

RESEARCH

Open Access



Novel human microbe-disease association prediction using network consistency projection

Wenzheng Bao, Zhichao Jiang and De-Shuang Huang*

From 16th International Conference on Bioinformatics (InCoB 2017)
Shenzhen, China. 20-22 September 2017

Abstract

Background: Accumulating biological and clinical reports have indicated that imbalance of microbial community is closely associated with occurrence and development of various complex human diseases. Identifying potential microbe-disease associations, which could provide better understanding of disease pathology and further boost disease diagnostic and prognostic, has attracted more and more attention. However, hardly any computational models have been developed for large scale microbe-disease association prediction.

Results: In this article, based on the assumption that microbes with similar functions tend to share similar association or non-association patterns with similar diseases and vice versa, we proposed the model of Network Consistency Projection for Human Microbe-Disease Association prediction (NCPHMDA) by integrating known microbe-disease associations and Gaussian interaction profile kernel similarity for microbes and diseases. NCPHMDA yielded outstanding AUCs of 0.9039, 0.7953 and average AUC of 0.8918 in global leave-one-out cross validation, local leave-one-out cross validation and 5-fold cross validation, respectively. Furthermore, colon cancer, asthma and type 2 diabetes were taken as independent case studies, where 9, 9 and 8 out of the top 10 predicted microbes were successfully confirmed by recent published clinical literature.

Conclusion: NCPHMDA is a non-parametric universal network-based method which can simultaneously predict associated microbes for investigated diseases but does not require negative samples. It is anticipated that NCPHMDA would become an effective biological resource for clinical experimental guidance.

Keywords: Microbe, Disease, Association prediction, Network consistency projection

Background

In the past few decades, accumulating evidence has demonstrated that human lives strongly rely on a diverse, complex and dynamic microbial community, including bacteria, protozoa, viruses, eukaryotes, archaea and so on [1]. Tremendous microorganisms inhabit a range of human organs such as skin, gut, mouth, stomach and vagina, where a commensal relationship between microbe and human host has been established after a long term adaptive co-evolution. Recently, more

and more reports have confirmed that microbiome could benefit human health by maintaining normal homeostasis, strengthening immune system, promoting host's metabolism, and modulating development of gastrointestinal tract [2]. Typically, it is reported that the number of bacterial cells in an adult intestine reaches 10^{14} , which is approximately 10 times as the number of total human cells [3]. More than 5,000,000 genes (out-numbering the human genetic potential by two orders of magnitude) are contained in the combined genomes of these bacteria, and tens of trillions of gene products are involved in a variety of biochemical and metabolic activities, providing important complement to host physiology [1, 4]. In a sense, it is reasonable to regard gut

* Correspondence: dshuang@tongji.edu.cn
Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Caoan Road 4800, Shanghai 201804, China

bacteria as an additional 'organ' for its equal metabolic capacity as the liver [5]. Essential gut bacteria could effectively promote nutrient absorption by assisting decomposition of indigestible polysaccharides and production of indispensable vitamins [3]. Furthermore, they provide important protection against invasion of food-borne pathogens by impacting on proliferation and differentiation of host intestinal epithelium [6, 7]. However, a system understanding of how these biochemical activities achieve still remain largely unknown.

According to recent reviews, microbiota in human bodies could be significantly influenced by both maternal genetics [8, 9] and environment variables including hygiene of food and residence [3], change of season [10], usage of antibiotics [11] and personal diet of host [12, 13]. These pivotal factors interact with each other and build a dynamic relationship system, modifications of which would lead to imbalance of microbial community and further impact on transcriptomic, proteomic and metabolic profiles of related microorganisms. With the rapid development of high-throughput sequencing techniques as well as newly developed computational tools, accumulating evidence has demonstrated that disorders of host's microbiota would increase the incidences of various complex human diseases such as liver diseases [14], diabetes [15], asthma [16], infectious colitis [17] and even cancers [18, 19]. For example, to identify action of microorganisms in asthmatic airways, Hilty et al. [20] studied 24 adult subjects composed of 11 patients with asthma, 5 patients with chronic obstructive pulmonary disease (COPD) and 8 healthy individuals, and found that adult asthma and COPD was inextricably related to high abundance of *Proteobacteria* and *Haemophilus* as well as low abundance of *Bacteroidetes* and *Prevotella*. Mondot et al. [21] analyzed DNA sequences extracted from fecal samples which are collected from 16 Crohn's disease (CD) patients and 16 healthy subjects. As a result, they observed decrease of *Faecalibacterium prausnitzii* abundance as well as increase of *Escherichia coli* abundance in CD-patients' fecal samples compared with the controls'. In addition, Chen et al. [22] discovered a shift in composition of liver microbiota when comparing healthy and liver cirrhosis samples. In this study, liver cirrhosis was observed to be related with increase in the abundance of *Bacilli*, *Enterobacteriaceae*, *Fusobacteriaceae*, *Pasteurellaceae*, *Proteobacteria*, *Streptococcaceae* and *Veillonellaceae* as well as decrease in the abundance of *Bacteroidaceae* and *Lachnospiraceae*.

As mentioned above, identifying potential associations between microbes and diseases has a long-term theoretical and practical significance not only for better understanding of disease formation and development mechanisms but also for discovery of novel medical solutions for disease prevention, diagnosis, treatment and prognosis [23].

However, current amount and quality of known microbe-disease associations are far from satisfying the requirements of medical research. In traditional way, researchers attempt to obtain new associations between microbes and diseases by biological or clinical experiments, which demand a large quantity of time and cost. With the rapid development of computer technology, more and more computational models have been developed to predict potential miRNA-disease associations [24, 25], potential lncRNA-disease associations [26] and potential drug-target interactions [27], where machine learning-based and similarity measure-based models have shown their outstanding prediction ability. It is essential to logically extend these prediction methods into microbe-disease association prediction field. Recently, Ma et al. [23] manually collected experimentally verified microbe-disease associations from published clinical research reports and built the first Human Microbe-Disease Association Database (HMDAD). Based on the records from HMDAD, powerful computational models could be developed to prioritize candidate microbes for investigated diseases in large-scale.

In this paper, based on the assumption that microbes with similar functions tend to share similar association or non-association patterns with similar diseases, we developed the model of Network Consistency Projection for Human Microbe-Disease Association prediction (NCPHMDA) to uncover potential microbe-disease associations. By taking advantages of known microbe-disease association network and Gaussian interaction profile kernel similarity network for microbes and diseases, NCPHMDA achieved reliable prediction performance. NCPHMDA could be applied to new microbes without any known associated diseases as well as new diseases without any known associated microbes. As a non-parametric network-based prediction method, NCPHMDA demonstrated obvious advantages when the known experimentally verified microbe-disease associations are insufficient. Three validation frameworks, global leave-one-out cross validation (global LOOCV), local leave-one-out cross validation (local LOOCV) and 5-fold cross validation (5-fold CV), have been implemented to evaluate the performance of NCPHMDA. As a result, NCPHMDA achieved AUCs of 0.9093, 0.7953, and 0.8918 in global LOOCV, local LOOCV, and 5-fold CV, respectively. Moreover, colon cancer, asthma and type 2 diabetes were taken as three independent case studies, where 9, 9, 8 out of top 10 predicted microbes were successfully confirmed by recent experimental and clinical reports, respectively.

Methods

Human microbe-disease associations

Human Microbe-Disease Association Database (HMDAD, <http://www.cuilab.cn/hmdad>) [23] integrated

483 high-quality microbe-disease entries, which were mainly collected from 16S RNA sequencing-based microbial literature. After removing the duplicate association records, 450 distinct microbe-disease associations were finally obtained, including 292 microbes and 39 diseases. Adjacency matrix A was adopted to quantify the relationship between diseases and microbes, where binary element $A(i,j)$ denotes the presence or absence of association between disease $d(i)$ and microbe $m(j)$ ('0' represents absence while '1' represents presence). Furthermore, to represent the number of microbes and diseases investigated in this article, variables nm and nd are respectively defined.

Gaussian interaction profile kernel similarity for diseases

Gaussian interaction profile kernel similarity for diseases was calculated based on the assumption that diseases with similar phenotypes always share similar association and non-association pattern with functionally similar microbes. We defined binary vector $IP(d(i))$ to denote the interaction profile of disease $d(i)$, which could be obtained by observing whether $d(i)$ has known association with each microbe or not (i.e. the i th row of adjacency matrix A). Then, Gaussian interaction profile kernel similarity matrix KD could be constructed after calculation of similarity value between each disease pair.

$$KD(d(i), d(j)) = \exp(-\gamma_d \|IP(d(i)) - IP(d(j))\|^2) \tag{1}$$

$$\gamma_d = \gamma'_d / \left(\frac{1}{nd} \sum_{i=1}^{nd} \|IP(d(i))\|^2 \right) \tag{2}$$

where value of parameter γ_d controls the bandwidth of Gaussian kernel. As presented in eq. (2), γ_d could be further calculated by dividing a new bandwidth parameter γ'_d by average number of associations with microbes for all the diseases. Here, we set $\gamma'_d = 1$ according to previous studies [28].

Gaussian interaction profile kernel similarity for microbes

Adopting the same approach, Gaussian interaction profile kernel similarity between microbe $m(i)$ and $m(j)$ could be obtained as follows.

$$KM(m(i), m(j)) = \exp(-\gamma_m \|IP(m(i)) - IP(m(j))\|^2) \tag{3}$$

$$\gamma_m = \gamma'_m / \left(\frac{1}{nm} \sum_{i=1}^{nm} \|IP(m(i))\|^2 \right) \tag{4}$$

where $IP(m(i))$ represents the interaction profile of microbe $m(i)$ (i.e. the i th column of adjacency matrix A). Normalized kernel bandwidth parameter γ_m could be calculated in the similar way as γ_d , where we select $\gamma'_m = 1$ according to Van et al. [28].

NCPHMDA

As shown in Fig. 1, NCPHMDA is a network-based prediction model which measures the relevance between microbes and diseases by calculating the nodes'

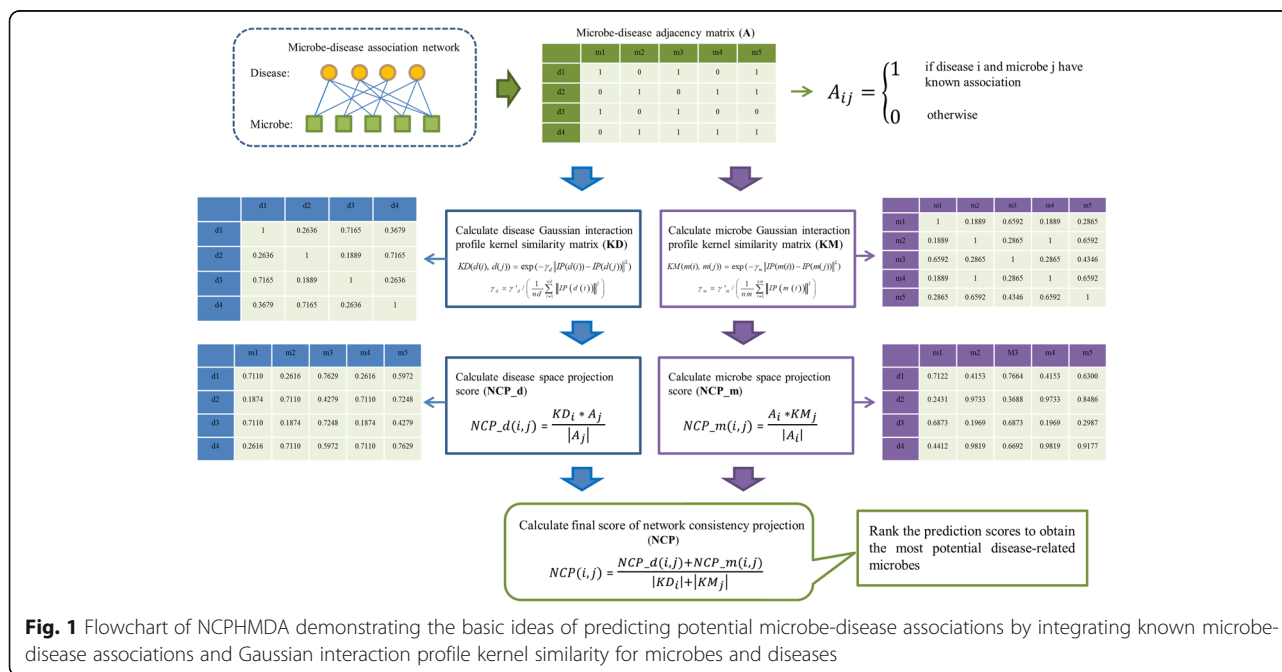


Fig. 1 Flowchart of NCPHMDA demonstrating the basic ideas of predicting potential microbe-disease associations by integrating known microbe-disease associations and Gaussian interaction profile kernel similarity for microbes and diseases

similarity in heterogeneous networks. Here, heterogeneous networks consist of microbe-disease association network constructed based on records from HMDAD [23] database, Gaussian interaction profile kernel similarity for diseases, and Gaussian interaction profile kernel similarity for microbes.

NCPHMDA first calculates two network consistency projection scores, disease space projection score and microbe space projection score, separately. The disease space projection score is calculated as follows.

$$NCP_d(i, j) = \frac{KD_i * A_j}{|A_j|} \tag{5}$$

where KD_i is the i th row of matrix KD and the vector represents the similarities between disease i and all other diseases. A_j is the j th column of matrix A and the vector represents the associations of microbe j and all diseases. $|A_j|$ is the norm of vector A_j . Matrix NCP_d is the projection score of disease Gaussian interaction profile kernel similarity network (represented as matrix KD) on the known microbe-disease association network (represented as matrix A), where the element $NCP_d(i, j)$ in row i and column j is the network projection of KD_i and A_j . Notably, the more similar diseases and disease i are, the more diseases associated with microbe j , and the smaller angle between KD_i and A_j , the greater network consistency projection score $NCP_d(i, j)$ is. The microbe space projection score could be combined and normalized in the similar way as follows.

$$NCP_m(i, j) = \frac{A_i * KM_j}{|A_i|} \tag{6}$$

where A_i is the i th row of matrix A , which consists of associations of disease i and all microbes. KM_j is the j th column of matrix KM , which comprises the similarities of microbe j and all other microbes. Matrix NCP_m is the projection score of microbe Gaussian interaction profile kernel similarity network (represented as matrix KM) on the known microbe-disease association network (represented as matrix A), where the element $NCP_m(i, j)$ in row i and column j is the network projection of KM_j and A_i . Remarkably, the more similar microbes and microbe j are, the more microbes associated with disease i , and the smaller angle between KM_j and A_i , the greater network consistency projection score $NCP_m(i, j)$ is.

Finally, we could combine and normalize NCP_d and NCP_m as follows.

$$NCP(i, j) = \frac{NCP_d(i, j) + NCP_m(i, j)}{|KD_i| + |KM_j|} \tag{7}$$

where $NCP_d(i, j)$ and $NCP_m(i, j)$ are the projection scores in disease space and microbe space of disease i and

microbe j , respectively. KD_i is the i th row of matrix KD , KM_j is the j th column of matrix KM , and $|\cdot|$ is the normalization operation. NCP is the final score matrix of network consistency projection, which measures the association probability between each microbe-disease pair.

Results and discussion

Performance evaluation

We implemented LOOCV and 5-fold CV on the experimentally verified microbe-disease associations recorded in HMDAD database to evaluate the prediction performance of NCPHMDA. In validation frameworks of LOOCV, we left out each known microbe-disease association in turn for model testing while adopted other known microbe-disease associations as training samples. According to whether all the diseases were investigated simultaneously or not, LOOCV could be further split into global LOOCV and local LOOCV. When global LOOCV was implemented, all the microbe-disease pairs without known supporting evidence in HMDAD were adopted as candidate samples, while when local LOOCV was implemented, we only took microbes without known confirmed relevance with investigated disease as candidate samples. In the framework of 5-fold CV, we randomly divided all the known microbe-disease associations into 5 average groups, 4 of which were used as training samples for model learning and the remaining one was used as testing samples for model evaluation. It needs to be emphasized that we repeated 5-fold CV for 100 times to reduce the potential deviations caused by random sample divisions. Each testing sample was ranked with all candidate samples, where the model was considered to achieve a successful prediction if the rank of the testing sample exceeds the given threshold. After setting a series of thresholds, corresponding true positive rates (TPR, sensitivity) were calculated by counting percentages of the test samples with higher ranks than investigated thresholds. Meanwhile, false positive rates (FPR, 1-specificity), which denote the percentages of the negative samples exceeding the given thresholds, were also obtained. To visualize the prediction ability, receiver-operating characteristics (ROC) curves were then drawn by plotting TPR against FPR at different thresholds. Area under ROC curve (AUC) was finally calculated as an essential performance evaluation criterion.

In this paper, we compared NCPHMDA with KATZHMDA [29], which has achieved excellent performance in potential microbe-disease association prediction. Two other previously proposed prediction methods (i.e. Regularized Least Squares [30] and Random Walk with Restart (RWR) [31]) were also applied to evaluate the prediction ability of NCPHMDA. To be clear, RWR algorithm only could predict associated microbes for given diseases and could not infer all

the missing associations for all the diseases simultaneously. Therefore, global LOOCV couldn't be implemented for RWR. In global LOOCV framework, NCPHMDA reached AUC of 0.9039 which had 0.0657, 0.3455 increase compared with KATZHMDA and Regularized Least Squares (See Fig. 2). In addition, NCPHMDA achieved AUC of 0.7953, which had 0.0977, 0.1141 and 0.1413 increase compared with Regularized Least Squares, KATZHMDA and Random Walk with Restart. Furthermore, 5-fold CV was also implemented. As a result, NCPHMDA yielded a reliable performance of 0.8918 +/- 0.0105. In conclusion, NCPHMDA has reliable performance in the framework of cross validations.

Case studies

NCPHMDA was implemented to prioritize candidate microbes of all investigated diseases in this study. For further prediction ability evaluation, three kinds of complex human diseases (i.e. colon cancer, asthma and type 2 diabetes) were taken as three independent case studies. Based on recent published clinical and biological reports, predicted microbes ranked in top 10 of these three complex diseases were validated respectively. Importantly, it should be noted that only microbe-disease pairs without known evidence collected in HMDAD database were classed into validation datasets, which guaranteed the absolute independence between validation candidates and known associations used for model training.

According to the well-known global cancer statistics report [32], colon cancer occupied the third leading cause of cancers in males and the second leading cause

of cancers in females in the past few decades. With the improved treatment and increased awareness, death rates of colon cancer patients have been decreasing in several developed countries. However, survival rates in developing countries are still far from meeting requirements because of the low detection rates in early stage. Recently, accumulating evidence have demonstrated that imbalance of microbial community has a close connection with occurrence and development of colon cancer. For example, Moore et al. [33] compared fecal floras of polyp patients (at high risk of colon cancer), Japanese-Hawaiians (at high risk), rural native Japanese (at low risk), rural native Africans (at low risk) and North American Caucasians (have a flora composition intermediate between two groups) and identified 15 colon cancer-related bacterial taxa. Surprisingly, they found that concentrations of *Bacteroides* and *Bifidobacterium* were positively related with colon cancer risks while concentrations of *Lactobacillus* and *Eubacterium aerofaciens* were negatively correlated with colon cancer risks. We implemented NCPHMDA on colon cancer for potential microbe-disease association prediction, and 9 out of the top 10 predicted microbes were successfully confirmed by biological literature (See Table 1). Typically, it is reported that colon cancer patients who have undergone preoperative insertion of a metallic stent and are aged sixty and older years are identified as risk factors for *Clostridium difficile* (1st in the prediction list) infection [34]. *Helicobacter pylori* (2nd in the prediction list) infection was found to be associated with risk increase of left-sided colorectal cancer [35]. By sequencing of 16S rRNA gene

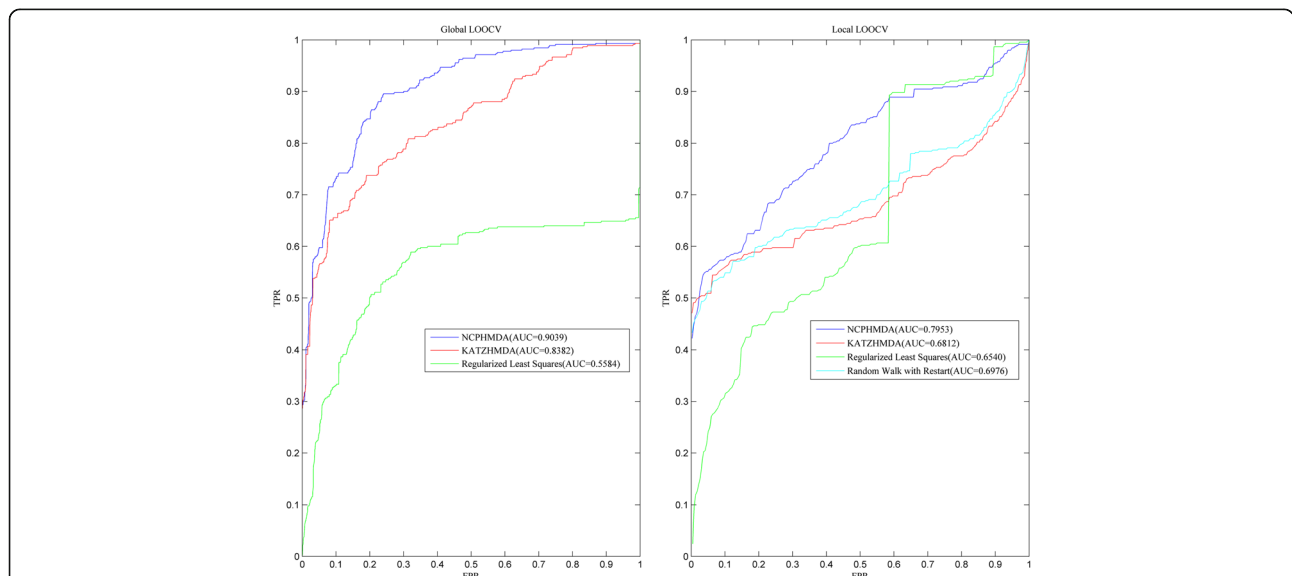


Fig. 2 Performance comparisons between NCPHMDA and three state-of-art prediction models (KATZHMDA, Regularized Least Squares and Random Walk with Reastart) in terms of ROC curve and AUC. As a result, NCPHMDA achieved AUCs of 0.9039 and 0.7953 based on global and local LOOCV, significantly outperforming previous classification models

Table 1 For further prediction performance evaluation, NCPHMDA was implemented on colon cancer to identify potential associated microbes. As a result, 9 out of the top 10 predicted microbes have been verified based on recent experimental literature

Rank	Microbe	Evidence
1	<i>Clostridium difficile</i>	PMID:21152135
2	<i>Helicobacter pylori</i>	PMID:22294430
3	<i>Proteobacteria</i>	PMID:25699023
4	<i>Prevotella</i>	PMID:25699024
5	<i>Staphylococcus aureus</i>	unconfirmed
6	<i>Clostridium coccooides</i>	PMID:18237311
7	<i>Firmicutes</i>	PMID:25699024
8	<i>Bacteroidetes</i>	PMID:25699024
9	<i>Actinobacteria</i>	PMID:26811603
10	<i>Clostridia</i>	PMID:19807912

V3 region, abundance of *Proteobacteria* (3rd in the prediction list) was discovered under-represented in sporadic colorectal carcinoma patients [36].

Asthma is a common chronic inflammatory disease of the airways of the lungs, which is generally believed to be caused by a combination of genetic and environmental factors [37]. Recent statistics indicated that incidence of asthma has been in the increasing trend in the past few decades, and the number of asthma patients grew from 183 million in 1990 to 242 million in 2013 [38]. Infection of pathogenic microorganisms (especially virus, chlamydia, mycoplasma and mold) is one of the leading causes of severe asthma. For example, Huang et al. [39] have discovered that differences in the bronchial airway microbial composition were correlated with the manifestation of clinical asthma features. They pointed out the direct link between abundance of *Sphingomonadaceae*, *Comamonadaceae*, *Oxalobacteraceae* and degree of bronchial hyperresponsiveness among asthmatic patients. By implementing NCPHMDA to prioritize candidate microbes, 9 out of the top 10 predicted microbes were successfully verified by recent clinical evidence (See Table 2). As for top 5 confirmed asthma-related microbes, concentrations of *Clostridium difficile* and *Staphylococcus aureus* (1st, 5th in the prediction list) were discovered increased in asthma patients' airways, while concentrations of *Firmicutes* and *Actinobacteria* were found decreased [40–42]. Importantly, *Clostridium coccooides* (3rd in the prediction list) subcluster XIVa species were proved serving as early indicators of possible asthma later in life, which could help prevent and diagnose asthma and provide guidance for clinical treatment [43].

According to recent disease statistic reports [44], diabetes mellitus represents 8.3% of the adult population and occupies the eighth leading cause of deaths annually. Type 2 Diabetes Mellitus (T2DM) makes up

Table 2 We implemented NCPHMDA on asthma to prioritize candidate microbes. As a result, 9 out of the top 10 predicted microbes have been confirmed based on recent experimental literature

Rank	Microbe	Evidence
1	<i>Clostridium difficile</i>	PMID:25974301
2	<i>Firmicutes</i>	PMID:23265859
3	<i>Clostridium coccooides</i>	PMID:21477358
4	<i>Actinobacteria</i>	PMID:23265859
5	<i>Staphylococcus aureus</i>	PMID:12743582
6	<i>Lactobacillus</i>	PMID:20592920
7	<i>Clostridia</i>	PMID:21477358
8	<i>Burkholderia</i>	PMID:24451910
9	<i>Lachnospiraceae</i>	PMID:17433177
10	<i>Enterococcus</i>	unconfirmed

approximate 90% of all diabetes mellitus cases and can lead to chronic complications including cardiovascular diseases, stroke and diabetic retinopathy. Increasing evidences have shown that formation and development of T2DM are closely related to low-grade inflammation and microbial infection [45]. Compositional changes in intestinal microbiota such as *Bacilli*, *Bacteroidetes*, *Betaproteobacteria*, *Clostridia*, *Clostridium*, *Firmicutes*, *Lactobacillus* and *Proteobacteria* were discovered in T2DM patient feces [46]. We took T2DM as a case study for potential T2DM-related microbe prediction, 8 out of the top 10 predicted microbes were confirmed by experimental reports (See Table 3). *Helicobacter pylori* (1st in the prediction list) infection was found to be involved in pathogenesis of insulin resistance in T2DM patients, which could be regarded as important biomarker for early detection of high blood glucose and prevention of high-risk T2DM communities [47]. Zhou

Table 3 NCPHMDA was implemented on type 2 diabetes to identify potential related microbes. As a result, 8 out of the top 10 predicted microbes have been confirmed based on recent experimental literature

Rank	Microbe	Evidence
1	<i>Helicobacter pylori</i>	PMID:24782613
2	<i>Clostridium difficile</i>	PMID:23734349
3	<i>Prevotella</i>	PMID:23613868
4	<i>Actinobacteria</i>	PMID:23613868
5	<i>Staphylococcus aureus</i>	PMID:16495627
6	<i>Lachnospiraceae</i>	unconfirmed
7	<i>Staphylococcus</i>	PMID:24385898
8	<i>Haemophilus</i>	unconfirmed
9	<i>Bacteroides</i>	PMID:20140211
10	<i>Enterobacteriaceae</i>	PMID:25759592

et al. [48] attempted to investigate the potential effect of T2DM on subgingival plaque of periodontal patients, and the results indicated that the abundance of *Prevotella* (3rd in the prediction list) was significantly different between diabetics and non-diabetics in subjects with healthy periodontium while populations of *Actinobacteria* (4rd in the prediction list) were significantly different between diabetics and their non-diabetic counterparts in subjects with periodontitis. Evidence of dysregulation of *Clostridium difficile* and *Staphylococcus aureus* (2nd and 5th in the prediction list) could be concluded from these clinical reports [49, 50].

Case studies on above three complex human diseases have confirmed the outstanding prediction ability of NCPHMDA. For further biological and clinical experiment validation, we prioritized and publicly released the prediction of all the unknown microbe-disease pairs (See Additional file 1). It is anticipated that the candidate microbe-disease pairs with higher ranks could offer valuable clues and would be confirmed by experimental observation in the near future.

Conclusions

With the rapid development of high-throughput sequencing techniques, increasing literature have demonstrated that imbalance of microbial community has critical impacts on host’s health and disease. Identifying potential microbes associated with investigated disease for better understanding of disease pathology and novel discovery of drugs has attracted more and more attention in recent years. However, few computational models have been developed for potential microbe-disease association prediction, which could significantly reduce experimental time and cost that traditional clinical researches suffer. In this study, based on the assumption that microbes with similar functions tend to share similar association or non-association patterns with similar diseases, we presented a novel computational model named NCPHMDA to prioritize candidate microbe-disease pairs for further experiment validation. NCPHMDA achieved outstanding AUCs of 0.9039, 0.7953 and average AUC of 0.8918 in global LOOCV, local LOOCV and 5-fold CV, respectively. In addition, case studies of colon cancer, asthma and type 2 diabetes mellitus were implemented for further prediction ability evaluation. As a result, 9, 9 and 8 out of the top 10 predicted microbes of these three complex diseases were confirmed by recent literature evidence. It is anticipated that NCPHMDA could serve as an important resource providing essential supports for further clinical or biological researches.

In conclusion, the following factors drove the excellent prediction performance of NCPHMDA. First of all, known microbe-disease associations collected in

HMDAD database are reliable as a basic information resource. Furthermore, Gaussian interaction profile kernel similarity for microbe and disease were integrated in NCPHMDA, which effectively improved the data completeness and further reduced model prediction bias. NCPHMDA could be implemented on new microbes without any known associated diseases as well as new diseases without any known associated microbes. In addition, NCPHMDA is a global ranking computational method and could prioritize all the candidate microbe-disease pairs for all investigated diseases in a large-scale.

It should be noted that some limitations still exist in the model design of NCPHMDA. Firstly, microbe-disease association network is sparse, which would limit the prediction accuracy of proposed model. This problem could be solved with collection of high-quality experimental microbe-disease associations in the future. Moreover, since calculation of Gaussian interaction profile kernel similarity was strongly relied on the known microbe-disease associations, the diseases with more known associated microbes are possibly predicted to be related with more potential microbes. Integrating more biological heterogeneous networks, such as disease phenotypic similarity network, disease semantic similarity network and microbe functional similarity network, could help improved the quality of existing networks and prediction performance of NCPHMDA. Establishing new similarity measures without dependence on the topological features of known microbe-disease association network is another improving direction which should never be ignored.

Additional file

Additional file 1: Table S1. Prediction of all the unknown microbe-disease pairs. (XLSX 232 kb)

Acknowledgements

The authors would like to thank all the guest editors and anonymous reviewers for their constructive comments.

Funding

The publication costs were funded by the grants of the National Science Foundation of China, Nos. 61520106006, 31571364, U1611265, 61532008, 61672203, 61402334, 61472282, 61472280, 61472173, 61572447, 61373098 and 61672382, China Postdoctoral Science Foundation Grant, Nos. 2016M601646.

Availability of data and materials

All raw data used for case studies and comparison in the present article are publicly available and can be obtained through their respective publication references. The results from study have been provided in the tables and Additional file 1.

Authors’ contributions

WB & ZJ conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript. DH designed, performed and analyzed experiments and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 18 Supplement 16, 2017: 16th International Conference on Bioinformatics (InCoB 2017): Bioinformatics. The full contents of the supplement are available online at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-18-supplement-16>.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflicts of interest regarding the publication of this paper.

Publisher’s Note

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

Published: 28 December 2017

References

1. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*. 2012;486(7402):215–21.
2. Ventura M, O’Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, van Sinderen D, O’Toole PW. Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat Rev Microbiol*. 2009;7(1):61–71.
3. Sommer F, Backhed F. The gut microbiota—masters of host development and physiology. *Nat Rev Microbiol*. 2013;11(4):227–38.
4. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207–14.
5. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JL, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. *Science (New York, NY)*. 2006;312(5778):1355–9.
6. Smith K, McCoy KD, Macpherson AJ. Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota. *Semin Immunol*. 2007;19(2):59–69.
7. Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JL. Host-bacterial mutualism in the human intestine. *Science (New York, NY)*. 2005;307(5717):1915–20.
8. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. A core gut microbiome in obese and lean twins. *Nature*. 2009;457(7228):480–4.
9. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhan R, Beaumont M, Van Treuren W, Knight R, Bell JT, et al. Human genetics shape the gut microbiome. *Cell*. 2014;159(4):789–99.
10. 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.
11. Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, Clardy J, Linington RG, Fischbach MA. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*. 2014;158(6):1402–14.
12. Walker AW, Ince J, Duncan SH, Webster LM, Holtrop G, Ze X, Brown D, Stares MD, Scott P, Bergerat A, et al. Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J*. 2011;5(2):220–30.
13. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science (New York, NY)*. 2011;334(6052):105–8.
14. Henao-Mejia J, Elinav E, Thaiss CA, Licona-Limon P, Flavell RA. Role of the intestinal microbiome in liver disease. *J Autoimmun*. 2013;46:66–73.
15. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, Hu C, Wong FS, Szot GL, Bluestone JA, et al. Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature*. 2008;455(7216):1109–13.
16. Rivas MN, Crother TR, Arditi M. The microbiome in asthma. *Curr Opin Pediatr*. 2016;

17. Sokol H, Seksik P, Rigottier-Gois L, Lay C, Lepage P, Podglajen I, Marteau P, Dore J. Specificities of the fecal microbiota in inflammatory bowel disease. *Inflamm Bowel Dis*. 2006;12(2):106–11.
18. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*. 2012;22(2):299–306.
19. Schwabe RF, Jobin C. The microbiome and cancer. *Nat Rev Cancer*. 2013;13(11):800–12.
20. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A, Poulter L, Pachter L, et al. Disordered microbial communities in asthmatic airways. *PLoS One*. 2010;5(1):e8578.
21. Mondot S, Kang S, Furet JP, Aguirre de Carcer D, McSweeney C, Morrison M, Marteau P, Dore J, Leclerc M. Highlighting new phylogenetic specificities of Crohn’s disease microbiota. *Inflamm Bowel Dis*. 2011;17(1):185–92.
22. Chen Y, Yang F, Lu H, Wang B, Chen Y, Lei D, Wang Y, Zhu B, Li L. Characterization of fecal microbial communities in patients with liver cirrhosis. *Hepatology (Baltimore, Md)*. 2011;54(2):562–72.
23. Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, Yang J, Kong W, Zhou X, Cui Q. An analysis of human microbe-disease associations. *Brief Bioinform*. 2016; bbw005.
24. Chen H, Zhang Z. Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med Genet*. 2013;6(1):12.
25. Gu C, Liao B, Li X, Li K. Network consistency projection for human miRNA-disease associations inference. *Sci Rep*. 2016;6:36054.
26. Yang X, Gao L, Guo X, Shi X, Wu H, Song F, Wang B. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS One*. 2014;9(1):e87797.
27. Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics (Oxford, England)*. 2010;26(12):i246–54.
28. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics (Oxford, England)*. 2011;27(21):3036–43.
29. 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 (Oxford, England)*. 2016;
30. Rifkin R, Klautau A. In defense of one-vs-all classification. *J Mach Learn Res* 2004, 5(Jan):101-141.
31. Guo W, Shang DM, Cao JH, Feng K, He YC. Identifying and analyzing novel epilepsy-related genes using random walk with restart algorithm. 2017, 2017:6132436.
32. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin*. 2011;61(2):69–90.
33. Moore WE, Moore LH. Intestinal floras of populations that have a high risk of colon cancer. *Appl Environ Microbiol*. 1995;61(9):3202–7.
34. Yeom CH, Cho MM, Baek SK, Bae OS. Risk factors for the development of Clostridium Difficile-associated colitis after colorectal cancer surgery. *J Korean Soc Coloproctol*. 2010;26(5):329–33.
35. Zhang Y, Hoffmeister M, Weck MN, Chang-Claude J, Brenner H. Helicobacter pylori infection and colorectal cancer risk: evidence from a large population-based case-control study in Germany. *Am J Epidemiol*. 2012;175(5):441–50.
36. Gao Z, Guo B, Gao R, Zhu Q, Qin H. Microbiota disbiosis is associated with colorectal cancer. *Front Microbiol*. 2015;6:20.
37. Martinez FD. Genes, environments, development and asthma: a reappraisal. *Eur Respir J*. 2007;29(1):179–84.
38. Vos T, Barber RM, Bell B, Bertozzi-Villa A, Biryukov S, Bolliger I, Charlson F, Davis A, Degenhardt L, Dicker D. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *Lancet*. 2015;386(9995):743–800.
39. Huang YJ, Nelson CE, Brodie EL, Desantis TZ, Baek MS, Liu J, Woyke T, Allgaier M, Bristow J, Wiener-Kronish JP, et al. Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J Allergy Clin Immunol*. 2011;127(2):372–81. **e1-3**
40. Fujimura KE, Lynch SV. Microbiota in allergy and asthma and the emerging relationship with the gut microbiome. *Cell Host Microbe*. 2015;17(5):592–602.
41. Marri PR, Stern DA, Wright AL, Billheimer D, Martinez FD. Asthma-associated differences in microbial composition of induced sputum. *J Allergy Clin Immunol*. 2013;131(2):346–52. **e1-3**

42. Bachert C, Gevaert P, Howarth P, Holtappels G, Van Cauwenberge P, Johansson S. IgE to Staphylococcus Aureus enterotoxins in serum is related to severity of asthma. *J Allergy Clin Immunol.* 2003;111(5):1131–2.
43. Vael C, Vanheirstraeten L, Desager KN, Goossens H. Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol.* 2011;11:68.
44. Atlas D: International diabetes federation. Press Release, Cape Town, South Africa 2006, 4.
45. Furet JP, Kong LC, Tap J, Poitou C, Basdevant A, Bouillot JL, Mariat D, Corthier G, Dore J, Henegar C, et al. Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes.* 2010;59(12):3049–57.
46. Larsen N, Vogensen FK, van den Berg FW, Nielsen DS, Andreasen AS, Pedersen BK, Al-Soud WA, Sorensen SJ, Hansen LH, Jakobsen M. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One.* 2010;5(2):e9085.
47. He C, Yang Z, NH L. Helicobacter pylori infection and diabetes: is it a myth or fact? *World J Gastroenterol.* 2014;20(16):4607–17.
48. Zhou M, Rong R, Munro D, Zhu C, Gao X, Zhang Q, Dong Q. Investigation of the effect of type 2 diabetes mellitus on subgingival plaque microbiota by high-throughput 16S rDNA pyrosequencing. *PLoS One.* 2013;8(4):e61516.
49. Hassan SA, Rahman RA, Huda N, Wan Bebakar WM, Lee YY. Hospital-acquired Clostridium Difficile infection among patients with type 2 diabetes mellitus in acute medical wards. *J R Coll Physicians Edinb.* 2013;43(2):103–7.
50. Tamer A, Karabay O, Ekerbicer H. Staphylococcus Aureus nasal carriage and associated factors in type 2 diabetic patients. *Jpn J Infect Dis.* 2006;59(1):10–4.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

