Application of Random Walk for Disease Prediction

Jixiang Zhang

College of Science, Northeastern University, Boston, United States of America

zhang.jixi@northeastern.edu

Abstract. Disease is a great challenge to human civilization. In the long history, human beings have been fighting against diseases. In addition to finding drugs to treat diseases, predicting diseases is also the focus of human research. However, many research methods now ignore the randomness of disease. As people realize the randomness of disease, random walk technology is gradually applied to predict disease-related problems. This paper studies the characteristics of applying random walk method to predict disease-related problems. This article first introduces the concept of random walk, and then discusses the characteristics of random walk applied to problems related to genetic diseases, random walk applied to problems related to cancer, and random walk applied to problems related to the relationship between microorganisms and diseases. Specifically, it includes lncrnas, pathogenic genes, cancer genes, cancer classification, microorganisms, and diseases. The research of this article hopes to make human beings more aware of the importance of the randomness of diseases in predicting diseases. It is hoped that the characteristics of the method of applying random walk presented in this study can inspire people's future research.

Keywords: Random Walk, Predict, Disease, Stochastic.

1. Introduction

Diseases can affect people's physical health, mental health, and even life span. Disease has always been a major challenge for mankind. In the history of the development of human civilization, there have been diseases such as the black death that have had a significant impact on mankind and brought disasters. Human beings have been fighting against diseases and looking for ways to deal with them. In addition to exploring drugs to treat diseases, prevention and prediction have always been the focus of human research. Good prediction methods for diseases can have a significant impact on society. There are many methods for predicting diseases, such as regression forecasting, time series forecasting, grey forecasting, etc. However, a part of it is based on statistics of what has happened [1]. Some diseases belong to random events, and some diseases do not have complete statistics, so sometimes these prediction methods based on events that have occurred are not necessarily reliable. Therefore, people use methods about stochastic processes to predict disease. A stochastic process in probability theory is a collection of random variables. Specifically, if there is a process consisting of a set of random variables that depend on a real parameter t, and t has the meaning of time, then the process is a random process. For example, the number of customers a restaurant receives from time t0 to time t1 is a set of random variables that depend on time t. There are many methods, formulas, and techniques for random processes, such as Gaussian processes. This article will discuss the characteristics of random walk algorithms when applied to predicting disease-related problems. Algorithms of random walks are not based on past events, which is useful for predicting some disease-related problems with incomplete statistics and diseases that occur randomly [2]. For example, the prediction of the number of people infected with Covid-19. A random walk can be compared to a route consisting of several steps, where each step in the route is randomly generated, regardless of the past, just an event that happened at that moment. This article will discuss the advantages and disadvantages of using random walks to predict disease-relate problems. Specifically, this paper will discuss the advantages and disadvantages of applying random walk's method in predicting genetic disease-related problems, cancer-related problems, microorganisms and disease-related problems.

2. The Concept of Random Walk

A random walk is a mathematical model, which is a series of trajectories [3]. The shapes of these trajectories are random. Specifically, these trajectories consist of random steps. Generations of each step of the trajectory is random, not related to the event that has already happened, it is the random event that happened at this time. It can be used to represent irregular changing forms [4]. This is somewhat like Brownian motion. The Fig.1 is random walk image segmentation algorithm. Random walks are used in many fields, such as computer science, biology, chemistry, psychology, etc. There are also a variety of different applications in these areas, such as Internet link analysis and financial stock markets. This article will discuss the application of random walks to predicting disease-related problems. Some of the methods discussed are based on random walks, and some are research methods created by combining random walks with other mathematical models. This work will discuss these different methods used in predicting different disease-related problems later [4-12].

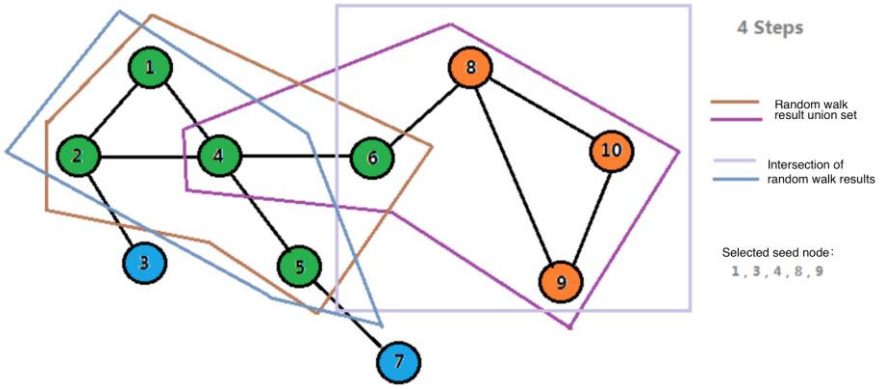


Figure 1. Random walk image segmentation algorithm

3. Application of Random Walk to Predict Different Diseases

3.1. Application of Random Walk to Predict Genetic Disease

Genetic diseases are one of the major challenges facing humanity. There are many hazards of genetic diseases, such as Down syndrome, which can affect the intellectual development and low self-esteem of offspring, congenital heart disease and affect human physical function. These problems have brought great challenges to human society. There are still many uncertainties in the study of genetic diseases. Finding the causes of genetic diseases, such as disease-causing genes, has great implications for improving medical care and understanding gene function [5]. When faced with challenges such as imperfect statistics, prediction methods combined with random walks can play some unique roles. This part will discuss the application of random walk in several problems related to genetic diseases. lncRNAs have been proved to be closely related to many diseases that are seriously harmful to human health. However, there are still many unknowns about the relationship between lncrnas and disease. Li et.al proposed a new prediction model based on local random walk, called lrwhlda, to infer the potential association between human lncrna and disease [6]. Lrwhlda contains a new heterogeneous network, which enables lrwhlda to predict in the absence of known lncrna disease associations. Then, this model is combined with an improved local random walk method to predict the new lncrna disease association. The Fig.2 is directed acyclic graph (DAG) corresponding to each disease. Zhao et.al also introduced another predictive model called Random Walk for lncRNA-Protein Associations Prediction (RWLPAP) [7]. It belongs to semi-supervised learning algorithms [7]. The method in Zhao et.al is used for lncRNA-protein associations prediction. Finally, RWLPAP successfully avoids the difficulty of extracting negative data sets and features. Passing through a crossover test, Zhao et.al compared the progress of RWLPAP and previous methods, and the conclusion was that RWLPAP's AUC was 0.88. This expression RWLPAP is

effective in directional interaction between lncRNAs and proteins. This also shows the high accuracy of the RWLPAP. Moreover, the method in Li et.al, which combined with random walk has achieved good success in prediction, and this success also reflects some characteristics of this method. About advance, this method improves the accuracy of prediction and has good performance. About disadvantage, this method has low time complexity. In general, the application of random walk is effective in the prediction of this problem.

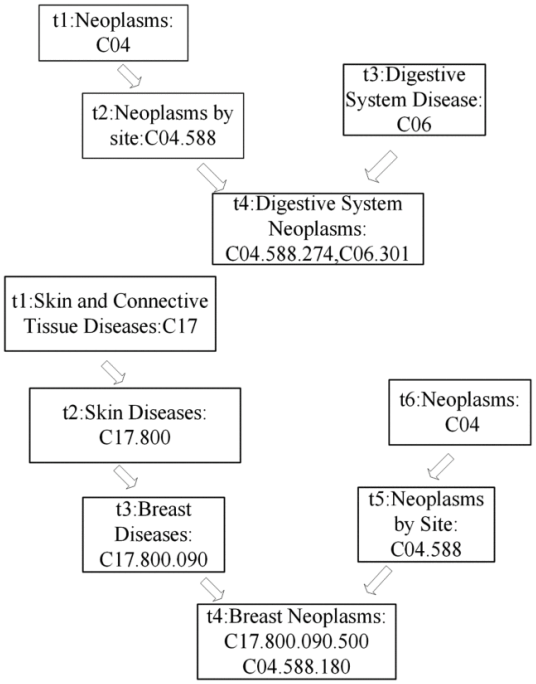


Figure 2. Directed acyclic graph (DAG) corresponding to each disease [6]

The prediction of pathogenic genes of genetic diseases is still a major focus of human health research, but also contains many challenges. Studies in systems biology have shown that many diseases with similar phenotypes are caused by functionally related or closely interacting proteins, which is called the modular nature of human genetic diseases [5]. In order to better explore the pathogenic genes and the methods of predicting genetic diseases, Chunshui Wei studied an algorithm based on the above theory. This is a global network distance analysis algorithm based on the idea of random walk. It is an algorithm for approximate calculation of PageRank vector, which can be completed in linear time [5]. Xu et.al proposed a method called Multipath2vec to predict the pathogenic genes [8]. The Fig.3 shows this method's process. The process is that, first, they construct the human gene-phenotype heterogeneous network. Then in order to embedding the network better, they design the multi-path to guide random walk in GP−network. Finally, they calculate the similarities and then rank the candidate genes. Compared with the four baseline methods CATAPULT, PRINCE, Deepwalk and Metapath2vec, the prediction results of Multipath2vec show higher accuracy. Furthermore, Chunshui Wei's final prediction result of the algorithm shows the following characteristics: first, the phase ratio is high, and the prediction speed is relatively fast. Second, high accuracy. In general, this algorithm combined with random walk performs well [8].

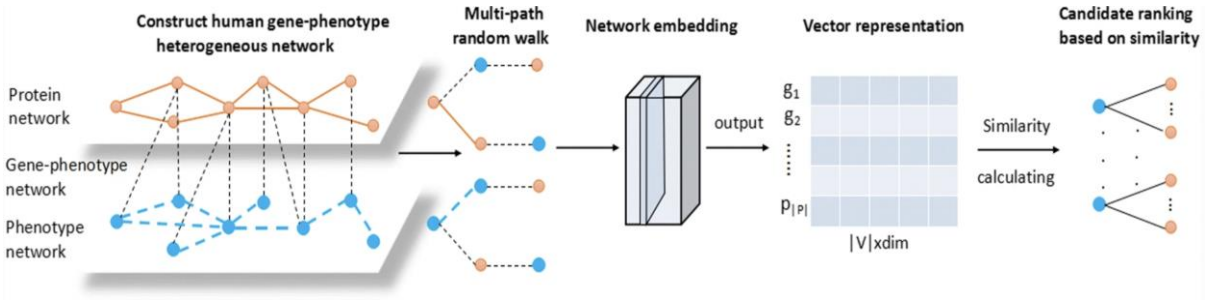


Figure 3. The process of Multipath2vec

Chunshui's research has demonstrated a novel approach based on random walks. Research by Yongjin Li and Jinyan Li presents an alternative approach to random walks in predicting genes. Their approach is achieved by random walk on multigraphs merging heterogeneous genomic and phenotype data [9]. This method includes a data platform that can give priority to disease genes and has strong noise tolerance. Then the idea of random walk is applied to make the modified step process multiple graphs to calculate the transfer matrix [9]. They also face the challenge of bias, incompleteness and noise in the prediction of a single source of genomic data. However, their experimental results show the characteristics of high accuracy.

Through the above several cases, it can be found that the prediction method combined with random walk shows the following characteristics when applied to genetic disease-related problems.

Positive Characteristics:

First of all, the accuracy is high. Several of the above studies have applied random walk methods, such as the research method built on top of heterogeneous networks in Li et.al or the method proposed by Chunshui Wei, which reflects this feature. The method in Zhao et.al and the method in Xu et.al also show the feature too. For instance, the accuracy of the method proposed by Zhao et.al is 0.88.

Besides, the efficiency is better. The result of the prediction method of the global network distance analysis algorithm based on the idea of random walk proposed by Chunshui Wei reflects this feature. This method can be completed in linear time, and reflects the characteristics of high efficiency.

Negative characteristic:

It is the low time complexity. Li et.al designed an improved local random walk method to predict novel lncRNA disease associations, which embodies this feature.

In general, the application of random walk method in predicting genetic disease-related problems shows many positive characteristics, and only one negative characteristic. It can be seen from these cases that the application of random walk has more advantages than disadvantages on these prediction methods. In general, the application of random walk method in predicting genetic disease-related problems shows many positive characteristics, and only one negative characteristic. It can be seen from these cases that the application of random walk has more advantages than disadvantages on these prediction methods.

3.2. Application of random walk to predict cancer

In addition to genetic diseases, cancer has always been a major enemy of mankind. Different cancers have many different types of negative effects on human health, such as weakened resistance, weakened body function, and even death. Human beings still lack good coping methods for many cancers, such as blood cancer. If human beings can make a breakthrough in the research of cancer prediction, society may make great progress. This paragraph will discuss some methods related to random walk to predict cancer-related problems [9-14]. Chenye Wang et al. studied the problem of discovering cancer driver genes through random walks on a gene mutation hypergraph. Cancer genes have caused great trouble to human beings. Many types of cancer are caused by cancer genes or gene mutations. For example, skin cancer, lung cancer, breast cancer. Predicting cancer genes has therefore become a major topic of human research. However, many calculation methods do not make full use of individual co-occurrence mutation information. Cooccurrence mutation information of individuals

is very important in tumorigenesis and tumor progression, resulting in a high false positive rate [10-13]. They proposed a random walk algorithm called DriverRWH on a weighted genetic mutation hypergraph model. This algorithm uses somatic mutation data and molecular interaction network data to prioritize candidate driver genes. The Fig.4 is the random walk on hypergraph. Wei et.al proposed another method based on an improved random walk method [14]. They propose a novel approach named Driver_IRW (Driver genes discovery with Improved Random Walk method). The key idea of Driver_IRW is to assign different transition probabilities for different edges of a constructed cancer-related network in accordance with the degree of the nodes' neighbors [14]. The Fig.5 shows the framework of this method. In the end, this algorithm showed good results. Ultimately, experimental results of this approach suggest that it can help prioritize cancer-related genes with greater accuracy. Furthermore, the method in Chenye Wang et al. is also effective for predicting the sequence of oncogenes in various types of cancers. The experimental results show high accuracy and more balanced characteristics.

Predicting cancer genes is crucial to human health, and the progress of such research will also promote the progress of human society. At the same time, predicting type of cancer is also of great significance to mankind. Predicting type can help people find more suitable treatments. Specifically, people can know their problems, avoid delaying the best treatment time or receiving the wrong treatment, and get more timely and effective treatment. Several methods are now available to predict the type or status of cancer, such as microarray-based genetic biomarkers that have been identified for predicting outcome. However, these methods are indeed stable and accurate.

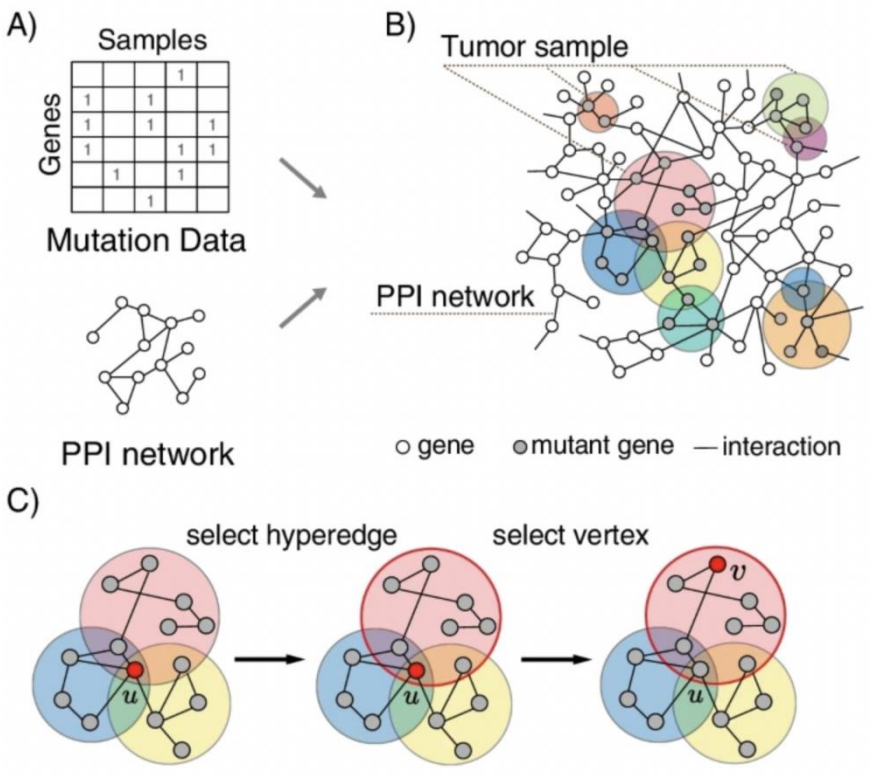


Figure 4. Random walk on hypergraph

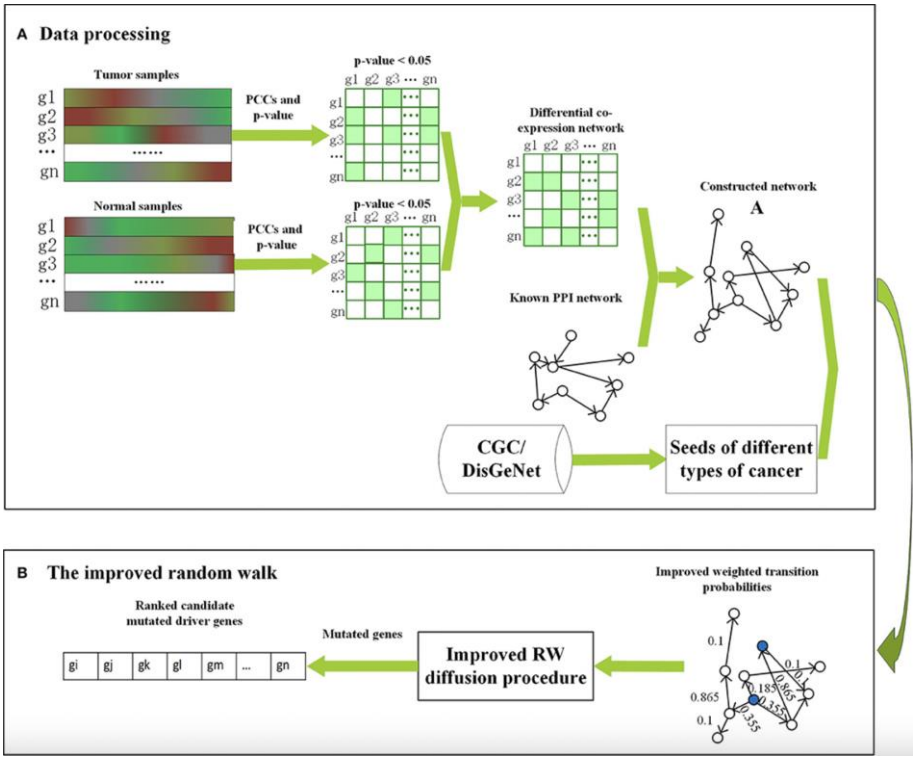


Figure 5. The framework of Driver_IRW for identifying cancer driver genes

For example, machine learning methods predict by inputting statistical data with different weights. Wei Liu et al. proposed a new method. This approach is based on a directed random walk (DRW) approach to infer pathway activity [9]. DRW assesses the topological importance of each gene by capturing structural information embedded in a network of directed pathways. The prediction effect of this method finally shows good results. By experiments on 18 cancer datasets, this method shows that it is relatively more accurate and robust performance compared to several existing gene- and pathway-based classification methods. In general, this method exhibits high accuracy and high stability.

The above methods use random walks. These methods present the following characteristics of random walk application and prediction of cancer-related problems:

The first point is high stability. The directed random walk (DRW)-based method proposed by Wei Liu et al. exhibits this feature when used to infer path activity. Experiments on the 18 cancer datasets mentioned in the paper show that the results are more stable compared to several existing gene- and pathway-based classification methods.

The second point is high accuracy. Several methods [8-12] reflect this feature. In addition to the above 18 experiments in Liu et al., DriverRWH applied to tumor samples of different cancer types in the Cancer Genome Atlas also showed good results. DriverRWH shows significantly better performance genes than the most advanced prioritization methods in terms of the area under the curve score and the cumulative number of known driver genes recovered from the top ranked candidates.

The third point is high balance. This feature is embodied in DriverRWH. The method using random walks has shown more accurate and balanced characteristics in the research of Chenye Wang et al. Overall, the application of random walks in predicting cancer-related problems is effective.

3.3. Application of random walk to predict Microorganism and Disease

Microorganisms refer to those tiny organisms, including bacteria, fungi, viruses, etc. There are many microorganisms in human living environment, and microorganisms are also inextricably linked with diseases. For example, microorganisms can cause meningitis, hepatitis, and other diseases. Many diseases caused by microorganisms also have great harm. The relationship between microorganisms

and diseases is also one of the topics of human research. Nowadays, there are many methods to predict the association between microorganisms and diseases, such as the prediction method based on linear model. Yao Chen proposed a method using random walk in heterogeneous networks. This method forms "microbial communities" through Markov clustering algorithm, and then constructs heterogeneous microbial social networks combined with disease data [14]. It uses heterogeneous network random walk method and correlation ranking to predict disease-related microorganisms. This method makes full use of the connection between microorganisms and disease data. The prediction results of this method show that the prediction results of pathogenic microorganisms of type 2 diabetes, asthma and psoriasis show that the prediction of disease microorganisms based on heterogeneous microbial social networks is more accurate and effective [14]. This method using random walk shows the characteristics of high accuracy. In general, this is an effective application.

4. Conclusion

This paper discusses some applications of random walk for predicting disease-related problems. This includes problems related to predicting genetic diseases, cancer, and the relationship between microorganisms and diseases. The methods of random walk show the following common characteristics: first, high accuracy. Second, high stability. At the same time, the application of random walk in predicting different disease-related problems also shows some different characteristics. There are two characteristics in predicting genetic disease related problems. First, high efficiency. Second, the time complexity is low. It shows a balanced feature in predicting cancer-related problems. In general, random walk plays a positive role in predicting disease-related problems.

Disease is a major problem of mankind today. In the long history of human civilization, disease has always been a hot topic. If the methods of predicting diseases can be improved, human health will make great progress. Nowadays, there are many methods to predict disease-related problems, but many of them ignore the randomness of disease. This also leads to the lack of accuracy and stability of the prediction results. The appropriate application of stochastic process related technology may improve the performance of some prediction methods. This paper also reflects many positive aspects of the application of random walk for predicting disease-related problems. There are many related technologies or formulas of stochastic process, random walk is not necessarily the most suitable technology for predicting disease-related problems, but we hope that humans will notice the benefits of random process or random walk for predicting diseases in the future. It is hoped that human beings can create more methods of disease prediction integrated with random process related technology in the future. It is hoped that the accuracy and stability of human disease prediction can be improved, and more people can prevent diseases in advance or receive timely and effective treatment. The quality of human life may therefore be improved, and human civilization will therefore be more advanced.

References

[1] Yingying Yuan, Jiancheng Dong. A comparative study of disease prediction methods based on mathematical models. Software Guide, 2009, 8(5): 108-110.

[2] Moritz Kohls, Tanja Hernandez. Expected Coverage of Random Walk Mobility Algorithm. German: Technische Universität Dortmund, Department of Electrical Engineering and Information Technology, 2016.

[3] Wirth, E. ; Szabó, G. ; Czinkóczky, A. Measure Landscape Diversity with Logical Scout Agents. ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016, 2: 491-495.

[4] Karl Pearson. The Problem of the Random Walk. Nature, 1995, 72: 294-301.

[5] Chunshui Wei. Pathogenic gene prediction algorithm based on random walk. Xidian University, 2013, (3):1-11.

[6] Jiechen Li, Haochen Zhao, Zhanwei Xuan et.al. A Novel Approach for Potential Human LncRNA-Disease Association Prediction Based on Local Random Walk. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(3): 1049-1059.

[7] Yongjin Li, Jinyan Li. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. BMC Genomics, 2012, 13 Supplement 7.

[8] Chenye Wang, Junhan Shi, Jiansheng Cai et.al. DriverRWH: discovering cancer driver genes by random walk on a gene mutation hypergraph. BMC Bioinformatics, 2022, 23: 275-281.

[9] Wei Liu, Chunquan Li, Yanjun Xu et.al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics, 2013, 29(17): 2169-2177.

[10] Han Zheng. From random walk to bacterial foraging optimization algorithm (BFOA), and then to random walk Graph Segmentation Algorithm. Developer's online home, 2022.

[11] Qi Zhao, Dan Liang, Huan Hu et.al. RWLPAP: Random Walk for lncRNA-protein Associations Prediction. Bentham Science Publishers, 25(9): 830-837.

[12] Bo Xu, Yu Liu, Shuo Yu. A network embedding model for pathogenic genes prediction by multi-path random walking on heterogeneous network. BMC Medical Genomics, 2019, 12.

[13] Pijing Wei, Fangxiang Wu, Junfeng Xia. Prioritizing Cancer Genes Based on an Improved Random Walk Method. Front. Genet, 2020, 2: 388-397.

[14] Yao Chen. Microbial disease association prediction based on microbial social network and random walk strategy. Central China Normal University, 2018, (2).