Hindawi
BioMed Research International
Volume 2020, Article ID 3978702, 8 pages
https://doi.org/10.1155/2020/3978702

Research Article

Human Microbe-Disease Association Prediction by a Novel Double-Ended Random Walk with Restart

Di Wang,[1] Yan Cui,[2] Yuxuan Cao,[2] Yuehan He,[3] and Hui Chen[2]

[1]*Department of Nuclear Medicine, Harbin Medical University Cancer Hospital, Harbin, China*
[2]*Department of Urology, Harbin Medical University Cancer Hospital, Harbin, China*
[3]*Harbin Medical University College of Bioinformatics Science and Technology, Harbin, China*

Correspondence should be addressed to Hui Chen; hmuhuichen@sina.com

Received 12 May 2020; Revised 7 July 2020; Accepted 20 July 2020; Published 11 August 2020

Guest Editor: Jialiang Yang

Microorganisms in the human body play a vital role in metabolism, immune defense, nutrient absorption, cancer control, and prevention of pathogen colonization. More and more biological and clinical studies have shown that the imbalance of microbial communities is closely related to the occurrence and development of various complex human diseases. Finding potential microbial-disease associations is critical for understanding the pathology of a few diseases and thus further improving disease diagnosis and prognosis. In this study, we proposed a novel computational model to predict disease-associated microbes. Specifically, we first constructed a heterogeneous interconnection network based on known microbe-disease associations deposited in a few databases, the similarity between diseases, and the similarity between microorganisms. We then predicted novel microbe-disease associations by a new method called the double-ended restart random walk model (DRWHMDA) implemented on the interconnection network. In addition, we performed case studies of colon cancer and asthma for further evaluation. The results indicate that 10 and 9 of the top 10 microorganisms predicted to be associated with colorectal cancer and asthma were validated by relevant literatures, respectively. Our method is expected to be effective in identifying disease-related microorganisms and will help to reveal the relationship between microorganisms and complex human diseases.

1. Introduction

Microorganisms include bacteria, archaea, protozoa, fungi, and viruses. There are different types of microorganisms on the human body and in the cavity connected to the outside world, such as the oral cavity, respiratory tract, intestinal tract, and urogenital tract [1, 2]. Microbes play important roles in human health, metabolism, immune defense, nutrient absorption, cancer control, and prevention of colonization of pathogens [3]. Microorganisms of the human body are mainly distributed on the body surface, intestine, and oral cavity, and the types and numbers of microorganisms are different. Among them, the number of microorganisms in the intestine is about ten times that of the body's own cells. In nature, the density of microorganisms isolated from the colon is the highest, and 60% of the dry weight of human feces is bacteria [4].

Numerous studies have shown that many diseases are related to changes in microorganisms. For example, patients with type 2 diabetes have been found to have moderate intestinal microecological disorders and lack of butyric acid-producing bacteria [5]. Intestinal microbial diseases lead to intestinal immune system dysfunction. For patients with irritable bowel syndrome (IBS), the number of chronic inflammatory cells in the colonic mucosa of increases, a large number of T cells are activated, and the expression of inflammatory reactions is accelerated [6]. In addition, epidemiological studies have shown that common mental illnesses such as autism and schizophrenia are associated with perinatal pathogen infections [7–11].

As mentioned above, discovering the potential links between microbes and diseases allows us to better understand the mechanisms by which disease is formed and developed. By regulating the microbial environment, medical solutions

for disease prevention, diagnosis, treatment, and prognosis can be provided to some extent. In previous biological or clinical experimental research methods, it took a lot of time and cost to obtain a new connection between microorganisms and disease. In recent years, many computational biology methods have provided new and effective tools for identifying the key links between microorganisms and disease. Ma et al. constructed a microbe-disease association data pants called HMDAD, which can help study the relationship between microbes and diseases [12], and provide data support for various calculation methods to predict new associations

In recent years, machine learning algorithms have achieved good performance in various fields [13]. At present, various machine learning algorithms have been used in the prediction of the association between microorganisms and diseases and have achieved good performance. As such, Chen et al. established a microorganism-human disease association network and further developed a new KATZ metric calculation model for the prediction of human microorganism-disease association (KATZHMDA) under the premise that similar-function microorganisms tend to the following assumptions [14]. Huang et al. [15] proposed a path-based human microorganism-disease association prediction (PBHMDA) method that integrates the identified nuclear-similarities of disease-microbe relationships and Gaussian interaction spectra into a heterogeneous network of diseases and microorganisms. The model traverses all possible pathways between microbes and diseases. A novel depth-first search algorithm is used to predict the microorganisms most likely to be associated with the disease. In addition, Wang et al. [16] proposed a new computational model of Laplace regularized least squares to reveal potential disease-related microorganisms. LRLSHMDA applies a semisupervised learning framework. In this model, a microbial similarity network and a disease similarity network are constructed based on the Gaussian interaction spectrum kernel similarity calculated from known disease-disease associations, and the cost function in the microbial space and disease space is then constructed and optimized integrating the optimal classifier function to calculate the correlation probability of microbial disease pairs. Although the reliable prediction performance of LRLSHMDA has been verified, the model still has some shortcomings and needs further improvement. For example, the number of proven microbial associations is too small, and a sparse network of known associations may affect the predictive performance of the model. Shen et al. [17] combined the known similarity of microbe-disease association with the nuclear similarity of the Gaussian interaction spectrum; a collaborative matrix decomposition calculation model was established for the microbial-disease association prediction (CMFHMDA) of humans and diseases. A special matrix decomposition algorithm is proposed to update the correlation matrix between microorganisms and diseases and infers the microorganisms most likely to be related to diseases. However, the performance of this model needs improvement.

In summary, though the tremendous progress made in computing predictions of microbial-disease associations, there are still some limitations. In order to better reveal the association between microbial diseases, based on the known heterogeneous network consisting of microbial-disease association and Gaussian interaction contour kernel similarity, we propose a computational model based on a double-ended restart random walk to predict disease-related microorganisms. To prove the superiority of the DRWHMDA algorithm, we applied the 5-fold CV and global LOOCV to evaluate the prediction performance of DRWHMDA. In addition, we used DRWHMDA for case studies of two diseases.

2. Materials and Methods

2.1. Materials. The general workflow of DRWHMDA is shown in Figure 1. First, we need to preprocess the data. The original data comes from a microbe-disease association dataset named HMDAD constructed by Ma et al. [13]. HMDAD contains 483 artificially planned microbiological associations involving 39 diseases and 292 microorganisms. Because there are multiple evidences for some associations, we extracted 450 different disease-microbial associations. Secondly, based on these known microbial-disease associations, we constructed disease networks, microbial networks, and microbial-disease related networks, respectively. Here, $N_d = 39$ indicates the number of diseases, and $N_m = 292$ indicates the number of microorganisms. Finally, a two-terminal random walk is performed through a heterogeneous network. Combine different prediction scores into the final associated prediction probability according to the linear combination.

2.2. Symptom-Based Disease Similarity (SDM). In the field of information retrieval, text documents or concepts are usually represented by feature vectors. Here, we describe the vector d_j of each disease j through symptoms.

$$d_j = \left(w_{1,j}, w_{2,j}, \cdots, w_{n,j}\right), \tag{1}$$

where $w_{i,j}$ quantifies the strength of the association between symptom i and disease j. The prevalence of different symptoms and diseases is very different. In order to solve this heterogeneity, we do not use absolute co-occurrence $w_{i,j}$ to measure the strength of the association between symptom i and disease j, but the term frequency and the reciprocal of the document frequency $w_{i,j}$.

$$w_{i,j} = w_{i,j} \log \frac{N}{n_i}, \tag{2}$$

where N represents the number of all diseases in the data set and n_i represents the number of diseases with symptom i.

Therefore, the similarity between the vectors d_x and d_y of the two diseases x and y is calculated as follows:

$$\cos\left(d_x, d_y\right) = \frac{\sum_i d_{x,i}, d_{y,i}}{\sqrt{\sum_i {d_{x,i}}^2}\sqrt{\sum_i {d_{y,i}}^2}}. \tag{3}$$

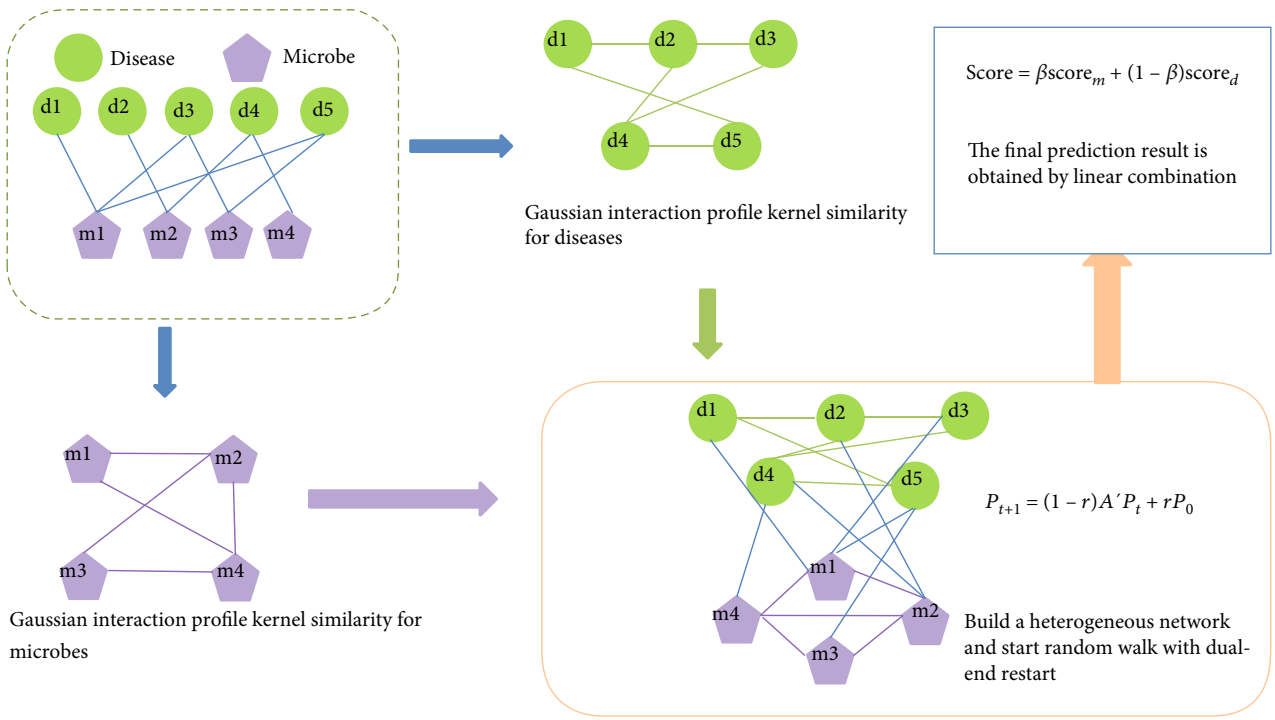

Figure 1: The workflow of DRWHMDA for inferring potential microbe-disease associations.

The cosine similarity ranges from 0 (no shared symptoms) to 1 (identical symptoms).

2.3. Effect of Gaussian Interaction Spectroscopy Nuclear Similarity on Disease. Based on the assumption that diseases with similar phenotypes always share similar associations and nonassociative patterns with functionally similar microorganisms, the Gaussian interaction distribution kernel similarity between disease and disease can be further calculated. We define the binary vector $\mathrm{VP}(d_i)$ to represent the interaction curve of disease d_i, which can be obtained by observing whether d_i is known to be associated with each microorganism (i.e., the ith row of the adjacency matrix). Then, after calculating the similarity value between disease pairs, the Gaussian interaction distribution kernel similarity matrix (KD) can be constructed.

$$\mathrm{KD}(d_i, d_j) = \exp\left(-\gamma_d \left|\left|\mathrm{VP}(d_i) - VP(d_j)\right|\right|^2\right) \tag{4}$$

$$\gamma_d = \gamma'_d / \left(\frac{1}{nd}\sum_{i=1}^{nd} \left|\left|\mathrm{VP}(d_i)\right|\right|^2\right) \tag{5}$$

The parameter value γ_d controls the bandwidth of the Gaussian kernel. As shown in (5), γ'_d can be further calculated by dividing the new bandwidth parameter γ'_d. The average of each disease is associated with microorganisms. Here, we $\gamma'_d = 1$ according to previous research [18].

From the above, we can see that the similarity of the Gaussian interaction spectrum kernel is only based on adjacency matrix A. If we want to effectively and scientifically predict potential disease-related microorganisms, it is necessary to incorporate other data sets similar to the Gaussian interaction spectrum kernel, recorded in PubMed bibliography based on disease and corresponding symptoms. Zhou et al. (2014) calculated similarities between diseases and established a symptom-based human disease network (HSDN). Here, we synthesize the Gaussian interaction spectrum kernel similarity of disease KD and symptom-based disease similarity SDM to obtain symptom-based disease similarity SD, and SD is calculated as follows:

$$\mathrm{SD} = \frac{\mathrm{KD} + \mathrm{SDM}}{2}. \tag{6}$$

2.4. Gaussian Interaction Spectrum Nuclear Similarity for Microbes. In the same way, the Gaussian interaction similarity m_i and m_j between microorganisms can be obtained as the Gaussian kernel similarity matrix (KM) between microorganisms.

$$\begin{gathered}\mathrm{KM}(m_i, m_j) = \exp\left(-\gamma_{\mathrm{m}} \left|\left|\mathrm{VP}(m_i) - \mathrm{VP}(m_j)\right|\right|^2\right). \\ \gamma_m = \frac{\gamma'_m}{1/nm\sum_{i=1}^{nm}\left|\left|\mathrm{VP}(m_i)\right|\right|^2}.\end{gathered} \tag{7}$$

where γ'_m is usually set to 1.

2.5. Building a Heterogeneous Network. A heterogeneous network can be expressed as $G = (D, E)$, where D represents 331

of all diseases and microorganisms and E represents the interaction of microorganisms and microorganisms, diseases and diseases, and diseases and microorganisms. The heterogeneous network is represented by $n * n$ adjacency matrix A, where n represents the number of diseases and microorganisms. By the similarity between the microorganisms (KM) and the similarity between the diseases (KD), the coefficients of similarity can construct a heterogeneity network. Then, for each adjacency matrix A, if there is an interaction between A_i and A_j, the i-th row and j-th column are set to 1, otherwise set to zero. Normalize adjacency matrix A:

$$A'_{[i,j]} = \frac{A_{[i,j]}}{\sum_{k=1}^{n} A_{[k,j]}}. \tag{8}$$

2.6. Restart Random Walk Algorithm in Both Directions. Through heterogeneous networks, random walks are used to find potential genetic association data between diseases or microorganisms. By randomly walking to convergence, you can get the probability of a disease or microbe at every point in the heterogeneous network. The relationship between microorganisms and diseases is indicated by calculating the correlation between the probability distributions of disease and microorganisms.

For a disease, we list all relevant diseases and microorganisms in our known data set, and then our related diseases and microbial collections are the seeds of the disease. Among them,

$$P_{\text{dis}} = [\psi_{\text{dis}^-1}, \psi_{\text{dis}^-2}, \cdots, \psi_{\text{dis}^-n}]^T. \tag{9}$$

Among them, the disease-related diseases and microbial aggregates ψ_{dis^-i} were set to 1, and the others were set to zero. Normalize P_{dis}:

$$P'_{\text{dis}[k]} = \frac{P_{\text{dis}[k]}}{\sum_{k=1}^{n} P_{\text{dis}[k]}}. \tag{10}$$

Similarly, we list all relevant diseases and microorganisms for a certain microbe, we know the data set, and then, we related diseases and microbial collections as P_{mic}.

Begin random walks and randomly access adjacent genes in each time scale ($t \longrightarrow t+1$). State probability P_{t+1} at time $t+1$:

$$P_{t+1} = (1-r)A' P_t + rP_0, \tag{11}$$

where P_t is the probability of time t and r is the probability of restart. According to previous studies, we set r to 0.7 [19]. If the difference between P_t and P_{t+1} is less than 10^{-6} used in the previous study, the process will reach a steady state [20, 21]. By using the mapped set P_{dis} as the seed of the disease and the mapped set P_{mic} as the seed of the microbe, we implemented a two-way random walk algorithm to obtain the association probability score_d with disease as the random seed and the association probability score_m with microorganism as the random seed. The association probability score between the disease d_i and the microorganism m_i is finally obtained by linearly combining the two predicted probabilities.

$$\text{Score} = \beta \text{score}_m + (1-\beta)\text{score}_d, \tag{12}$$

where β represents the parameter of the linear combination; we set the default value to 0.7.

3. Results

3.1. Performance Evaluation. To verify the predictive performance of DRWHMDA, we implemented 5-fold CV and global LOOCV on the model based on the HMDAD database. In each 5-fold CV, the known correlation matrix Y is divided into 5-folds; then each fold is taken as a test set, and the remaining 4 folds are treated as a training set. On the other hand, in the global LOOCV, each known microbial-disease association is sequentially excluded from the test, and other microbial-disease associations are used as training samples for model learning. Specifically, all microbial-disease pairs without known evidence of correlation will be considered candidate samples. Further obtain the rank of each missing test sample relative to the candidate sample. Test samples with a prediction level above a given threshold will be considered to have successfully predicted. We evaluated the predictive performance of the model based on the AUC value of the area under the curve of the receiver. Specifically, only test samples ranked above a certain threshold can be considered correct predictions. We then set the true-positive rate (TPR, sensitivity) and false-positive rate (FPR, 1 − specificity) as the horizontal axis and the vertical axis, respectively. Therefore, we can draw a receiver operating characteristic (ROC) curve composed of points corresponding to different thresholds and then obtain the area (AUC) under the ROC curve. A model with an AUC value equal to 0.5 is equivalent to random prediction. When the AUC takes a maximum of 1, the model has excellent prediction performance. In other words, when the value of AUC is greater than 0.5 and less than 1, the larger the value is, the better the prediction performance of the model.

As shown in Figure 2, the 5-fold CV value of DRWHMDA was 0.8676, which was significantly larger than those of KATZHMDA (0.8382), LRLSHMDA (0.8493), and ABHMDA (0.8571). What was more, the global LOOCV value of our model reached 0.8897, which was also obviously better than those of KATZHMDA (0.8644), LRLSHMDA (0.8843), and ABHMDA (0.8861). These results confirmed the superior prediction performance of DRWHMDA.

To investigate the selection of restart probability r for the performance of DRWHMDA, we set various values of r ranging from 0.1 to 0.9 and calculated AUC in the framework of 5-fold CV. As shown in Table 1, as the restart probability r gradually increases, the prediction performance obtained through DRWHMDA increases first and then decreases.

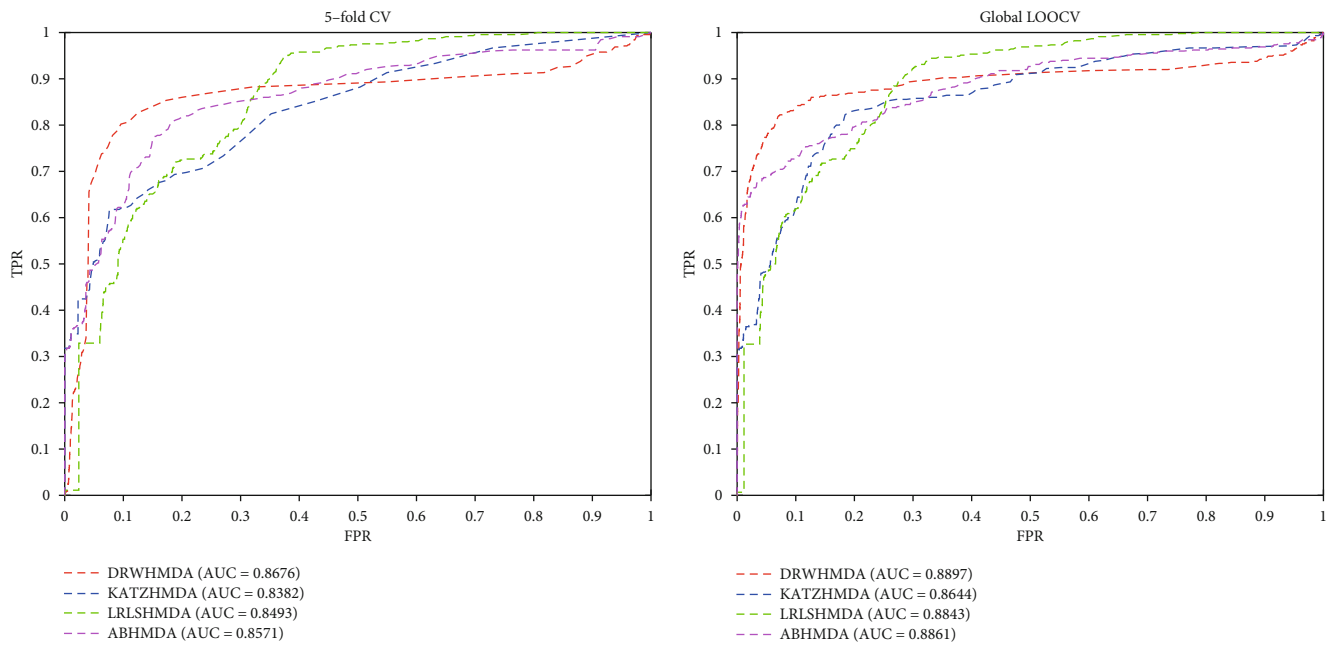

Figure 2: The ROC curves for DRWHMDA and other approaches in microbe-disease association prediction for 5-fold cross-validation and global LOOCV.

Table 1: Prediction AUCs of DRWHMDA at different choices of restart probability r.

DRWHMDA	AUC	DRWHMDA	AUC
$r = 0.1$	0.8511	$r = 0.6$	0.8684
$r = 0.2$	0.8513	$r = 0.7$	0.8674
$r = 0.3$	0.8525	$r = 0.8$	0.8666
$r = 0.4$	0.8597	$r = 0.9$	0.8590
$r = 0.5$	0.8695		

3.2. Case Study. In the present study, double-ended random walks were used to screen candidate microorganisms for all the investigated diseases. To further evaluate the predictive performance of DRWHMDA, we included 10,038 unknown samples in HMDAD, involving 39 diseases and 292 microorganisms. The corresponding unknown samples are classified and ranked by the DRWHMDA algorithm, and it is verified whether the relevant literature has verified the association between the top ten microorganisms and the disease under study. Among them, an independent case analysis was performed on colon cancer and asthma.

3.3. Relationship between Colon Cancer and Microorganisms. According to previous research, the intestinal microflora is the most complex, and it is most closely related to various behavioral diseases in humans. Imbalance of the human intestinal microbial flora can lead to autoimmune diseases [22], obesity [23, 24], inflammatory bowel disease (IBD) [25], diabetes [26], and even cancer [27, 28]. According to the world's leading cancer statistics report, colon cancer has been a high-risk area for men and women over the past few decades [29]. Therefore, it is necessary to study the pathogenesis of colon cancer in order to explore new treatment methods. More and more evidences show that the imbalance of microbial community is closely related to the occurrence and development of colon cancer. For example, in the sequence analysis of 16S rRNA gene V3 region in patients with sporadic colorectal cancer, the protein bacteria are insufficient [30]; Staphylococcus produces tannase; its activity may be related to the development of colon cancer [31]. Compared with noncancerous tissues, Lactococcus and Fusarium are more abundant in cancerous tissues, and Pseudomonas and Escherichia coli are less abundant [32]. We applied DRWHMDA to the first case study of colon cancer. Of the top 10 predicted microorganisms, 9 have been validated based on recent experimental literature (see Table 2). Evidence suggests that Clostridium difficile- (first in the

Table 2: The 10 microbes predicted to be most likely to be associated with colon cancer.

Microbe	Evidence
Clostridium difficile	PMID:21152135
Helicobacter pylori	PMID:22294430
Protein bacteria	PMID:25699023
Prevotella	PMID:25699024
Staphylococcus aureus	Unconfirmed
Clostridium globosum	PMID:18237311
Fermicket	PMID:25699024
Bacteroides	PMID:25699024
Actinomycetes	PMID:26811603
Clostridium	PMID:19807912

Table 3: The 10 microbes predicted to be most likely to be associated with asthma.

Microbe	Evidence
Clostridium difficile	PMID:25974301
Fermicket	PMID:23265859
Clostridium globosum	PMID:21477358
Actinomycetes	PMID:23265859
Staphylococcus aureus	PMID:12743582
Lactobacillus	PMID:20592920
Clostridium	PMID:21477358
Burkholder	PMID:24451910
Gracilariaceae	PMID:17433177
Lachnospiraceae	PMID: 27433177

prediction list) associated colitis is a known complication of colon and rectal surgery and can increase morbidity and mortality during surgery, thereby increasing hospital stay time and costs [33].

3.4. The Relationship between Asthma and Microbes. Asthma is a common chronic inflammatory disease of the lungs and is generally thought to be caused by a combination of genetic and environmental factors. According to the latest statistics, the incidence of asthma has been rising in recent decades, with the number of asthma patients increasing from 183 million in 1990 to 242 million in 2013. Infection by pathogenic microorganisms (especially viruses, chlamydia, mycoplasma, and mold) is one of the main causes of severe asthma [34–37]. For example, studies have shown that Proteus accounts for a higher proportion of microorganisms in asthma patients and that Firmicutes are reduced in asthma patients compared to normal people. Moreover, there is evidence that when the hypopharyngeal area of a newborn is infected with Streptococcus pneumoniae, the risk of developing asthma is increased compared to uninfected [38, 39]. Therefore, research on asthma-related microorganisms is crucial and will help us to gain a deeper understanding of the pathogenesis and treatment of asthma. Prioritizing candidate microorganisms by implementing DRWHMDA, recent clinical evidence successfully validated 9 of the top 10 predicted microorganisms (see Table 3). As for the top five confirmed asthma-associated microorganisms, Clostridium difficile and Staphylococcus aureus (No. 1 and No. 5 in the prediction table) were found to be increased in number in airway concentrations in asthma patients, while Firmicutes and Actinomycetes were found to be reduced [40]. Importantly, the XIVa subclass of Clostridium globosum (No. 3 in the prediction table) has been proven to be an early indicator of future asthma, help prevent and diagnose asthma, and provide guidance for clinical treatment.

For clarity, we illustrate in Figure 3 the association network of the top 10 predicted microbial candidates for two diseases. It is worth noting that some top candidates were found to be related to several diseases. For example, both Fermicket and Clostridium have been documented to prove that they are related to the occurrence of asthma and colon cancer at the same time.

4. Discussion

Over the years, a lot of evidence has shown that microorganisms living in the human body are closely related to human life activities and human diseases. Abnormal levels of specific microorganisms are closely related to the development of various human diseases. Microbial disease-related knowledge can provide valuable insights into understanding complex disease mechanisms and preventing, diagnosing, and treating various diseases. However, little work has been done to predict microbial candidates for large-scale human complex diseases. Therefore, in this paper, a computational model based on known microbial-disease correlation is proposed. A microbial similarity network and a disease similarity network are constructed using Gaussian kernel similarity. Using the existing experimentally validated associations, we connected the two networks. The double-ended restart random walk method is used to walk on the network, and the correlation probability order representing the candidate microorganism-disease association is obtained. The construction network with different correlations is applied to the optimization of prediction performance, and the optimal prediction parameters are obtained. The results show that DRWHMDA achieved average AUC reliability performance of 0.8676 and 0.8897 in the 5-fold cross-validation and LOOCV framework, respectively. Given its good predictive performance, we believe that the model can be used as one of the effective tools to accelerate biomedical identification of underlying disease-related microorganisms.

Although DRWHMDA has achieved satisfactory results, this method still has some limitations. First, we only use Gaussian kernel similarity to construct a similarity network which is too simplistic. Improving the predictive performance of DRWHMDA by integrating disease or microbe similarity from multiple data sources (such as sequence similarity) may help. Secondly, as more and more microbes and disease associations are identified, collecting more validated data will help us conduct further research. Finally, we experimentally verify candidate microbes related to the disease, and some have not been verified in the literature, because the verification of these candidate microbes through biological wet experiments will also be one of the important directions for our subsequent research.

5. Conclusion

The main goal of the current research is to predict the microorganisms that may be related to the disease through the calculation method, thereby reducing the verification cost of the biological wet experiment, so that people can more deeply explore the impact of microorganisms on human complex diseases. Therefore, this paper proposes a calculation model of microbial disease correlation based on double-ended random walk. The results show that DRWHMDA has achieved more reliable and stable

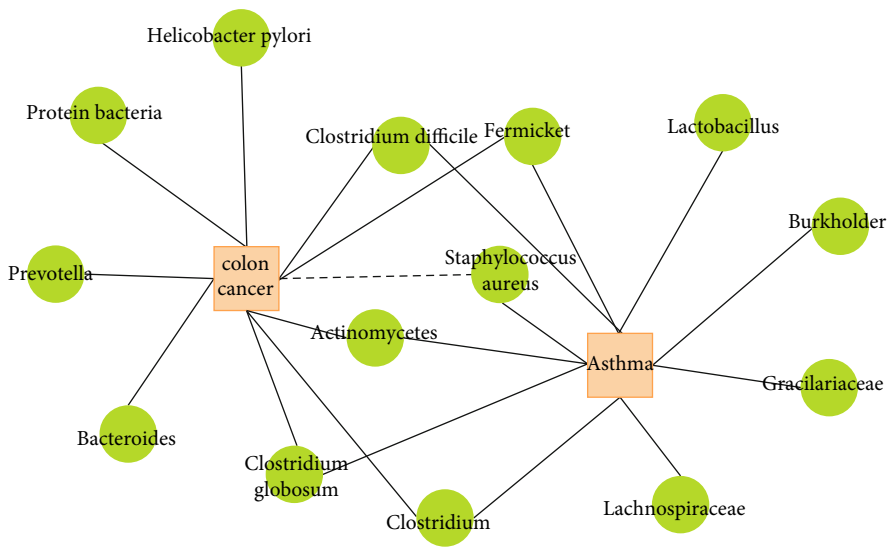

Figure 3: The network of the top 10 predicted associations for the two diseases via DRWHMDA. The dotted line indicates that it has not been confirmed by the literature.

prediction performance than other algorithms. We believe that the model can be used as one of the effective tools for accelerating biomedical identification of potential disease-related microorganisms.

Data Availability

The database used in this study was downloaded from the Human Microbe-Disease Association Database (HMDAD, http://www.cuilab.cn/hmdad).

Conflicts of Interest

The authors declare no competing interests.

Authors' Contributions

Di Wang and Hui Chen conceived the concept of the work. Yan Cui, Yuxuan Cao, and Yuehan He performed the experiments. Yan Cui and Di Wang wrote the paper. All authors approved the final version of this manuscript. Di Wang and Yan Cui contributed equally to this work.

Acknowledgments

This research was funded by the Harbin Applied Technology Research and Development Project (grant no. 2017RAXXJ073), Hai Yan Foundation of Harbin Medical University Cancer Hospital (grant no. JJMS2016-04), Wu Jieping Medical Foundation Technology Research Project (grant no. 320.6750.17107), and Heilongjiang Postdoctoral Financial Assistance (LBH-Z16216).

References

[1] M. Ventura, S. O'Flaherty, M. J. Claesson et al., "Genome-scale analyses of health-promoting bacteria: probiogenomics," *Nature Reviews. Microbiology*, vol. 7, no. 1, pp. 61–71, 2009.

[2] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight, "Bacterial community variation in human body habitats across space and time," *Science*, vol. 326, no. 5960, pp. 1694–1697, 2009.

[3] B. Das and G. B. Nair, "Homeostasis and dysbiosis of the gut microbiome in health and disease," *Journal of Biosciences*, vol. 44, no. 5, p. 44(5), 2019.

[4] J. M. Pickard, M. Y. Zeng, R. Caruso, and G. Núñez, "Gut microbiota: role in pathogen colonization, immune responses, and inflammatory disease," *Immunological Reviews*, vol. 279, no. 1, pp. 70–89, 2017.

[5] J. Qin, Y. Li, Z. Cai et al., "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, 2012.

[6] M. Pimentel, E. J. Chow, and H. C. Lin, "Eradication of small intestinal bacterial overgrowth reduces symptoms of irritable bowel syndrome," *The American Journal of Gastroenterology*, vol. 95, no. 12, pp. 3503–3506, 2000.

[7] V. A. Mittal, L. M. Ellman, and T. D. Cannon, "Gene-environment interaction and covariation in schizophrenia: the role of obstetric complications," *Schizophrenia Bulletin*, vol. 34, no. 6, pp. 1083–1094, 2008.

[8] S. M. Finegold, D. Molitoris, Y. Song et al., "Gastrointestinal microflora studies in late-onset autism," *Clinical Infectious Diseases*, vol. 35, Suppl 1, pp. S6–s16, 2002.

[9] Y. Chen, F. Yang, H. Lu et al., "Characterization of fecal microbial communities in patients with liver cirrhosis," *Hepatology*, vol. 54, no. 2, pp. 562–572, 2011.

[10] J. Xu, L. Cai, B. Liao et al., "Identifying potential miRNAs–disease associations with probability matrix factorization," *Frontiers in Genetics*, vol. 10, p. 1234, 2019.

[11] J. Xu, W. Zhu, L. Cai et al., "LRMCMDA: predicting miRNA-disease association by integrating low-rank matrix completion with miRNA and disease similarity information," *IEEE Access*, vol. 8, pp. 80728–80738, 2020.

[12] W. Ma, L. Zhang, P. Zeng et al., "An analysis of human microbe-disease associations," *Briefings in Bioinformatics*, vol. 18, no. 1, pp. 85–97, 2017.

[13] J. Xu, L. Cai, B. Liao, W. Zhu, and J. L. Yang, "CMF-impute: an accurate imputation tool for single-cell RNA-seq data," *Bioinformatics*, vol. 36, no. 10, pp. 3139–3147, 2020.

[14] X. Chen, Y. A. Huang, Z. H. You, G. Y. Yan, and X. S. Wang, "A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 34, no. 8, p. 1440, 2018.

[15] Z.-A. Huang, X. Chen, Z. Zhu et al., "PBHMDA: path-based human microbe-disease association prediction," *Frontiers in Microbiology*, vol. 8, p. 233, 2017.

[16] F. Wang, Z. A. Huang, X. Chen et al., "LRLSHMDA: Laplacian regularized least squares for human microbe-disease association prediction," *Scientific Reports*, vol. 7, no. 1, p. 7601, 2017.

[17] Z. Shen, Z. Jiang, and W. Bao, "CMFHMDA: Collaborative Matrix Factorization for Human Microbe-Disease Association Prediction," *Intelligent Computing Theories and Application*, 2017.

[18] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[19] H. Shi, J. Xu, G. Zhang et al., "Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes," *BMC Systems Biology*, vol. 7, no. 1, p. 101, 2013.

[20] M. Hofree, "Network-based stratification of tumor mutations," *Nature Methods*, vol. 10, no. 11, pp. 1108–1115, 2013.

[21] X. Zhong, H. Yang, S. Zhao, Y. Shyr, and B. Li, "Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes," *BMC Genomics*, vol. 16, Suppl 7, p. S7, 2015.

[22] Y. K. Lee and S. K. Mazmanian, "Has the microbiota played a critical role in the evolution of the adaptive immune system?," *Science*, vol. 330, no. 6012, pp. 1768–1773, 2010.

[23] M. Vijay-Kumar, J. D. Aitken, F. A. Carvalho et al., "Metabolic syndrome and altered gut microbiota in mice lacking toll-like receptor 5," *Science*, vol. 328, no. 5975, pp. 228–231, 2010.

[24] L. Abenavoli, E. Scarpellini, C. Colica et al., "Gut microbiota and obesity: a role for probiotics," *Nutrients*, vol. 11, no. 11, p. 2690, 2019.

[25] N. O. Kaakoush, A. S. Day, K. D. Huinao et al., "Microbial dysbiosis in pediatric patients with Crohn's disease," *Journal of Clinical Microbiology*, vol. 50, no. 10, pp. 3258–3266, 2012.

[26] L. Wen, R. E. Ley, P. Y. Volchkov et al., "Innate immunity and intestinal microbiota in the development of type 1 diabetes," *Nature*, vol. 455, no. 7216, pp. 1109–1113, 2008.

[27] M. Castellarin, R. L. Warren, J. D. Freeman et al., "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma," *Genome Research*, vol. 22, no. 2, pp. 299–306, 2012.

[28] M. Eslami, S. Sadrifar, M. Karbalaei, M. Keikha, N. M. Kobyliak, and B. Yousefi, "Importance of the microbiota inhibitory mechanism on the Warburg effect in colorectal cancer cells," *Journal of Gastrointestinal Cancer*, 2019.

[29] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: a Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[30] Z. Gao, B. Guo, R. Gao, Q. Zhu, and H. Qin, "Microbiota disbiosis is associated with colorectal cancer," *Frontiers in Microbiology*, vol. 6, p. 20, 2015.

[31] Q. Zhu, Z. Jin, W. Wu et al., "Analysis of the intestinal lumen microbiota in an animal model of colorectal cancer," *PLoS One*, vol. 9, no. 3, article e90849, 2014.

[32] N. Noguchi, M. Fukuzawa, T. Wajima et al., "Specific clones of Staphylococcus lugdunensis may be associated with colon carcinoma," *Journal of Infection and Public Health*, vol. 11, no. 1, pp. 39–42, 2018.

[33] J. P. Furet, L. C. Kong, J. Tap et al., "Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers," *Diabetes*, vol. 59, no. 12, pp. 3049–3057, 2010.

[34] M. E. Dumas, R. H. Barton, A. Toye et al., "Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 33, pp. 12511–12516, 2006.

[35] F. Backhed, H. Ding, T. Wang et al., "The gut microbiota as an environmental factor that regulates fat storage," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 44, pp. 15718–15723, 2004.

[36] L. Ma, Y. P. Pan, and J. Q. Zhang, "Study of putative periodontal pathogens colonies in type 2 diabetes with chronic periodontitis," *Shanghai Kou Qiang Yi Xue*, vol. 19, no. 6, pp. 611–615, 2010.

[37] R. Hu, F. Zeng, L. Wu et al., "Fermented carrot juice attenuates type 2 diabetes by mediating gut microbiota in rats," *Food & Function*, vol. 10, no. 5, pp. 2935–2946, 2019.

[38] R. Murphy, P. Tsai, M. Jüllig, A. Liu, L. Plank, and M. Booth, "Differential changes in gut microbiota after gastric bypass and sleeve gastrectomy bariatric surgery vary according to diabetes remission," *Obesity Surgery*, vol. 27, no. 4, pp. 917–925, 2017.

[39] J. Long, Q. Cai, M. Steinwandel et al., "Association of oral microbiome with type 2 diabetes risk," *Journal of Periodontal Research*, vol. 52, no. 3, pp. 636–643, 2017.

[40] C. Vael, L. Vanheirstraeten, K. N. Desager, and H. Goossens, "Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma," *BMC Microbiology*, vol. 11, no. 1, p. 68, 2011.