Huang *et al. BMC Genomics* (2021) 22:916
https://doi.org/10.1186/s12864-022-08423-w

BMC Genomics

RESEARCH Open Access

GBDR: a Bayesian model for precise prediction of pathogenic microorganisms using 16S rRNA gene sequences

Yu-An Huang[1*†], Zhi-An Huang[2†], Jian-Qiang Li[3*], Zhu-Hong You[1], Lei Wang[4], Hai-Cheng Yi[5] and Chang-Qing Yu[1]

From Fifteenth International Conference on Intelligent Computing (ICIC 2019) Nanchang, China. 3-6 August 2019

Abstract

Background: Recent evidences have suggested that human microorganisms participate in important biological activities in the human body. The dysfunction of host-microbiota interactions could lead to complex human disorders. The knowledge on host-microbiota interactions can provide valuable insights into understanding the pathological mechanism of diseases. However, it is time-consuming and costly to identify the disorder-specific microbes from the biological "haystack" merely by routine wet-lab experiments. With the developments in next-generation sequencing and omics-based trials, it is imperative to develop computational prediction models for predicting microbe-disease associations on a large scale.

Results: Based on the known microbe-disease associations derived from the Human Microbe-Disease Association Database (HMDAD), the proposed model shows reliable performance with high values of the area under ROC curve (AUC) of 0.9456 and 0.8866 in leave-one-out cross validations and five-fold cross validations, respectively. In case studies of colorectal carcinoma, 80% out of the top-20 predicted microbes have been experimentally confirmed via published literatures.

Conclusion: Based on the assumption that functionally similar microbes tend to share the similar interaction patterns with human diseases, we here propose a group based computational model of Bayesian disease-oriented ranking to prioritize the most potential microbes associating with various human diseases. Based on the sequence information of genes, two computational approaches (BLAST+ and MEGA 7) are leveraged to measure the microbe-microbe similarity from different perspectives. The disease-disease similarity is calculated by capturing the hierarchy information from the Medical Subject Headings (MeSH) data. The experimental results illustrate the accuracy and

*Correspondence: yahuang1991@gmail.com; lijq@szu.edu.cn

†Yu-An Huang and Zhi-An Huang wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[1] Department of Information Engineering, Xijing University, Xi'an 710123, China

[3] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518000, China

Full list of author information is available at the end of the article

effectiveness of the proposed model. This work is expected to facilitate the characterization and identification of promising microbial biomarkers.

Keywords: Pathogenic microorganisms, Computational prediction model, 16S rRNA sequence analysis, Microbe-disease association network, Bayesian ranking

Background

Researchers are increasingly aware of the critical effects of the human microorganisms on our physical condition. Microorganisms, a.k.a. microbes, are referred to viruses, bacteria, archaea, and eukaryotes (protozoa and fungi) [1]. They can inhabit and thrive in almost each kind of natural environments, of course including the human body. Human microbial communities locate in different parts of the human body, including the external (e.g. skin) and the internal (e.g. the mucosal epithelia of vagina and intestine). In the adult gut, the large majority of intestinal microbes (10^{13}–10^{14}) inhabiting the gastrointestinal tract can approach the population quantity of human cells [2]. Recently, accumulating evidences [3–5] show that the onset of human disorders could be attributed to the dysfunction of human microbiota.

The symbiotic relationship between the human microbiota and its host has been demonstrated to get involved in multiple important biological activities. The human microbiota can be influenced by multiple factors of its host such as the genetics, lifestyle, body site, age, health status and others (e.g. antibiotics and smoking) [6–9]. The resident microbial flora can also affect human physical conditions via multiple microbial genome-encoded metabolic functions. Such metabolic functions can strengthen the metabolic capacity of its host. Therefore, human microbes play a key role in many important biological processes, e.g., by defending against pathogens, enhancing the immune system, getting access to nutrients, as well as degrading toxic compounds [6]. Identifying the latent relationships between microbes and human diseases can provide valuable insights into understanding the pathology of human diseases. For example, butyrate as the primary energy source of intestinal epithelial cells can also function as a key component to suppress the signal transduction pathways of expressing proinflammatory cytokines. Individuals with inflammatory bowel disease (IBD) have been found to have population declines of butyrate-producing microbes, such as *Clostridium leptum* and *Clostridium coccoides* groups [10]. This phenomena could also lead to decreasing butyrate utilization [11], which implies the fact that the restoration of host-microbe equilibrium could cure or prevent human complex diseases.

Thanks to the high volume of genomic data by high-throughput techniques, increasing bioinformatics tools and databases have been proposed for downstream analysis and data management [12–14]. For examples, a 16S rDNA analysis toolkit named W.A.T.E.R.S [15]. can be used for sequence alignment, operational taxonomic units (OTUs) determination, phylogenetic tree construction, and etc. Moreover, hundreds of microbe-disease associations are publicly available in Human Microbe-Disease Association Database (HMDAD) [16]. However, the known microbe-disease associations are just the tip of the iceberg and far from enough towards a complete picture for clinical medicine. The routine wet-lab experiments for biomarker discovery are easy to fail in clinical trials after considerable effort, money, and time have been already invested. In recent years, computational models have been proposed to prioritize seminal biomarker candidates using heterogeneous biological information in several fields including risk gene-disease association prediction [17–19], protein-protein interaction (PPI) prediction [20, 21], drug-target interaction prediction [22], and etc. [14]. The successful applications of these studies motivate us to devise an effective computational model for prioritizing potential pathogenic microbes.

16S ribosomal RNA (rRNA) gene encodes the 30S small subunit of the ribosomal RNA molecules of ribosomes. The low resolution of 16S rRNA gene enables the rapid and accurate identification to establish taxonomic relationships between microbes [23]. The major difference between 16S rRNA gene sequences (~ 1500 bp) tends to fall in the nine hypervariable regions (V1-V9), representing dramatic variations for alignments. Therefore, 16S rRNA sequencing analysis is widely used to capture natural species-specific “fingerprints” for phylogenetic comparisons.

Recently, increasing effective computational prediction models like KATZHMDA [24] and PBHMDA [25] were developed to explore the potential microbe-disease associations using the known microbe-disease association network as well as the calculated homologous similarity matrices. In light of a social network prediction algorithm called KATZ [26], KATZHMDA proposes a new proxy measure index to calculate the probabilities of unknown microbe-disease associations by considering the number of walks within the network and their own lengths. Moreover, PBHMDA is developed as a path-based prediction model to perform a restricted depth-first search by traversing all possible paths between microbes and diseases. However, some limitations could limit their usage and effectiveness, for example, by introducing the systematic bias of the predicted similarity matrices, merely focusing on the known diseases and microbe, and applying the global scoring schemes. Under the hypothesis that the functionally similar microbes tend to share the similar interaction patterns with pathologically similar human diseases, the ultimate goal of this work is to facilitate the discovery of validated biomarkers for helping the early diagnosis, risk assessment, tracking progression, and drug development. Here, we present a Group based computational model of Bayesian Disease-oriented Ranking (GBDR) for identification of potential microbe-disease associations based on the HMDAD database. Heterogeneous biological information is leveraged to compute similarity matrices, including disease semantic similarity, microbe similarity based on BLAST+ scores and microbe similarity based on MEGA7 evolutionary distance scores. The proposed model obtains the supreme prediction accuracy via leave-one-out cross validation (LOOCV) and *k*-fold cross validation (*k*-fold CV) in comparison with other state-of-the-art methods. Experiment results demonstrate that the group-based collaboration filtering and inferred similarity matrices can contribute to the improvement of prediction performance. We conduct a case study for an important disease to manually validate those predicted pathogenic microbes ranked in the top-20 list via published literatures. As a result, the reliable performance of the proposed model is fully demonstrated. It is anticipated that GBDR could be an effective computational tool to accelerate the identification of pathogenic microorganisms.

Results

Cross validation and case study

Under the frameworks of LOOCV and k-fold CV, the performance of GBDR is thoroughly evaluated. Since GBDR is devised as a disease-oriented ranking computational model, it aims to prioritize the most potential microbes for each disease. As such, we adopt a local scoring scheme for performance evaluation. As for LOOCV, each known microbe-disease association is used to test the model in turns while the rest are used for training until all counterparts are selected. Similarly, in the simulations of *k*-fold CV, the whole set of known microbe-disease associations are randomly split into *k* groups, where (k-1) groups form a training set while the remainder is a testing set until each group is tested in truns. To reduce the bias of random divisions, 50 times *k*-fold CVs are conducted to then achieve the average results.

The microbe-disease association prediction is actually a binary classification problem. The receiver operating characteristic (ROC) curve is extensively used to evaluate the performance of binary classification models. It is plotted by the true positive rate (sensitivity) versus false positive rate (1-specificity). In this work, sensitivity/specificity represents what a high probability of a predicted result can be told to make a positive/negative prediction correctly. The area under ROC curve (AUC) is a numerical evaluation coefficient between 0 and 1. For a disease *d*, the AUC value can be defined accordingly as:

$$\mathrm{AUC}_d = \frac{1}{\left|\mathcal{R}^{te}(d)\right|}\sum\nolimits_{(i,j)\in\mathcal{R}^{te}(d)}\delta\left(\hat{r}_{di} > \hat{r}_{dj}\right) \qquad (1)$$

where $\mathcal{R}^{te}(d) = \left\{\left(i,j\right)|\left(d,i\right) \in \mathcal{R}^{te}, \left(d,j\right) \notin \mathcal{R} \cup \mathcal{R}^{te}\right\}$. $\mathcal{R}^{te}(d)$ is a test dataset of d, $\hat{r}_{di}$ and $\hat{r}_{dj}$ are predicted values, and $\delta()$ is a binary indicator. If the equation within the brackets is true, $\delta()$ =1, otherwise 0. The final AUC value can be averaged as follows:

$$\mathrm{AUC} = \frac{\sum_{u\in D^{te}} AUC_d}{\left|D^{te}\right|} \qquad (2)$$

Here, D^{te} is a disease set on testing sets. Normally, AUC = 1 represents a perfect prediction and AUC = 0.5 represents a completely random one.

First of all, GBDR is compared with PBHMDA and KATZHMDA based on the known microbe-disease associations from HMDAD database via LOOCV (see Fig. 1). For a fair comparison, all compared models employ the same data resources, i.e. disease semantic similarity, microbe similarity based on BLAST+ scores and microbe similarity based on MEGA7 evolutionary distance scores. GBDR, PBHMDA and KATZHMDA achieve AUC values of 0.9456, 0.6087 and 0.6185, respectively. The proposed model performs better than the other two state-of-the-art models. PBHMDA and KATZHMDA have similar prediction performance in terms of local LOOCV. Since both of them are proposed to globally predict the most potential microbe-disease associations using the global scoring schemes, the class imbalance could lead to degrade their prediction performance to some extent.

Second, the proposed model is also compared with the other representative algorithms of recommender system via LOOCV (see Fig. 2), including singular value decomposition (SVD) based model, latent factor model (LFM), microbe-based collaborative filtering (CF), disease-based CF and neighbor-based CF models. Since GBDR is originally proposed as a recommendation algorithm, the fundamental assumption of pairwise association is adopted to resolve the limitations of the pointwise association assumption where the unknown (unlabeled) microbe-disease association are irrelevant. It is interesting to know the performance difference between the GBDR

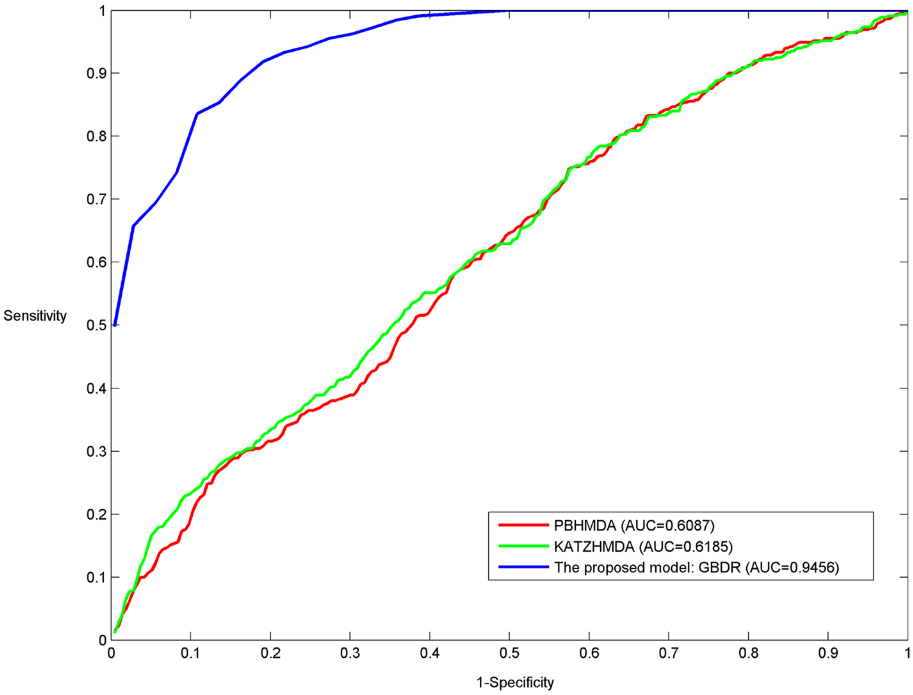


Fig. 1 The proposed model is compared with PBHMDA and KATZHMDA based on HMDAD database via LOOCV

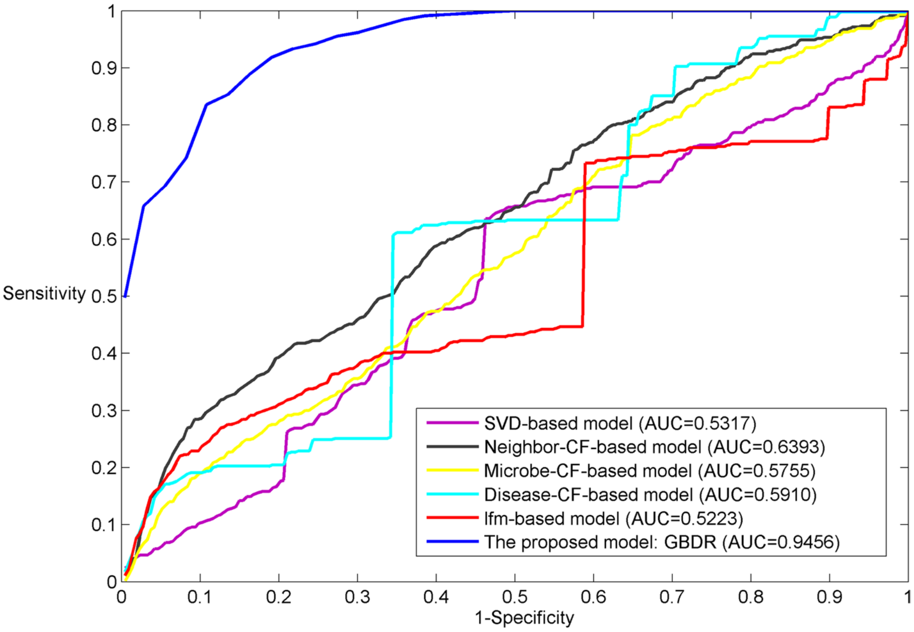


Fig. 2 The comparison result between GBDR and other recommendation algorithms via LOOCV

and other representative recommendation algorithms. Since the purpose of this work is to “recommend” the most possible microbes to a certain disease, it is intuitive and meaningful to use these recommendation algorithms for the performance comparison. As we can see in Fig. 2, GBDR also achieves the highest AUC of 0.9456. These representative recommendation algorithms tend to show a moderate predictive power in this case. Among these classical recommendation algorithms, the neighbor-based CF model obtains the best performance achieving the AUC of 0.6393. The result suggests that the disease-oriented ranking model with Bayesian filtering is capable of capturing latent relationships between microbes and diseases. The new and improved assumption in Bayesian disease-oriented ranking is more effective for prediction by introducing richer interactions among microbes. Particularly, the unified effect of group preference and individual preference is linearly combined to naturally maximize the overall likelihood.

Finally, we also implement *k*-fold CV for further evaluation (see Table 1). As a result, the proposed model yields average AUCs of 0.8266 ± 0.0805, 0.8866 ± 0.0270 and 0.8926 ± 0.0167 in 2-fold CV, 5-fold CV and 10-fold CV, respectively. Both LOOCV and *k*-fold CV can demonstrate the effectiveness of GBDR. Furthermore, colorectal carcinoma (CRC) is selected as an important human disease for a case study. As a result, 9 out of top-10 and 16 out of top-20 predicted microbes have been experimentally confirmed to have associations with the development of CRC. Detailed information is provided in Additional file 1.

Table 1 The proposed model is evaluated in 2-fold, 5-fold and 10-fold CV, respectively

	2-fold CV	**5-fold CV**	**10-fold CV**
AUC	0.8266+/−0.0805	0.8866+/−0.0270	0.8926+/−0.0167

Effectiveness evaluation of group-based collaboration filtering

In this section, we conduct LOOCV to evaluate the prediction performance with or without the group-based preference strategy (as shown in Fig. 3A). Without the group preference, the proposed model suffers a nearly 16.2% decline in prediction accuracy with an AUC value of 0.7925. It shows that the group preference strategy is efficient to aggregate the group preference for the disease-oriented ranking through injecting richer interactions among microbes. The linear combination of pairwise preference and group preference is more effective than the simple pairwise preference.

Moreover, we further evaluate the prediction performance of the proposed model with the disease-based or microbe-based group preference respectively via LOOCV (see Fig. 3B). Firstly, without integrated similarity of microbe or disease used in Eq. (18), the proposed model only leverages the known microbe-disease associations for prediction. In this scenarios, combined with microbe-based group preference, the proposed model obtains an improved performance with AUC of 0.8793. And the counterpart with disease-based group preference yields an AUC value of 0.5130. This result supports our assumption that the coordinated functions of microbial groups may pathologically influence the susceptibility to human diseases whereas the human diseases fail to have a significant group trend to affect microbial communities. Secondly, based on the known microbe-disease associations, the proposed model is carried out with both microbe and disease similarity matrices using the microbe-based group preference. In this way, the proposed model achieves the best prediction performance with AUC of 0.9456. On the other hands, the proposed model with the disease-based

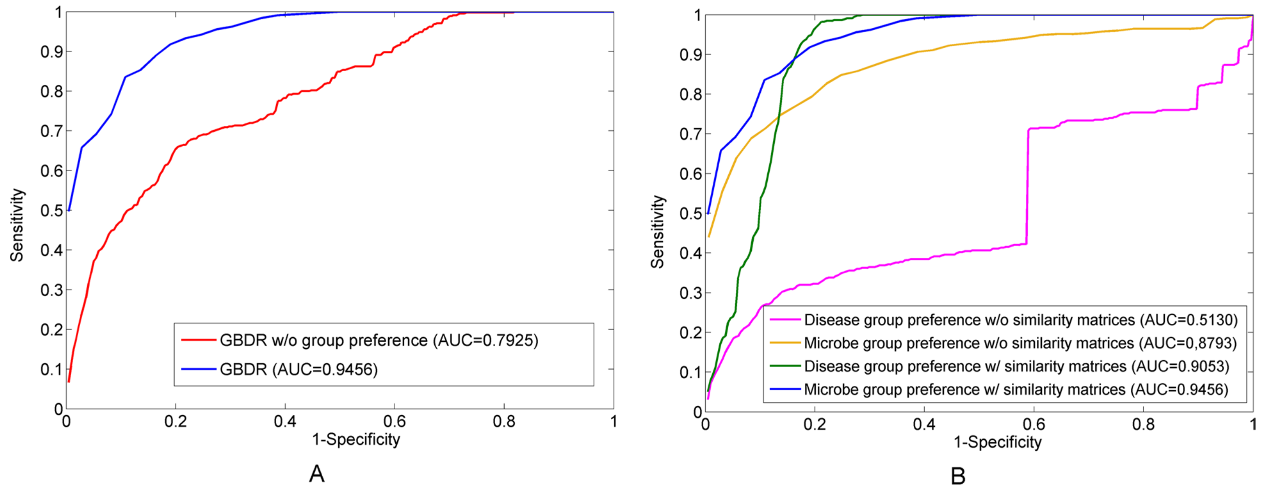


Fig. 3 The prediction performance with and without the group preference is evaluated by LOOCV in part **A**. And the proposed model based on disease group preference or microbe group preference is evaluated by LOOCV respectively in part **B**

group preference shows a significant increase (39.23%) in the prediction accuracy achieving the AUC of 0.9053. This result suggests that the inferred similarity matrices provide useful heterogeneous information to effectively discriminate seminal biomarker candidates.

Effectiveness evaluation of combining different types of similarities

As mentioned at the end of the above section, the inferred similarity matrices can effectively improve the prediction performance of our model based on the disease-based group preference. It motivates us to conduct the performance effect analysis of different types of similarities proposed in our model via 5-fold CV. The results are shown in Table 2. As a baseline, the GBDR without using any similarity matrices shows the average AUC of 0.6189 with standard deviation of 0.0561. Using the disease semantic similarity, the prediction accuracy achieves average AUC value of 0.6796+/−0.0468 with 6.27% improvement. Moreover, when the GBDR is integrated with both microbe similarity based on BLAST+ scores and microbe similarity based on MEGA7 evolutionary distance scores, the prediction accuracy is improved by 10.02% achieving the average AUC value of 0.7171+/−0.0427. Finally, the GBDR obtains the best average AUC value of 0.8081+/−0.0284 using the disease semantic similarity and the integrated microbe similarity. The result demonstrates that the sequence information of gene exploited by computational approaches (BLAST+ and MEGA 7) enables the precise measure of the biological homology between microbes. Furthermore, the disease semantic similarity based on the Medical Subject Headings (MeSH) descriptors can reflects the molecular relatedness between hereditary diseases. Although the known microbe-disease association network is sparse as its links are limited in number, applying diverse similarities to the proposed model is useful to provide discriminative biological information, and therefore enabling the precise prediction of pathogenic microbes.

Effect analysis of key parameters in GBDR

There are several key parameters in GBDR, e.g., regularization weights $\boldsymbol{\alpha}_u$, $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$, learning rate $\boldsymbol{\gamma}$, number of later features z, and group sizes $|\mathcal{G}|$. Based on 5-fold CV, we conduct an effect analysis to explore the potential of parameter tuning. As we can see in Fig. 4, GBDR reaches the peak of AUC values when regularization weights $\boldsymbol{\alpha}_u$, $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ are to 0.01 and learning rate $\boldsymbol{\gamma}$ is set to 0.001. By evaluating all possible combinations via grid search, the achieved AUC values vary from 0.8536 to 0.8866. For better demonstration, we combine the results of z and $|\mathcal{G}|$ as a whole in Fig. 5. Regarding of the different number of latent features z, no significant change is observed in terms of AUC values. Based on the results, we set 30 as the default value of z. Moreover, GBDR achieves the highest AUC value when $|\mathcal{G}|$ is set to 5. As a baseline, GBDR without group preference setting (i.e., $|\mathcal{G}|=1$) suffers from degradation by 6.5% as expected. In short, the performance of GBDR is not sensitive to the key parameters tuning.

Discussion

Accumulating evidence show that different types of microbiota are associated with the mechanism of human diseases, forming a complex causal network. Although there are a number of methods having been proposed for predicting such important associations, 16s rRNA gene sequences, the information most easily obtained in microbe research, haven't been utilized for this task. To bridge this gay, we proposed a Bayesian prediction model called GBDR, using various types of information including 16s rRNA sequences. GBDR is based on the computation of disease/microbe similarity assuming that similarity microbes tend to be involved in similar disease mechanism. The experimental results show that the prediction based on such an assumption is feasible and effective. We anticipate that GBDR can help the researchers find the relevant diseases for a specific type of microbe given its 16s rRNA sequence.

Conclusion

The human microbiota has attracted the increasing attention thanks to its key role playing in human biological activities. It has even been deemed as the "forgotten organ" in the human body. Recent researches show the dysfunction of host-microbiota homeostatic balance can

Table 2 When combined with different similarity matrices, the proposed model is evaluated via 5-fold CV based on the disease-based group preference

Combined similarity matrices	5-fold CV
No integrated similarity of microbe or disease via Eq. (18)	0.6169+/− 0.0561
Disease semantic similarity	0.6796+/−0.0468
Microbe sequence similarity and microbe evolutionary distance-based similarity	0.7171+/−0.0427
Disease semantic similarity, microbe sequence similarity and microbe evolutionary distance-based similarity	0.8081+/−0.0284

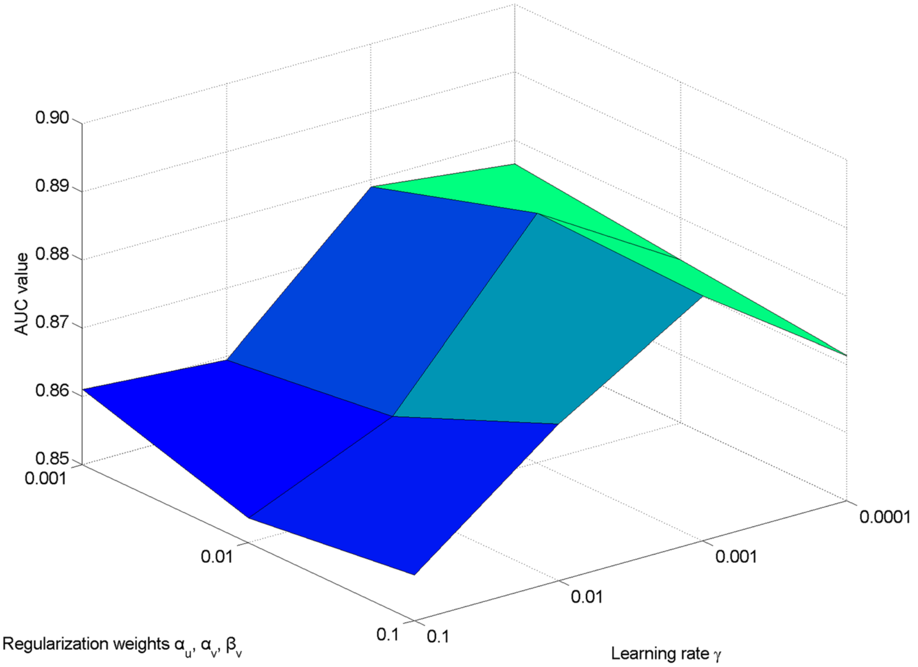


Fig. 4 The parameter analysis of regularization weight α_u, α_v and β_v versus learning rate γ

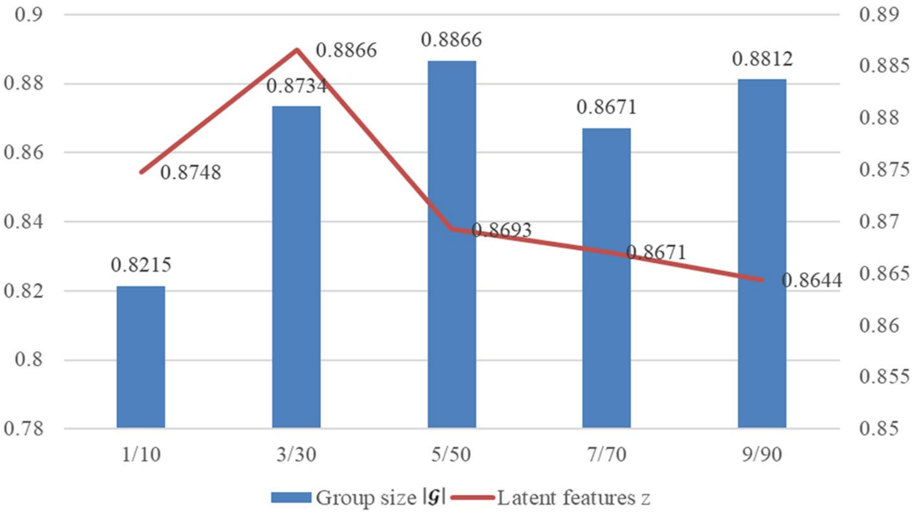


Fig. 5 The effect influence of group size |𝒢| and number of latent features z

result in the onset of various human diseases. The great advance of technology, especially PCR amplification and next-generation sequencing, allows to generate high volumes of sequences, providing a new window for the follow-up downstream analysis. In this work, we leveraged 16S rRNA gene to infer microbe similarity based on BLAST+ scores and microbe similarity based on MEGA7 evolutionary distance scores. The framework of GBDR is proposed to identify the most seminal disease-specific microorganisms on a large scale. Based on the results of simulation experiments, GBDR is demonstrated to achieve higher performance than the two state-of-the-art models and representative recommendation algorithms via LOOCV. Furthermore, reliable predictive capability of GBDR is validated by *k*-fold CV and a case study. GBDR is expected to provide valuable insights into advancing the identification of potential microbes as ideal biomarkers for evaluating and measuring human complex diseases. The prediction list of the most seminal pathogenic microbes is released in Additional file 2 sorted by various specific diseases. Based on the assumption that functionally similar microbes tend to share the similar interaction patterns with human diseases, the main goal of this work is to prioritize the most potential disease-related microbes. That is, the involved microbes could be affected by the involved human diseases and/or the human diseases could be caused by the involved microbes. Identifying the seminal disease-microbe association is the first key step to develop the full potential for further in vitro tests in therapeutics and clinical research.

Several factors can be summarized to improve the effectiveness of GBDR. Firstly, the integrated microbe similarity holds a significant potential to characterize the remarkable feature patterns between microbes. The hierarchical relevance is leveraged to measure disease semantic similarity. Secondly, the group-based pairwise strategy is capable of extracting the fruitful information based on the group preference of microbes associating with a certain disease. Thirdly, the Bayesian approach with disease-oriented ranking is quite suitable for the local prediction of microbe-disease associations. However, GBDR adopts a local scoring scheme for disease-oriented ranking without global normalization process. GBDR is inapplicable to globally predicting the most potential microbe-disease associations like PBHMDA and KATZHMDA.

Methods

Materials

In this work, three types of biological information are utilized, i.e. the known microbe-disease associations derived from HMDAD database (http://www.cuilab.cn/hmdad) [16], 16S rRNA partial or complete gene sequences downloaded from the Nucleotide Database of the National Centre for Biotechnology Information (NCBI) [27] and MeSH descriptors provided by the Nation Library of Medicine (NLM) [28]. It is noted that, Ma et al. [16] searched articles regarding human microbiome-related research published before July 2014. The HMDAD database provides 450 non-repetitive known microbe-disease associations including 292 microbes and 39 human diseases (see Additional file 3). The numbers of microbes and diseases are denoted as *nm* and *nd*, respectively. We note that all the symbols that used in the Methods section are summarised in the Table 3. All known microbe-disease associations are converted into an adjacency binary matrix as variable $\mathcal{R}$ of size nm × nd representing their association relationships. Namely, $\mathcal{R}(\boldsymbol{m_i}, \boldsymbol{d_j}) = \mathbf{1}$ indicates microbe *mi* is known to be associated with disease *dj*, otherwise $\mathcal{R}(\boldsymbol{m_i}, \boldsymbol{d_j}) = \mathbf{0}$.

Table 3 Summary of the symbols used

# of microbes: *nm*	**# of diseases: *nd***
known microbe-disease associations: $\mathcal{R}$	# of latent features: *z*
Disease similarity: S_d	Microbe similarity: S_m
Predicted probability: $\hat{\boldsymbol{r}}$	Bernoulli distribution: $\boldsymbol{\delta}$ ()
Group preference: $\mathcal{G}$	Objective function: $\mathcal{F}$
Regularization weights: $\boldsymbol{\alpha_u}$, $\boldsymbol{\alpha_v}$ and $\boldsymbol{\beta_v}$	Microbe latent feature vector: $\boldsymbol{U}$
Disease latent feature vector: V	Bias value: $\boldsymbol{b}$
Model parameters: $\boldsymbol{\Theta}$	Learning rate: $\boldsymbol{\gamma}$

Disease semantic similarity

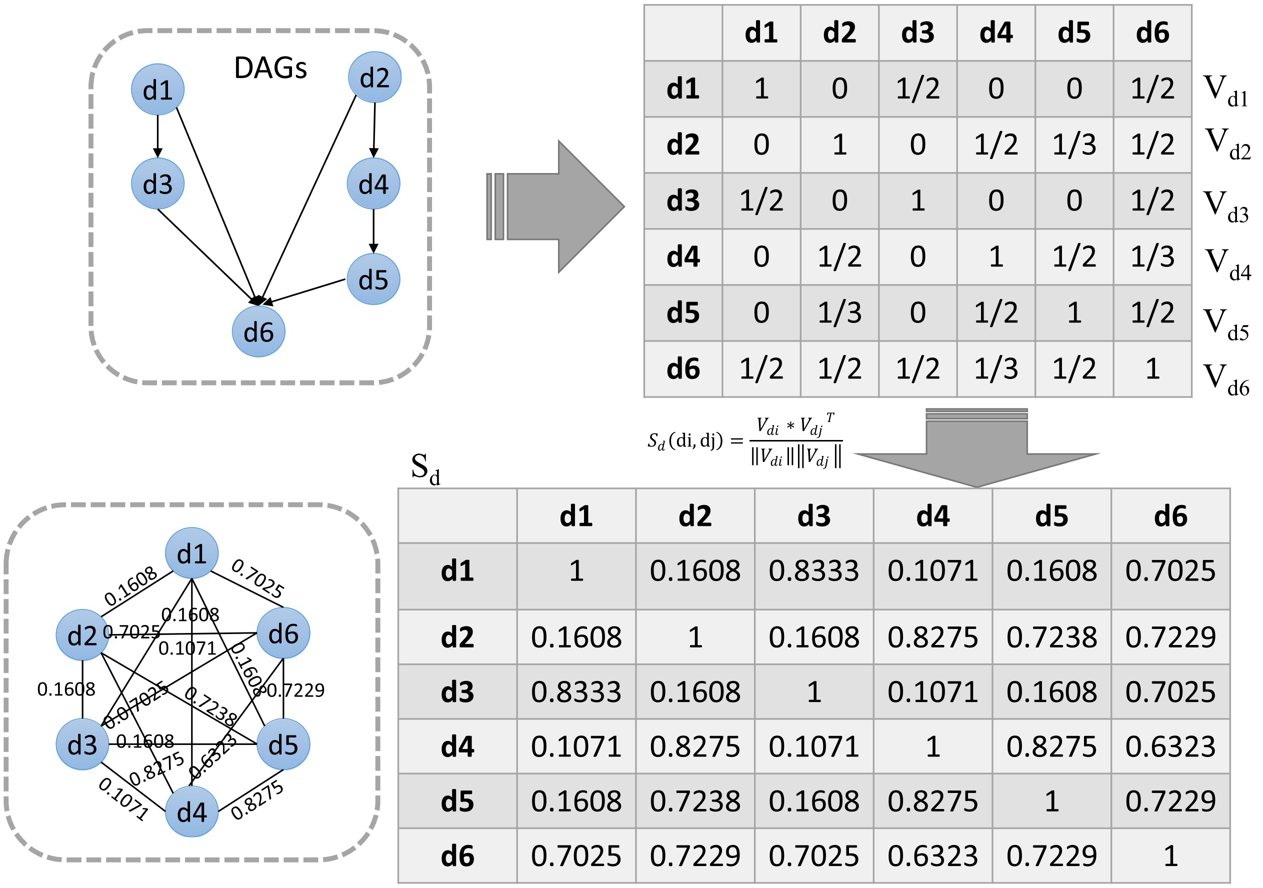
The hierarchy system of MeSH descriptors is informative to offer semantic-based taxonomic categorization for various human diseases. For example, the MeSH ID of overnutrition (C18.654.726) shares the same prefix with its subtype obesity's (C18.654.726.500). Accordingly, the relationships between any disease and others can be established by respective Directed Acyclic Graphs (DAGs) using the hierarchy of MeSH IDs [29]. Each disease has at least one MeSH ID which numerically represents its location in DAGs. Figure 6 illustrates the calculation process of disease semantic similarity. Empirically, the shorter path between the ancestor node *d* and the target node *t*, the higher weight value should be given. It can be formulated as follows:

$$V_d(t) = \begin{cases} 1, & if\ t = d \\ \frac{1}{len(d,t)+1}, & if\ t \in \{the\ descendant\ node\ of\ d\} \\ 0, & otherwise \end{cases} \tag{3}$$

where $len(d,t)$ is the shortest path length between the ancestor node d and one of its descendant node t. For example in Fig. 6, d_1 is the ancestor node of d_3 and d_6. $V_{d1}(d_3) = 1/(len(d_1,d_3)+1) = 1/2$ where $len(d_1,d_3) = 1$. Likewise, $V_{d1}(d_6) = 1/(len(d_1,d_6)+1) = 1/2$ where $len(d_1,d_6) = 1$. In this way, a feature vector for d_1 can be computed as $V_{d1} = (1,0,1/2,0,0,1/2)$. Then, we further calculate the semantic similarity of any two diseases d_i and d_j by cosine similarity measure:

$$S_d(\mathrm{di},\mathrm{dj}) = \frac{V_{di} * {V_{dj}}^T}{\|V_{di}\| \|V_{dj}\|} \tag{4}$$

where V_{di} and V_{dj} are the feature vectors of d_i and d_j, respectively.



	d1	d2	d3	d4	d5	d6	
d1	1	0	1/2	0	0	1/2	V_{d1}
d2	0	1	0	1/2	1/3	1/2	V_{d2}
d3	1/2	0	1	0	0	1/2	V_{d3}
d4	0	1/2	0	1	1/2	1/3	V_{d4}
d5	0	1/3	0	1/2	1	1/2	V_{d5}
d6	1/2	1/2	1/2	1/3	1/2	1	V_{d6}

$$S_d(\mathrm{di},\mathrm{dj}) = \frac{V_{di} * V_{dj}^{\ T}}{\|V_{di}\|\|V_{dj}\|}$$

S_d

	d1	d2	d3	d4	d5	d6
d1	1	0.1608	0.8333	0.1071	0.1608	0.7025
d2	0.1608	1	0.1608	0.8275	0.7238	0.7229
d3	0.8333	0.1608	1	0.1071	0.1608	0.7025
d4	0.1071	0.8275	0.1071	1	0.8275	0.6323
d5	0.1608	0.7238	0.1608	0.8275	1	0.7229
d6	0.7025	0.7229	0.7025	0.6323	0.7229	1

Fig. 6 The calculation process of disease semantic similarity S_d

Microbe similarity based on BLAST+ scores and MEGA7 evolutionary distance scores

Sequence similarity and evolutionary distance-based similarity are two effective measurements to examine the relatedness among microbes from different perspectives. The former reflects the degree of likeness between any two sequences while the latter refers to the divergence of their common ancestral sequence. Although the calculation of both are based on the same information source, i.e., 16S rRNA gene sequences, they do not have to be similar as a necessary condition. Basic Local Alignment Search Tool (BLAST) is a specific sequence similarity search program (http://www.ncbi.nlm.nih.gov/blast) [30]. We use its variant BLAST+ [31] to compare a targeted 16S rRNA gene sequence of each target microbe against the sequences of other microbes as a nucleotide sequence database in turns. *Identity* is an important glossary of BLAST+ to measure the extent to which two (nucleotide or amino acid) sequences are invariant in an alignment. In this work, *identity* is used for measuring microbe sequence similarity. In this way, we define a matrix as *Iden* of size $nm \times nm$ to store the *identity* values yielded by the alignment. Then the microbe sequence similarity denoted as *MSS* is calculated by normalizing *Iden* matrix as follows:

$$\mathrm{MSS}(\mathrm{mi},\mathrm{mj}) = \frac{Iden(mi, mj) - \mathrm{Min.}(Iden)}{\mathrm{Max.}(Iden) - \mathrm{Min.}(Iden)} \tag{5}$$

It is noted that, among 292 investigated microbes, five microbe have no available 16S rRNA gene sequences in NCBI (denoted as “unavailable” to “FASTA filename” in Additional file 3). We simply set their sequence similarities as the mean of the rest available.

The calculation of microbe evolutionary distance-based similarity is mainly based on the molecular evolutionary genetics analysis of MEGA 7 [32] (http://www.megasoftware.net/). The evolutionary distance between any two sequences is measured by the number of nucleotide substitutions involved. First, Clustal W is used to perform multiple sequences alignment [33]. To reduce the disturbance caused by the gaps, all sequences are trimmed down to the shortest size by removing terminal redundancy at 5′ and 3′ terminus. Then the option of complete deletion option is set to address the issues of gaps and missing data. We utilized *p*-distance model to measure the evolutionary distances based on substitutions (including transitions and transversions). *p*-distance [34] for nucleotide sequences is written as:

$$\hat{p} = \frac{n_d}{n} \tag{6}$$

where n_d refers to the number of different nucleotides between two tested sequences and *n* is the total number of nucleotides examined. The higher value of evolutionary distances denotes the higher evolutionary diversity. The evolutionary distances are subtracted from 1 and the result is denoted by a matrix as *ED* of size *nm*×*nm*. Similarly, the microbe evolutionary distance-based similarity (*MES*) is also normalized as follows:

$$\mathrm{MES}(\mathrm{mi},\mathrm{mj}) = \frac{ED(mi,mj) - \mathrm{Min.}(ED)}{\mathrm{Max.}(ED) - \mathrm{Min.}(ED)}. \tag{7}$$

For those microbes without available 16S rRNA gene sequences, their evolutionary distance-based similarities are also set to the overall mean level. If the unavailable microbe list increases, performance degeneration is inevitable to happen. We can address this problem based on the know microbe-disease associations by exploiting the implicit information from the topological network structure. According to the previous works [22, 35], Gaussian interaction profile kernel similarity and local similarity-based methods (e.g., the Jaccard index and Salton index) can be applied to calculate the biological function-based similarity of those microbes without available 16 s rRNA gene sequences. Finally, we empirically merge *MSS* and *MES* to represent the final microbe similarity S_m:

$$S_m(m_i,m_j) = \frac{MSS(m_i,m_j) + MES(m_i,m_j)}{2} \tag{8}$$

Group preference based Bayesian disease-oriented ranking

Based on the previous work [36, 37] in recommender system, the pointwise association assumption, i.e., considering all known microbe-disease associations as "interactions" and unknown ones as "no interactions", could mislead the learning process. However, the pairwise association assumption over two microbes could relax the pointwise preference assumption by treating that a disease *d* is more probably related to a microbe *i* than a microbe *j* represented as $\hat{\boldsymbol{r}}_{\boldsymbol{di}} > \hat{\boldsymbol{r}}_{\boldsymbol{dj}}$ where *i* belongs to the known association with *d* whereas *j* does not. Empirically, this assumption generates better prediction results than the pointwise assumption. Inspired by this idea [38, 39], we present group pairwise preference based Bayesian disease-oriented ranking for prioritizing the most potential pathogenic microbes (see Fig. 7).

Based on the known microbe-disease associations for a typical disease *d*, we first define the overall likelihood of pairwise preferences (LPP) among the whole set of microbe (denoted as $\mathcal{M}$):

$$\begin{aligned}\mathbf{LPP}(\boldsymbol{d}) &= \prod\nolimits_{i,j\in\mathcal{M}} \Pr\left(\hat{\boldsymbol{r}}_{\boldsymbol{di}} > \hat{\boldsymbol{r}}_{\boldsymbol{dj}}\right)^{\delta((\boldsymbol{d},\boldsymbol{i})\succ(\boldsymbol{d},\boldsymbol{j}))} \times \left[1-\Pr\left(\hat{\boldsymbol{r}}_{\boldsymbol{di}} > \hat{\boldsymbol{r}}_{\boldsymbol{dj}}\right)\right]^{[1-\delta((\boldsymbol{d},\boldsymbol{i})\succ(\boldsymbol{d},\boldsymbol{j}))]}\\ &= \prod\nolimits_{(\mathbf{d},\mathbf{i})\succ(\mathbf{d},\mathbf{j})} \Pr\left(\hat{\boldsymbol{r}}_{\boldsymbol{di}} > \hat{\boldsymbol{r}}_{\boldsymbol{dj}}\right)\left[1-\Pr\left(\hat{\boldsymbol{r}}_{\boldsymbol{di}} > \hat{\boldsymbol{r}}_{\boldsymbol{dj}}\right)\right]\end{aligned} \tag{9}$$

where $(\boldsymbol{d},\boldsymbol{i})\succ(\boldsymbol{d},\boldsymbol{j})$ means that disease *d* is more potentially associated with microbe *i* than microbe *j*. And $\boldsymbol{\delta}((\boldsymbol{d},\boldsymbol{i})\succ(\boldsymbol{d},\boldsymbol{j}))$ is Bernoulli distribution over the binary random variable. To better approximate the disease-oriented pairwise preference over two microbes, the Bayesian disease-oriented ranking method is adopted to simplify the term **LPP(*d*)** as follows [38]:

$$\mathbf{BDR}(\boldsymbol{d}) = \prod\nolimits_{i\in\mathcal{M}_d^{tr}}\prod\nolimits_{j\in\mathcal{M}^{tr}\backslash\mathcal{M}_d^{tr}} \Pr\left(\hat{\boldsymbol{r}}_{\boldsymbol{di}} > \hat{\boldsymbol{r}}_{\boldsymbol{dj}}\right)\left[1-\Pr\left(\hat{\boldsymbol{r}}_{\boldsymbol{di}} > \hat{\boldsymbol{r}}_{\boldsymbol{dj}}\right)\right] \tag{10}$$

here $\boldsymbol{i} \in \mathcal{M}_{\boldsymbol{d}}^{\boldsymbol{tr}}$ indicates the known microbe-disease association pair (*i*, *d*) in training data and $\boldsymbol{j} \in \mathcal{M}^{\boldsymbol{tr}}\backslash\mathcal{M}_{\boldsymbol{d}}^{\boldsymbol{tr}}$ represents the microbe-disease association pair (*j*, *d*) is unknown. We assume that the group preference is an overall preference score of a microbe group on a disease. If microbe-disease pair (*i*, *d*) is a known association but (*i*, *b*) is not, the group preference can be represented as:

$$(\mathcal{G},\boldsymbol{d}) \succ (\mathcal{G},\boldsymbol{b}),\textbf{where } \boldsymbol{i} \in \mathcal{G} \textbf{\textit{ and }} \mathcal{G} \subseteq \mathcal{M}_{\boldsymbol{d}}^{\boldsymbol{tr}} \tag{11}$$

It can assume that the group preference of $\mathcal{G} \subseteq \mathcal{M}_{\boldsymbol{d}}^{\boldsymbol{tr}}$ on a disease *d* is probably stronger than the individual preference of microbe *i* on disease *b*. To learn the unified effect of both individual preference and group preference, we linearly combine them as follows:

$$(\mathcal{G},\boldsymbol{d}) + (\boldsymbol{i},\boldsymbol{d}) \succ (\boldsymbol{i},\boldsymbol{b}) \text{ or } \hat{\boldsymbol{r}}_{\mathcal{G}\boldsymbol{id}} > \hat{\boldsymbol{r}}_{\boldsymbol{ib}} \tag{12}$$

where $\hat{\boldsymbol{r}}_{\mathcal{G}\boldsymbol{id}} = \boldsymbol{\rho}\hat{\boldsymbol{r}}_{\mathcal{G}\boldsymbol{d}} + (\mathbf{1}-\boldsymbol{\rho})\hat{\boldsymbol{r}}_{\boldsymbol{id}}$ is the combined preference of individual preference $\hat{\boldsymbol{r}}_{\boldsymbol{id}}$ and group preference $\hat{\boldsymbol{r}}_{\mathcal{G}\boldsymbol{d}}$ and parameter $\boldsymbol{\rho}$ controls the weight of two different preferences from 0 to 1. We equally set $\boldsymbol{\rho}$ to 0.5 in this study. In this way, we can define group Bayesian disease-oriented ranking (GBDR) analogously to how we define in Eq. (10) as follows:

$$\mathbf{GBDR}(\boldsymbol{i}) = \prod\nolimits_{d\in\mathcal{D}_i^{tr}}\prod\nolimits_{b\in\mathcal{D}^{tr}\backslash\mathcal{D}_i^{tr}} \Pr\left(\hat{\boldsymbol{r}}_{\mathcal{G}\boldsymbol{id}} > \hat{\boldsymbol{r}}_{\boldsymbol{ib}}\right)\left[1-\Pr\left(\hat{\boldsymbol{r}}_{\boldsymbol{ib}} > \hat{\boldsymbol{r}}_{\mathcal{G}\boldsymbol{id}}\right)\right] \tag{13}$$

where $\mathcal{G} \subseteq \mathcal{M}_{\boldsymbol{d}}^{\boldsymbol{tr}}$. $\boldsymbol{d} \in \mathcal{D}_{\boldsymbol{i}}^{\boldsymbol{tr}}$ means disease *d* has an interaction with microbe *i* in training data. Likewise, $\boldsymbol{b} \in \mathcal{D}^{\boldsymbol{tr}}\backslash\mathcal{D}_{\boldsymbol{i}}^{\boldsymbol{tr}}$ means disease *b* has an unknown interaction with microbe *i*. For given two microbes *i* and *j*, the joint likelihood is simply approximated via multiplication like

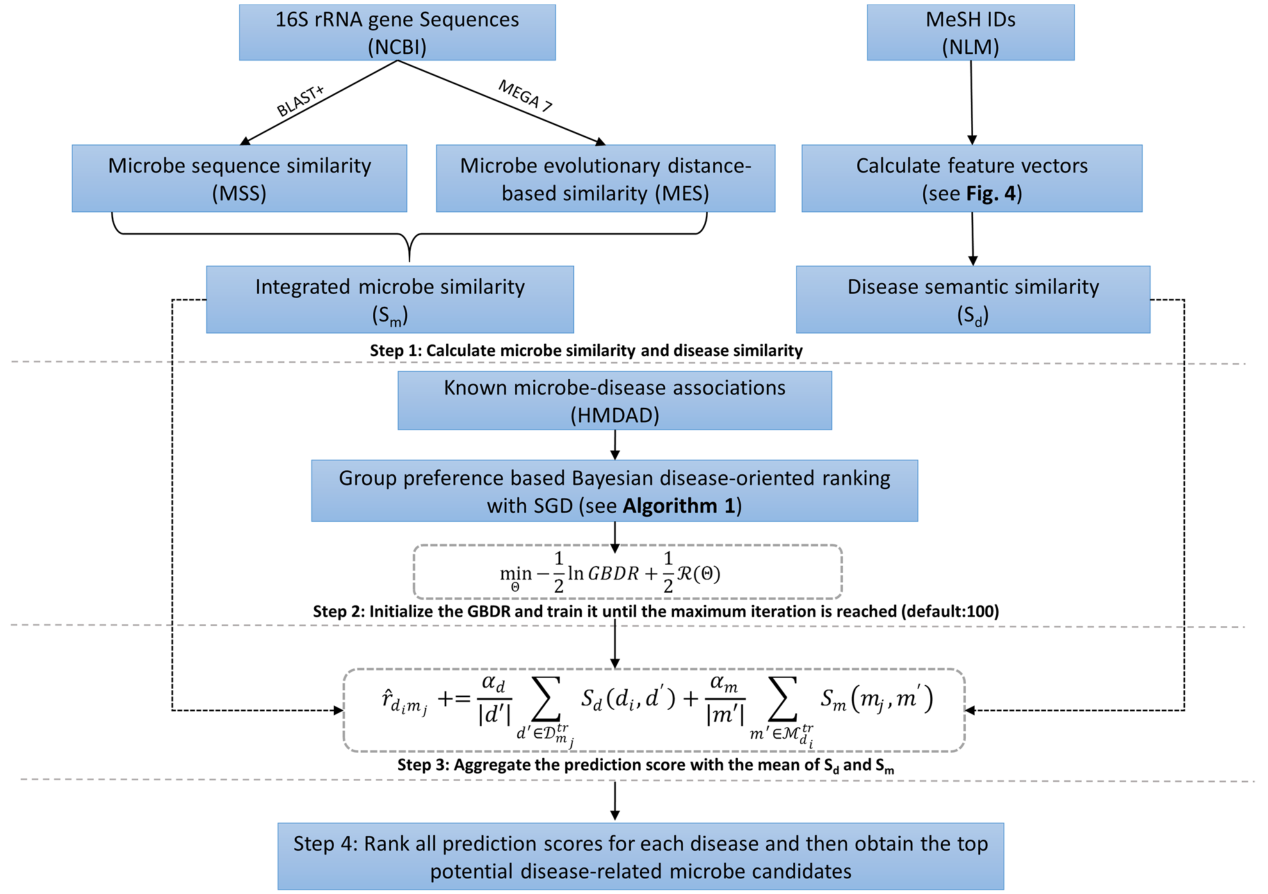

16S rRNA gene Sequences (NCBI)
MeSH IDs (NLM)
BLAST+
MEGA 7
Microbe sequence similarity (MSS)
Microbe evolutionary distance-based similarity (MES)
Calculate feature vectors (see **Fig. 4**)
Integrated microbe similarity (S_m)
Disease semantic similarity (S_d)
Step 1: Calculate microbe similarity and disease similarity
Known microbe-disease associations (HMDAD)
Group preference based Bayesian disease-oriented ranking with SGD (see **Algorithm 1**)
$\min_{\Theta} -\frac{1}{2}\ln GBDR + \frac{1}{2}\mathcal{R}(\Theta)$
Step 2: Initialize the GBDR and train it until the maximum iteration is reached (default:100)
$\hat{r}_{d_i m_j} += \frac{\alpha_d}{|d'|}\sum_{d'\in\mathcal{D}^{tr}_{m_j}} S_d(d_i,d') + \frac{\alpha_m}{|m'|}\sum_{m'\in\mathcal{M}^{tr}_{d_i}} S_m(m_j,m')$
Step 3: Aggregate the prediction score with the mean of S_d and S_m
Step 4: Rank all prediction scores for each disease and then obtain the top potential disease-related microbe candidates

Fig. 7 The flowchart of GBDR

$\mathbf{GBDR(i,j)}\approx\mathbf{GBDR(i)}\times\mathbf{GBDR(j)}$. Therefore, the overall likelihood for all microbes and all diseases can be formulated as:

$$\mathbf{GBDR} = \prod_{i\in\mathcal{M}^{tr}}\prod_{d\in\mathcal{D}^{tr}_i}\prod_{b\in\mathcal{D}^{tr}\setminus\mathcal{D}^{tr}_i} \Pr\left(\hat{\boldsymbol{r}}_{\mathcal{G}id} > \hat{\boldsymbol{r}}_{ib}\right)\left[1-\Pr\left(\hat{\boldsymbol{r}}_{ib} > \hat{\boldsymbol{r}}_{\mathcal{G}id}\right)\right] \tag{14}$$

where $\mathcal{G}\subseteq\mathcal{M}^{\boldsymbol{tr}}_{\boldsymbol{d}}$. Given $\boldsymbol{\Theta} = \{\boldsymbol{U}_i\in\mathbb{R}^{1\times nd}, \boldsymbol{V}_d\in\mathbb{R}^{1\times nd}, \boldsymbol{b}_d\in\mathbb{R}, \boldsymbol{i}\in\mathcal{M}^{tr}, \boldsymbol{d}\in\mathcal{D}^{tr}\}$ is a set of model parameters to be learned, one common way to estimate the model parameters is to minimize the log-likelihood function of GBDR as follows,

$$\min_{\boldsymbol{\Theta}} -\frac{\mathbf{1}}{\mathbf{2}}\ln\boldsymbol{GBDR} + \frac{\mathbf{1}}{\mathbf{2}}\mathcal{R}(\boldsymbol{\Theta}). \tag{15}$$

We use stochastic gradient descent (SGD) algorithm to optimize the object function in Eq.(15). Before using the algorithm of SGD, a subset of microbes is randomly sampled to form a microbe group $\mathcal{G}$. In this way, for each random sampling, it includes a microbe i, a disease d, a disease b and a microbe group $\mathcal{G}$ where $\mathbf{i}\in\mathcal{G}$. The objective function in Eq. (15) can be written as:

$$\begin{aligned}\mathcal{F}(\mathcal{G},\boldsymbol{i},\boldsymbol{d},\boldsymbol{b}) &= -\ln\left(\hat{\boldsymbol{r}}_{\mathcal{G}id}-\hat{\boldsymbol{r}}_{ib}\right) + \tfrac{\alpha_u}{2}\textstyle\sum_{j\in\mathcal{G}}\left\|\boldsymbol{U}_j\right\|^2 + \tfrac{\alpha_v}{2}\left\|\boldsymbol{V}_d\right\|^2\\ &\quad + \tfrac{\alpha_v}{2}\left\|\boldsymbol{V}_b\right\|^2 + \tfrac{\beta_v}{2}\left\|\boldsymbol{b}_d\right\|^2 + \tfrac{\beta_v}{2}\left\|\boldsymbol{b}_b\right\|^2\\ &= \ln\left[\mathbf{1}+\exp\left(-\hat{\boldsymbol{r}}_{\mathcal{G}id;ib}\right)\right] + \tfrac{\alpha_u}{2}\textstyle\sum_{j\in\mathcal{G}}\left\|\boldsymbol{U}_j\right\|^2 + \tfrac{\alpha_v}{2}\left\|\boldsymbol{V}_d\right\|^2\\ &\quad + \tfrac{\alpha_v}{2}\left\|\boldsymbol{V}_b\right\|^2 + \tfrac{\beta_v}{2}\left\|\boldsymbol{b}_d\right\|^2 + \tfrac{\beta_v}{2}\left\|\boldsymbol{b}_b\right\|^2\end{aligned} \tag{16}$$

where $\hat{\boldsymbol{r}}_{\mathcal{G}id;ib} = \hat{\boldsymbol{r}}_{\mathcal{G}id} - \hat{\boldsymbol{r}}_{ib}$, and $\boldsymbol{\alpha_u}$, $\boldsymbol{\alpha_v}$ and $\boldsymbol{\beta_v}$ are the regularization weights ranging from 0.0001 to 0.1. $\boldsymbol{U_j}\in R^{1\times z}$ is the latent feature vector for microbe j, where z is the number of latent features. $\boldsymbol{V_d}\in R^{1\times z}$ and $\boldsymbol{b_d}$ are disease d's latent feature vector and bias values, respectively. We can then update the model parameters $\boldsymbol{\Theta}$ as:

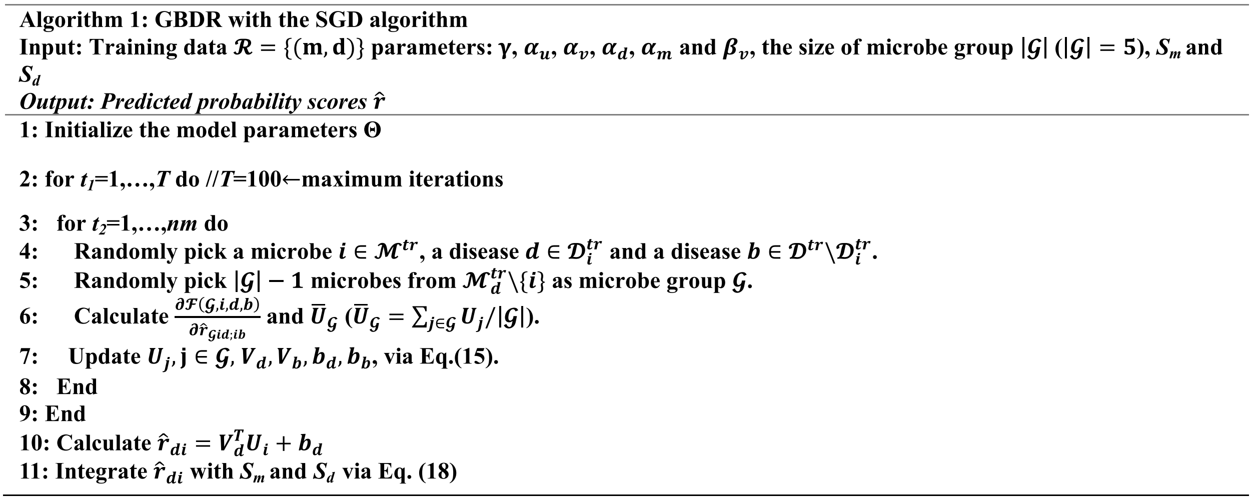
$$\boldsymbol{\Theta} = \boldsymbol{\Theta} - \boldsymbol{\gamma}\frac{\boldsymbol{\partial}\mathcal{F}(\mathcal{G},\boldsymbol{i},\boldsymbol{d},\boldsymbol{b})}{\boldsymbol{\partial\Theta}} \tag{17}$$

where the learning rate $\boldsymbol{\gamma}$ is set to 0.01 in this study via parameter tuning. The learning process is repeatedly trained until it reaches the maximum iterations (default: 100). The predicted score of microbe i on disease d is calculated via $\hat{\boldsymbol{r}}_{\boldsymbol{di}} = \boldsymbol{V}^{\boldsymbol{T}}_{\boldsymbol{d}}\boldsymbol{U}_{\boldsymbol{i}} + \boldsymbol{b}_{\boldsymbol{d}}$. The calculation procedure of GBDR is described by the pseudo-code in Algorithm 1.

Then we calculate $\hat{\boldsymbol{r}}_{\boldsymbol{di}}$ with the integrated microbe similarity S_m and disease semantic similarity S_d. For an unknown disease-microbe pair (d_i, m_j), $\boldsymbol{d}' \in \mathcal{D}^{\boldsymbol{tr}}_{\boldsymbol{m_j}}$ means a set of diseases having associations with microbe m_j in training data and $\boldsymbol{m}' \in \mathcal{M}^{\boldsymbol{tr}}_{\boldsymbol{d_i}}$ indicates a set of microbes having associations with disease d_i. Finally, the final prediction score of d_i on m_j could be calculated by adding the mean values as follows:

$$\hat{\boldsymbol{r}}_{\boldsymbol{d_i m_j}} += \frac{\boldsymbol{\alpha_d}}{|\boldsymbol{d}'|} \sum\nolimits_{\boldsymbol{d}' \in \mathcal{D}^{\boldsymbol{tr}}_{\boldsymbol{m_j}}} \boldsymbol{S_d}\left(\boldsymbol{d_i}, \boldsymbol{d}'\right) + \frac{\boldsymbol{\alpha_m}}{|\boldsymbol{m}'|} \sum\nolimits_{\boldsymbol{m}' \in \mathcal{M}^{\boldsymbol{tr}}_{\boldsymbol{d_i}}} \boldsymbol{S_m}\left(\boldsymbol{m_j}, \boldsymbol{m}'\right) \quad (18)$$

where parameters $\boldsymbol{\alpha_d}$ and $\boldsymbol{\alpha_m}$ control the weights of S_m and S_d respectively. In this way, $\hat{\boldsymbol{r}}_{\boldsymbol{d_i m_j}}$ is the predicted probability score of the unknown disease-microbe pair (d_i, m_j) ranging from -1 to 1. The higher value $\hat{\boldsymbol{r}}_{\boldsymbol{d_i m_j}}$, the higher probability of the potential association between disease d and microbe j. Then the model calculate $\hat{\boldsymbol{r}}$ for each unknown microbe-disease association. Finally, the potential microbe-disease associations can be predicted by ranking the predicted probability scores. The total time complexity of the proposed model is $O(Tnm|\mathcal{G}|z)$.



Algorithm 1: GBDR with the SGD algorithm
Input: Training data $\mathcal{R} = \{(\mathrm{m}, \mathrm{d})\}$ parameters: $\boldsymbol{\gamma}$, $\boldsymbol{\alpha_u}$, $\boldsymbol{\alpha_v}$, $\boldsymbol{\alpha_d}$, $\boldsymbol{\alpha_m}$ and $\boldsymbol{\beta_v}$, the size of microbe group $|\mathcal{G}|$ ($|\mathcal{G}| = 5$), S_m and S_d
Output: Predicted probability scores $\hat{\boldsymbol{r}}$

```
Initialize the model parameters $\Theta$

for $t_1$=1,…,$T$ do //$T$=100←maximum iterations

  for $t_2$=1,…,$nm$ do
     Randomly pick a microbe $i \in \mathcal{M}^{tr}$, a disease $d \in \mathcal{D}^{tr}_i$ and a disease $b \in \mathcal{D}^{tr} \backslash \mathcal{D}^{tr}_i$.
     Randomly pick $|\mathcal{G}| - 1$ microbes from $\mathcal{M}^{tr}_d \backslash \{i\}$ as microbe group $\mathcal{G}$.
     Calculate $\frac{\partial \mathcal{F}(\mathcal{G},i,d,b)}{\partial \hat{r}_{\mathcal{G}id;ib}}$ and $\bar{U}_{\mathcal{G}}$ ($\bar{U}_{\mathcal{G}} = \sum_{j \in \mathcal{G}} U_j / |\mathcal{G}|$).
     Update $U_j$, j $\in \mathcal{G}$, $V_d$, $V_b$, $b_d$, $b_b$, via Eq.(15).
  End
End
Calculate $\hat{r}_{di} = V_d^T U_i + b_d$
Integrate $\hat{r}_{di}$ with $S_m$ and $S_d$ via Eq. (18)
```

Abbreviations

rRNA: ribosomal RNA; IBD: inflammatory bowel disease; OTUs: operational taxonomic units; LOOCV: leave-one-out cross validation; k-fold CV: k-fold cross validation; ROC: receiver operating characteristic; AUC: area under ROC curve; SVD: singular value decomposition; CF: collaborative filtering; CRC: colorectal carcinoma; NCBI: National Centre for Biotechnology Information; MeSH: Medical Subject Headings; NLM: Nation Library of Medicine; DAGs: Directed Acyclic Graphs; BLAST: Basic Local Alignment Search Tool; SGD: stochastic gradient descent.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08423-w.

Additional file 1. The case study of colorectal carcinoma.

Additional file 2. The prediction list of most likely pathogenic microbes is sorted by final prediction scores.

Additional file 3. Names of all investigated microbes and diseases, and known human microbe-disease associations obtained from HMDAD database.

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of BMC Genomics Volume 22 Supplement 1, 2021: Proceedings of the 2019 International Conference on Intelligent Computing (ICIC 2019): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-22-supplement-1.

Authors' contributions

ZAH & YAH conceived the algorithm, developed the program, and wrote the manuscript. ZHY and JQL carried out analyses, designed and performed experiments. LW, CQY & HCY prepared the data sets, carried out analyses, helped with manuscript editing and program design. All authors read and approved the final manuscript.

Authors' information

Yu-An Huang obtained his Ph.D. degree in Department of Computing at the Hong Kong Polytechnic University. He is currently an Associate Professor with Department of Information Engineering, Xijing University, Xi'an, China. Zhi-An Huang obtained his PhD degree in the City University of Hong Kong. His current research interests mainly focus on medical image analysis,

machine learning, and medical data mining. He is currently a Research Assistant Professor with Center for Computer Science and Information Technology, City University of Hong Kong Dongguan Research Institute, Dongguan, China.
Jian-Qiang Li obtained his PhD degree in South China University of Technology. He is now working as the vice dean of College of Computer & Software Engineering, Shenzhen University, Shenzhen, China.
Zhu-Hong You obtained his Ph.D. degree in control science and engineering from University of Science & Technology of China (USTC). He is currently a professor with Department of Information Engineering, Xijing University, Xi'an, China.
Lei Wang received the Ph.D. degree from China University of Mining and Technology. He is currently working as a Postdoc research fellow in the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi, China.
Hai-Cheng Yi is currently a PhD student at the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences. He works on machine learning, network analysis with applications in the field of bioinformatics.
Chang-Qing Yu is now an associate professor with Department of Information Engineering, Xijing University, Xi'an, China. His research interests include machine learning, network analysis.

Funding

YAH and JQL were supported by the National Key RD Program of China under Grants 2020YFA0908700. ZAH was supported by the National Natural Science Foundation of China under Grant No. 61702424. ZHY was supported by the National Natural Science Foundation of China under Grant No. 61572506.

Availability of data and materials

The datasets used and/or analysed in this study can be downloaded from the following public databases: the dataset of 16S rRNA partial or complete gene sequences from the National Center for Biotechnology Information (NCBI) repository (https://www.ncbi.nlm.nih.gov/); the dataset of MeSH descriptors from the Nation Library of Medicine (NLM) repository (https://www.nlm.nih.gov/mesh/meshhome.html); the dataset of known microbe-disease associations from the Human Microbe-Disease Association Database (HMDAD) repository (http://www.cuilab.cn/hmdad).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Information Engineering, Xijing University, Xi'an 710123, China. [2]Center for Computer Science and Information Technology, City University of Hong Kong Dongguan Research Institute, Dongguan, China. [3]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518000, China. [4]Guangxi Academy of Science, Nanning 530000, China. [5]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Ürümqi 830000, China.

Received: 26 December 2021 Accepted: 25 February 2022
Published online: 16 March 2022

References

1. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258–70.
2. Sender R, Fuchs S, Milo R. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. Cell. 2016;164(3):337–40.
3. Wang B, Yao M, Lv L, Ling Z, Li L. The human microbiota in health and disease. Engineering. 2017;3(1):71–82.
4. Tilg H, Kaser A. Gut microbiome, obesity, and metabolic dysfunction. J Clin Invest. 2011;121(6):2126–32.
5. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13(9):R79.
6. Savitz LD. The human microbiota: the role of microbial communities in health and disease. Acta Biologica Colombiana. 2016;21(1):5–15.
7. Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell. 2014;158(6):1402–14.
8. Davenport ER, Mizrahi-Man O, Michelini K, Barreiro LB, Ober C, Gilad Y. Seasonal variation in human gut microbiome composition. PLoS One. 2014;9(3):e90731.
9. Mason MR, Preshaw PM, Nagaraja HN, Dabdoub SM, Rahman A, Kumar PS. The subgingival microbiome of clinically healthy current and never smokers. ISME J. 2015;9(1):268–72.
10. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006;55(2):205–11.
11. Thibault R, Blachier F, Darcy-Vrillon B, de Coppet P, Bourreille A, Segain JP. Butyrate utilization by the colonic mucosa in inflammatory bowel diseases: a transport deficiency. Inflamm Bowel Dis. 2010;16(4):684–95.
12. Huang ZA, Wen Z, Deng Q, Chu Y, Sun Y, Zhu Z. LW-FQZip 2: a parallelized reference-based compression of FASTQ files. BMC Bioinformatics. 2017;18(1):179.
13. Huang Z-A, Huang Y-A, You Z-H, Zhu Z, Sun Y. Novel link prediction for large-scale miRNA-lncRNA interaction network in a bipartite graph. BMC Med Genet. 2018;11(6):113.
14. Huang Z-A, Huang Y-A, You Z-H, Zhu Z, Yu C-Q, Huang W, Guo J: Predicting lncRNA-miRNA interaction via graph convolution auto-encoder. Front Genetics 2019, 10:758.
15. Hartman AL, Riddle S, McPhillips T, Ludascher B, Eisen JA: Introducing W.A.T.E.R.S.: a workflow for the alignment, taxonomy, and ecology of ribosomal sequences. BMC Bioinformatics 2010, 11:317.
16. Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, et al. An analysis of human microbe-disease associations. Brief Bioinform. 2017;18(1):85–97.
17. Huang Y-A, You Z-H, Li L-P, Huang Z-A, Xiang L-X, Li X-F, et al. EPMDA: an expression-profile based computational model for microRNA-disease association prediction. Oncotarget. 2017;8(50):87033.
18. Huang Z-A, Zhang J, Zhu Z, Wu EQ, Tan KC: Identification of Autistic Risk Candidate Genes and Toxic Chemicals via Multilabel Learning. IEEE Transactions on Neural Networks and Learning Systems 2020.
19. Sun Y, Zhu Z, You Z-H, Zeng Z, Huang Z-A, Huang Y-A. FMSM: a novel computational model for predicting potential miRNA biomarkers for various human diseases. BMC Syst Biol. 2018;12(9):121.
20. Huang YA, You ZH, Gao X, Wong L, Wang L. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. Biomed Res Int. 2015;2015:902198.
21. Huang YA, You ZH, Chen X, Yan GY. Improved protein-protein interactions prediction via weighted sparse representation model combining continuous wavelet descriptor and PseAA composition. BMC Syst Biol. 2016;10(Suppl 4):120.
22. Huang Y, You Z, Chen X. A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences. Curr Protein Pept Sci. 2018;19(5):468–78.
23. Coenye T, Vandamme P. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol Lett. 2003;228(1):45–9.
24. Chen X, Huang YA, You ZH, Yan GY, Wang XS. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2017;33(5):733–9.
25. Huang ZA, Chen X, Zhu Z, Liu H, Yan GY, You ZH, et al. PBHMDA: path-based human microbe-disease association prediction. Front Microbiol. 2017;8:233.
26. Katz L. A new status index derived from sociometric analysis. Psychometrika. 1953;18(1):39–43.
27. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(Database issue):D61–5.

28. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88(3):265–6.
29. Wang D, Wang J, Lu M, Song F, Cui Q: Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics (Oxford, England) 2010, 26(13):1644–1650.
30. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL: NCBI BLAST: a better web interface. Nucleic acids research 2008, 36(Web Server issue):W5–9.
31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
32. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–4.
33. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947–8.
34. Thomas RH. Molecular evolution and Phylogenetics. Heredity. 2001;86(86):385–5.
35. Huang Z-A, Huang Y-A, You Z-H, Zhu Z, Sun Y. Novel link prediction for large-scale miRNA-lncRNA interaction network in a bipartite graph. BMC Med Genet. 2018;11(6):17–27.
36. Hu Y, Koren Y, Volinsky C: Collaborative Filtering for Implicit Feedback Datasets. In: Eighth IEEE International Conference on Data Mining: 2009. 263–272.
37. Pan R, Zhou Y, Cao B, Liu NN, Lukose R, Scholz M, Yang Q: One-Class Collaborative Filtering. In: Eighth IEEE International Conference on Data Mining: 2008. 502–511.
38. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L: BPR: Bayesian personalized ranking from implicit feedback. In: Conference on Uncertainty in Artificial Intelligence: 2009. 452–461.
39. Pan W, Chen L: GBPR: group preference based Bayesian personalized ranking for one-class collaborative filtering. In: International Joint Conference on Artificial Intelligence: 2013. 2691–2697.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.