Shi et al. BMC Bioinformatics 2018, 19(Suppl 9):281
https://doi.org/10.1186/s12859-018-2274-3

BMC Bioinformatics

RESEARCH Open Access

BMCMDA: a novel model for predicting human microbe-disease associations via binary matrix completion

Jian-Yu Shi[1*], Hua Huang[2], Yan-Ning Zhang[3], Jiang-Bo Cao[1] and Siu-Ming Yiu[4]

From 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017)
Honolulu, Hawaii, USA. 30 May - 2 June 2017

Abstract

Background: Human Microbiome Project reveals the significant mutualistic influence between human body and microbes living in it. Such an influence lead to an interesting phenomenon that many noninfectious diseases are closely associated with diverse microbes. However, the identification of microbe-noninfectious disease associations (MDAs) is still a challenging task, because of both the high cost and the limitation of microbe cultivation. Thus, there is a need to develop fast approaches to screen potential MDAs. The growing number of validated MDAs enables us to meet the demand in a new insight. Computational approaches, especially machine learning, are promising to predict MDA candidates rapidly among a large number of microbe-disease pairs with the advantage of no limitation on microbe cultivation. Nevertheless, a few computational efforts at predicting MDAs are made so far.

Results: In this paper, grouping a set of MDAs into a binary MDA matrix, we propose a novel predictive approach (BMCMDA) based on Binary Matrix Completion to predict potential MDAs. The proposed BMCMDA assumes that the incomplete observed MDA matrix is the summation of a latent parameterizing matrix and a noising matrix. It also assumes that the independently occurring subscripts of observed entries in the MDA matrix follows a binomial model. Adopting a standard mean-zero Gaussian distribution for the nosing matrix, we model the relationship between the parameterizing matrix and the MDA matrix under the observed microbe-disease pairs as a probit regression. With the recovered parameterizing matrix, BMCMDA deduces how likely a microbe would be associated with a particular disease. In the experiment under leave-one-out cross-validation, it exhibits the inspiring performance (AUC = 0.906, AUPR =0.526) and demonstrates its superiority by ~ 7% and ~ 5% improvements in terms of AUC and AUPR respectively in the comparison with the pioneering approach KATZHMDA.

Conclusions: Our BMCMDA provides an effective approach for predicting MDAs and can be also extended to other similar predicting tasks of binary relationship (e.g. protein-protein interaction, drug-target interaction).

Keywords: Microbe-disease association, Matrix completion, Prediction, Machine learning

* Correspondence: jianyushi@nwpu.edu.cn
[1]School of Life Sciences, Northwestern Polytechnical University, Xi'an 70072, China
Full list of author information is available at the end of the article

Background

Human intestine provides a nutrient-rich and temperature-constant habitat for microbes, such that the microbes have a mutualistic association with their host [1]. Diverse communities of microbes, especially bacteria, are found by sequencing techniques (e.g. 16S ribosomal RNA sequencing) in human bodies [2]. It is surprising that the number of genes in human microbiome is up to 5 million [3]. Both these genes and their products are participating in a diverse range of biological activities, such as metabolic capabilities, pathogens, immune system, and gastrointestinal development [4]. It can be said that they somehow serve as a physiological complement in the human body. Meanwhile, both communities and populations of microbes can be significantly influenced by their dynamic habitat in the human body. Diverse environmental variables, such as season [5], host diet [6], smoking [7], hygiene [3] and use of antibiotics [8], may change the habitat of microbes frequently. This kind of mutualistic associations between human host and its microbiota would cause the modifications of transcriptomic, proteomic and metabolic profiles in the human body. However, some of the modifications could be harmful.

Beyond the fact that microbe is the main player in the pathogenic mechanism of infectious diseases, an increasing number of clinical studies have demonstrated that the microbiota in human body is strongly associated with a wide range of human non-infectious diseases, such as cancer [9], obesity [10, 11], diabetes [12, 13], kidney stones [14] and systemic inflammatory response syndrome [15]. Nevertheless, people have only a limited understanding of what microbes cause the diseases and how they do.

Fortunately, the increasing number of experimentally validated associations between human non-infectious diseases and microbes enable us to perform a systematic analysis on microbe-disease associations (MDAs). For example, Ma et al. recently published the first database of MDA, Human Microbe-Disease Association Database (HMDAD), by collecting a large number of MDAs from previously published literature [16]. The MDA entries in HMDAD mainly focuses on experimentally supported associations between diverse microbes and non-infectious diseases, and all of them are experimentally supported with sufficient samples. The systematic analysis on a large scale of MDAs provides a new insight to discover the mechanism of microbe-related non-infectious diseases [17]. As one of the most important steps towards that goal, the identification of MDA is helpful to understand how non-infectious diseases develop and exploit novel methods for disease diagnosis and therapy. However, traditional experiment-based approaches for discovering MDAs are time-consuming and costly. Even worse, many bacteria cannot be cultivated at all by current culturing bio-techniques [18].

As the complement of biological experiment-based approaches, computational approaches are promising to rapidly screen MDA candidates, such that the further biological validation reduces the cost and time significantly. More importantly, they are expected to output the MDA candidates involving uncultivable microbes. A few efforts have been made to develop computational models for the large-scale MDA prediction. Recently, a pioneering work developed an approach, KATZHMDA, for predicting potential MDAs on a large scale [19]. After constructing an MDA network based on HMDAD, KATZHMDA models MDA prediction as link prediction on the network.

In this work, by modeling MDA prediction as a problem of matrix completion (Fig. 1), we propose a new predictive approach based on Binary Matrix Completion (BMCMDA) to predict potential MDAs on a large scale by only using a set of approved microbe-disease associations. The following sections are organized as follows. Section Method first introduces the basic idea to model MDA prediction, then represents the algorithm of binary matrix completion. Section Experiments briefly describes the benchmark dataset of MDA, shows how to tune the parameters in the proposed model, and demonstrates the ability of BMCMDA by the comparison with other state-of-the-art approaches. The final section draws our conclusion. In addition, human non-infectious diseases are termed as 'diseases' and their microbes in the body are termed as 'microbes' in the following texts for concision.

Methods

Problem formulation

Given p kinds of microbes $M = \{m_i\}$, q types of diseases $D = \{d_j\}$, and a set of associations between them, we aim to deduce or predict new potential associations among them. Those microbe-disease associations can be organized into a $p \times q$ binary adjacent matrix $\mathbf{A} = \{a_{ij}\}$, where $a_{ij} = +1$ and $a_{ij} = -1$ account for whether m_i is associated with d_j or not respectively, and $a_{ij} = ?$ if the association between m_i and d_j is NOT observed. Our problem is to deduce how likely those unobserved entries are MDAs (Fig. 1).

Matrix completion is one of the popular techniques to deduce the relationship between two types of objects (i.e. users and items) in recommendation system. However, the standard algorithms of matrix completion working on real-valued or categorical observations fail to infer the binary relationship between the objects [20], such as MDA prediction. Therefore, we adopted a different technique in the next section.

Binary matrix completion

We state the problem as a matrix completion with 1-bit observation, in which each observed entry represents a positive (yes) or negative (no) response to MDA. Such a binary matrix completion can be defined as a generalized linear model,

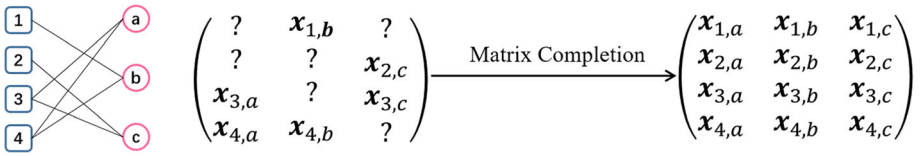

Fig. 1 A Toy MDA Example for Matrix Completion. The left matrix is an observed matrix, in which x_{ij} are the observed MDA entries and '?'s denote the unobserved microbe-disease pairs. The right matrix is the expected matrix with fully observed entries

$$a_{ij} = \begin{cases} +1 & x_{ij} + z_{ij} \geq 0 \\ -1 & x_{ij} + z_{ij} < 0 \end{cases} \tag{1}$$

where only a subset Ω of entries of **A** is observed, $\mathbf{X} = \{x_{ij}\}$ is a low-rank parameterizing distribution matrix of **A**, and $\mathbf{Z} = \{z_{ij}\}$ is a stochastic matrix containing noise. The recovery of matrix **X** is usually transformed to another form to solve as follows [21].

Given an incomplete observed MDA matrix $\mathbf{A} \in \mathbb{R}^{p \times q}$, a subset of its observed entry subscripts $\Omega \subset [p] \times [q]$ and a differentiable function $f: \mathbb{R} \to [0, 1]$, we observe

$$a_{ij} = \begin{cases} +1 & \text{with the probability } f(x_{ij}) \\ -1 & \text{with the probability } 1 - f(x_{ij}) \end{cases} \text{ for } \forall (i,j) \in \Omega \tag{2}$$

where [d] denotes the set of integers {1,..,d}. In other words, the entries of **A** depend on a $p \times q$ underlying low-rank preference matrix $\mathbf{X} = \{x_{ij}\} \in \mathbb{R}^{p \times q}$ somehow (Fig. 2).

We assume that the subscript subset Ω follows a binomial model, in which the subscript $(i, j) \in [p] \times [q]$ of each observed entry in A occurs with probability m/(pq) independently, where m is the cardinality (the number of observed entries) of Ω. The assumption reflects p×q independent experiments, of which each determines microbe-disease associations with m/(pq) success probability.

In addition, if we suppose that the entries of the underlying noising matrix **Z** are independently and identically drawn from the distribution, whose cumulative distribution function (CDF) is given by $F_Z(x) = P(z \leq x) = 1 - f(-x)$, then the model in Formula (2) reduces to its special case in Formula (1). In such a sense, the selection of CDF f is equivalent to that of Z. Thus, X can be also viewed as a parameter of a distribution.

Since our aim is to determine the likelihood that a microbe would be associated with a particular disease, we naturally model MDA prediction as the problem that recovers the latent low-rank matrix **X**.

When defining the CDF $f(x_{ij}) = 1 - \Phi(-x_{ij}/\sigma) = \Phi(x_{ij}/\sigma)$, where Φ is the cumulative distribution function of a standard Gaussian (a standard mean-zero Gaussian with variance σ^2 for the noising matrix **Z**), Formula (2) captures a probit regression model. Thus, the recovery of **X** can be achieved by solving the following optimization problem [21],

$$\begin{aligned} &\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} \\ &F_{\Omega,\mathbf{A}}(\mathbf{X}) F_{\Omega,\mathbf{A}}(\mathbf{X}) = \sum_{(i,j)\in\Omega} \left(B(a_{ij} = +1) \right. \\ &\left. \log\left(f(x_{ij})\right) + B(a_{ij} = -1) \log\left(1 - f(x_{ij})\right) \right) \\ &f(x_{ij}) = 1 - \Phi(-x_{ij}/\sigma) = \Phi(x_{ij}/\sigma) \\ &s.t. \quad \|\mathbf{X}\|_* \leq \sqrt{rpq} \end{aligned} \tag{3}$$

where $B(\varepsilon)$ is the binary indicator function for an event ε(i.e. $B(\varepsilon) = 1$ if ε occurs and 0 otherwise), $\Phi(x_{ij}/\sigma) \in \mathbb{R} \to [0, 1]$ is the cumulative distribution function of a standard Gaussian distribution with variance σ^2, and r is the expected rank of **X**.

Consider that Formula (3) is just a special instance of the general formulation

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \boldsymbol{C} \tag{4}$$

where $f(\mathbf{x})$ is a smooth convex function from $\mathbb{R}^n \to \mathbb{R}$, and $\boldsymbol{C}$ is a closed convex set in $\mathbb{R}^n$. In particular, defining $\boldsymbol{V}$ as the bijective linear mapping that vectorizes $\mathbb{R}^{p \times q}$ to $\mathbb{R}^{pq}$, we

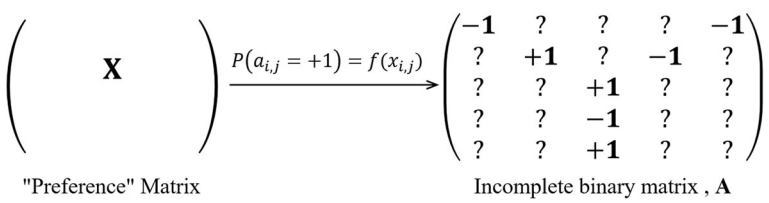

Fig. 2 Binary matrix completion. The left matrix is the latent preference matrix. The right matrix is the observed matrix, in which the observed entries are labelled with '+ 1' if an MDA is found, with '-1' if a non-MDA is found, and '?' if the entry is not observed

have $f(\mathbf{x}) = -F_{\Omega,\ \mathbf{A}}(\boldsymbol{V}^{-1}\mathbf{x})$ and $\boldsymbol{C} = \boldsymbol{V}(\{\mathbf{X} : \|\mathbf{X}\|_* \le \tau\})$. Therefore, non-monotone Spectral Projected Gradient (SPG) can be applied to solve the above optimization [22]. It is an iterative algorithm, which requires at each iteration the evaluation of $f(\mathbf{x})$, its gradient $g(\mathbf{x}) = \nabla f(\mathbf{x})$ and an orthogonal projection $\boldsymbol{P}_{\boldsymbol{C}}(\mathbf{v})$ onto $\boldsymbol{C}$, $\boldsymbol{P}_{\boldsymbol{C}}(\mathbf{v}) = \arg\min \|\mathbf{x} - \mathbf{v}\|_2$ subject to $\mathbf{x} \in \boldsymbol{C}$. Since the orthogonal projection onto the nuclear-norm ball $\boldsymbol{C}$ amounts to singular-value soft thresholding [23], the projection is equivalent to

$$\boldsymbol{P}_{\boldsymbol{C}}(\mathbf{X}) = \boldsymbol{S}_{\lambda}(\mathbf{X}) := \mathbf{U}\ \max\{\boldsymbol{\Sigma} - \lambda\mathbf{I}, 0\}\mathbf{V}^T \tag{5}$$

where $\mathbf{X} \overset{SVD}{=} \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, $\boldsymbol{\Sigma} = diag(\sigma_1, \ldots, \sigma_n)$, the maximum operation is taken entry-wise and $\lambda \ge 0$ is the smallest value for which $\sum_{i=1}^{n} \max\{\sigma_i - \lambda\} \le \tau$.

Cross validation

As a standard technique, cross-validation (CV) is popularly adopted to evaluate the performance of machine learning models and estimate their power of generalization on future samples. Usually, there are two kinds of CV, k-fold cross-validation (k-CV) and leave-one-out cross-validation (LOOCV).

In the scheme of k-CV, all the observed samples are randomly split into k subsets of approximately equal size. Among them, one subset is taken as the testing set, in which the samples are masked as unobserved. Meanwhile, the remaining k-1 subsets are merged as the training set, in which the observed samples are used to train a predicting model. Once the training is done, the predicting model is performed on the testing set and outputs the confidence scores of being observed samples for all the masked samples. This procedure repeats k times by taking each subset as the testing set in turn. In each round of k-CV, the performance of the predicting model is measured and recorded. Its final performance is defined as the average of the performance in all the rounds.

LOOCV can be regarded as an extreme case of k-CV, where k is equal to the number of observed samples. In each step of LOOCV, each observed sample is blinded as an unobserved one and the remaining observed samples are used to build the predicting model. The procedure of LOOCV takes each of the observed samples as the testing sample in turn. When the number of samples is enough large, the results of k-CV and LOOCV have no significant difference in statistics.

The performance of MDA prediction is measured by Receiver Operating Characteristic (ROC) curve as well as Precision-Recall (PR) curve. Two measuring metrics adopted are both the Area Under ROC curve (AUC) and the Area Under PR curve (AUPR). One could easily obtain other metrics, such as true positive rate (TPR, Recall, or Sensitivity) and false positive rate (FPR, 1-Specificity), by setting thresholds on ROC or PR curves.

Results and discussion

Dataset

We adopted the same dataset of MDAs as that in [19]. The dataset was originally collected from the Human Microbe-Disease Association Database (HMDAD, http://www.cuilab.cn/hmdad), which was built in 2016 and published in 2017 [16]. HMDAD collected MDA entries from 61 publications in microbiome studies based on 16s RNA sequencing. Each entry is an experimentally supported association between diverse microbes and non-infectious diseases with sufficient samples. HMDAD provides a benchmark source for developing prediction model [19].

Originally, there are 483 MDAs, including 292 microbes and 39 human diseases in the dataset. After removing the duplicate MDAs, which come from different experiments, Chen et al. [19] give 450 distinct MDAs among those microbes and diseases, and organizes them into a 292×39 association matrix. The corresponding MDA network is shown in Fig. 3.

Parameter tuning

In this section, we investigated the influence of two important parameters in Formula 2, the standard derivation σ and the estimated rank r. First, we tuned it from the list {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. Since the maximum rank r_{max} of the underlying matrix is equal to $\min(p, q)$, we then tuned r from the ratio list of $\{\frac{1}{10}, \frac{1}{9}, \frac{1}{8}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1\}$ w.r.t r_{max} and searched the best values on the 10 × 10 grid expanded by both σ and r.

Considering that AUPR is a better metric than AUC when the number of positive samples is significantly less than that of negative samples [24], we recorded the performance of BMCMDA for each pairwise value of (σ, r) under 5-CV in terms of AUPR (Fig. 4). When running BMCMDA, all the parameters (e.g. the number of iterations and the tolerance of stopping iteration) in SPG were set to their default values.

Finally, we picked up the pair of $(\sigma^*, r^*) = (0.2, \frac{1}{3}r_{max})$, which achieves the highest one among 100 values of AUPR, as the best value of (σ, r), and further applied them in all the following experiments.

Comparison with the state-of-the-art approach

With the best pair (σ^*, r^*), we compared BMCMDA with three approaches, including one baseline approach and two state-of-the-art approaches, RKNNMDA [25] and KATZHMDA [19]. The

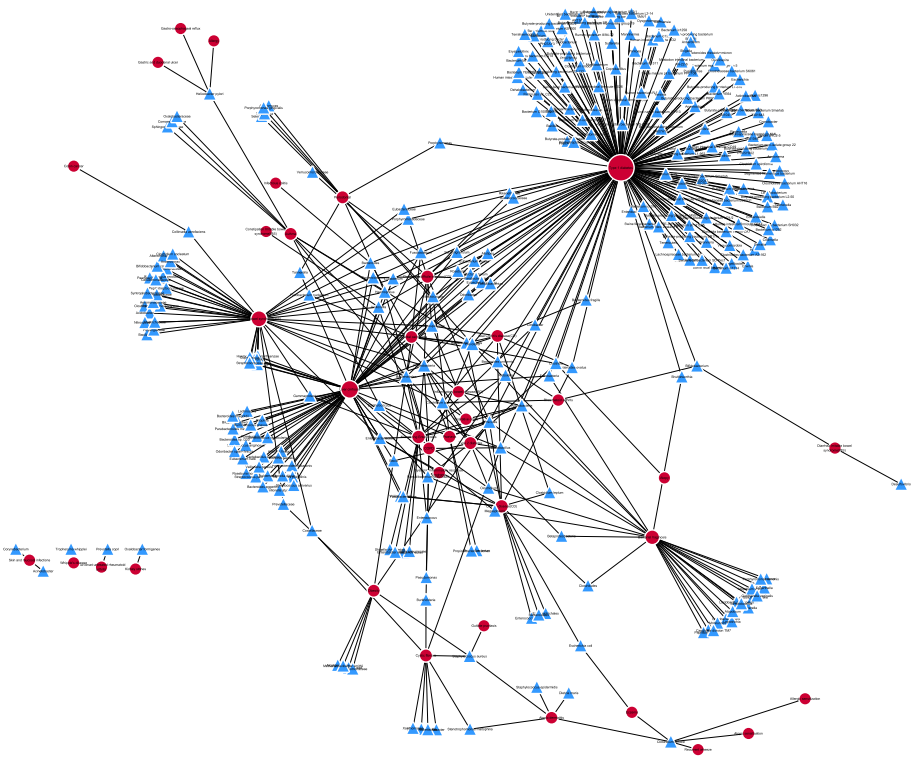

Fig. 3 The Network of Microbe-Disease Associations. Blue triangles and red circles denote microbes and diseases respectively. Lines between nodes are the associations between them. The minimum, the median, the mean, and the maximum of microbe degrees are 1, 1, 1.54 and 11, while those of disease degrees are 1, 3, 11.54 and 167 respectively

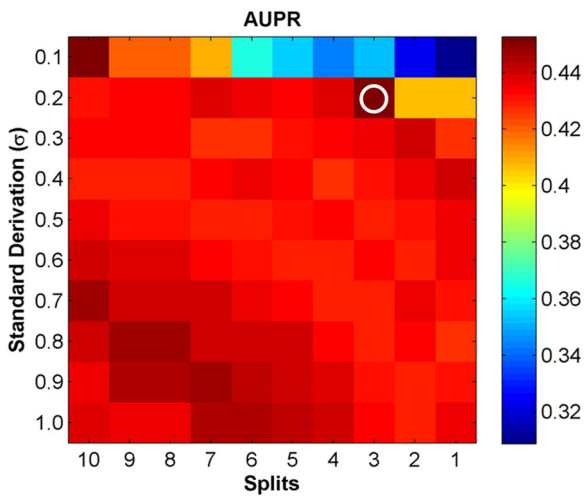

Fig. 4 Illustration of Determining the Best Value Pair of (σ, r). The position w.r.t (σ^*, r^*) is highlighted by a white circle

baseline approach directly applies singular value decomposition (SVD) on the MDA adjacency matrix with missing entries and uses the product of two unitary matrices and the rectangle diagonal matrix to recover the missing values. RKNNMDA was originally designed for miRNA-disease associations [25]. It performs MDA prediction by directly applying a ranking-based KNN on the MDA prediction [19]. KATZHMDA also constructs a heterogeneous network, which consists of the known MDA network and two MDA-induced networks [19]. The first MDA-induced network indicates a microbe similarity network, while the second one accounts for a disease similarity network. Both of them are derived from the MDA network by Gaussian interaction profile kernel. By leveraging KATZ index to calculate similarities between microbe nodes and disease nodes in the heterogeneous network, KATZHMDA infers the potential association between a microbe node and a disease node if the value of their KATZ index is large. The comparison was performed with the exactly same dataset under LOOCV as mentioned in [19]. The results in Fig. 5. show that BMCMDA wins the best and outperforms those approaches significantly.

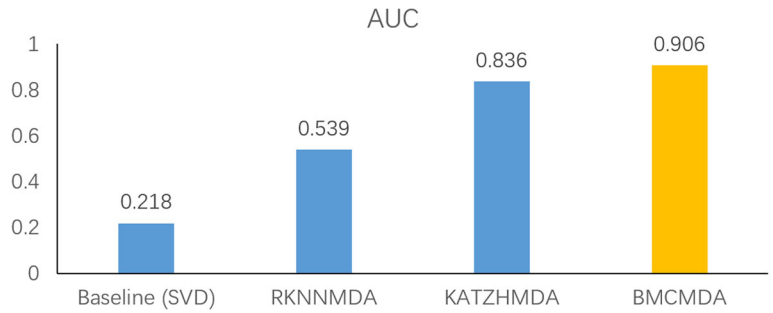

Fig. 5 Comparison with state-of-the-art approaches

Furthermore, we selected the second best approach KATZHMDA to make a detailed comparison. Considering the fact that AUPR is a better metric than AUC when the number of positive samples is significantly less than that of negative samples [24], we measured the prediction by not only ROC curves but also PR curves. The results illustrated in Fig. 6 show that BMCMDA, compared with KATZHMDA, achieves a significant improvement of both ~ 7% increment in terms of AUC and ~ 5% increment in terms of AUPR.

Conclusions

As the complement of biological experiments, computational methods have a potential to be a promising approach, which predicts MDA candidates rapidly among a plenty of microbe-disease pairs with the advantage of no limitation on microbe cultivation.

In this paper, we have modeled MDA prediction in a novel sight, which utilizes an underlying real-valued matrix to reflect the magnitude of MDAs and regards the binary MDA adjacent matrix as its incomplete and noisy observation. Upon this model, we have proposed a new approach based on Binary Matrix Completion (BMCMDA) to predict potential MDAs among a large scale of microbe-disease pairs. The comparison with other state-of-the-art approaches demonstrates the superiority of BMCMDA for predicting microbe-disease associations on a large scale and also validates that the assumption we adopted is reasonable. Obviously, BMCMDA can be directly applied to other similar forms of problems in bioinformatics, including the inference of the binary relationship between mono-partite objects (e.g. protein-protein interaction, drug-drug interaction [26, 27] and drug combination [28]) or that between bi-partite objects (e.g. drug-target interaction [29, 30], gene-disease association, RNA-disease association [31]).

In addition, we consider the possible improvement of BMCMDA. First, we may enhance the MDA prediction by integrating additional and independent microbe/disease similarities or features with BMCMDA. Secondly, as suggested in [31], we may generalize BMCMDA to be appropriate in more predicting scenarios, including the prediction of the associations between newly-found microbes (having no known MDA) and existing diseases, the prediction of the associations between existing microbes and newly-concerned diseases

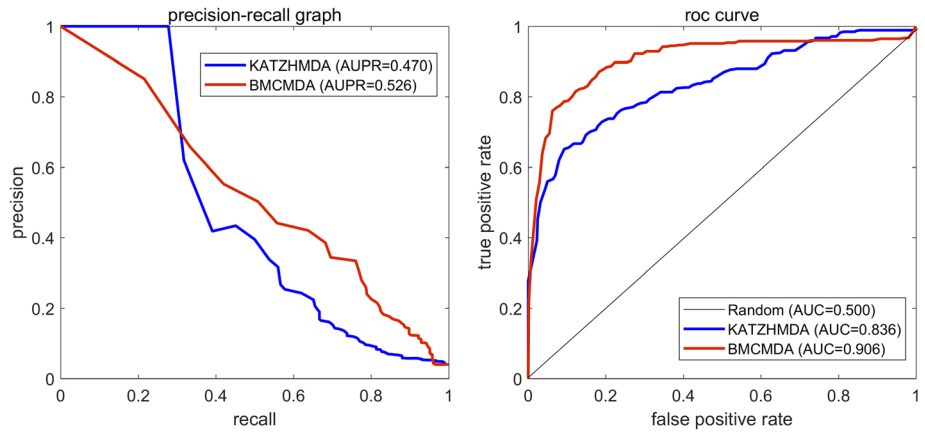

Fig. 6 Comparison between BMCMDA and KATZHMDA in terms of ROC curve and PR curve

(having no known MDA), and the prediction of the associations between newly-found microbes and newly-concerned diseases.

Abbreviations

AUC: The area under the receiver operating characteristic curve; AUPR: The area under precision-recall curve; BMCMDA: Binary Matrix Completion for predicting human Microbe-Disease Associations; CDF: Cumulative distribution function; CV: Cross-validation; HMDAD: Human Microbe-Disease Association Database; LOOCV: Leave-one-out cross-validation; MDA: Microbe-disease association; SPG: Spectral Projected Gradient

Acknowledgements

The abridged 2-page abstract of this work was previously published in the Proceedings of the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017), Lecture Notes in Computer Science: Bioinformatics Research and Applications [32].

Funding

This work was supported by RGC Collaborative Research Fund (CRF) of Hong Kong (C1008-16G), National High Technology Research and Development Program of China (No. 2015AA016008), the Fundamental Research Funds for the Central Universities of China (No. 3102015ZY081), the Program of Peak Experience of NWPU (2016), and China National Training Programs of Innovation and Entrepreneurship for Undergraduates (No. 201710699330). The publication charge was funded by China National Training Programs of Innovation and Entrepreneurship for Undergraduates (No. 201710699330).

Availability of data and materials

The dataset of MDA used in this work can be download from https://github.com/JustinShi2016/ISBRA2017

About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 9, 2018: Selected articles from the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017): bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-9.

Authors' contributions

JYS and YNZ conceived, designed and carried out the experiments. JYS and SMY drafted the manuscript. HH collected the heterogeneous data. JYS performed the experiments. JBC answers the final round of textual comments. JYS and SMY analysed the data. JYS and HH developed the codes used in the analysis. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Life Sciences, Northwestern Polytechnical University, Xi'an 70072, China. [2]School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an 70072, China. [3]School of Computer Science, Northwestern Polytechnical University, Xi'an 70072, China. [4]Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China.

Published: 13 August 2018

References

1. Hsiao EY, Mcbride SW, Hsien S, Sharon G, Hyde ER, Mccue T, Codelli JA, Chow J, Reisman SE, Petrosino JF. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 2013;155:1451–63.
2. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla A, Creasy HH, Earl AM, Fitzgerald M, Fulton RS. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
3. Sommer F, Bäckhed F. The gut microbiota–masters of host development and physiology. Nat Rev Microbiol. 2013;11:227–38.
4. Ventura M, O'Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, Van SD, O'Toole PW. Genome-scale analyses of health-promoting bacteria: probiogenomics. Nat Rev Microbiol. 2009;7:61–72.
5. Davenport ER, Mizrahiman O, Michelini K, Barreiro LB, Ober C, Gilad Y. Seasonal variation in human gut microbiome composition. PLoS One. 2014;9:e90731.
6. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505:559–63.
7. 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;9:268–72.
8. Donia MS, Cimermancic P, Schulze CJ, Brown LCW, 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;158:1402–14.
9. Moore WE, Moore LVH. Intestinal floras of populations that have a high risk of colon cancer. Appl Environ Microbiol. 1995;61:3202–7.
10. Ley RE, Backhed F, Turnbaugh PJ, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A. 2005;102:11070–5.
11. Zhang H, Dibaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, Parameswaran P, Crowell MD, Wing RA, Rittmann BE. Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci U S A. 2009;106:2365–70.
12. Brown CT, Davisrichardson AG, Giongo A, Gano KA, Crabb DB, Mukherjee N, Casella G, Drew JC, Ilonen J, Knip M. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. PLoS One. 2011;6:e25792.
13. Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G, Drew JC, Ilonen J, Knip M, Hyoty H. Toward defining the autoimmune microbiome for type 1 diabetes. ISME J. 2011;5:82–91.
14. Hoppe B, Groothoff JW, Hulton S, Cochat P, Niaudet P, Kemper MJ, Deschenes G, Unwin RJ, Milliner DS. Efficacy and safety of Oxalobacter formigenes to reduce urinary oxalate in primary hyperoxaluria. Nephrol Dial Transplant. 2011;26:3609–15.
15. Mshvildadze M, Neu J, Shuster JJ, Theriaque DW, Li N, Mai V. Intestinal microbial ecology in premature infants assessed with non-culture-based techniques. J Pediatr. 2010;156:20–5.
16. 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;
17. Nathan C. Fresh approaches to anti-infective therapies. Sci Transl Med. 2012;4:140sr2.
18. Stewart EJ. Growing Unculturable Bacteria. J Bacteriol. 2012;194:4151–60.
19. 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.
20. Lin Z, Liu R, Su Z. Linearized alternating direction method with adaptive penalty for low-rank representation. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, editors. Advances in Neural Information Processing System. Granada: Curran Associates; 2011. p. 612–20.
21. Davenport MA, Plan Y, van den Berg E, Wootters M. 1-bit matrix completion. Information and Inference: A Journal of the IMA. 2014;3:189–223.
22. Birgin EG, Martinez JM, Raydan M. Nonmonotone spectral projected gradient methods on convex sets. SIAM J Optim. 1999;10:1196–211.
23. Cai J, Candes EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20:1956–82.
24. Jiao Y, Du P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quantitative. Biology. 2016;4:320–30.

25. Chen X, Wu Q-F, Yan G-YRKNNMDA. Ranking-based KNN for MiRNA-disease association prediction. RNA Biol. 2017;14:952–62.
26. Yu H, Mao K-T, Shi J-Y, Huang H, Chen Z, Dong K, Yiu S-M. Predicting and understanding comprehensive drug-drug interactions via semi-nonnegative matrix factorization. In: The sixteenth Asia Pacific bioinformatics conference Yokohama, Japan; 2018.
27. Shi J-Y, Huang H, Li J-X, Lei P, Zhang Y-N, Yiu S-M. Predicting comprehensive drug-drug interactions for new drugs via triple matrix factorization. In: IWBBIO: 2017; Spain. Lecture notes in computer science: bioinformatics and biomedical engineering. Granada: Springer; 2017. p. 108–17.
28. Shi J-Y, Li J-X, Gao K, Lei P, Yiu S-M. Predicting combinative drug pairs towards realistic screening via integrating heterogeneous features. BMC Bioinformatics. 2017;18(12):409.
29. Shi J-Y, Li J-X, Lu H-M. Predicting existing targets for new drugs base on strategies for missing interactions. BMC Bioinformatics. 2016;17(8):282.
30. Shi J-Y, Liu Z, Yu H, Li Y-J. Predicting drug-target interactions via within-score and between-score. Biomed Res Int. 2015;2015:350983. 9 pages
31. Shi J-Y, Huang H, Zhang Y-N, Long YX, Yiu SM. Predicting binary, discrete and continued lncRNA-disease associations via a unified framework based on graph regression. BMC Med Genet. 2017;10(4):65.
32. Shi J-Y, Huang H, Zhang Y-N, Yiu S-M. Microbe-Disease Associations via Binary Matrix Completion. In: Cai ZP, Daescu O, Li M, editors. Lecture Notes in Bioinformatics, vol. 10330: Hawaii: Springer; 2017. p. XV-XVI.