MDADP: A Webserver Integrating Database and Prediction Tools for Microbe-Disease Associations

Lei Wang, Hao Li, Yuqi Wang, Yihong Tan, Zhiping Chen, Tingrui Pei, and Quan Zou, *Senior Member, IEEE*

***Abstract*—More and more evidence has demonstrated that microbiota play important roles in the life processes of the human body. In recent years, various computational methods have been proposed for identifying potentially disease-associated microbes to save costs in traditional biological experiments. However, prediction performances of these methods are generally limited by outdated and incomplete datasets. And moreover, until now, there are limited studies that can provide visual predictive tools for inferring possible microbe-disease associations (MDAs) as well. Hence, in this manuscript, a novel webserver called MDADP will be proposed to identify latent MDAs, in which, a new MDA database together with interactive prediction tools for MDAs studies will be designed simultaneously. Especially, in the newly constructed MDA database, 2019 known MDAs between 58 diseases and 703 microbes have been manually collected first. And then, through adopting the average ranking method and the co-confidence method respectively, eight representative computational models have been integrated together to identify potential disease-related microbes. As a result, MDADP can provide not only interactive features for users to access and capture MDAs entities, but also effective tools for users to identify candidate microbes for different diseases. To our knowledge, MDADP is the first online platform that incorporates a new MDA database with comprehensive MDA prediction tools. Therefore, we believe that it will be a valuable source of information for researches in microbiology and disease-related fields. MDADP can be accessed at http://mdadp.leelab2997.cn.**

***Index Terms*—Association database, association prediction tool, disease, microbe.**

I. Introduction

MICROORGANISMS in human bodies consist mainly of bacteria, archaea, fungi, viruses and protozoa, which are usually parasitized in various human organs such as the gastrointestinal tract, respiratory tract, oral cavity, stomach, skin and genitourinary tract [1]. Since microbiota are ubiquitous in human bodies, they are also known as another vital organ of the human body [2]. In recent years, more and more evidence has proven that microbiota play important roles in certain physiological processes of human bodies, such as improving metabolism, enhancing immunity and maintaining the ecological balance of the body [3], [4]. In addition, they can as well be central or causative agents of many diseases [5]. For instance, studies showed that microorganisms are associated with about 20% of human malignancies [6]. Up to now, with rapid advances in clinical biotechnologies and sequencing technologies, researches on microbiome have experienced exponential growth, which lead to mounting microbe-disease associations (MDAs) being uncovered [7], [8]. Mining potential MDAs can reveal more useful biomedical information in disease-related areas (e.g., disease-causing genes and drugs) and is expected to provide new strategies for disease diagnosis and treatment [9]. For example, in the field of drug repurposing, it has been hypothesized and verified that drugs used to treat type 2 diabetes can also be used to treat colorectal cancer, due to the strong microbe correlation between these two diseases [10], [11]. Thus, understanding microbe-disease associations may be very useful for the diagnosis and treatment of complex diseases such as gastrointestinal inflammation, diabetes, and even cancer.

However, using traditional wet experimental methods to identify MDAs is quite expensive and time-consuming [12], [13]. For the past few years, with rapid developments of complex network technologies, machine learning and artificial intelligence techniques, in order to reduce the time, labor and cost of traditional biological experiments methods have been successively

Manuscript received August 13, 2021; revised January 20, 2022; accepted February 27, 2022. Date of publication March 7, 2022; date of current version July 4, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61873221 and in part by the Natural Science Foundation of Hunan Province under Grant 2019JJ70010. *(Corresponding authors: Hao Li; Quan Zou.)*

Lei Wang is with the College of Computer Engineering & Applied Mathematics, Changsha University, Changsha 410005, China, and also with the Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411100, China (e-mail: wanglei@xtu.edu.cn).

Yihong Tan, Zhiping Chen, and Tingrui Pei are with the College of Computer Engineering & Applied Mathematics, Changsha University, Changsha 410005, China (e-mail: yhtan@ccsu.edu.cn; zpchen@ccsu.edu.cn; peitingrui@xtu.edu.cn).

Hao Li and Yuqi Wang are with the Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411100, China (e-mail: leehao@smail.xtu.edu.cn; wangyuqi@smail.xtu.edu.cn).

Quan Zou is with the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610000, China (e-mail: zouquan@nclab.net).

Digital Object Identifier 10.1109/JBHI.2022.3156166

proposed to infer potential MDAs. For instance, Ma *et al.* constructed the first human microbe-disease association database (HMDAD) in 2017 based on known microbe-disease associations in publications of microbe-related studies, in which, contained 483 associations between 39 diseases and 292 microbes were selected from 61 publications [9]. Based on these known MDAs from HMDAD, Chen *et al.* proposed a microbe-disease association prediction model called KATZHMDA [14], and since then, lots of computational methods have been developed in succession by adopting diverse strategies. However, by far, few studies can provide visualized MDA prediction tools, which are not conducive to subsequent researches of MDA. In addition, almost all studies can only validate their prediction performances based on the HMDAD database, so that predictive performances of these models may to some degree be unreliable. Furthermore, limited by the number of combinations between microbes and diseases in HMDAD (there are only 39 diseases and 292 microbes in HMDAD), the difference between those potentially associated microbe-disease pairs identified by different models will be very small, even if the prediction performances differ significantly between them, which may severely restrict the application capabilities of those superior computational models. In addition, considering that the data in the HMDAD database are very limited and to some extent obsolete, it is possible that potential MDAs recommended by these predictive models may no longer be time-sensitive. Hence, in order to address the above-mentioned issues, in this study, a novel webserver called MDADP was designed by integrating a new MDA database with effective MDA prediction tools. In MDADP, we first manually collected 2019 known associations between 58 diseases and 703 microbes to construct the new MDA database. And then, through screening and integrating eight representative MDA prediction algorithms, we proposed two kinds of different predictive tools based on the average ranking method (MDADP _ AR) and the co-confidence method (MDADP _ CC) separately to recommend more reliable potential MDAs. To the best of our knowledge, MDADP is the first webserver that integrates a new MDA database and visualized MDA prediction tools. Hence, we believe that it may become a useful tool for future researches in microbiology and disease-related fields. The major contributions of this paper are as follows:

- A new MDA database consisting of 2019 known associations between 58 diseases and 703 microbes was constructed based on which, a novel MDA dataset with a scale of 1767 non-redundant known associations for identification of potential MDAs was built.
- By integrating and analyzing eight representative MDA prediction models, two kinds of effective identification models were designed to recommend reliable potential MDAs separately.
- Based on the newly constructed MDA database and identification models, a visualized platform was provided, in which, lots of functions including searching, sorting, filtering, visualization, and downloading of MDAs are implemented simultaneously. To our knowledge, MDADP is the first online platform that can provide visualized tools together with a new database for prediction of potential MDAs, which may be a useful tool for future research in microbiology and disease-related fields.

II. Materials and Methods

A. Construction of the New MDA Database

For constructing the new MDA database, a series of keywords to search the Pubmed database for human microbe-related publications, including but not limited to "Human", "Microbiome", "Disease", "Microbe", "Neoplasms", "Cancer", etc., were adopted to search for human microbe-related publications in the Pubmed database. After preliminary screening of publications published before the start of our study (September 2020), in final, more than 500 candidate publications were extracted. Subsequently, 261 publications were refined by reading the abstracts and results of them. All information of these selected studies was recorded into tabular files by meticulous manual management according to following rules: (1) All diseases will be named and classified in a standardized way according to the vocabulary provided by the MeSH database, (2) All microbes will be classified according to the NCBI taxonomy, (3) Regulatory relationships between microbes and diseases will be recorded (positive or negative) in the new database (4) Information on experimental methods and samples used in these publications will be recorded in the new database. Ultimately, 2019 validated MDAs between 58 different diseases and 703 different microbes were collected from 261 publications, among them, there were 1012 positive associations and 1007 negative associations. And besides, all these 58 diseases would be classified into 15 categories, such as Cardiovascular Diseases, Chemically-Induced Disorders, Digestive System Diseases, Endocrine System Diseases, Immune System Diseases, Infections, Mental Disorders, Mental Disorders, Nervous System Diseases, Nervous System Diseases, Pathological Conditions, Signs and Symptoms, Pathological Conditions, Signs and Symptoms, Skin and Connective Tissue Diseases, Stomatognathic Diseases and Urogenital Diseases. Meanwhile, according to the NCBI taxonomy, microbes would be further classified into phylum, class, order, family, genus, species, and no rank. As a result, the numbers of different disease-associated microbes and different microbe-associated diseases in MDADP are statistically presented in Fig. 1, while the 30 most common diseases and microbes in MDADP are shown in Fig. 2. As can be seen in Fig. 2(a), the most common disease is Type 1 Diabetes Mellitus, with more than 170 microbes associated with it. Additionally, the second and third common diseases are Breast Neoplasms and Lung Neoplasms separately, with 156 and 99 associated microbes. Besides, other kinds of human diseases recorded by MDADP can be found in different human organs, such as colon (Colorectal Neoplasms), oral cavity (Mouth Neoplasms), intestine (Irritable Bowel Syndrome), pancreas (Pancreatic Neoplasms), kidney (Kidney Disease), and so on. As can be seen in Fig. 2(b), it is obvious that the most common microbe is Bacteroides. Besides, all these 30 most common microbes have more than 13 related diseases. Obviously, these experimentally supported microbe-disease associations not only can serve as a source of data for predictive models but also can inspire bioinformatics researchers to mine more useful

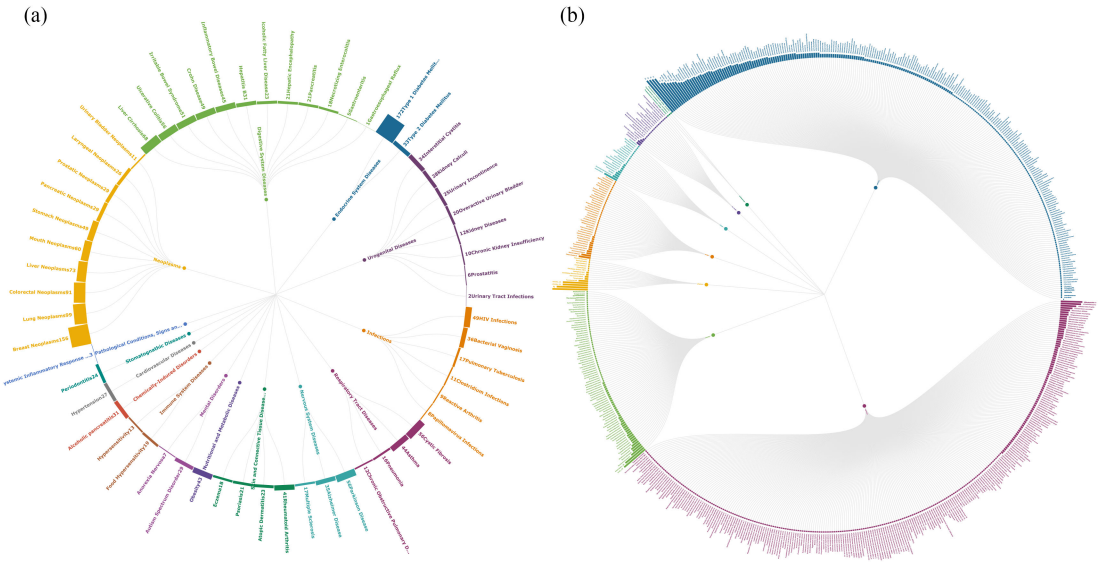

Fig. 1. Statistical chart of MDADP. (a) Statistics of the number of different disease related microbes in MDADP. (b) Statistics of the number of different microbe related diseases in MDADP.

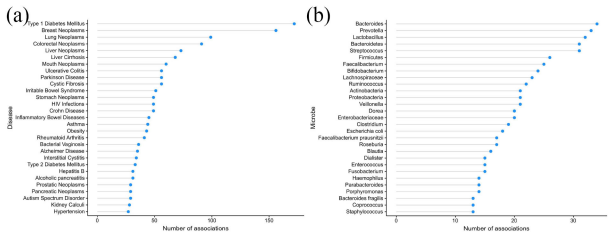

Fig. 2. (a) The 30 most common diseases in MDADP. (b) The 30 most common microbes in MDADP.

biological information in future studies. For example, Zhang *et al.* analyzed associations between autism and other diseases based on data downloaded from HMDAD [15]. Long *et al.* designed a computational method to identify microbe-drug associations based on HMDAD [16] as well.

B. MDA Dataset

According to above description, it is easy to know that the MDADP database contains 2019 validated MDAs between 58 diseases and 703 microbes. After removing duplicated associations, we finally obtained 1767 non-redundant known MDAs. And for convenience, we refer to the dataset of 58 human diseases as SD, and the i-th disease in SD as d_i, and similarly, we refer to the dataset of 703 microbes as SM, and the j-th microbes in SM as m_j. Thereafter, a 58 $\times$ 703 dimensional association matrix A can be constructed as follows: for any given disease d_i and microbe m_j, if and only if there is a known association between them, there is $A(i, j) = 1$, otherwise, there is $A(i, j) = 0$. Fig. 3 shows the bipartite network graph consisting of microbe nodes, disease nodes and edges (associations) in MDADP respectively.

In order to avoid random results caused by a single dataset, during experiments, HMDAD will be utilized as the alternative data source to verify the reliability of competitive computational

TABLE I
COMPARISON BETWEEN THE MDADP DATASET AND THE HMDAD DATASET

Dataset	Number of microbes	Number of diseases	Number of MDAs	Average number of associations per disease	Average number of associations per microbe
MDADP	703	58	1767	30.47	2.51
HMDAD	292	39	450	11.54	1.54

models. HMDAD is a MDA database constructed by Ma *et al.* in 2017, which contains 483 associations between 39 diseases and 292 microbes. After removing duplicated data, there are 450 non-redundant MDAs in HMDAD. Table I shows the comparison between datasets of MDADP and HMDAD. The scatter plots of the interaction distribution for these two datasets are drawn in Fig. 4. It is obvious that comparing with the MDADP dataset, the HMDAD dataset has less data and is obviously sparser.

C. Construction of Effective Models for Potential MDA Prediction

Recently, researchers have developed a variety of sophisticated MDAs prediction models based on the HMDAD database. Zhao *et al.* classified current MDAs computational models into four types, namely score function-based models, network algorithm-based models, machine learning-based models and experimental analysis-based models [13]. These models aim to identify potential MDAs by adopting various computational algorithms and further recommend top-k candidate MDAs. Although considerable successes have been achieved, the performances and applications of these models in different databases still deserve further investigation.

In this section, we will first select eight state-of-the-art MDA computational models to compare their performances based on the MDADP. Through comprehensive consideration of predictive performances, algorithm characteristics, code availability and reproducibility of all existing state-of-the-art models, the following eight models including two score function-based

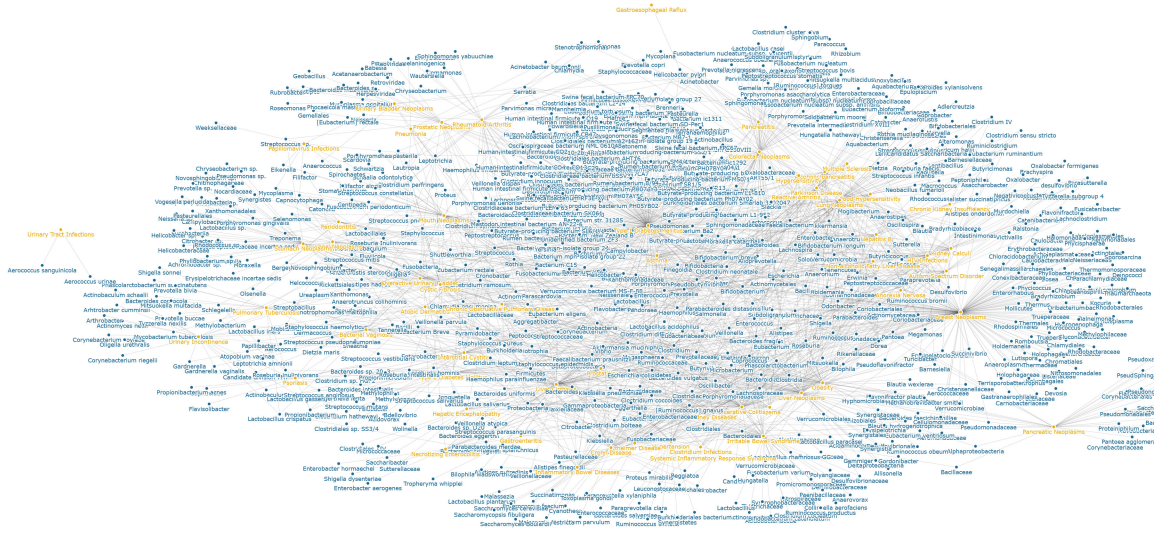

Fig. 3. Bipartite network graph based on the MDADP database.

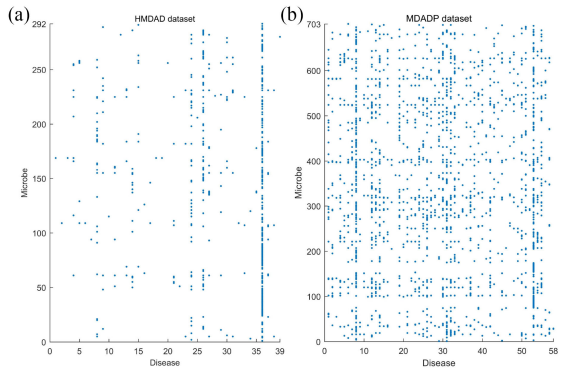

Fig. 4. Scatterplots of the HMDAD dataset and the MDADP dataset.

models (BWNMHMDA [17], KATZHMDA [14]), four network algorithm-based models (BiRWHMDA [17], NBLPHMDA [18], BiWMP [19] and HMDA-Pred [20]) and two machine learning-based models (LRLSHMDA [21], BPNNHMDA [22]) are selected as our final competitive models. And for the sake of fairness, while calculating potential similarities between diseases and microbes, all parameters of these competitive models will be assigned to default values given by their proposers. In addition, for convenience, we adopt the sequence numbers 1 to 8 to represent models of BWNMHMDA, KATZHMDA, BiRWHMDA, NBLPHMDA, BiWMP, LRLSHMDA, BPNNHMDA and HMDA-Pred separately. Thus, we can define the 58×703 dimensional predicted score matrix obtained by the k-th competitive model as S_k. And then, it is obvious that $S_k(i,j)$ represents the predicted score of potential association between a pair of given disease d_i and microbe m_j obtained by the k-th model. Thereafter, by ranking all predicted scores in S_k in terms of diseases, we can obtain a new ranking matrix R_k, where $R_k(i,j)$ denotes the ranking of the given microbe m_j in all candidate microbes relating to the given disease d_i. Specifically, for a given unknown microbe-disease pair, the smaller its ranking value, the higher its potential relevance.

Considering that for any given candidate MDA, the predicted scores obtained by different models may vary greatly, hence, it is problematic to achieve better predictive performance by simply summing together these predicted scores obtained by all competitive models. Hence, in this section, we will adopt the following two kinds of strategies to combine these eight representative models together to further construct two kinds of novel predictive models called MDADP _ AR and MDADP _ CC respectively:

(1) Average Ranking strategy based Model (named MDADP _ AR): In MDADP _ AR, the top K competitive models with the best predictive performances will be selected out first according to experimental results. And then, let M denote the set of sequence numbers of these K selected models, an average ranking matrix R_{av} will be obtained according to the following (1):

$$R_{av}(i,j) = \frac{\sum_{k \in M} R_k(i,j)}{K} \tag{1}$$

Obviously, the parameter K has key impact on the predictive performance of MDADP _ AR. According to experimental results, MDADP _ AR can achieve the best predictive performance while K is set to 3, and correspondingly, the top 3 competitive models with the best predictive performances are BWNMHMDA, NBLPHMDA and HMDA-Pred separately.

(2) Co-confidence strategy based Model (named MDADP _ CC): In MDADP _ CC, for a candidate MDA between any given disease d_i and microbe m_j, its co-confidence value will be obtained according to the following (2):

$$R_{cc}(i,j) = \sum_{k=1}^{8} W_k(i,j) \tag{2}$$

Where,

$$W_k(i,j) = \begin{cases} 1: & if\ R_k(i,j) \geqslant \delta \\ 0: & otherwise \end{cases} \tag{3}$$

Obviously, values of elements in above matrix R_{cc} will vary from 1 to 8, and the parameter δ affects the strictness of the co-confidence strategy. To ensure a stricter synergy confidence, δ will be set to 100 (about %14 of the total number of microbes) in MDADP _ CC. Specifically, for a given unknown microbe-disease pair, the larger its co-confidence value, the higher its potential relevance.

Through analyzing above two strategies, it is easy to know that the advantage of MDADP _ AR is that it can achieve better error tolerance by combining results obtained by top K competitive models with the best predictive performances, while the advantage of MDADP _ CC is that it can make models with poor prediction performances work well by obtaining co-confidence values of candidate MDAs.

In MDADP webserver, above eight representative predictive models and two strategies for ranking potential MDAs have been integrated for researchers to query more conveniently.

D. Implementation of the Server

xMDADP webserver with Model-View-Controller architecture is realized by using Python language and Flask front-end framework, and the platform has been deployed on Alibaba Cloud Elastic Computing Service. In addition, data of microbe-disease associations is stored in MDADP by MySQL database. Moreover, in the webserver, lots of functions including searching, sorting, filtering, visualization, and downloading of MDAs are implemented simultaneously. MDADP can be accessed at http://mdadp.leelab2997.cn. All MDAs datasets curated in this paper can be downloaded at https://github.com/HaoLeextu/MDADP.

III. Result

A. Assessment of Predictive Performance

The leave-one-out cross validation (LOOCV) and 5-fold cross validation (5-Fold CV) are two widely used frameworks for assessing performance of predictive models, in this section, we will adopt these two kinds of frameworks to evaluate performances of above competitive methods based on the MDADP dataset and the HMDAD dataset respectively. In LOOCV, all unknown MDAs are considered as candidate samples. Each known MDA will be excluded in turn as a test sample, and the remaining known associations are used as training samples. After executing the computational model, the predicted values of test samples are ranked against the predicted values of all candidate samples, and the test sample with ranking above a given threshold will be considered as a successful prediction. Obviously, different true positive rates (TPR, sensitivity) and false positive rates (FPR, 1-specificity) can be obtained when different thresholds are set. Here, the TPR refers to the percentage between the number of test samples with rankings above a given threshold and the number of known MDAs. Meanwhile, the FPR indicates the percentage of candidate samples ranked above a given threshold. Using the FPRs and TPRs under different thresholds as the x-axis and y-axis, respectively, the receiver operating characteristic (ROC) can be further plotted. Thereafter, the area under curve (AUC) can be taken to evaluate the prediction performance,

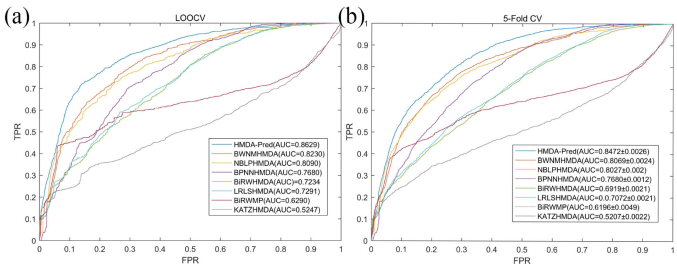

Fig. 5. (a) Performances achieved by eight candidate methods under frameworks of LOOCV based on the MDADP. (b) Performances achieved by eight candidate methods under frameworks of 5-Fold CV based on the MDADP dataset.

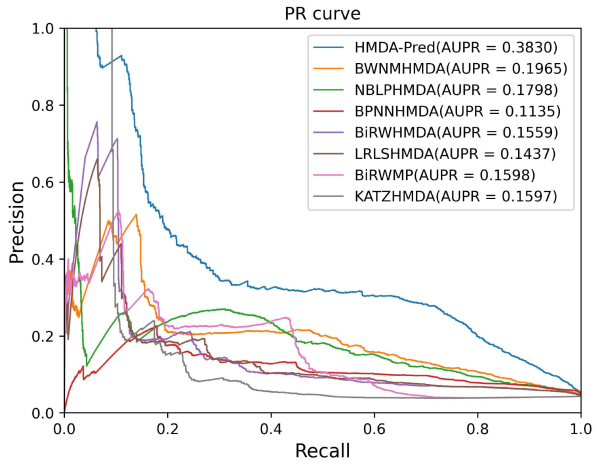

Fig. 6. The PR curve of eight candidate methods in LOOCV.

where the closer the AUC value is to 1 indicates the better prediction performance. 5-Fold CV is similar to LOOCV, except that it differs in dividing the samples. In the 5-Fold CV, all samples are divided equally into 5 parts. One part is selected as the testing set in each round, and the other 4 parts are used as the training set. Since the process of dividing samples is random, the 5-Fold CV is executed 100 times, and the stability of the prediction model is determined by calculating the average of all AUCs as the final result and calculating the standard deviation (STD). As another essential evaluation metric, the area under the PR curves(AUPR) can show the balance of recall and accuracy, and is therefore suitable for evaluating the prediction performance of different methods with unbalanced datasets [20]. We plotted the PR curves of eight methods and calculated their AUPR values by LOOCV.

ROC and PR curves for the eight candidate methods based on the MDADP dataset are illustrated in Figs. 5 and 6, respectively. Obviously, the first three models with the best prediction performances are BWNMHMDA, NBLPHMDA, and HMDA-Pred. For comparing the performance of these candidate methods in recovering known MDAs, we statistically counted the number of true MDAs among the top 200, 500, 1000, 2000, and 5000 associations predicted by each model in LOOCV, as illustrated in Fig. 7. It can be seen from Fig. 7 that HMDA-Pred and BWNMHMDA achieved better results at all thresholds. NBLPHMDA performed poorly at the threshold of 200, but performed well at

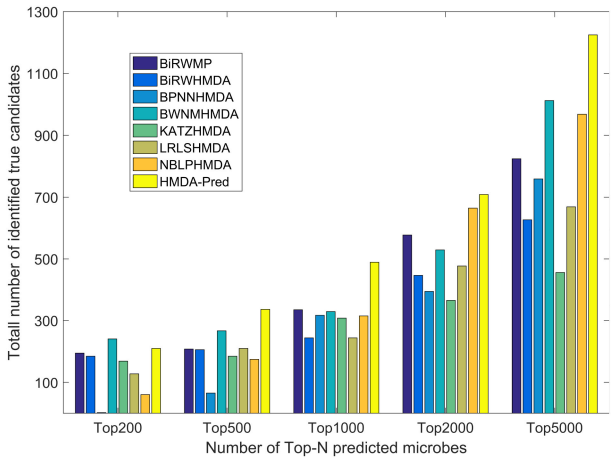

Fig. 7. Performances achieved by eight candidate methods in recovering known associations.

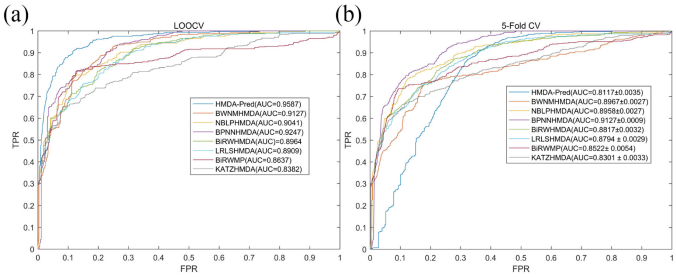

Fig. 8. (A) Performances achieved by seven candidate methods under frameworks of LOOCV based on the HMDAD dataset. (B) Performances achieved by seven candidate methods under frameworks of 5-Fold CV based on the HMDAD dataset.

remaining thresholds. It is remarkable that BiRWMP achieved relatively low AUCs but performed relatively well in this metric.

In order to further verify the performances of these candidate models, we present ROC curves and AUC values in LOOCV and 5-Fold CV for all candidate methods based on the HMDAD dataset as well. Through observing Figs. 5 and 8, it is easy to find that all candidate methods can achieve higher AUCs in HMDAD. Among them, the AUCs of BPNNHMDA, BWNMHMDA and NBLPHMDA are relatively higher than other four candidate methods. It is noteworthy that KATZHMDA performs poorly in MDADP. The main reason for this disparity is that prediction results of KATZHMDA are biased towards those well-investigated diseases and microbes. For instance, KATZHMDA can achieve excellent prediction performance (AUC value of 0.8382) for Type 1 Diabetes Mellitus in HMDAD, since the associations relating to Type 1 Diabetes Mellitus account for more than one-third of the total records in HMDAD (167 in 483 associations). However, this biased advantage is not manifested in MDADP where the data are relatively evenly distributed.

B. Case Studies of MDADP _ AR and MDADP _ CC

To further confirm the validity of MDADP in predicting potential MDAs, Alzheimer disease would be selected as the case disease. During experiments, we selected and compared the top 5 microbes associated with the case disease predicted by MDADP _ AR and MDADP _ CC for investigation. The Alzheimer disease is one of the most common neurodegenerative diseases, affecting more than 50 million people worldwide. It is extremely troublesome to human health and is the fifth major cause of death [23]. Numerous studies have suggested that the Alzheimer disease may begin in the gut and is closely linked to dysbiosis of intestinal flora [24]. Top 5 microbes relating to the Alzheimer disease predicted by MDADP based on average ranking method and co-confidence method are shown in Table II. Four of the top five microbes identified by MDADP _ AR and MDADP _ CC as being associated with the Alzheimer disease are confirmed by publications. Guo *et al.* performed 16S ribosomal RNA sequencing on stool samples from patients newly diagnosed with the Alzheimer disease and healthy controls, and showed that Prevotella was significantly increased at the genus level in AD patients [25]. Ivakhniuk *et al.* explored the relationship between composition of the intestinal microflora and Alzheimer's disease, and they found that patients with Alzheimer's disease expressed significantly decreased Lactobacillus in their gut microflora [26].

IV. Discussion and Future Work

What role microorganisms play in the physiological and pathological states of the human body is a current hot research topic [27]. A growing number of studies show a close relationship between microbes and human diseases. In recent years, researchers around the world have been devoting to studying the complex relationship between microbes and diseases to provide new strategies for disease prevention, diagnosis and treatment [24], [28], [29]. The utilization of computational models to uncover MDAs can provide new perspectives to reveal disease mechanisms, by playing a role in areas such as the discovery of disease-causing genes and drug therapies [9]. Recent studies have found high similarities between type 2 diabetes and colorectal cancer, and based on their association with microorganisms, it was hypothesized that both diseases could be treated with the same drug [12]. This hypothesis was successfully tested and validated by biomedical scientists [11], [30]. Hence, the high-performing MDA computational model by leveraging the similarity feature and diverse biomedical data is expected to reduce the time, effort and cost of wet labs' projects by precisely narrowing the potential search space for MDA for researchers. Therefore, the establishment of a systematic MDA database and the provision of reliable MDA prediction tools are of great significance to scientists working in related fields.

In this paper, we present an online platform called MDADP that incorporates a new MDA database with comprehensive MDA prediction tools. In contrast to the recently proposed Disbiome database [31] and HMDAD database [9], our MDADP database uses structured criteria to organize disease terms. When data collection was performed, the complex aliases, extended descriptions (e.g. "new-onset untreated rheumatoid arthritis) and ambiguity between symptoms and disease, allowing ambiguity in disease nomenclature [12]. This is not only detrimental to database expansion, but also the integration of different

TABLE II
TOP 5 ALZHEIMER DISEASE-ASSOCIATED MICROBES PREDICTED BY MDADP_AR AND MDADP_CC

Ranks in MDADP_AR	Microbe	Evidence	Ranks in MDADP_CC	Microbe	Evidence
1	Prevotella	PMID: 33523001	1	Staphylococcus	unconfirmed
2	Lactobacillus	PMID: 34103438	2	Prevotella	PMID: 33523001
3	Roseburia	PMID: 32593306	3	Lactobacillus	PMID: 34103438
4	Clostridium	PMID: 29857583	4	Roseburia	PMID: 32593306
5	Faecalibacterium	unconfirmed	5	Dorea	PMID: 31781354

databases. We organized disease nomenclature using disease terms from the Mesh database and hence allow to retrieve standardized disease terms from different disease repositories in a consistent way. Moreover, microbes in the MDADP database were classified and mapped with NCBI taxonomy. This will facilitate the designers of predictive models to predict potential microbe-disease associations based on classification levels to improve the reliability of predictions.

Certainly, in the current version of MDADP, there are many aspects that need to be improved. For example, although these selected MDA computational models could achieve high AUCs on the HMDAD dataset, it did not perform well enough on the MDADP dataset. We hope that researchers can design more models with better predictive performance based on the MDADP database in the future. In addition, parameters of in these models were set to the optimal values given by their proposers based on the HMDAD database. It is obvious that the parameters could be optimized based on the MDADP database as well. Moreover, these models could introduce more diverse prior information about diseases and microbes, such as disease symptom similarity [32], disease semantic similarity [33], [34], and microbe functional similarity [35]. We observed that some researchers have proposed models for predicting MDAs based on graph convolutional neural networks (NinimHMDA [36]) recently, which have achieved impressive prediction performances. In the future, we will introduce more excellent computational models such as NinimHMDA to MDADP, which will be particularly beneficial for MDADP _ CC to infer potential MDAs with higher co-confidence. Furthermore, we will keep expanding both the diseases and their associated microbes for the MDADP database. With the expansion of MDAs entries and prediction methods, MDADP can rank and select potential microbe-disease pairs on a larger scale for validation experiments in downstream laboratories.

Since MDADP is a systematic online platform integrating a database and prediction tools for MDAs, we believe it will be a valuable source of information for scientists in microbiology and disease-related fields. Furthermore, MDADP promises to be a useful and effective platform in biomedical research, which may inspire bioscientists and computational scientists to form new research hypotheses about microbe-disease interactions, refine existing experimental methods, and validate their conclusions.

V. AUTHOR'S CONTRIBUTIONS

Conceptualization, L.W. and H.L.; Methodology, H.L., Q.Z. and L.W.; Validation, Y.W., Z.C. and T.P.; Formal Analysis, L.W. and H.L.; Investigation, Q.Z. and Y.W.; Resources, Z.C. and Q.Z.; Data Curation, Y.W. and T.P.; Writing-Original Draft Preparation, L.W. and H.L; Writing-Review and Editing, L.W. and Q.Z.; Supervision, L.W. and Q.Z.; Project Administration, L.W. and Q.Z.; Funding Acquisition, L.W. All authors read and approved the final manuscript.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for suggestions that helped improve the paper substantially.

REFERENCES

[1] F. Sommer and F. Bäckhed, "The gut microbiota masters of host development and physiology," *Nature Rev. Microbiol.*, vol. 11, no. 4, pp. 227–238, Feb. 2013.
[2] S. R. Gill *et al.*, "Metagenomic analysis of the human distal gut microbiome," *Science*, vol. 312, no. 5778, pp. 1355–1359, 2006.
[3] M. Ventura *et al.*, "Genome-scale analyses of health-promoting bacteria: Probiogenomics," *Nature Rev. Microbiol.*, vol. 7, no. 1, pp. 61–71, Nov. 2008.
[4] D. Bouskra *et al.*, "Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis," *Nature*, vol. 456, no. 7221, pp. 507–510, 2008.
[5] Y. Yuichiro, "Gut microbiota in health and disease," *Ann. Nutr. Metab.*, vol. 71, pp. 242–246, 2017.
[6] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: A synthetic analysis," *Lancet Glob. Health*, vol. 4, no. 9, pp. e609–e616, Sep. 2016.
[7] R. F. Schwabe and C. Jobin, "The microbiome and cancer," *Nature Rev. Cancer*, vol. 13, no. 11, pp. 800–812, Oct. 2013.
[8] N. Team, "A review of 10 years of human microbiome research activities at the US National Institutes of Health, fiscal years 2007-2016," *Microbiome*, vol. 7, no. 1, Feb. 2019, Art. no. 31.
[9] W. Ma *et al.*, "An analysis of human microbe-disease associations," *Brief. Bioinf.*, vol. 18, no. 1, pp. 85–97, 2017.
[10] C. Richard and J. Blay, "Thiazolidinedione drugs down-regulate CXCR4 expression on human colorectal cancer cells in a peroxisome proliferator activated receptor γ-dependent manner," *Int. J. Oncol.*, vol. 30, pp. 1215–1222, May 2007.
[11] S. Svacina, "Colorectal cancer and diabetes," *Vnitrni Lekarstvi*, vol. 57, no. 4, pp. 378–380, 2011.
[12] Z. Wen, C. Yan, G. Duan, S. Li, F.-X. Wu, and J. Wang, "A survey on predicting microbe-disease associations: Biological data and computational methods," *Brief. Bioinf.*, vol. 22, no. 3, Aug. 2020, Art. no. bbaa157.
[13] Y. Zhao, C.-C. Wang, and X. Chen, "Microbes and complex diseases: From experimental results to computational models," *Brief. Bioinf.*, vol. 22, no. 3, 2021, Art. no. bbaa158.
[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. 33, no. 5, pp. 733–739, 2017.
[15] M. Zhang, W. Ma, J. Zhang, Y. He, and J. Wang, "Analysis of gut microbiota profiles and microbe-disease associations in children with autism spectrum disorders in China," *Sci. Rep.*, vol. 8, no. 1, Sep. 2018, Art. no. 13981.
[16] Y. Long, M. Wu, Y. Liu, C. K. Kwoh, J. Luo, and X. Li, "Ensembling graph attention networks for human microbe-drug association prediction," *Bioinformatics*, vol. 36, no. Suppl_2, pp. i779–i786, 2020.
[17] H. Li *et al.*, "A novel human microbe-disease association prediction method based on the bidirectional weighted network," *Front. Microbiol.*, vol. 10, Apr. 2019, Art. no. 676.

[18] L. Wang, Y. Wang, H. Li, X. Feng, D. Yuan, and J. Yang, "A bidirectional label propagation based computational model for potential microbe-disease association prediction," *Front. Microbiol.*, vol. 10, Apr. 2019, Art. no. 684.
[19] X. Shen, H. Zhu, X. Jiang, X. Hu, and J. Yang, "A novel approach based on bi-random walk to predict microbe-disease associations," in *Proc. Int. Conf. Intell. Comput.*, 2018, pp. 746–752.
[20] Y. Fan, M. Chen, Q. Zhu, and W. Wang, "Inferring disease-associated microbes based on multi-data integration and network consistency projection," *Front. Bioeng. Biotechnol.*, vol. 8, 2020, Art. no. 831.
[21] F. Wang *et al.*, "LRLSHMDA: Laplacian regularized least squares for human microbedisease association prediction," *Sci. Rep.*, vol. 7, no. 1, Aug. 2017, Art. no. 7601.
[22] H. Li *et al.*, "Identifying microbe-disease association based on a novel back-propagation neural network model," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2502–2513, Nov./Dec. 2021.
[23] R. Hodson, "Alzheimer's disease," *Nature*, vol. 559, no. 7715, p. S1, 2018.
[24] J. Durack and S. V. Lynch, "The gut microbiome: Relationships with disease and opportunities for therapy," *J. Exp. Med.*, vol. 216, no. 1, pp. 20–40, 2019.
[25] M. Guo, J. Peng, X. Huang, L. Xiao, F. Huang, and Z. Zuo, "Gut microbiome features of Chinese patients newly diagnosed with Alzheimer's disease or mild cognitive impairment," *J. Alzheimer's Dis.*, vol. 80, pp. 299–310, 2021.
[26] T. Ivakhniuk and Y. Ivakhniuk, "Intestinal microbiota in Alzheimer's disease," *Georgian Med. News*, vol. 313, pp. 94–98, 2021.
[27] A. B. Shreiner, J. Y. Kao, and V. B. Young, "The gut microbiome in health and in disease," *Curr. Opin. Gastroenterol.*, vol. 31, no. 1, pp. 69–75, Jan. 2015.
[28] A. C. Wong and M. Levy, "New approaches to microbiome-based therapies," *MSystems*, vol. 4, no. 3, pp. e00122–19, 2019.
[29] R. Schlaberg, "Microbiome diagnostics," *Clin. Chem.*, vol. 66, no. 1, pp. 68–76, Dec. 2019.
[30] C. L. Richard and J. Blay, "Thiazolidinedione drugs down-regulate CXCR4 expression on human colorectal cancer cells in a peroxisome proliferator activated receptor γ-dependent manner," *Int. J. Oncol.*, vol. 30, no. 5, pp. 1215–1222, 2007.
[31] Y. Janssens *et al.*, "Disbiome database: Linking the microbiome to disease," *BMC Microbiol.*, vol. 18, no. 1, pp. 1–6, 2018.
[32] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms-disease network," *Nature Commun.*, vol. 5, no. 1, Jun. 2014, Art. no. 4212.
[33] L. M. Schriml *et al.*, "Human disease ontology 2018 update: Classification, content and workflow expansion," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D955–D962, 2019.
[34] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microrna functional similarity and functional network based on microrna-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.
[35] Y. Long and J. Luo, "WMGHMDA: A novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network," *BMC Bioinf.*, vol. 20, no. 1, Nov. 2019, Art. no. 541.
[36] Y. Ma and H. Jiang, "Ninimhmda: Neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe-disease association," *Bioinformatics*, vol. 36, no. 24, pp. 5665–5671, 2020.