Identification of Microbe Disease Association

Abishek T P[1], Athira K[2], Meera Susan Cherian[3], Prof. Joby George[4]

[1,2,3] Department of Computer Science and Engineering Mar Athanasius College of Engineering Ernakulam, India
[4]Associate Professor, Department of Computer Science and Engineering Mar Athanasius College of Engineering

---

Abstract - *There are countless microbes in the human body which are closely related to human life activities and human diseases and they play various roles in the physiological process. There is growing evidence that microbes are closely associated with human diseases. Research of disease-related microbes helps us understand the mechanisms of diseases and provides new strategies for diseases diagnosis and treatment. However, traditional biological experiments are time-consuming and expensive, so it has become a research topic in bioinformatics to predict potential microbe-disease associations by adopting computational methods. The dataset used is Human Microbe Disease Association Database (HMDAD). The model is divided into data preparation, integrating the data, eigenvalue decomposition, training the model and sorting the samples. The computation model uses node similarity information, eigen decomposition, k - means clustering, decision tree classifier to score the microbe disease association. To assess the prediction ability of the model leave -one -out cross validation method is used. The results indicate the reliable capability for inferring the most possible microbedisease associations. Therefore, the model achieves a superior performance compared with other approaches.*

Key Words: **Microbe, disease, microbe-disease association, node-information, link propagation, association prediction, decision tree**

1. INTRODUCTION

Microorganisms have been widely found in the oceans, soils, human bodies and other places, and their existences have profound impacts on human life. With the rapid development of high-through sequencing technologies and modern bio-informatics, researches on microbiology have attracted increasing attention from the scientific and medical communities. Meanwhile, microbes participate in different levels of metabolic activities in the human body and are interdependent with the host. Therefore, the health of human microbiome in human body is an important factor for human health. On the surface, human life activities depend on microorganisms, but the hosts and their living environment always affect the survival of microorganisms. For example, the use of antibiotics and western-style high-fat diets may alter microbial composition.

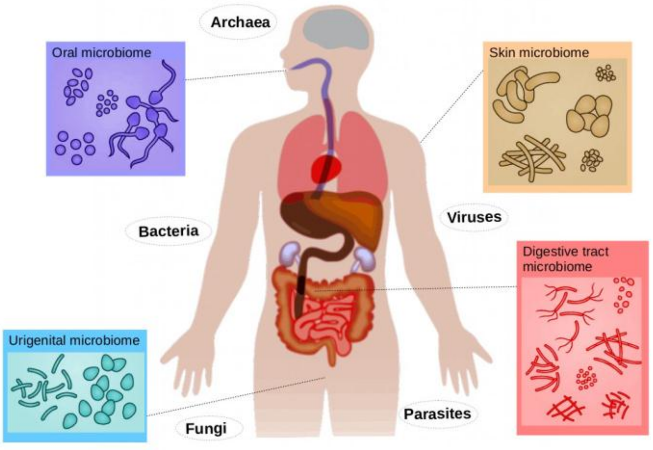


Fig-1: Human Microbiome

Microbiota is all microbes existing on human body surfaces and cavity mucous membrane connected with the outside world. Generally, microbes are divided into the following categories: bacteria, fungi, archaea, viruses and others. Microbes are widespread in our bodies and body surfaces, having important effects on human metabolism, behavior, development, adaptation and even evolution. There are rich and diverse microbes in the intestinal tract, skin, oral cavity and genitourinary tract of the human body. It has been confirmed that the number of microbes that survive and reproduce in body and on the body surfaces is 100 billion, 10 times the number of human cells. Microbiota and human body are mutually beneficial symbiotic relationship. Microbes involved in human metabolism, such as using polysaccharides and nitrogen compounds in diet, participating in drug metabolism and affecting drug efficacy; participating in the regulation of the immune system, endocrine system and nervous system. Scientists realized that simply focusing on the human body and the human genome does not fully grasp the key issues of human disease and health. Clinical studies show that the disorder of the microbial population is related to multiple system diseases like digestive system diseases such as irritable bowel syndrome, inflammatory bowel disease immune system diseases such as allergy, asthma, multiple sclerosis, metabolism, endocrine system diseases such as obesity, diabetes, and neuropsychiatric disorders such as depression, autism, and so on. Microorganisms have an important effect on infectious diseases and non-infectious diseases. The human body is possible to get sick when foreign microorganisms invade or a microbial community is imbalanced. For example, there are more abundant Fusobacterium in asthmatic patients than healthy people. Lecithinase-negative Clostridium and Lactobacillus are much more in colorectal carcinoma patients. Increased Lactobacillus can result in tertiary lymphoid. All the above reports suggested that there are close associations between microbes and human diseases. Therefore, finding new Microbe-Disease Associations helps to provide diagnostic and therapeutic clues for clinical researches. As mentioned, 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. 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 MDAs 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, potential lncRNA-disease association and potential drug-target interactions, where machine learning⍰based and similarity measure-based models have shown their outstanding prediction ability. So, it is essential to logically extend these prediction methods into microbe⍰disease association prediction field. Therefore, a deeper understanding of microbe related pathological relationships may provide new ideas for the study of new treatment and prevention strategies for diseases, as well as promote global human health.

2. RELATED WORKS

2.1 Back - Propagation Neural Network Model

The Back Propagation Neural Network is one of the most popular models and is widely applied to prediction and classification problems. The aim of this method is to design a novel BPNN model for microbe-disease association prediction. Therefore, a 3-layer BPNN with 292 nodes per layer is constructed[4], in which, the information of known associations between each microbe and all diseases would be used as the input signals of corresponding nodes in the input layer. Next, a new activation function to activate the hidden layer and the output layer based on the hyperbolic tangent function is designed. Thereafter, the weights and node biases of the BPNN would be continuously updated during the back - propagation process until the whole network reaches convergent state. The proposed model consists of three layers including the input layer, the output layer and the hidden layer. The dataset consists of 292 microbes and 39 diseases which are obtained from HMDAD database. Hence, the input data to the network is 292 ∗ 39 dimensional matrix. An adjacency matrix A represents the known microbe - disease association. The method for controlling a neural network is to set and adjust its connection weights and node biases. Traditionally, initial weights are set to random values which will have a direct effect to the training efficiency and convergence speed. Hence, Gaussian Interaction Profile kernel similarity is used to optimize the initial weight and to obtain similarity matrix S_M.

$$SM(i,j) = exp(-\gamma_m \| IP(m_i) - IP(m_j) \|^2)$$

where m_i and m_j are ith and jth microbe in SM. IP(m_i) and IP(m_j) are the ith and jth row in the adjacency matrix. where γ ′ m is the parameter for Gaussian kernel bandwidth.

$$\gamma_m = \gamma'_m / \left(\frac{1}{N_m} \sum_{i=1}^{N_m} \| IP(m_i) \|^2 \right)$$

Based on similarity matrix SM, two identical initial weight matrices $W^{[1]}$ and $W^{[2]}$ can be constructed.

$$W^{[1]}(i,j) = W^{[2]}(i,j) = \frac{SM(i,j) - min(SM)}{max(SM) - min(SM)} - 0.5.$$

Hence, $W^{[1]}$(i,j) is the weight between the ith node of the input layer and the jth node of the hidden layer and $W^{[2]}$(i,j) is the weight between the ith node of the hidden layer and the jth node of the output layer. The bais to each node for its higher activity is set to a random value between [−ρ, ρ] initially.

A new activation function to activate the hidden layer and the output layer based on the hyperbolic tangent function is designed. Thereafter, the weights and node biases of the BPNN would be continuously updated during the backpropagation process until the whole network reached convergent state. And then, based on the outputs of the convergent network, a microbe-disease correlation score matrix could be obtained.

The GIP similarity was used to optimize initial connection weights of BPNNHMDA, although it could improve the training process, but it fixed the convergent direction of the neural network which means that some poorly optimized weights may not be conducive to effective prediction of some potential microbe-disease associations. In addition, the learning rate of BPNN was fixed in BPNNHMDA, which was not conducive to the training process either. Conservative and small learning rate would reduce the training speed, while large learning rate may make the neural network miss the optimal error. Therefore, it is necessary to improve the adaptive change ability of learning rate to reduce the training time and improve the prediction performance of BPNNHMDA in future.

2.2 Bipartite Network Model

A bipartite network to predict potential microbe-disease interactions based on known microbe-disease associations only. The assumption considered in this method is two nodes are similar if they have common neighbors or are connected to similar nodes. Three datasets from the LncRNADisease database, Lnc2Cancer database and MNDR database were collected.

Let L be a set of lncRNAs and D be a set of diseases then, the L-D network can be described as a bipartite graph G (L, D, E) where E is set of edges[14]. An adjacency matrix A is constructed to show the similarity between lncRNA and diseases.

Using the power law distribution, the similarity between microbes nodes are obtained if they have common nieghbors and disease nodes as well. If they do not have common neighbors then if the nodes are connected to similar nodes using SimRank method their similarity can be obtained. Finally, by integrating both the methods and adjacency matrix a recommendation matrix can be calculated for both microbe and disease, then again by integrating two recommendation matrix the microbe-disease association can be inferred.

2.3 Laplacian Regularized Least Squares Method

Similar diseases tend to be associated with functionally similar lncRNAs. Based on this, selecting lncRNAs that are not related to the given disease is difficult or even impossible, so a computational model of Laplacian Regularized Least Squares for LncRNA–Disease Association in the semisupervised learning framework was developed. This method prioritizes the entire lncRNAome for disease of interest by integrating known phenome-lncRNAome network obtained from the database of LncRNADisease, disease similarity network and lncRNA similarity network. It is a global approach that can rank candidate disease–lncRNA pairs for all the diseases simultaneously.

2.4 Inductive Matrix Completion Model

A novel matrix completion-based model named IMCMDA for miRNA-disease associations prediction. This model of IMCMDA[6] was implemented based on the known miRNAdisease associations, disease semantic similarity, miRNA functional similarity, Gaussian interaction profile kernel similarity for miRNAs and diseases.

The data of known human miRNA-disease associations were retrieved from the HMDD V2.0 database and a nd × nm adjacency matrix A was defined as A(d(i),m(j)) = 1 if disease d(i) has association with miRNA m(j) else A(d(i),m(j)) = 0. The miRNA functional similarity was calculated based on assumption that functionally similar miRNAs tend to cause similar diseases. The dataset was taken from cuilab and a nm * nm matrix FS to represent the miRNA functional similarity constructed in which element FS(m(i), m(j)) denotes the functional similarity between miRNA m(i) and m(j).

In this model, the known microbe-disease associations and the integrated microbe similarity and disease similarity are combined to calculate the prediction score of each microbe-disease pair. IMCMDA predicts the microbedisease associations by using the low-rank inductive matrix completion algorithm. A crucial advantage of IMC is that it utilizes disease similarity and microbe similarity as the feature of disease and microbe to complete the missing microbe-disease association. It means that using the feature vector of a new disease without any known related microbes to predict the relevance-scores between this new disease and all microbes. In addition, searching the optimal solution with an alternating gradient descent algorithm made sure the reliability of the disease eigenvectors and the microbe eigenvectors. Thus, the model is a semi- supervised model. The advantage of semi-supervised model is that it doesn't rely on negative samples. It only needs positive samples and unlabeled samples, which greatly reduces the difficulty of building models.

3. BACKGROUND

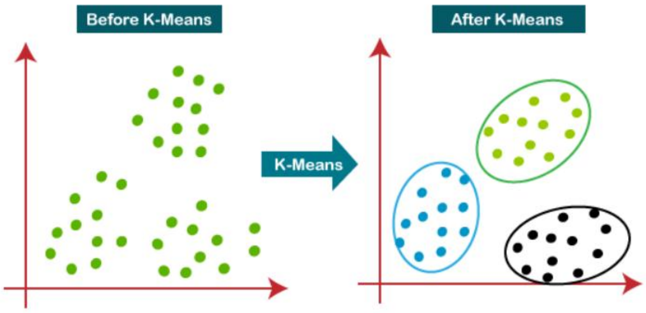
3.1 HMDAD Dataset

The Human Microbe-Disease Association Database is a resource which collected and curated the human microbe-disease association data from microbiota studies. After data processing, 450 high-quality known associations including 292 microbes and 39 diseases have been obtained. The dataset contains various parameters but for the purpose of determining the association, microbes and disease are encoded to numbers and known associations is replaced with respective numbers of microbes and disease, hence parameters are discarded.

3.2 K - means Clustering

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It groups the unlabeled dataset into different clusters. t allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in the algorithm.

The working of k – means algorithm include selection of the number K to decide the number of clusters, select random K points or centroids it can be other from the input dataset. Assign each data point to their closest centroid, which will form the predefined K clusters, then calculate the variance and place a new centroid of each cluster. Repeat these, which means reassign each datapoint to the new closest centroid of each cluster. If any reassignment occurs, then again reassign each datapoint to closet centroid else finish and the model is ready.



3.3 Decision Tree Classifier

Decision Tree Learning is supervised learning approach used in statistics, data mining and machine learning. In this formalism, a classification or regression decision tree is used as a predictive model to draw conclusions about a set of observations. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but the resulting classification tree can be an input for decision making. So, the ultimate goal is to create a model that predicts the value of a target variable based on several input variables.

3.4 Eigenvalue Decomposition

In linear algebra, eigen decomposition is the factorization of a matrix into a canonical form, whereby the matrix is represented in terms of its eigenvalues and eigenvectors. Only diagonalizable matrices can be factorized in this way. When the matrix being factorized is a normal or real symmetric matrix, the decomposition is called "spectral decomposition", derived from the spectral theorem. In other words, this technique decomposes a complex matrix into its eigen vectors and values, thus reduces the complexity in computation.

4. Implementation

In this section, the proposed method is implemented on the HMDAD dataset. The studies were carried out using the Google colab environment that supports python programming language with various modules such as numpy, math.Scikit - learn library which provides a selection of efficient tools for machine learning and statistical modeling including classification, clustering.

5. DATA PREPARATION

An adjacency matrix A can be obtained from HMDAD dataset by extracting 450 known associations. If there is a known association between disease d(i) and microbes m(j), then the value of the element A(d(i),m(j)) is 1 else 0. nd and nm represents the number of disease and microbes considered This serves as input for further processing.

6. Node Similarity Information Measurement

Considering the assumption that if two similar diseases were associated with two microbes, respectively, the two microbes were likely to be similar, and there were similar interaction and non-interaction pattern between diseases and microbes, Gaussian interaction profile kernel similarity for disease KD was constructed to indicated the similarities between diseases based on the known associations of disease-microbe pairs.

$$\gamma_{\mathrm{d}} = \frac{\gamma'_{\mathrm{d}}}{\frac{\sum_1 nd \| IP(d(i)) \|^2}{nd}}$$

where γ ′ d = 1

$$KD\left(d(i), d(j)\right) = \exp\left(-\gamma_{\mathrm{d}} \| IP\left(d(i)\right) - IP\left(d(i)\right) \|^2\right)$$

To effectively and scientifically predict potential microbe-disease association, it is necessary to introduce other datasets with the Gaussian interaction profile kernel similarity. For this, symptom-based human disease dataset from HSDN taken and integrated the Gaussian interaction profile kernel similarity for disease KD and the symptom-based disease similarity SDM to obtain the Integrated symptom-based disease similarity SD. SD can be calculated using

SD = (KD + SDM)/2

The similarity matrix S_m between microbes is also find the same GIP kernel similarity as used for disease similarity.

7. Eigenvalue Decomposition

Eigenvalue decomposition technique is used to reduce the large amount of memory and time overhead is required in the inverse operation of the matrix. S_m and S_d is a real symmetric matrix, the eigenvalue decomposition technique is used here to improve the computational efficiency. $S_m = R_m\Lambda_m R_m T$ and $S_d = R_d\Lambda_d R_d T$ are eigenvalue decomposition of S_m and S_d where R_m and R_d are eigen vectors; Λ_m and Λ_d are eigen values.

8. TRAINING THE MODEL

The main idea of method is to train different classifiers which are weak classifiers for the same training samples, and then grouped these weak classifiers with different ratios to form a stronger classifier to score and sort samples. Decision tree is used as weak classifier.

The sample with confirmed association as a positive sample, else unknown sample. Since there are more unknown samples than positive samples, and it was unreasonable to directly train such unbalanced datasets. A novel method is used to balance the datasets which uses k-means clustering to divide the unknown sample into k parts, and then randomly extract some samples from each part as negative samples, while positive samples kept unchanged. In order to make the dataset more balanced for effective training, the number of the unknown samples randomly selected must be approximately equal to the positive sample. Thus, the negative and positive samples together form the training samples. Each training sample was weighted with an initial weight of 1/n, where n is the total number of training samples. The main purpose of the training process is to calculate the proportion of each weak classifier in the final strong classifier and update the weight of each training sample according to whether it was classified correctly by the last classifier and the overall classification accuracy of the last classifier. After updating, the new training sample set with modified weight values, it is sent to the next weak classifier for training.

Three lists D_I, h(i), Y with n elements is built. The value of each element in D_I is the weight of the corresponding sample when the ith weak classifier trained the sample. The value of i was 0, 1, 2, , , , 29. In other words,D_0 was a list with all elements being 1/n . The elements in list h(i) and Y can take the value either 0 or 1 that depends on the prediction of the ith weak classifier and the corresponding sample is positive or not.

The error function ϵ_i is given by

$$\epsilon_i = \sum_{j=1}^{n} D_i 1_{h(i)j \neq Y_j}$$

The error function ϵ_i is equal to the sum of the weights of the samples, whose label predicted by the weak classifier h(i)j is different from the known label Yj i.e; єi is equal to the sum of the weights of all the samples that were predicted wrong and the prediction matrix P contains the strength between disease i and microbe j.

The proportion of the ith weak classifier in the strong classifier is defined as

$$\alpha_i = \frac{\log \frac{1-\epsilon_i}{\epsilon_i}}{2}$$

Hence, it is inferred that the smaller the error function is , the larger will be the proportion of the weak classifier in the strong classifier.

The variate Zi is calculated using the error function.

$$Z_i = 2\left[\epsilon_i\ (1-\epsilon_i)\right]^2$$

The weight of the sampleD_{i+1}(j) is updated using the previous weight

$$D_{i+1}\ (j) = \frac{1}{Z_i} D_i\ (j)\ e^{-\alpha_i Y_j h(i)_j}$$

After the weights of samples being updated, the samples with the new weights were sent to the next weak classifier to start the next training until all the weak classifiers completed the training. Here 30 weak classifiers are used for reducing the prediction time and to increase the accuracy.

The score of the jth sample

$$s\ (j) = \sum_{i=0}^{29} \alpha_i H\ (i)_j$$

where H(i)j is the score obtained by the ith weak classifier for the jth sample i.e; the score of the sample is equal to the sum of the product of the sample's goal scored by weak classifier and the corresponding weight.

9. RESULT

The proposed model uses node-information based Link Propagation for Human Microbe-Disease Association prediction to prioritize the most possible disease-related microbes. Node similarity information that contains Gaussian profile kernel similarity and characterstics of disease symptom, has been integrated to promote strong associations between the most likely nodes through link propagation. The model is trained using k- means clustering and Decision tree classifier. The technology of Eigenvalue transformation have been adopted to simplify the solving process of the model.Leave - one -out cross validation method is applied to assess the prediction ability of the model.The model accurately calculates the score between microbe disease association which helps to determine the most potential microbe that causes the disease.

The accuracy of the proposed model is found to be 87.69 %.The LOOCV value of KATZHMDA and LRLSHMDA are 0.6998 and0.7508 respectively. These results confirmed the superior prediction performance of the proposed model.

10. FUTURE RESEARCH

In future, the relationships between diseases and other molecule can be discovered, make a comprehensive analysis of the diseases. The improving prediction performance of the proposed method compared to previous approaches, it is expected that the prediction ability will be further improved if a more comprehensive similarity calculation method is taken into consideration.

11. CONCLUSION

The prediction of microbe-disease association serves as a biomarker detection and drug discovery for disease diagnosis, treatment, prognosis and prevention. The datasets used by the model were relatively reliable. The potential similarities for diseases and microbes through Gaussian interaction profile kernel similarity were extracted, eigen decomposition reduces the computational complexity and multiple weak classifiers combined into one strong classifier according to different weights to score the samples. The high-precision weak classifiers accounted for a high proportion which improve the accuracy of the strong classifier.

REFERENCES

[1] Prioritizing Human Microbe-Disease Associations Utilizing a Node⍰Information-Based Link Propagation Method LI PENG 1 , DONGZHOU 1 , WEI LIU2 , LIQIAN ZHOU3 , LEI WANG 4 , BIHAI ZHAO 4 , AND JIALIANG YANG 5, IEEE Access, Volume8, 2020.

[2] Human Microbe-Disease Association Prediction Based on Adaptive Boosting Li-Hong Peng1† , Jun Yin2† , Liqian Zhou1 , Ming-Xi Liu3* and Yan Zhao2 *.

[3] A Novel Method for LncRNA-Disease Association Prediction Based on an lncRNA-Disease Association Network Pengyao Ping, Lei Wang , Linai Kuang , Songtao Ye , Muhammad Faisal Buland Iqbal , and Tingrui Pei, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 16, NO. 2, MARCH/APRIL 2019 .

[4] Identifying Microbe-Disease Association Based on a Novel Back Propagation Neural Network Model,Hao Li , Yuqi Wang , Zhen Zhang , Yihong Tan , Zhiping Chen , Xiangyi Wang, Tingrui Pei , and Lei Wang,IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 18, NO. 6, NOVEMBER/DECEMBER 2021 .

[5] Novel human lncRNA–disease association inference based on lncRNA expression profiles Xing Chen1,2,* and Gui-Ying Yan1,2.

[6] Predicting miRNA-disease association based on inductive matrix completion Xing Chen1, * , Lei Wang1 , Jia Qu1 , Na-Na Guan2 , Jian-Qiang Li2 .

[7] Link Propagation: A Fast Semi-supervised Learning Algorithm for Link Prediction Hisashi Kashima Tsuyoshi Kato† Yoshihiro Yamanishi‡ Masashi Sugiyama§ Koji Tsuda¶.

[8] A New Link Prediction Algorithm: Node Link Strength Algorithm Yin Guisheng1 , Yin Wansi2 , Dong Yuxin3 Department of Computer Science and Technology Harbin Engineering University Harbin, China

[9] An eigenvalue transformation technique for predicting drugtarget interaction Qifan Kuang1 , Xin Xu3 , Rong Li2 , Yongcheng Dong3 , Yan Li1 , Ziyan Huang1 , Yizhou Li1 Menglong Li1.

[10] Fast and Scalable Algorithms for Semi-supervised Link Prediction on Static and Dynamic Graphs Rudy Raymond1 and Hisashi Kashima2.

[11] RNMFMDA: A Microbe-Disease Association Identification Method Based on Reliable Negative Sample Selection and Logistic Matrix Factorization With Neighborhood Regularization Lihong Peng, Ling Shen, Longjie Liao, Guangyi Liu and Liqian Zhou*.

[12] Integrating random walk and binary regression to identify novel miRNA-disease association Ya-Wei Niu1 , Guang-Hui Wang1*, GuiYing Yan2 and Xing Chen3*.

[13] A Robust Algorithm Based on Link Label Propagation for Identifying Functional Modules from Protein-protein Interaction Networks Hao Jiang, Fei Zhan, Congtao Wang, Jianfeng Qiu, Yansen Su, Chunhou Zheng, Xingyi Zhang, Senior Member, IEEE, and Xiangxiang Zeng.

[14] BPLLDA: Predicting lncRNA-Disease Associations Based on Simple Paths With Limited Lengths in a Heterogeneous Network Xiaofang Xiao1†, Wen Zhu2†, Bo Liao1,2 *, Junlin Xu1 , Changlong Gu1 , Binbin Ji 2 , Yuhua Yao2 , Lihong Peng3 and Jialiang Yang2,4