学校编码：10384　　分类号＿＿＿密级
学　号：　23020151154781　　UDC

厦门大学

硕　士　学　位　论　文

人类疾病相关的 microRNA 预测研究

Computational Prediction of Human Disease-Related MicroRNAs by Path-Based Random Walk

MUGUNGA ISRAEL

指导教师姓名：曾湘祥副教授

专　业　名　称：计算机科学与技术

论文提交日期：2017 年 5 月　日

论文答辩时间：2017 年 5 月　日

学位授予日期：2017 年　5 月　日

答辩委员会主席：

评　阅　人：

2017 年　　5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　月　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　月　日解密，解密后适用上述授权。

（　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　月　日

Contents

List of Tables

List of Figures

Abbreviations

MiRNAs: MicroRNAs

C. elegans: Caenorhabditis elegans

RDnet: MicroRNA–Disease network

SVM: Support Vector Machines

DAG: Directed Acyclic Graph

MISIM: MiRNA Similarity

PPI: Protein-Protein Interaction

ROC: Receiver operating characteristic

AUC: Area Under the Curve

NGS: Generation Sequencing

DNA: Deoxyribonucleic Acid

RLSMDA: Regularized Least Square for MiRNA–Disease Association

UTR: Untranslated Region

MIDP: MicroRNA associated with Disease Prediction

HMDD: Human MicroRNAs Disease Database

摘 要

MicroRNAs（miRNAs）是一类非编码且有调控功能的小分子 RNA。它主要在后转录阶段调控基因的表达，同时，对疾病发病机理的研究有着重要意义。识别与疾病相关的 miRNA 是研究疾病致病机理的基础。生物实验是一种识别 miRNA 是否与疾病相关的主要的方法。由于生物方法的一些缺点（如实验成本高、周期长），现已经有许多研究者通过不同的计算方法去发现 miRNA 与疾病的关联，从而辅助生物实验研究。

使用计算方法去预测潜在的疾病与 miRNA 关联是一种直接且有效地方法。但现在只有少量预测算法可以达到较好的预测准确率，这也是面临的主要问题。因此，本文提出了基于路径的随机游走算法预测潜在的 miRNA 与疾病关联的方法，从而去克服计算方法的研究挑战。除此之外，本文还从 miRNA 与疾病的数据网络中提取相似性，并将其作为 miRNA 与疾病的特征值。基于随机游走的预测算法，游走点通过计算 MiRNA 与疾病向量的相似性，从疾病向量游走到 miRNA 向量。从而对于任意给定疾病，可计算 miRNA 与该疾病的相似性得分，且得分越高的 miRNA 是越可能与该疾病相关。

最后，我们将所提出的算法应用在两种肿瘤数据集中进行 miRNA 的预测。我们将乳腺肿瘤和胰腺肿瘤作为我们的研究案例，从而检验算法的预测效果。跟已有的模型相比，我们的预测模型取得了更优的预测效果。

关键词：疾病相关的 miRNA;计算预测方法；随机游走；相似性得分；系统生物学

Abstract

MicroRNAs (miRNAs) have been discovered as important genetic regulation of genes expression in animals and plants. MiRNAs are a class of very little non-coding regulatory RNA molecules that modulate the expression of several genes at the post-transcriptional level and play a critical role in disease pathogenesis. Discovering the relationship between diseases and miRNAs is fundamental for understanding the pathological process of diseases. Therefore, biological examination is a major method of recognizing whether miRNAs are related to any disease. However, this method presented bottlenecks (e.g. time-consumption, and high cost) due to big data from different databases that render it complex. Beforehand, many researchers performed different computational methods to discover relationship between diseases and miRNAs to assist in biological tests. Computational techniques to predict potential disease-related miRNAs is the only immediate way to overcome the presented difficulties. Nevertheless, one main problem for computational methods is the lack of enough bioinformatics methods that predict potential miRNA–disease associations with a high degree of accuracy.

Therefore, in this thesis, we propose a computational prediction method of human disease-related miRNAs by path-based random walk to predict potential candidates of disease-related miRNAs to overcome challenges stated in this research area. Moreover, construction of disease–miRNA networks to come up with the similarity between diseases and miRNAs evolved as features extraction between miRNAs and diseases. Based on random walk, the walker moves from each disease's vertex of the disease's network to miRNA vertices by calculating the similarity score between disease–miRNA vertices. As a result, for a given disease with miRNAs, the similarity score has been calculated and the miRNAs with higher similarity scores were confirmed as the potential candidates for a given disease.

Lastly, we apply our method to two different types of cancer datasets; Breast Neoplasms, and Pancreatic Neoplasms as our case study to assess the performance of our method. Path-based random walk gave a better prediction accuracy compared to previous models which will contribute to biological investigations and help future researchers to overcome the major issues in this research area.

Keywords: disease-related miRNA, computational prediction method, random walk, similarity score, system biology.

CHAPTER 1: INTRODUCTION

1.1 Background

MiRNAs are a category of little, endogenous RNAs that are 21~25 nucleotides long. MiRNAs can be originate in plants, animals, various infections, and function in RNA quieting and post-transcriptional directive of organic phenomenon [1, 2]. MiRNAs are concerned in several different biological manners, like apoptosis, growth, differentiation, and virus-related infection [3]. Increasing confirmation implicates miRNAs in human cancer growth, development, prognosis, diagnosis, and appraisal of treatment response [4-6]. Since the first miRNA (lin-4) was discovered from *C. elegans* twenty years ago, many miRNAs are annotated in varied species with experimental and computational methods. An identification of miRNAs that underlie human diseases is a crucial goal of medical specialty researchers. However, the discovery of disease-related miRNA via existing biological methods is costly and time-consuming [7]. Therefore, computational prediction models are significant techniques for identifying the most likely miRNA–disease associations prior to additional experimental examinations. On the other hand, one main challenge with miRNA researches could be the absence of enough bioinformatics techniques with efficiency accuracies to forecast prospective disease-related miRNAs.

To overcome these major issues, we propose a computational predictive method to predict potential disease–miRNA relationships. Our proposed technique is built on path-based feature and random walk to obtain a relevancy score between disease-related miRNAs. For this reason, several dataset have been combined to build (RDnet) network between miRNAs and diseases where the path-based to random walk applied to rank every miRNA of a given disease. Thus, the walker starts walking from each known disease-related miRNAs with equal probability and the walker endures until it reaches unknown disease-related miRNAs. Finally, the walker stays at the unknown miRNA vertices to measure the similarity score. The obtained similarity score values were ranked. The highly ranked scores are potential candidates of disease–miRNA associations.

1.2 Motivation

In previous years, biological conducting tests are the most methods accustomed to recognize whether or not miRNA is connected to a given disease. Through growing of

biological data and the discovered of new miRNAs each year, experimental methods to discover a disease-related miRNAs presents significant challenges (e.g. amount of time and cost) [7]. Due to these difficulties of experimental method to identify an association between diseases and related miRNAs, computational prediction approaches have suggested [7]. Many computational methods have presented to predict the relationship of miRNA–disease associations. Generally, the prediction of possible disease-related miRNA built on the networks is to figure out the similarity between disease and miRNA to the networks [7]. These problems have been motivated us to contribute on this research field.

Therefore, in this thesis, we propose a computational prediction method of human disease-related miRNAs by path-based random walk to conquer these challenges. Based on random walk, RDnet network was constructed, the walker starts walking from disease nodes with equal probability to disease-related miRNAs. In addition, form known miRNA nodes, the walker is walking to its neighboring unknown nodes until converged. A given disease used as a query to set the relationship score of all known and unknown miRNAs. Then, the similarity score was computed as the time the walker spends to unknown miRNAs. If the node is a disease-related miRNA, the walker continues, stays otherwise. Therefore, we extracted all unknown miRNAs to a given disease for ranking and the higher similarity score ranks have confirmed as the potential candidates of disease-related miRNAs.

1.3 Contribution

From the best of our knowledge a lot of work has done to predict disease–miRNA associations in the field of Data Mining and Bioinformatics in the last two decades. However, there are some drawbacks in the previous presented methods such that, there is a low quality of the datasets in different models which results a critical performance. In addition, there are methods evaluated by using only disease–miRNA similarity, miRNA similarities, and disease similarities by utilizing disease information only, which possibly effect in an unfairness for disregarding the miRNA features. Therefore, our contribution is to combine different datasets for a disease–miRNA network which is helpful in an integration and extraction of useful features. As a result, we present a computational predictive technique to predict possible disease–miRNA associations by path-based random walk which achieves a great accuracy performance compared to previous methods. Our computational predictive method will help in future research as

there are a lot of miRNAs with no known related disease as well as a few number of diseases that have no any related miRNA.

1.4 Problem definition and research objectives

After discovering that miRNAs have impacts to immune system as positive or negative effects, discovering new and predicting potential miRNAs candidate's related disease have been a research topic for last two decades. Previously, to predict new miRNAs has been conducted by biological test. Nevertheless, a number of new miRNAs have been discovered each year to be associated with disease which affect this technique to face with more challenges, such as a time and cost used to discover disease-related miRNAs [7]. Therefore, computational methods have been proposed to support biological tests. Although computational methods have suggested, there is still some demanding; lack of enough computational methods has shown to be one the most challenging issue with these methods.

Therefore, our research objective is to come up with computational predictive method for disease-related miRNAs with efficiency and high degree of accuracy. We propose a method which is concerned on graph feature and path-based random walk to calculate the similarity scores between diseases to miRNAs. In addition, all unknown miRNAs for a given disease have been extracted for prediction. We therefore ranked all miRNAs' similarity scores and the highly ranked scores have been confirmed by public databases as potential candidates of disease–miRNA associations.

1.5 Structure of the thesis

Chapter 1 presents the background of our thesis to build a computational predictive method of miRNAs and diseases associations which is a path-based random walk and graph theory as our research topic. Furthermore, we discuss what motivated us to work on this topic and our contribution to this research area. In addition, we describe our problem and research objectives of research work. The remaining of this thesis is structured as follow.

Chapter 2 provides a review about miRNAs, the discovery of the first two miRNAs and the role of miRNAs in human genes as well as their effects to human diseases. Besides, we discuss biological investigations as the only technique was used to determine the association between diseases and miRNAs. However, biological method has presented challenges such as the time taken to examine disease-related miRNAs and resources to

discover them. Lastly in this chapter, we discuss previous computational prediction methods of disease–miRNA associations in general as the basis of this thesis to overcome the challenges that have been presented by biological investigations.

Chapter 3 presents the methodology used in our thesis; we first describe all terms used in our thesis to make a better understanding of the research work such us; graph theory and random walk. In this chapter moreover, we deliberate the datasets utilized to perform our method. On the other hand, we define feature selection as one of the most important aspects of designing a classifier for computational prediction methods. Additionally, in this section, we define our miRNAs and diseases network (RDnet) derived from various datasets and we finally describe our method which is a path-based random walk and graph theory as the core contribution of our work to this research area.

Chapter 4 discusses supervised machine learning algorithms for classification, such as SVM (Support Vector Machines) [86] and Naïve Bayes classifiers. Furthermore, these algorithms helps in classification and prediction to various useful datasets. In this chapter also, we discuss the experimental results of our method by applying these classifiers to the extracted datasets by our method, to classify and predict miRNAs and diseases associations. To conclude this section therefore, we discuss about a case study by applying two different types of cancer to our method to verify its performances.

Chapter 5 presents the conclusion of our research work and last but not least, we discuss the challenges that we have faced in our research as well as the future research work.

Lastly in this thesis, we present our academic achievements and we acknowledge all people who helped us in this research work.