Bioinformatics, 33(5), 2017, 733–739
doi: 10.1093/bioinformatics/btw715
Advance Access Publication Date: 14 December 2016
Original Paper

OXFORD

Systems biology

A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases

Xing Chen[1,*,†], Yu-An Huang[2,†], Zhu-Hong You[3,*], Gui-Ying Yan[4] and Xue-Song Wang[1,*]

[1]School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China, [2]Department of Computing, Hong Kong Polytechnic University, Hong Kong, [3]Chinese Academy of Science, Xinjiang Technical Institute of Physics and Chemistry, Ürümqi 830011, China and [4]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on July 12, 2016; revised on November 3, 2016; editorial decision on November 4, 2016; accepted on November 9, 2016

Abstract

Motivation: Accumulating clinical observations have indicated that microbes living in the human body are closely associated with a wide range of human noninfectious diseases, which provides promising insights into the complex disease mechanism understanding. Predicting microbe–disease associations could not only boost human disease diagnostic and prognostic, but also improve the new drug development. However, little efforts have been attempted to understand and predict human microbe–disease associations on a large scale until now.

Results: In this work, we constructed a microbe-human disease association network and further developed a novel computational model of KATZ measure for Human Microbe–Disease Association prediction (KATZHMDA) based on the assumption that functionally similar microbes tend to have similar interaction and non-interaction patterns with noninfectious diseases, and vice versa. To our knowledge, KATZHMDA is the first tool for microbe–disease association prediction. The reliable prediction performance could be attributed to the use of KATZ measurement, and the introduction of Gaussian interaction profile kernel similarity for microbes and diseases. LOOCV and k-fold cross validation were implemented to evaluate the effectiveness of this novel computational model based on known microbe–disease associations obtained from HMDAD database. As a result, KATZHMDA achieved reliable performance with average AUCs of 0.8130 ± 0.0054, 0.8301 ± 0.0033 and 0.8382 in 2-fold and 5-fold cross validation and LOOCV framework, respectively. It is anticipated that KATZHMDA could be used to obtain more novel microbes associated with important noninfectious human diseases and therefore benefit drug discovery and human medical improvement.

Availability and Implementation: Matlab codes and dataset explored in this work are available at http://dwz.cn/4oX5mS.

Contacts: xingchen@amss.ac.cn or zhuhongyou@gmail.com or wangxuesongcumt@163.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Lives of higher animals including humans are closely related to a diverse microbial community which is mainly composed of bacteria, as well as archea, viruses, fungi and protozoa (HMP Consortium, 2012; Sommer and Bäckhed, 2013). It can be considered as an essential 'organ' for humans since the microbiome benefits the host with improved metabolic capabilities, protection from pathogens, enhancement of immune system, and modulation of gastrointestinal development (Ventura *et al.*, 2009). For example, essential bacteria could boost the metabolism of indigestible polysaccharides and produce necessary vitamins, thus they are important elements for the development and differentiation of intestinal epithelium and immune system, which could help the host maintain tissue homeostasis (Consortium, 2012; Kau *et al.*, 2011). Mammalian intestine provides a nutrient-rich and temperature constant environment for microbiota to survive, which forms a mutualistic association (Hsiao *et al.*, 2013). There are about 10^{14} bacterial cells existing in an adult intestine, which is roughly ten times the number of human cells (Sommer and Bäckhed, 2013). The number of microbiome genes can reach 5 million, which outnumbers the human genetic potential by two orders of magnitude (Sommer and Bäckhed, 2013). This tremendous amount of gene product can lead a diverse range of biochemical and metabolic activities, serving as a host physiology complement. However, a systems understanding of microbiome remains limited.

The microbiota can be greatly influenced by their dynamic habitat which undergoes constant changes owing to different environmental variables like host diet (David *et al.*, 2014; Muegge *et al.*, 2011), season (Davenport *et al.*, 2014), smoking (Mason *et al.*, 2015), hygiene (Sommer and Bäckhed, 2013) and use of antibiotics (Donia *et al.*, 2014). These modifications in microbial community can further modify transcriptomic, proteomic and metabolic profiles, which can leads to harmful interactions between the host and microbiota. High-throughput sequencing techniques and newly developed software tools are revolutionizing analysis researches on microbiome, including 16S ribosomal RNA (rRNA) gene sequence (16S) and taxonomic profiles and whole-genome shotgun (WGS) (HMP Consortium, 2012). An increasing number of clinic reports have indicated that the body microbiota is closely associated with a diverse range of human noninfectious diseases such as cancer (Moore and Moore, 1995), diabetes (Brown *et al.*, 2011; Giongo *et al.*, 2011), obesity (Ley *et al.*, 2005; Zhang *et al.*, 2009), kidney stones (Hoppe *et al.*, 2011) and systemic inflammatory response syndrome (Mshvildadze *et al.*, 2010). For example, Penders *et al.* (2007) have discovered that differences in the gut microbiota composition are associated with the manifestation of atopic symptoms and atopic sensitization. They have pointed out the direct link between *Clostridium difficile* and all atopic symptoms and sensitization. Gevers *et al.* (2014) have reported a series of microbiomes associated with the new-onset Crohn's disease. Specifically, they found the status of Crohn's disease was strongly related to an increased abundance in bacteria including *Veillonellaceae*, *Pasteurellacaea*, *Enterobacteriaceae* and *Fusobacteriaceae*, and a decreased abundance in *Bacteroidales*, *Erysipelotrichales* and *Clostridiales*. Liu *et al.* have observed a shift in the composition of the oral microbiota when comparing healthy and periodontal disease samples (Liu *et al.*, 2012). In this article, periodontal disease was found to be associated with significantly lower abundance in a set of gram-positive genera including *Streptococcus*, *Granulicatella* and *Actinomyces*. Skov *et al.* showed that toxins from *Staphylococcus aureus* bacteria could function as superantigen which bypasses the normal control of T-cell activation, activate all T-cell clones, and further result in vigorous T-cell activation and cytokine release (Skov and Baadsgaard, 2000).

There are a growing number of clinic studies discovering new associations between microbes and human diseases, and these provide potential possibility of constructing microbe–disease association network. As known for more than a century, microbe is the main player in the pathogenic mechanism of infectious diseases. However, biological observations also demonstrate the role of microbes playing in the mechanism of non-infective diseases including diabetes (Brown *et al.*, 2011), obesity (Ley *et al.*, 2005; Zhang *et al.*, 2009) and cancers (Moore and Moore, 1995). Recently, Ma *et al.* have built the first Human Microbe–Disease Association Database (HMDAD) by curating large-scale microbe–disease associations from previously published studies (Ma *et al.*, 2016). Specifically, HMDAD database mainly focuses on experimentally supported associations between diverse microbes and non-infective-diseases. Furthermore, they discovered that microbe-based loops are significantly coherent. The procedure for microbes to influence disease development can be very complicated because they coexist in a community influencing each other and can act as co-causes in diverse diseases (Nathan, 2012). For example, phages can regulate the switches of bacteria through Iysogeny, and viral-bacterial coinfection has become a common clinical manifestation in lung diseases (Feiner *et al.*, 2015; Jamieson *et al.*, 2013). In addition, virulence is not an intrinsic property of a microbe but dependent on its specific context which is closely associated with the resistance and tolerance ability of hosts as well as the level of pathogen burden (Medzhitov *et al.*, 2012; Råberg *et al.*, 2009; Råberg *et al.*, 2007; Schneider and Ayres, 2008). Microbe-associated disease morbidity and mortality are usually ascribed to either high pathogen virulence or low host resistant (Jamieson *et al.*, 2013). HMDAD database collected and curated the human microbe–disease association data from microbiota studies based on sufficient samples and therefore can provide valuable information source for prediction model to an extensive extent.

Microbe–disease associations could provide great insight into understanding complex disease mechanism. For example, *gastric and duodenal ulcers* and *Whipple's disease* were once considered as noninfectious disease in origin, but were reclassified as infectious when the associated microbes were detected (Nathan, 2012). Discovering and predicting new microbe–disease associations is of great significance for the understanding of noninfectious disease formation and development mechanism, and the development of novel methods for disease diagnosis and therapy. Traditional treatments for bacterial infection diseases usually consider the use of antibiotics. However, antibiotic-based treatment can be a double-edged sword by suppressing both the pathogen and protective microbiota, and the chronic use of antibiotics can cause a negative impact on intestinal flora (Donia *et al.*, 2014). Restoring the damaged microbial communities would be a promising alternative approach to the disease treatment. Therefore, predicting the microbes associated with a specific disease can offer valuable information for the therapeutic regimen. For example, fecal microbiota transplantation is confirmed as a safe and effective treatment option for clostridium difficile infection (CDI), which reintroduces normal flora via donor feces, corrects the imbalance and reestablishes the normal bowel function (Bakken *et al.*, 2011). Conventional experiment-based methods for discovering microbe–disease associations are time-consuming and costly. And some bacterium even cannot be cultivated by current culturing techniques in the laboratory (Stewart, 2012). To date, little efforts have been made to the development of computational models for large-scale microbe–disease association prediction, which could effectively select most potential microbe–disease association candidates for experimental validation, and therefore reduce the cost and time of experimental researches.

In this work, we developed the model of KATZ measure for Human Microbe–Disease Association prediction (KATZHMDA) for predicting potential microbe–disease associations. To our knowledge, it is the first computational model proposed for predicting microbe–disease associations. KATZHMDA can predict new microbe–disease associations in a large scale by combining the known microbe–disease associations and Gaussian interaction profile kernel similarity for microbes and diseases. The proposed model only relies on the topology information of known microbe–disease association network as information source. To evaluate the prediction performance of the proposed model, evaluation frameworks of leave-one-out cross validation (LOOCV) and 5-fold cross validation were adopted based on the microbe–disease associations in HMDAD database. A series of comparison experiments were also implemented to evaluate the influence of the number of walks on prediction performance. As a result, KATZHMDA achieved its best performance when the number of walks was set as 2. Specifically, the proposed model achieved an average value of AUCs of 0.8301 ± 0.0033 for 5-fold cross validation, and AUC of 0.8382 for LOOCV. The prediction results fully demonstrated that the model of KATZHMDA is feasible and effective for predicting large-scale microbe–disease associations by only considering the topology information of known microbe–disease association network.

2 Materials

Known microbe–disease associations were downloaded from the Human Microbe–Disease Association Database (HMDAD, http://www.cuilab.cn/hmdad) in April, 2016 (Ma *et al.*, 2016). The microbe–disease associations were mainly collected from the 16s RNA sequencing-based microbiome studies which only give out genus-level information. For those microorganism names which are presented higher than genus-level, the database creators kept them in original names. As a result, there are 483 microbe–disease associations (including 39 human disease and 292 microbes) collected from 61 publications. After removing the same microbe–disease associations from different evidences, we obtained 450 distinct associations, which further constructed an adjacency matrix A of size 292×39 as the information source.

3 Methods

3.1 Gaussian interaction profile kernel similarity for microbes

Based on the assumption that functionally similar microbes share the similar interaction and non-interaction pattern with diseases, the Gaussian interaction profile kernel similarity for microbes was constructed from known microbe–disease association network (Chen and Yan, 2013). The procedure of Gaussian interaction profile kernel similarity mainly consists of two steps. Firstly, the interaction profile of each microbe is represented as a binary vector which encodes the presence or absence of associations between the microbe and each disease in the known microbe–disease association network ('0' and '1' represent absence and presence, respectively). For a given microbe m_i, its interaction profile $IP(m_i)$ would be defined as the *ith* column of the adjacency matrix A (if there is known association between disease d_i and microbe m_j, $A(i,j)$ is 1; otherwise 0) . In the second step, the Gaussian interaction profile kernel similarity between each microbe pair (say m_i and m_j) was computed based on their interaction profiles as follow:

$$KM(m_i, m_j) = \exp\left(-\gamma_m \|IP(m_i) - IP(m_j)\|^2\right) \tag{1}$$

$$\gamma_m = \gamma'_m / \left(\frac{1}{n_m}\sum_{k=1}^{n_m} \|IP(m_k)\|^2\right) \tag{2}$$

Here, γ_d denotes the normalized kernel bandwidth based on the new bandwidth parameter γ'_d; n_d denotes the number of diseases; Entity $KD(i,j)$ denotes the Gaussian interaction profile kernel similarity between disease d_i and d_j.

3.2 Gaussian interaction profile kernel similarity for diseases

Disease Gaussian interaction profile kernel similarity matrix was constructed based on the assumption that similar diseases would have similar interaction and non-interaction pattern with microbes in the known microbe–disease interaction network. Specifically, Gaussian interaction profile kernel similarity matrix for diseases, KD, was computed in a similar way as microbes as follow:

$$KD(d_i, d_j) = \exp\left(-\gamma_d \|IP(d_i) - IP(d_j)\|^2\right) \tag{3}$$

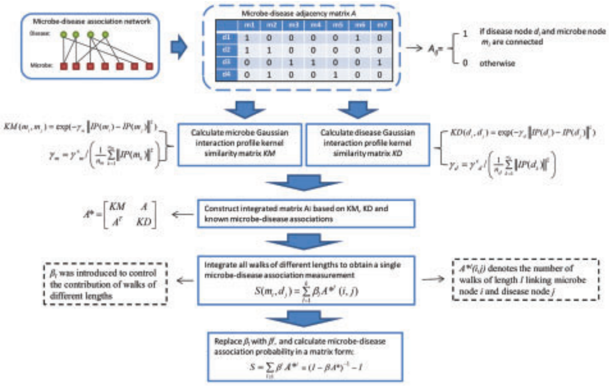
$$\gamma_d = \gamma'_d / \left(\frac{1}{n_d}\sum_{k=1}^{n_d} \|IP(d_k)\|^2\right) \tag{4}$$

Here, γ_d denotes the normalized kernel bandwidth based on the new bandwidth parameter γ'_d; n_d denotes the number of diseases; Entity $KD(i,j)$ denotes the Gaussian interaction profile kernel similarity between disease d_i and d_j. Motivated by previous works (Vanunu *et al.*, 2010), a logistics function was then implemented to regulate disease-disease similarity.

3.3 KATZHMDA

Inspired by the successful application of KATZ measure in the fields of social network prediction (Katz, 1953), disease-gene association prediction (Yang *et al.*, 2014), and lncRNA-disease association prediction (Chen, 2015a, b), we explored the KATZ measure by developing a new computational model for predicting microbe–disease associations (see Fig. 1). KATZ is a network-based measurement method which calculates the nodes' similarity in a heterogeneous network to solve the problem of link prediction. This method considers the number of walks between nodes and walk lengths in a graph as effective similarity metrics. Therefore, we transformed the problem of measuring microbe–disease association into counting the number of walks of connections between microbe and disease nodes in the heterogeneous network. In addition, the influence of walk length was also considered as effective information by integrating the number of walks and their lengths for measuring the correlation of microbe–disease pairs. Here, heterogeneous network consists of microbe similarity network based on microbe Gaussian interaction profile kernel similarity, disease similarity network based on disease Gaussian interaction profile kernel similarity, and known microbe–disease association network constructed based on recorded associations from HMDAD database.

KATHMDA first calculates the number of walks between microbe nodes (say m_i) and disease nodes (say d_j) in the known microbe–disease association network. We constructed adjacency matrix A based the known microbe–disease associations, and the number of l-length walks between m_i and d_j could be obtained by


Microbe-disease association network
Disease:
Microbe:
Microbe-disease adjacency matrix A
1 if disease node d_i and microbe node m_j are connected
0 otherwise
$A_{ij}=$
Calculate microbe Gaussian interaction profile kernel similarity matrix KM
Calculate disease Gaussian interaction profile kernel similarity matrix KD
$KM(m_i,m_j)=\exp(-\gamma_m\|IP(m_i)-IP(m_j)\|^2)$
$\gamma_m=\gamma'_m/\left(\frac{1}{n_m}\sum_{i=1}^{n_m}\|IP(m_i)\|^2\right)$
$KD(d_i,d_j)=\exp(-\gamma_d\|IP(d_i)-IP(d_j)\|^2)$
$\gamma_d=\gamma'_d/\left(\frac{1}{n_d}\sum_{i=1}^{n_d}\|IP(d_i)\|^2\right)$
Construct integrated matrix Ai based on KM, KD and known microbe-disease associations
$A^*=\begin{bmatrix}KM & A\\ A^T & KD\end{bmatrix}$
Integrate all walks of different lengths to obtain a single microbe-disease association measurement
$S(m_i,d_j)=\sum_{l=1}^{k}\beta_l A^{*l}(i,j)$
β_l was introduced to control the contribution of walks of different lengths
$A^{*l}(i,j)$ denotes the number of walks of length l linking microbe node i and disease node j
Replace β_l with β^l, and calculate microbe-disease association probability in a matrix form:
$S=\sum_{l\geq1}\beta^l A^{*l}=(I-\beta A^*)^{-1}-I$

	m1	m2	m3	m4	m5	m6	m7
d1	1	0	0	0	0	1	0
d2	1	1	0	0	0	0	0
d3	0	0	1	1	0	0	1
d4	0	1	0	0	1	0	0

Fig. 1. Flowchart of KATZHMDA based on known microbe–disease association network (Color version of this figure is available at *Bioinformatics* online.)

calculating $A^l(i,j)$. In order to introduce Gaussian interaction profile kernel similarity for microbes and diseases into the computation model, we further integrated microbe similarity matrix *KM*, disease similarity matrix *KD* and adjacency matrix *A* as follow:

$$A^* = \begin{bmatrix} KM & A \\ A^T & KD \end{bmatrix} \tag{5}$$

In this way, a new integrated matrix *A** could be obtained and used for further microbe–disease association prediction, which allows KATZHMDA to be applied to new microbes and diseases with no known associations if we could obtain the further similarity information without rely on known microbe–disease network topology information. All walks of different lengths were integrated for obtaining a single measurement of each microbe–disease pair. Considering the different contribution from walks of different lengths (shorter walks tend to contribute more to the similarity measurement), we introduced nonnegative coefficient sequence β_l to dampen the contribution of longer walks. The corresponding coefficients of shorter walks in β_l would be larger than those of longer walks. For example, given two coefficients β_{l1} and β_{l2} ($l_1 < l_2$), the value of β_{l1} would be set to be larger than β_{l2}. Therefore, the potential association probability between microbe m_i and disease d_j could be calculated as the entity $S(m_i,d_j)$ of matrix *S*:

$$S(m_i, d_j) = \sum_{l=1}^{k} \beta_l A^{*l}(i,j) \tag{6}$$

Here, we replace β_l by β^l and transform this equation into the matrix form:

$$S = \sum_{l \geq 1} \beta^l A^{*l} = (I - \beta A^*)^{-1} - I \tag{7}$$

Here, matrix *S* with dimension $(39+292) \times (39+292)$ depicts the association possibilities of all the microbe–disease pairs. Furthermore, we could represent the matrix S in Equation (7) as the following partitioned matrix form similar to Equation (5):

$$S = \begin{bmatrix} S11 & S12 \\ S21 & S22 \end{bmatrix} \tag{8}$$

It could be easily inferred that matrix S12 is our final prediction result, which could provide the association probability between each microbe and disease. However, the microbe–disease association network is still sparse due to the limited data from HMDAD database. Considering walkers of long lengths in sparse network may be meaningless and therefore disturb the association prediction, we here set *k* to be 2, 3 and 4, and evaluated the influence of this parameter setting on the prediction performance. Specifically, when *k* was set as 2, 3 and 4, the potential association probability between m_i and d_j could be calculated by the following formulas based on aforementioned Equations (7) and (8), respectively. In this way, final prediction result matrix could be represented by matrix *A*, *KD* and *KM*.

$$S_{k=2} = \beta \cdot A + \beta^2 \cdot (KM \cdot A + A \cdot KD) \tag{9}$$

$$S_{k=3} = S_{k=2} + \beta^3 \cdot \left(A \cdot A^T \cdot A + KM^2 \cdot A + KM \cdot A \cdot KD + A \cdot KD^2\right) \tag{10}$$

$$\begin{aligned} S_{k=4} = S_{k=3} &+ \beta^4 \cdot (KM^3 \cdot A + A \cdot A^T \cdot KM \cdot A \\ &+ \mathrm{KM} \cdot A \cdot A^T \cdot A + A \cdot KD \cdot A^T \cdot A) \\ &+ \beta^4 \cdot (A \cdot A^T \cdot A \cdot KD + KM^2 \cdot A \cdot KD \\ &+ \mathrm{KM} \cdot A \cdot KD^2 + A \cdot KD^3) \end{aligned} \tag{11}$$

3.4 Integrating symptom-based disease similarity

In the model of KATZHMDA, Gaussian kernel-based similarity is used to measure the similarities of microbes and diseases based on their interaction profiles, providing an extensible prediction framework by combining other types of disease/microbe similarity. There are some computational methods having been proposed to measure disease similarity based on different kinds of data. For example, Zhou *et al.* previously proposed a method to construct the symptom-based human disease network (HSDN) in which disease similarity is computed based on co-occurrence of disease/symptom terms in PubMed bibliographic records (Zhou *et al.*, 2014). We here introduce the symptom-based disease similarity (*SDM*) into KATZHMDA model and evaluate the prediction performance of the combined model in the evaluation frameworks of LOOCV and 5-fold cross validation. Specifically, new disease similarity matrix (say *KD'*) was computed as a mean matrix of *SDM* and *KD*:

$$KD' = \frac{KD + SDM}{2} \tag{12}$$

4 Results

4.1 Leave-one-out cross validation

To evaluate the prediction performance for microbe–disease associations, KATZHMDA was evaluate by LOOCV based on the known associations in HMDAD database, which provides many confirmed microbe–disease associations manually collected from published biological reports. In the validation framework of LOOCV, each known microbe–disease association was left out in turn for testing and the other microbe–disease associations were used as training samples for model learning. Specifically, all the microbe–disease pairs without known relevance evidences would be regarded as candidate samples. The rank of each left-out testing sample relative to the candidate samples was further obtained. The testing samples with a prediction rank higher than the given threshold would be considered to be successfully predicted. We could obtained the corresponding true positive rates (TPR, sensitivity) and false positive rates (FPR, 1-specificity) by setting different thresholds. Here, sensitivity means the percentage of the test samples which were predicted with higher ranks than the given threshold, and specificity was computed as the percentage of negative samples with lower ranks than the threshold. The receiver-operating characteristics (ROC) curves could then be drawn by plotting TPR versus FPR at different thresholds. To evaluate the prediction performance of KATZHMDA, the areas under ROC curve (AUC) were further calculated. AUC value of 1 demonstrates a perfect prediction while the AUC value of 0.5 indicates purely random performance.

The number of walks k is the important parameter of KATZHMDA, and its value could influence the prediction performance for microbe–disease associations. Therefore, we implemented a series of comparison experiments to evaluate the influence of k. As a result, KATZHMDA achieved the best prediction performance when k was set as 2, and kept a decreasing trend in prediction performance when k increased from 2 to 4 (see Table 1 and Fig. 2). Specifically, when k was set as 2, KATZHMDA achieved AUC of 0.8382 in the framework of global LOOCV. This change of performance in comparison experiments could be due to the small number of known microbe–disease associations in HMDAD database. Since the known microbe–disease association network depicted by HMDAD database is sparse (only 483 recorded microbe–disease associations including 39 human diseases and 292 microbes), walks with longer lengths may be meaningless and disturb the prediction of KATZHMDA. Therefore, we finally set $k = 2$ for model development. The parameter values of γ' and β were selected by following previous works in which γ' and β were set as 1.0 and 0.01 respectively (Chen, 2015a, b; Chen *et al.*, 2016a, b, c, d, e; Chen and Yan, 2013). We also implemented a grid searching method to evaluate our parameter selecting by using LOOCV. As a result, the yielded AUC kept a decreasing trend from 0.8382 to 0.8065 when γ was set as 1.0, 1.5, 2.0 and 2.5, remained unchanged when β was set as 0.01, 0.05 and 0.10.

4.2 K-fold cross validation

K-fold cross validation was also implemented for the performance evaluation of KATZHMDA. In the framework of k-fold cross validation, all the known microbe–disease association samples were randomly equally divided into k parts. And $k-1$ parts were then used as training samples for model learning while the rest part was used as testing samples for model evaluation. Specifically, 2-fold and 5-fold cross validation were implemented to further evaluate the prediction performance of KATZHMDA. In a similar way as LOOCV, all the microbe–disease pairs without known relevance evidences would be considered as candidate samples. Considering the potential bias caused by random sample division for performance evaluation, we repeatedly divided the known microbe–disease associations 100 times, and the corresponding ROC curves and AUCs were obtained in the similar way as LOOCV. As a result, KATZHMDA achieved the best prediction performance with average AUCs of 0.8171 and 0.8301 with standard deviation of 0.0051 and 0.0033 when using the 2-fold and 5-fold cross validation (see Table 2).

Table 1. Performance comparison among different parameter settings ($k = 2$, 3 and 4) in the framework of 5-fold cross validation

$k = 2$	$k = 3$	$k = 4$
0.8301 ± 0.0033	0.8280 ± 0.0034	0.8184 ± 0.0033

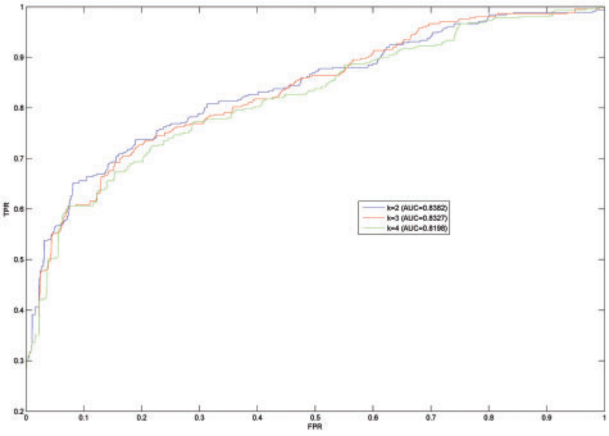


Fig. 2. Prediction performance of KATZHMDA with different parameter settings ($k = 2$, 3 and 4) in terms of ROC curve and AUC based on global LOOCV

The model of KATZHMDA has demonstrated its reliable and effective prediction performance in the LOOCV and k-fold cross validation. Therefore, we prioritized all the candidate microbes for the diseases recorded in HMDAD database by using the experimentally confirmed microbe–disease associations stored in HMDAD database and implementing the model of KATZHMDA. We publicly released the predicted of microbes for each disease, which may offer valuable information and clues for biological experiments (see Supplementary Table S1). It is anticipated that the microbe–disease associations with higher ranks would be confirmed by experimental observations in the future.

4.3 Comparison with other methods

To further evaluate the performance of the proposed model, we explored the baseline method which performs singular value decomposition (SVD) on the adjacency matrix A to recover the missing values as the possibility scores of associations between diseases and microbes. Two other previously proposed prediction models (i.e. RKHMDA and NETCBI-HMDA) were also implemented to compare with KATZHMDA (Shi *et al.*, 2013; van Laarhoven *et al.*, 2011). These two models were not proposed for inferring microbe–disease associations but they are also based on same bipartite graph prediction as KATZHMDA solves. As a result, KATZHMDA achieved the best performance among four explored methods with AUC of 0.8382 while NETCBI-HMDA and RKHMDA yielded AUCs of 0.1510 and 0.5387 respectively (see Fig. 3).

In addition, we also explored a microbe-based disease-disease similarity which was previously proposed by Ma *et al.* and mainly based on a three-node loop measure (Ma *et al.*, 2016). By comparing the result of this disease similarity measure (AUCs of 0.8308 and 0.8274 ± 0.0036 in LOOCV and 5-fold CV), we found that a simple Gaussian kernel function has more reliable performance (see Table 3 and Fig. 4). As mentioned in Section 3.4, KATZHMDA model provides an extensible framework to combine other kinds of disease and microbe similarity. Therefore, we further introduced symptom-based disease similarity and evaluate the extensibility ability to some extent (see Table 3 and Fig. 4). As a result, when combined with symptom-based disease similarity, the KATZHMDA model obtained an increased prediction performance with AUCs of

Table 2. Performance comparison among different evaluation frameworks (LOOCV, 2-fold and 5-fold cross validation)

LOOCV	2-Fold CV	5-Fold CV
0.8382	0.8171 ± 0.0051	0.8301 ± 0.0033

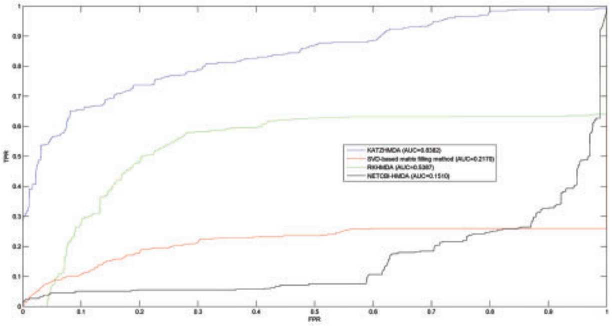


Fig. 3. Prediction performances of different models comparing with KATZHMDA in terms of ROC curve and AUC based on global LOOCV (Color version of this figure is available at *Bioinformatics* online.)

Table 3. Performance comparison among different disease similarity measures in 5-fold cross validation

Method	5-Fold CV
KATZHMDA with loop-based disease similarity	0.8274 ± 0.0036
KATZHMDA combined with additional symptom-based disease similarity	0.8566 ± 0.0033
KATZHMDA	0.8301 ± 0.0033

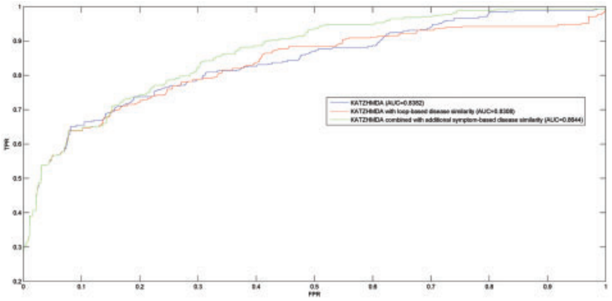


Fig. 4. Prediction performances of KATZHMDA with different disease similarity measures in terms of ROC curve and AUC based on global LOOCV (Color version of this figure is available at *Bioinformatics* online.)

0.8644 and 0.8566 ± 0.0033 in the frameworks of LOOCV and 5-fold cross validation, which accords with our hypothesis that additional data from independent information sources can further improve the performance of KATZHMDA.

5 Discussion and conclusion

Accumulating evidences show the development of noninfectious human diseases, especially gastrointestinal diseases, is closely associated with the involvement of microbes. And microbe–disease association network can offer valuable information for the understanding of disease mechanism and novel drug discovery. However, there is no computational model for predicting potential microbe–disease associations. To utilize the wealth of microbe–disease association data collected from previously published experimental reports, in this article, we proposed the first computational model for predicting microbe–disease associations, KATZHMDA, by integrating known microbe–disease associations and Gaussian interaction profile kernel similarity for microbes and diseases. In order to evaluate the prediction performance of KATZHMDA, the validation frameworks of LOOCV and 5-fold cross validation were implemented based on known microbe–disease associations in HMDAD database. The reliable prediction performance fully demonstrates the effectiveness of the proposed model. The microbe–disease associations with higher predicted ranks are expected to be confirmed by experimental observations in the future.

KATZHMDA was developed to predict potential disease-related microbes by measuring the correlation between nodes in the network, which is inspired by the successful use of KATZ method for predicting friends in social networks (Katz, 1953) (See Fig. 1). The proposed model is mainly based on the assumption that microbes of similar functions tend to get involved in similar disease association patterns and similar diseases are more possibly associated with the abnormal abundance of functionally similar microbes. KATZHMDA measures the correlations between candidate microbes and investigated diseases by integrating walks with different lengths in a heterogeneous network, which is constructed by combining the known disease-microbe network, disease similarity network and microbe similarity network. KATZHMDA, as a global computational method, could reconstruct potential microbe–disease associations for all diseases simultaneously in a large-scale network. It is anticipated KATZHMDA could become a useful and effective computational tool for biomedical researches.

There are still some limitations existing in the current version of KATZHMDA. First, since the optimal value of *k* (the number of walks) could be influenced by the sparsity of the known microbe–disease association network, its value would need to be adjusted when more newly records have been introduced into the database. In addition, Gaussian interaction profile kernel similarity was calculated greatly relying on the known microbe–disease associations, and therefore would cause inevitable bias towards those well-investigated diseases and microbes. In other words, the diseases with more related microbe records in database would be more possibly predicted to be associated with more potential microbes. And the same goes for the microbes with more related disease records. Furthermore, the prediction performance of KATZHMDA is still not very satisfactory, and additional data integration may benefit its predictive ability. There are various kinds of prior information of diseases and microbes which could be introduced to this computational model, such as disease phenotypic similarity, disease semantic similarity (Chen *et al.*, 2016a, b, c, d, e; Chen *et al.*, 2015; Chen and Yan, 2014; Huang *et al.*, 2016a, b; You *et al.*, 2014) in MeSH DAGs and various microbe-related interactions. Introducing other biological data could significantly improve the performance of the network-based prediction computational model (Chen, 2015a, b; Chen, 2016; Chen *et al.*, 2012a, b; Chen *et al.*, 2016a, b, c, d, e; Huang *et al.*, 2016a, b; Liu *et al.*, 2014; Wong *et al.*, 2015; You *et al.*, 2010; You *et al.*, 2013). Finally, KATZHMDA cannot be applied to new diseases and microbes which are poorly investigated without any known associations. Introducing similarity without rely on the topology information of known microbe–disease association network could solve this important limitation.

Acknowledgements

We thank anonymous reviewers for very valuable suggestions.

Funding

XC was supported by National Natural Science Foundation of China under Grant No. 11301517 and 11631014. ZHY was supported by National Natural Science Foundation of China under Grant No. 61572506. GYY was supported by National Natural Science Foundation of China under Grant

No. 11371355 and 11631014. XSW was supported by Fundamental Research Funds for the Central Universities (2014YC07).

Conflict of Interest: none declared.

References

Bakken,J.S. *et al.* (2011) Treating *Clostridium difficile* infection with fecal microbiota transplantation. *Clin. Gastroenterol. Hepatol.*, **9**, 1044–1049.

Brown,C.T. *et al.* (2011) Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PloS One*, **6**, e25792.

Chen,X. (2015a) KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.*, **5**, 16840.

Chen,X. (2015b) Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.*, **5**, 13186.

Chen,X. (2016) miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method. *Mol. Biosyst.*, **12**, 624–633.

Chen,X. *et al.* (2012a) Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst.*, **8**, 1970–1978.

Chen,X. *et al.* (2012b) RWRMDA: predicting novel human microRNA-disease associations. *Mo.l Biosyst.*, **8**, 2792–2798.

Chen,X. *et al.* (2015) Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.*, **5**, 11338.

Chen,X. *et al.* (2016a) FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. *Oncotarget*, **7**, 45948–45958.

Chen,X. *et al.* (2016b) NLLSS: Predicting Synergistic Drug Combinations based on semi-supervised learning. *PLOS Comput. Biol.*, **12**, e1004975.

Chen,X. *et al.* (2016c) Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinf.*, doi: 10.1093/bib/bbw060.

Chen,X. *et al.* (2016d) WBSMDA: within and between score for MiRNA-disease association prediction. *Sci. Rep.*, **6**, 21106.

Chen,X. *et al.* (2016e) HGIMDA: Heterogeneous Graph Inference for MiRNA-Disease Association prediction. *Oncotarget*, **7**, 65257–65269.

Chen,X. and Yan,G.Y. (2013) Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics*, **29**, 2617–2624.

Chen,X. and Yan,G.Y. (2014) Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.*, **4**, 5501.

Consortium,H.M.P. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Davenport,E.R. *et al.* (2014) Seasonal variation in human gut microbiome composition. *PloS One*, **9**, e90731.

David,L.A. *et al.* (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, **505**, 559–563.

Donia,M.S. *et al.* (2014) A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*, **158**, 1402–1414.

Feiner,R. *et al.* (2015) A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat. Rev. Microbiol.*, **13**, 641–650.

Gevers,D. *et al.* (2014) The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe*, **15**, 382–392.

Giongo,A. *et al.* (2011) Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J.*, **5**, 82–91.

Hoppe,B. *et al.* (2011) Efficacy and safety of *Oxalobacter formigenes* to reduce urinary oxalate in primary hyperoxaluria. *Nephrol. Dialysis Transplant.*, **26**, 3609–3615.

Hsiao,E.Y. *et al.* (2013) Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, **155**, 1451–1463.

Huang,Y.A. *et al.* (2016a) Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinf.*, **17**, 184.

Huang,Y. *et al.* (2016b) ILNCSIM: improved lncRNA functional similarity calculation model. *Oncotarget*, **7**, 25902–25914.

Jamieson,A.M. *et al.* (2013) Role of tissue protection in lethal respiratory viral-bacterial coinfection. *Science*, **340**, 1230–1234.

Katz,L. (1953) A new status index derived from sociometric analysis. *Psychometrika*, **18**, 39–43.

Kau,A.L. *et al.* (2011) Human nutrition, the gut microbiome and the immune system. *Nature*, **474**, 327–336.

Ley,R.E. *et al.* (2005) Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 11070–11075.

Liu,B. *et al.* (2012) Deep sequencing of the oral microbiome reveals signatures of periodontal disease. *PloS One*, **7**, e37919.

Liu,M.X. *et al.* (2014) A computational framework to infer human disease-associated long noncoding RNAs. *PLoS One*, **9**, e84408.

Ma,W. *et al.* (2016) An analysis of human microbe–disease associations. *Brief. Bioinf.*, bbw005.

Mason,M.R. *et al.* (2015) The subgingival microbiome of clinically healthy current and never smokers. *ISME J.*, **9**, 268–272.

Medzhitov,R. *et al.* (2012) Disease tolerance as a defense strategy. *Science*, **335**, 936–941.

Moore,W. and Moore,L.H. (1995) Intestinal floras of populations that have a high risk of colon cancer. *Appl. Environ. Microbiol.*, **61**, 3202–3207.

Mshvildadze,M. *et al.* (2010) Intestinal microbial ecology in premature infants assessed with non–culture-based techniques. *J. Pediatrics*, **156**, 20–25.

Muegge,B.D. *et al.* (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.

Nathan,C. (2012) Fresh approaches to anti-infective therapies. *Sci. Transl. Med.*, **4**, 140sr142–140sr142.

Penders,J. *et al.* (2007) Gut microbiota composition and development of atopic manifestations in infancy: the KOALA Birth Cohort Study. *Gut*, **56**, 661–667.

Råberg,L. *et al.* (2009) Decomposing health: tolerance and resistance to parasites in animals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **364**, 37–49.

Råberg,L. *et al.* (2007) Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. *Science*, **318**, 812–814.

Schneider,D.S. and Ayres,J.S. (2008) Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases. *Nat. Rev. Immunol.*, **8**, 889–895.

Shi,X. *et al.* (2013) A critical role for the long non-coding RNA GAS5 in proliferation and apoptosis in non-small-cell lung cancer. *Mol. Carcinog.*, **54**, E1–E12.

Skov,L. and Baadsgaard,O. (2000) Bacterial superantigens and inflammatory skin diseases. *Clin. Exp. Dermatol.*, **25**, 57–61.

Sommer,F. and Bäckhed,F. (2013) The gut microbiota—masters of host development and physiology. *Nat. Rev. Microbiol.*, **11**, 227–238.

Stewart,E.J. (2012) Growing unculturable bacteria. *J. Bacteriol.*, **194**, 4151–4160.

van Laarhoven,T. *et al.* (2011) Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, **27**, 3036–3043.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Ventura,M. *et al.* (2009) Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat. Rev. Microbiol.*, **7**, 61–71.

Wong,L. *et al.* (2015) Detection of interactions between proteins through rotation forest and local phase quantization descriptors. *Int. J. Mol. Sci.*, **17**, 21.

Yang,X. *et al.* (2014) A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS One*, **9**, e87797.

You,Z.H. *et al.* (2010) Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics*, **26**, 2744–2751.

You,Z.H. *et al.* (2013) Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinf.*, **14**, S10.

You,Z.H. *et al.* (2014) Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinf.*, **15**, S9.

Zhang,H. *et al.* (2009) Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 2365–2370.

Zhou,X. *et al.* (2014) Human symptoms–disease network. *Nat. Commun.*, **5**, 4212.