bioinformatics, it has been applied to disease-gene prediction [21] and inferring gene-phenotype relationship [22]. To emphasize the different weights of jumping between different networks, we extended the traditional RW method to predict microbe-disease association.

The walker can walk in disease network and microbe network, respectively and jump to the other network by disease-microbe network. Thus, at each step, the probabilities of reaching the nodes could be calculated. For the \((s+1)\)-th step, the probability was

\[
\begin{pmatrix}
p_{m_{i1}}^{s+1} \\
p_{m_{i2}}^{s+1} \\
\vdots \\
p_{m_{INm}}^{s+1}
\end{pmatrix} = (1 - \gamma) \ast (M^{\ast} \ast (\begin{pmatrix}
p^0_{m_{i1}} \\
p^0_{m_{i2}} \\
\vdots \\
p^0_{m_{INm}}
\end{pmatrix} + \gamma \ast \begin{pmatrix}
\eta \ast p_{d_{i1}}^1 \\
\eta \ast p_{d_{i2}}^1 \\
\vdots \\
\eta \ast p_{d_{INd}}^1
\end{pmatrix})),
\]

where \( p_{m_{is}} \) denoted the \( s \)-th iteration probability of reaching microbe \( m_i \), \( p^0_{m_{is}} \) denoted the initial probability of microbe \( m_i \), \( \gamma \in (0, 1) \) denoted the restart probability, \( \eta \) denoted the weights of the two networks. \( M^\ast \) defined by (11) denoted the transition matrix as

\[
M^\ast = \left[ \begin{array}{c}
SM^\ast \\
DM^\ast \\
SD^\ast
\end{array} \right].
\]

\( SM' \) that represented the transition probability from a microbe to a microbe was defined by

\[
SM' = \left( sm'(i, j) \right)_{Nm \times Nm},
\]

where

\[
sm'(i, j) = \frac{(1 - \lambda) \ast sm(m_i, m_j)}{\sum_i sm(m_i, m_j)},
\]

\( sm'(i, j) \) denoted the transition probability from microbe \( m_i \) to \( m_j \), \( \lambda \in (0, 1) \) denoted that the walker stayed in the microbe network with the weight \((1-\lambda)\).

\( MD' \) that represented the transition probability from microbe network to disease network was defined by

\[
MD' = \left( md'(i, j) \right)_{Nm \times Nd},
\]

where

\[
md'(i, j) = \frac{\lambda \ast md(m_i, d_j)}{\sum_i md(m_i, d_j)},
\]

\( md'(i, j) \) denoted the transition probability from microbe \( m_i \) to disease \( d_j \), \( \lambda \in (0, 1) \) denoted that the walker jumped from the microbe network to disease network with the weight \( \lambda \).

\( DM' \) that represented the transition probability from disease network to microbe network was defined by

\[
DM' = \left( dm'(i, j) \right)_{Nd \times Nm},
\]

where

\[
dm'(i, j) = \frac{\xi \ast md(m_i, d_j)}{\sum_i md(m_i, d_j)},
\]

\( dm'(i, j) \) denoted the transition probability from disease network to microbe network with the weight \( \xi \).
disease $d_i$ to microbe $m_j$, $\xi \in (0, 1)$ denoted the weight of jumping from disease to microbe network.

$SD'$ represented the transition probability from a disease to a disease, defined by

$$SD' = \left( sd'(i, j) \right)_{d_i,d_j}, \quad (15)$$

where

$$sd'(i, j) = \frac{(1 - \xi)sd(d_i, d_j)}{\sum \limits_{i} sd(d_i, d_i)},$$

$sd'(i, j)$ denoted the transition probability from disease $d_i$ to disease $d_j$, $(1 - \xi)$ denoted the weight of staying in disease network.

In (12)-(15), the value was set to be 0 if the denominator was 0. After several iterations, the difference between the vectors $p^{r+1}$ and $p^r$ become negligible such as $1e-10$. Thus, we got a stationary probability distribution which represented a probability distribution which represented a

Determination of parameters based on PSO with mutation

In the RW algorithm, the parameters $\lambda$, $\gamma$, $\eta$, and $\xi$ are important to the prediction performance. For example, in Figure 2, if with parameters 1 ($\lambda = 0.99$, $\gamma = 0.078$, $\eta = 0.44$, $\xi = 0.10$), the ROC of LOOCV is shown as the pink curve with AUC = 0.49. While with parameters 2 ($\lambda = 0.35$, $\gamma = 0.2$, $\eta = 0.25$, $\xi = 0.6$), the ROC of LOOCV is shown as the blue curve with AUC = 0.8996. In order to get the optimal parameters i.e. $\lambda$, $\gamma$, $\eta$, and $\xi$, we utilized the PSO method.

PSO is a population-based stochastic optimization technique, proposed by Kennedy and Eberhart (1995) [23]. In PSO, the solution was thought of as a bird, called a particle. In this study, all particles were searched in a four dimensional space with values between 0 and 1. All particles (parameters values) were evaluated by a fitness-function (16) to judge whether the current solution was good. Considering the accuracy and high AUC, the fitness function $F(x_i)$ was defined as

$$F(x_i) = N_{Actual} - N_{Predicted}(x_i), \quad (16)$$

where $N_{Actual} = 450$ denoted the number of actual disease-microbe associations, $N_{Predicted}(x_i)$ denoted the number of tested associations ranked top 10 with parameter $x_i = [\lambda, \gamma, \eta, \xi]^T$. The target was to get a suitable parameters vector to get the minimum value of $F(x_i)$.

In the canonical PSO, each particle had position and velocity. For example, the position $x_i$ and velocity $v_i$ of particle $i$ was updated at each iteration according to

$$v_i = w^* v_i + c1^* r1^* (p_{bi} - x_i) + c2^* r2^* (gb - x_i), \quad (17)$$

$$x_i = x_i + v_i, \quad (18)$$

where $v_i$ denoted the velocity of the $i$-th particle, $w$ denoted the weight of current velocity, $c1$ and $c2$ denoted two positive constants representing learning parameters, $r1$, $r2$ denoted a random number between 0 and 1, $p_{bi}$ denoted the optimal parameters value of the $i$-th particle had searched for, $gb$ denoted global best particle, $x_i$ denoted the current value of the $i$-th particle. In order to avoid the locally optimal solution, we allowed particles to mutate with a certain probability, i.e. they could reset their positions sometimes. Thus, each particle’s velocity was updated according to (17) and (18) until the convergence criterion was satisfied. After the convergence criterion was satisfied, we got global best particle to be the optimal parameters.

Figure 2: Prediction performance of PRWHMDA with different parameters.

Results and Discussion

Determination of parameters based on PSO with mutation

We set $c1 = 1.49445$, $c2 = 1.49445$, and set iteration times to be 27, the size of population to be 2, respectively. The iteration result is shown in Figure 3. The $x$-axis represented the iteration times, while the $y$-axis represented the fitness or called predicted error number of associations. Figure 3 showed that at the first iteration, there were 447 wrongly predicted associations, after 26 generations the number reduced to 125. Finally, the parameters calculated by PSO were $\lambda = 0.6$, $\gamma = 0.5$, $\eta = 0.76$, $\xi = 0.79$.

Cross validation

To effectively evaluate the prediction performance of PRWHMDA, LOOCV was implemented on the known disease-microbe associations from HMDAD database. Each known microbe-disease
association was taken out in turns as a test association and the others were used for training to reconstruct the prediction model. Each test association was ranked in all unverified microbe-disease associations with their final probabilities. Receiver-operating characteristics (ROC) curve is widely used in binary classification problems. By changing the thresholds, the true positive rate (TPR, sensitivity) versus false positive rate (FPR, 1-specificity) can be calculated to plot the ROC curves. In the proposed method, TPR refers to the percentage of the positive test samples with higher ranks than the specific threshold, and FPR refers to the percentage of negative test samples with higher ranks than the specific threshold. AUC was calculated to further evaluate the prediction performance. As a result, our model obtained the AUC of 0.915 (see Figure 4 in red line).

Furthermore, 5-fold CV was carried out. All the known microbe-disease associations were randomly divided into five disjoint groups, in which one group was for testing in turn and the other four groups were for training model. To reduce the bias brought by sample divisions, we executed the 5-fold CV 100 times to evaluate the robustness of PRWHMDA. As a result, the average AUC value was 0.8875 with the standard deviation of 0.0046.

**Comparison with other methods**

KATZHMD [14], RW3RHMDA [18] and BiRWHMDA [19] are state-of-the-art methods for predicting microbe-disease associations. All these methods are based on a heterogeneous network which was constructed by connecting the microbe similarity network and the disease similarity network via the known microbe-disease associations. To further validate the performance of the proposed method, we implemented these three methods using the same data sets and calculated the LOOCVs (see Figure 4). Comparing to other methods, the proposed method got a better ROC and higher AUC, while RWR3HMDA, KATZHMDA and BiRWHMDA had AUC values of 0.8112, 0.8332, and 0.8964, respectively.

**Case studies**

To further evaluate the prediction performance of PRWHMDA, we also implemented case studies of asthma, IBD, and T1D to predict novel associated microbes. The association strength score calculated by the proposed method and the benchmark was also presented, respectively. We used the RWR with three parameters as the benchmark.

**Asthma**

Asthma is a common chronic lung disease. To evaluate the prediction performance on asthma, we implemented a case study of asthma with our approach. In the prediction list, 10 of top-10 predicted microbes have been verified to have an impact on the asthmatic patients (see Table 1).

For example, *Actinobacteria*, *Firmicutes*, *Lachnospiraceae*, and *Veillonella* (1st, 2nd, 6th and 7th in the prediction list) were found to have a lower proportion in asthmatic patients when compared to non-asthmatic people [24, 26, 27]. *Clostridium coccoides* (3rd in the prediction list) was found significantly associated with a positive Asthma Predictive Index (API) [28]. Based on the effect of *Streptococcus pneumonia* (4th in the prediction list) on the development of asthma, effective immune-modulatory therapies were hoped to be presented [29]. *Lactobacillus* (5th in the prediction list) could inhibit airway inflammation in an ovalbumin (OVA)-induced murine model of asthma [30]. *Propionibacterium acnes* (9th in the prediction list) was more prevalent in asthma patients; therefore, *Propionibacterium* (8th in the prediction list) was also considered to be associated with asthma [31].
Fusobacterium (10th in the prediction list) was much more abundant in asthmatic patients relative to healthy people [25]. These predictions may give an insight into the potential prevention action of asthma.

**Inflammatory bowel disease (IBD)**

In the prediction list of IBD, 10 of top-10 predicted microbes have been validated by experimental literatures (see Table 2).

Clostridium cocoides (1st in the prediction list) were less represented in A-IBD patients, compared to healthy subjects [32]. In the IBD group, Firmicutes (2nd in the prediction list) was significantly increased, whereas Bacteroidetes (3rd in the prediction list) was decreased. At genus level Streptococcus (6th in the prediction list) and Veillonella (9th in prediction list) were increased [33]. The results showed that Bacteroidetes (3rd in the prediction list) was significantly increased in Proteobacteria in the salivary microbiota of IBD patients. The dominant genera Prevotella (4th in the prediction list), Haemophilus (7th in the prediction list) and Veillonella (9th in prediction list) were found to largely contribute to dysbiosis (dysbacteriosis) observed in the salivary microbiota of IBD patients [34]. Conventional culture showed that the counts of total obligate anaerobes and of obligate anaerobes such as Bacteroidaceae (10th in the prediction list) was decreased in the faeces of IBD patients. On the other hand, the counts of total facultative anaerobes such as Lactobacillus (8th in the prediction list) were increased [35]. Research showed that dysbiosis within deeper layers of the ileum of CD patients, there were specifically enrichment of enterotoxigenic Staphylococcus aureus. Staphylococcus (5th in the prediction list) was thus validated [36].

**Type 1 diabetes**

In the prediction list of T1D, 9 of predicted microbes in the top 10 have been validated by experimental literatures (see Table 3).

At the genus level, a significant increase in the number of Clostridium (Genus of 1st in the prediction list) was found in the children with diabetes [38]. It was found that Enterobacteriaceae (2nd in the prediction list) was increased in patients with type 1 diabetes compared to the control group. A concomitant increase in Enterobacteriaceae may lead to a disturbance in the ecological balance of intestinal flora, which could be a triggering factor in type 1 diabetes etiology [39]. Furthermore, subjects with autoantibodies, seronegative FDRs, and new-onset patients had different levels of the genera Firmicutes, Lactobacillus, and Staphylococcus (3rd in the prediction list) compared with healthy control subjects with no family history of autoimmunity [40]. An increase of Faecalibacterium prausnitzii (4th in the prediction list) might result in a disturbance in the ecological balance, which could cause type 1 diabetes [41]. At genus level, Mora Murri et al. [38] found a significant increase in the number of Clostridium (5th in the prediction list) in the children with diabetes. A study was performed on non-obese prediabetic (NOD) mice. Metabolic changes of NOD mice were accompanied by diminished gut microbial diversity of the Clostridium leptum group (6th in the prediction list) [42]. Two species in the genus Staphylococcus were abundantly presented in the conjunctiva of healthy rats, and were replaced by Enterococcus species (7th in the prediction list) and other bacterial species in diabetic rats [43]. Members of the genus Desulfovibrio (8th in the prediction list) were identified in the colon samples of the control rats, with <0.5% abundance. It was noted that the insulin treatment was accompanied by fundamental changes within the phylum Proteobacteria: their overall abundance decreased spectacularly, and the genus Klebsiella was basically eradicated, being replaced by the genera Desulfovibrio (8th in the prediction list) and Bifidobacterium, with a combined abundance of 4% [37]. GCC-fed NOD mice had the expected high incidence of hyperglycemia whereas NOD mice fed with a GFC had significantly reduced incidence of hyperglycemia. When the fecal microbes were compared, Tannerella species (9th in the prediction list) were increased in the microbes of GCC mice[44].

Moreover, as shown in Tables 1-3, the association strength score of the proposed method is higher than that of the benchmark. In summary, these case studies further demonstrate that the proposed approach is effective to predict novel microbe-disease associations. The top-10 potential associated microbes for all the 39 diseases are listed in Table s1 of Supplementary File.

**Table 1: Candidate microbes of Asthma**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Microbe</th>
<th>Evidence</th>
<th>Score of PRW-HMDA (%)</th>
<th>Score of RWRS-HMIDA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Actinobacteria</td>
<td>PMID: 23265859</td>
<td>0.23</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>Firmicutes</td>
<td>PMID: 23265859</td>
<td>0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>3</td>
<td>Clostridium cocoides</td>
<td>PMID: 21477358</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>Streptococcus</td>
<td>PMID: 17950502</td>
<td>0.13</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>Lactobacillus</td>
<td>PMID: 20592920</td>
<td>0.12</td>
<td>0.05</td>
</tr>
<tr>
<td>6</td>
<td>Lachnospiraceae</td>
<td>Validated by [4]</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>7</td>
<td>Veillonella</td>
<td>PMID: 25329665</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td>8</td>
<td>Propionibacterium</td>
<td>PMID: 27433177</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>9</td>
<td>Propionibacterium acetica</td>
<td>PMID: 27433177</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>10</td>
<td>Fusobacterium</td>
<td>Validated by [25]</td>
<td>0.09</td>
<td>0.03</td>
</tr>
</tbody>
</table>

http://www.ijbs.com
benefits from several major factors as the following. Progress of biomedical identification of potential PRWHMDA could be an effective tool accelerating the promising prediction performance, it is believed that T1D according to the literature evidences. Given the 10, 10, and 9 of top-10 inferred microbes have been average AUCs of 0.8875 ± 0.0046. In our case studies, LOOCV with AUC of 0.915 and 5-fold CV with probability, which represented candidate microbe-disease associations. PSO was incorporated into the training data set, PSO with mutation was adopted to get the optimal parameters. (4) We used the extended RW method that has adjustable parameters to emphasize the probabilities of transitions of different networks. There are some limitations in the performance of PRWHMDA. Because the experimentally verified microbe-disease associations used in our approach are relatively insufficient, the sparse association network could affect the predictive capability. It is anticipated that this problem will be solved when collecting more microbe-disease associations in the future. Furthermore, PRWHMDA cannot be applied to diseases and microbes which are not in HMDAD. It is believed that the network would be powerful with the introducing of other types of disease/microbe similarity based on different kinds of data, such as symptom-based disease similarity, etc. We think that this method can guide medical experiments to get the potential associations.

**Supplementary Material**
Table S1. http://www.ijbs.com/v14p0849s1.pdf

**Abbreviations**
PSO: Particle Swarm Optimization; RWR: random walk with restart; LOOCV: leave one out cross validation; AUC: The area under ROC curve; IBD: inflammatory bowel disease; T1D: type 1 diabetes; NOD: non-obese prediabetic.

**Acknowledgements**
This work was supported by the Natural Science Foundation of China [Grant Numbers 61473335, 61533011].

**Competing Interests**
The authors have declared that no competing interest exists.

**References**
5. Davis-Richardson AG, Ardissonne AN, Dias R, Simell V, Leonard MT, Kemppainen KM, et al. Bacteroides dorei dominates gut microbiome prior to...