Methods 124 (2017) 120–125

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Prioritizing disease-causing microbes based on random walking on the heterogeneous network

Xianjun Shen *, Yao Chen, Xingpeng Jiang, Xiaohua Hu, Tingting He, Jincai Yang

School of Computer, Central China Normal University, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 27 March 2017
Received in revised form 22 May 2017
Accepted 13 June 2017
Available online 16 June 2017

Keywords:
Disease network
Microbe network
Random walk
Heterogeneous network

ABSTRACT

As we all know, the microbiota show remarkable variability within individuals. At the same time, those microorganisms living in the human body play a very important role in our health and disease, so the identification of the relationships between microbes and diseases will contribute to better understanding of microbes interactions, mechanism of functions. However, the microbial data which are obtained through the related technical sequencing is too much, but the known associations between the diseases and microbes are very less. In bioinformatics, many researchers choose the network topology analysis to solve these problems. Inspired by this idea, we proposed a new method for prioritization of candidate microbes to predict potential disease-microbe association. First of all, we connected the disease network and microbe network based on the known disease-microbe relationships information to construct a heterogeneous network, then we extended the random walk to the heterogeneous network, and used leave-one-out cross-validation and ROC curve to evaluate the method. In conclusion, the algorithm could be effective to disclose some potential associations between diseases and microbes that cannot be found by microbe network or disease network only. Furthermore, we studied three representative diseases, Type 2 diabetes, Asthma and Psoriasis, and finally presented the potential microbes associated with these diseases by ranking candidate disease-causing microbes, respectively. We confirmed that the discovery of the new associations will be a good clinical solution for disease mechanism understanding, diagnosis and therapy.

1. Introduction

Metagenomics and 16S pyrosequencing have enabled the study of ecosystem structure and dynamics to great depth and accuracy. At the same time, the Human Microbiome Project (HMP) was built with the goal of characterizing and identifying the microorganisms which are found in association with both healthy and diseased human (the human microbiome), developing new technologies and tools for computational analysis and establishing a resource repository [1]. There is a huge number of microorganisms mainly composed of bacteria but also including viruses, fungi and so on that inhabit various human organs such as skin, mouth, hair, stomach and gastrointestinal tract [2]. These microorganisms rarely live in isolation, but coexist in complex interconnections with various relationships. It is generally that the commensal microbiota has extensive impact on physiology and they also contribute to the development of the immune system [3–5], protection against pathogens [6] and drug metabolism [7]. For example, the overgrowth of competitive pathogenic species can disrupt the relationships among the normal intestinal microbiota, which will finally lead to disease. Meanwhile the relationship between microbe and host is not a one-way relationship but a mutualistic one, for instance, a microbial species' metabolic by-products can alter the surrounding environment to the detriment of other microorganisms, and different taxonomic groups such as bacteria may cooperate to build a biofilm which could confer antibiotic resistance to its members. Identifying human disease-causing microorganisms is still a huge challenge. More recently, some computational methods have been proposed for studying microorganism and human diseases [8–10]. These studies included the identification of bacteria genes which are contributing to pathogenicity, the prediction of the impact of microbial proteins in human biological events, analysis of specific genes in the human gut microbiome. Accordingly, some microbial network construction methods gradually have been put forward that are key to understanding these microorganism function [11,12] including a large automated text-mining of microbial interactions such as co-occurrence and mutual exclusion

* Corresponding author.
E-mail addresses: xjshen@mail.ccnu.edu.cn (X. Shen), chenyao99912@163.com (Y. Chen), xpjiang@mail.ccnu.edu.cn (X. Jiang), xh29@drexel.edu (X. Hu), tthe@mail.ccnu.edu.cn (T. He), jcyang@mail.ccnu.edu.cn (J. Yang).

http://dx.doi.org/10.1016/j.ymeth.2017.06.014
1046-2023/

and the organization of bacterial network by mining scientific literature, respectively. The purpose of these methods is to help us more comprehensive, more systematic understanding of microbial communities.

Random walk is first put forward by Karl Pearson in 1905 [13], is a kind of statistical model and made up of continuous randomly motion trajectory, like a drunken man whose every step forward is a random and has nothing to do with running in front in the process of walk. The random walk on the figure refers to the starting point randomly selects a neighbor node and moves to the neighbor nodes, then the current node repeats above iteration process as a new starting point. In the field of biological information, Köhler et al. [14] improve random walk, and apply this method to PPI (protein-protein interaction) network for disease genes prediction. Li et al. [15] put forward a random walking on the heterogeneous network to infer gene-phenotype relationship based on the random walk with restart method which proposed by Köhler. They all achieve good prediction effect. We applied this algorithm to disease-microbe data and use related methods to validate its performance.

In this work, we proposed a new way to identify disease-microbe relationship. We first curated large-scale microbe-disease association data, a symptoms-based human disease data and microbe data, then we applied disease and microbe data to random walk with restart on the heterogeneous network (RWRH) to infer the potential disease-microbe relationships. We connected the disease network and microbe network by disease-microbe relationship and constructed a heterogeneous network. We extended the random walk with restart algorithm to the heterogeneous network, using the target disease and corresponding microbes as seed nodes. The RWRH algorithm ranks diseases and microbes at the same time, the top ranked microbe was selected as the most association to the disease. We used this algorithm to disclose the potential relationship between diseases and microbes.

2. Material and methods

In this section, we will introduce the method in details. The novel disease-causing microorganism prioritization approach. Firstly, we will describe the construction method of heterogeneous network by integrating disease data and microbe data; secondly, carrying on the random walk on the heterogeneous network; finally, candidate microorganism related to the diseases are ranked. The detail is as follows:

2.1. Construction of the heterogeneous network

Three types of data sources are represented by three networks, namely disease network, microbe network and disease-microbe relationships network. In the disease network, each node is one kind of disease, and the link between two diseases means that the symptoms-based similarity between the two diseases is not zero [16]. The microbe network is constructed by Spearman correlation [17], each node represents one microbe, and the edge between two microbes represents correlation is greater than a certain threshold. The microbe network is a microbial abundance matrix, which is the abundance data of different microorganisms in different sampling points. The disease-microbe network is represented by 0–1 matrix. If some literatures have proved that the occurrence of disease is associated with the increase or decrease of microorganisms, we assign value 1 to the correlation matrix. So e $(i, j) = 1$ represents the change of microbe i will cause disease j happen, e $(i, j) = 0$ represents that no literature proved disease j associated with microbe i. We constructed the heterogeneous network by connecting the disease network and microbe network using the known disease-microbe relationships network. A simple example of the heterogeneous network is shown in Fig. 1.

$A = \begin{bmatrix} A_M & B \\ B^T & A_D \end{bmatrix}$ represents an adjacency matrix of heterogeneous network, where $A_{M_{(n\times n)}}$ and $A_{D_{(m\times m)}}$ are microbe network and disease network, respectively. $B_{(n\times m)}$ is a microbe-disease network composed of n microbes and m diseases. B^T is the transpose of B. Let λ be the probability of the random walker jumping from microbe network to disease network or vice versa. Here, the transition matrix of the heterogeneous network is as shown in the following formula:

$$M = \begin{bmatrix} M_M & M_{MD} \\ M_{DM} & M_D \end{bmatrix} \tag{1}$$

here M_M and M_D are microbe network and disease network transition matrix, respectively. M_{MD} is a probability transition matrix from microbe network to disease network, accordingly, M_{DM} is the transpose of M_{MD}.

The transition probability from m_i to d_j can be described as follows:

$$(M_{MD})_{i,j} = p(d_j|m_i) = \begin{cases} \frac{\lambda B_{ij}}{\sum_j B_{ij}}, & if \sum_j B_{ij} \neq 0 \\ 0 & otherwise. \end{cases} \tag{2}$$

Similarity, the transition probability from d_i to m_j can be described as follows:

$$(M_{DM})_{i,j} = p(m_j|d_i) = \begin{cases} \frac{\lambda B_{ji}}{\sum_j B_{ji}}, & if \sum_j B_{ij} \neq 0 \\ 0 & otherwise. \end{cases} \tag{3}$$

The element of M_M (microbe network) at i-th row and j-th column is $p(m_j|m_i)$, the probability of the random walker transition from m_i to m_j. It is defined as follows:

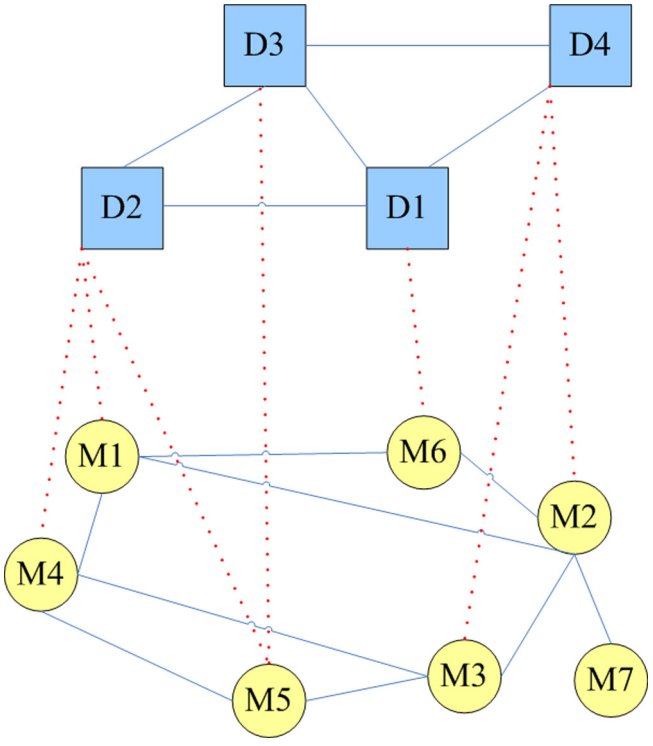

Fig. 1. Illustration of the heterogeneous network. The upper subnetwork is disease network, and the lower network is microbe network. They are connected by disease-microbe relationship.