Microbe-Disease Association Prediction

Yiren Wang, Yunuo Ma
Columbia University
Email: {yw3162, ym2774}@columbia.edu

February 14, 2020

1 Background

In many recent years' studies, both biological and clinical research has suggested that microbiota plays a crucial role in human health and diseases. The human body contains at least 1000 different species of known bacteria and carries 150 times more microbial genes than are found in the entire human genome [1]. Microbes have huge influences on the human body that any subtle aberrant behavior would possibly trigger many modern diseases, for instance, allergies, obesity, digestive disorders, and mental illness [2]. The study of the relationship between microorganisms and human disease pathological mechanisms could provide valuable medical information, which may promote disease diagnosis, prevention, and control.

Because of the cost of traditional experimental methods, computational methods are desirable in identifying microbe candidates for diseases[3]. There are several existing computational methods based on Heterogeneous Information Network (HIN), such as KATZHMDA, RWRHMDA, and NTSHMDA[4, 5, 6]. However, these models may bring bias to the well-studied microbes and diseases, and they failed to utilize the rich prior biological information. In 2017, Huang et al. proposed the computational model of PBHMDA (Research Path Based Human Microbe-Disease Association), which introduced the depth-first search algorithm to the HIN framework. Their model strongly depends on Gaussian kernel similarity to calculate the similarity for microbes and diseases and thus is not applicable to find the unknown association[3]. Similarly, BDSILP is a label propagation-based method for ranking candidate microbes for diseases and still could not make predictions for new diseases without known association[7].

Machine learning methods are applied more frequently in the MDA prediction field. For example, LRLSHMDA is a semi-supervised model, which employs Laplacian regularized least squares classifier to find the most related microbes for the objective diseases[8]. Most recently, in the work of WMGHMDA, the researchers implement a novel weighted meta-graphs model and improve the prediction performance for new diseases[9]. Meta-graph is also a conventional machine learning algorithm, which has been widely applied in the recommendation system.

2 Introduction

In this project, we aim to explore more machine learning algorithms for microbe-disease association prediction. Our goal is to develop a novel computational method to exploit structural information from the integrated MDA network and provide more accuracy and unbiased prediction about the pathology relationship between various diseases and microbes.

2.1 Heterogeneous Information Network

We are going to utilize the Bipartite Network Schema to construct interactions among microbes and diseases. Firstly, we will build the microbes similarity network and diseases similarity network, respectively, based on the integrated similarity for microbes and diseases.

After that, we will combine the HMDAD data to our HIN with each microbe and disease as a node, and then allocate the edge connecting the nodes among two integrated networks with the known microbe-disease association.

2.2 Potential Prediction Methods

2.2.1 K Nearest Neighbors

Since the size of existing data about microbes and diseases is limited, simple machine learning algorithm like K Nearest Neighbors could be applied as a starter to find k most potentially related types microbes based on one given type of microbe and similarly for diseases.

2.2.2 Collaborative Filtering

The item-item collaborative filtering algorithm computes the similarity between each pair of items that is the similarity between each microbe-disease pair. Therefore, we could prioritize the potential microbes that are highly-likely related to the disease. This algorithm takes the weighted sum of ratings of "item-neighbors". The prediction is given by:

$$P_{d,m} = \frac{\sum_N (s_{m,N} \cdot A_{d,N})}{\sum_N |s_{m,N}|},$$

where $s_{m,N}$ is the similarity between microbes and $A_{d,N}$ indicates the association between the disease d and microbes.

2.2.3 Deep Neural Network

Deep Autoencoders, with hidden layers and nonlinear activation layer are also our candidate method for prediction. Deep neural network could capture the latent feature of the objective microbes and diseases, and thus more powerful. However, because of the data limitation, training of the neural network is quite challenging.

2.3 Performance Evaluation

We plan to evaluate our model and compare it with other state-of-the-art computation methods in two aspects, recovering known microbe-disease association and inferring potential microbes for new diseases.

2.3.1 Cross-Validation

We intend to apply the LOOCV and k-fold CV to the model evaluation. Under the methodology of LOOCV, one observed microbe-disease pair is randomly assigned as the testing sample, and the remaining observed MDA pairs are the training data samples in each round of simulation. Then we will rank the prediction score of each test sample against all candidate samples to figure out whether the test microbe-disease pair could be inferred by our method successfully. Similarly, the k-fold CV algorithm randomly partitions all observed MDA pairs into k equal-sized groups, with one single group as holdout data and the rest of the data as training data. This process shall be repeated for 100 times to mitigate the bias introduced from data division.

3 Dataset

3.1 HMDAD

HMDAD (Human Microbe-Disease Association Database) provides the validated human microbe- disease association data from microbiota studies. Currently, HMDAD contains 483 disease-microbe entries, including 39 diseases and 292 microbes.[10]

3.2 MeSH

The NATIONAL LIBRARY OF MEDICINE produced the Mesh (Medical Subject Headings) database, which includes a plenty of vocabulary thesaurus about diseases. The terms naming descriptors are stored in a hierarchical structure.[11]

4 Plan

- Progress I (February 21): Data collection and cleaning; Applying the open-source semantic similarity neural network to MeSH data
- Milestone II (March 6): Similarity measurement among diseases
- Progress II (March 27): Similarity measurement among microbes; Completing the HIN construction
- Milestone III (April 10): Developing prediction model and tuning parameters
- Progress III (April 24): Model Evaluation

References

[1] Luke Ursell, Henry Haiser, Will Treuren, Neha Garg, Lavanya Reddivari, Jairam Vanamala, Pieter Dorrestein, Peter Turnbaugh, and Rob Knight. The intestinal metabolome: An intersection between microbiota and host. *Gastroenterology*, 146, 05 2014.

[2] A. Collen. *10% Human: How Your Body's Microbes Hold the Key to Health and Happiness*. Harper, 2015.

[3] Zhi-An Huang, Xing Chen, Zexuan Zhu, Hongsheng Liu, Gui-Ying Yan, Zhu-Hong You, and Zhenkun Wen. Pbhmda: Path-based human microbe-disease association prediction. *Frontiers in Microbiology*, 8, 2017.

[4] Xing Chen, Yu-An Huang, Zhu-Hong You, Gui-Ying Yan, and Xue-Song Wang. A novel approach based on katz measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics*, 2016.

[5] Xianjun Shen, Yao Chen, Xingpeng Jiang, Xiaohua Hu, Tingting He, and Jincai Yang. Predicting disease-microbe association by random walking on the heterogeneous network. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016.

[6] Jiawei Luo and Yahui Long. Ntshmda: Prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, page 1–1, 2018.

[7] Wen Zhang, Weitai Yang, Xiaoting Lu, Feng Huang, and Fei Luo. The bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access*, 6:38052–38061, 2018.

[8] Fan Wang, Zhi-An Huang, Xing Chen, Zexuan Zhu, Zhenkun Wen, Jiyun Zhao, and Gui-Ying Yan. Lrlshmda: Laplacian regularized least squares for human microbe–disease association prediction. *Scientific Reports*, 7(1), Aug 2017.

[9] Yahui Long and Jiawei Luo. Wmghmda: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network. *BMC Bioinformatics*, 20(1), Jan 2019.

[10] Hmdad. `http://www.cuilab.cn/hmdad`. (Accessed on 02/12/2020).

[11] Medical subject headings - mesh - ncbi. `https://www.ncbi.nlm.nih.gov/mesh/?term=Medical%20Subject%20Headings&utm_source=gquery&utm_medium=search`. (Accessed on 02/12/2020).