miRNA-disease 关联预测方法研究

李金金

指导教师 曾湘祥 副教授

厦门大学

学校编码：10384　　分类号_____密级_____

学　　号：23020131153186　　UDC_____

厦门大学

硕士学位论文

miRNA-disease 关联预测方法研究

The Research of MiRNA-Disease Correlation Prediction

李金金

指导教师：曾湘祥副教授

专业名称：计算机技术

论文提交日期：2016年5月

论文答辩日期：2016年5月

学位授予日期：2016年　月

答辩委员会主席：_____

评　　阅　　人：_____

2016年　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）：李金金

2016 年 5 月 17 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）：王金金

2016 年 5 月 17 日

摘　要

MiRNA 是一类重要的非编码 RNA，它在许多的生命活动中扮演着非常重要的角色。经研究证实，miRNA 与 disease 的产生有着密切的联系。本文主要基于生物异构网络和机器学习的方法对 miRNA-disease 关系进行分析与预测。主要包括以下几个方面的内容：

(1) 本文详细介绍了当前已提出的 miRNA-disease 关联预测的主要方法，并且对比了不同方法之间的差异，为生物信息研究人员后续对 miRNA-disease 关联分析和预测的研究提供参考。现已提出的 miRNA-disease 关联分析和预测方法，大多利用 miRNA 的生物特性构建 miRNA 之间的相似性以及利用文本挖掘的方式研究疾病之间的相似度，从而进一步研究 miRNA-disease 之间的关系。

(2) 引入并改造了成功应用在社交网络中预测社交关系的方法：KATZ 和 CATAPULT 方法，来解决 miRNA-disease 关联分析与预测的问题。KATZ 方法是基于生物异构网络的方法，充分利用异构网络中 miRNA 节点与 disease 节点之间不同步长的不同到达方式的预测方法。CATAPULT 是基于半监督学习的方法，利用偏向的支持向量机预测 miRNA-disease 的关系。然而，对于 miRNA-disease 关联对，正的关联关系是确定的，但是反的关联关系是不确定的。因此，对于问题的这种特性，引入了偏向的支持向量机来做支撑，以准确有效地预测 miRNA-disease 的关联关系。

(3) 提出了基于集成学习的 miRNA-disease 关联关系预测方法。对于 miRNA-disease 关系预测的研究已经涌现了多种方法，但是，较为流行的机器学习方法并没有在 miRNA-disease 关系预测上发挥作用。本文综合多种现有的预测 miRNA-disease 关系的方法作为 miRNA-disease 对的特征向量，再利用集成分类器对 miRNA-disease 对进行分类。

关键词：miRNA-disease 关系；生物异构网络；支持向量机

Abstract

MiRNA is an import class of noncoding RNA and plays a key role in regulating lift process. Researches have shown that miRNA plays an import role in the occurrence and development of disease. In this paper, we analysis and predict the associations between miRNAs and diseases based on the biological heterogeneous network and machine learning method. The main contents are as follows:

(1) In this paper, we introduce in detail the methods that have been proposed to predict the associations between miRNAs and diseases and compare the differences between the methods, which provide the important reference information for the follow-up study of biological information personnel who are interested in the miRNA-disease associations' research. The majority of methods use text mining to calculate the diseases similarity and biological characteristics of miRNA to obtain the miRNA similarity, then, miRNA-disease associations are calculated based on diseases similarity and miRNAs similarity. For example, family and cluster characteristics of miRNA are often used.

(2) We introduce and reform the methods, which succeed in predicting social network association. KATZ based on biological heterogeneous network and synthesizes the different step sizes and different modes of arrival between miRNA and disease in heterogeneous network. CATAPULT based on semi supervised learning method and uses the biased support vector machine to classifier miRNA-disease pairs. For the miRNA-disease pair's classification problems, the positive correlations are determined, but the negative correlations are uncertain. Taking into account the characteristics of the problem, we introduce biased support vector machine to analysis and predict the associations between miRNAs and diseases effectively and accurately.

(3) We proposed the prediction method based on ensemble learning. A number of approaches have emerged for miRNA-disease association problem, but more popular machine learning method does not play a role in the issue. In this method, we

integrated multiple methods predicting the miRNA-disease association to act as the features. This not only synthesizes the advantages of various proposed methods, but also introduces the machine learning methods.

Key words: miRNA-disease association; Biological heterogeneous network; SVM

目 录

CONTENTS

第一章 绪 论

1.1 课题研究背景及意义

1.1.1 课题研究背景

1985 年，美国科学家提出了一个大规模跨国跨学科的科学探索工程-人类基因组计划。人类基因组计划是生命科学中的一个重大课题，它的目的在于获得人类基因组完整的核苷酸序列，用于鉴别人类基因组中的所有基因。人类将借此重大的科学计划破译人类的生命，解读人类自身的奥秘。人类基因组计划于 1990 年 10 月 1 日正式启动，预计人类基因组计划 15 年内将耗资 30 亿美元解读人类基因。2000 年 6 月，人类基因组计划参与国的协作组宣布：人类生命蓝图已基本完成。公告意味着人类从自己基因的角度对人类生老病死的客观规律、疾病诊断和治疗有了更深入的了解和掌握，对人类生命的起源有了更准确的认识。同时，标识着人类在摸索生命奥秘的道路上树立了新的里程碑。

人类基因组计划的完成，意味着全新的“后基因时代”即将开启。迄今为止，人类已经拥有了大量的生物数据，加上大数据时代的带来，生物数据仍然在迅猛地增长。从海量的生物数据中挖掘出生命的信息是人类基因组计划的真正目的。后基因组时代即将遇到的问题主要包括：基因组的表达与调控、蛋白质产物的功能、基因组的多样性、模式生物基因组的研究等。后基因组时代的研究将让人类更加深入地了解遗传语言功能的逻辑架构；基因的功能与结构之间的关系；人类生长、发育、衰老和死亡的原理；脑功能的表现与神经活动之间的关系；疾病的发生与发展；信息传递和作用机理；基因后机理以及关于人类生命科学的各种问题等。人类基因组研究的目的就是从根底开始探索人类生命的起源；人类物种间以及不同个体之间的差别；导致疾病发生与发展的原因；长寿与衰老的原因等。美国国立卫生院和能源部实施了两个重大科研计划是：基因组到生命(Genomes to Life, GTL)[1]与 Roadmap。GTL 科研计划的基础是准确地刻画出生命系统的所有“分子机器”，认清“分子机器”在生命体中是如何协调工作的。其目标是：“分子机器”的辨别，其中，分子机器是蛋白质的复合物产物，其扮演着执行生命系统的功能；充分地认识与理解调控“分子机器”的原理；发展计算机技术以