学校编码：10384　　分类号＿＿＿密级

学号：23020141153176　　UDC＿＿＿

硕　士　学　位　论　文

基于生物异构网络的致病基因预测方法

Prediction Method of Pathogenic Genes on Heterogeneous Biological Network

丁宁翔

指导教师姓名：曾湘祥副教授

专　业　名　称：计算机技术

论文提交日期：2017 年 5 月

论文答辩时间：2017 年 5 月

学位授予日期：2017 年　月

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）：

20　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）：

20　　年　　月　　日

摘 要

预测生物异构网络中的致病基因，是生物信息学领域中的一个重要研究课题。如何更准确地预测遗传疾病的致病基因，对于检测药物靶基因、改善医疗措施、延长病人存活周期与实施生物实验都具有重大的指导意义。

随着计算机和人工智能技术的飞速发展，生物信息学的研究也迈入新阶段，不同类型的生物数据被收集公开，并被广泛应用于致病基因预测问题。在当今的生物学领域，对于人类的遗传疾病，探究致病基因的主要方法有以下三种：定位克隆、定位候选和非定位候选，随着人类基因组计划的完成和海量生物数据的出现，定位候选策略逐渐成为探究致病基因的主要方式。随着大规模的基因和疾病相关数据的出现，我们可以利用链接预测的算法来完成致病基因的预测。

本文主要针对生物异构网络中的致病基因预测问题进行数据挖掘研究。具体地，我们做了如下两项工作和创新：

（1）通过非人类同源基因数据，重新构建了生物异构网络，并提出了基于概率分布的协同过滤预测模型。模型假设在相同的特征空间里，如果两个节点之间的欧几里得距离越近，则两个节点越相似。基于这个假设，将基因与人类疾病之间的关系预测通过概率分布转化为二分类概率模型。为了提高预测的准确性，我们还加入了更多的先验信息，增加了两项不同的约束，形成两个改进的模型。为了验证模型预测的准确性，我们在真实的生物数据上进行了多组实验，并与现有的预测算法进行对比，分析算法性能。

（2）针对生物异构网络中已知负样本缺乏的问题，我们提出了基于 PU Learning 的协同过滤预测模型。该模型将生物异构网络中的致病基因预测问题转换为推荐系统中的推荐问题，在使用归纳型矩阵填补方法进行预测模型学习的基础上，加入了 PU Learning 的方法。该方法解决了生物网络中负样本缺失的问题，并能够对于训练过程中未出现的人类疾病与基因进行预测。

关键词：生物异构网络；致病基因预测；协同过滤

Abstract

It is an important research topic in bioinformatics to predict the pathogenic genes on heterogeneous biological network. The prediction of genes related to genetic diseases is of great significance for the discovery of drug target genes, the improvement of medical care, prolonging the life cycle of patients and the implementation of biological experiments.

With the development of computer technology and artificial intelligence, the research of bioinformatics have entered a new stage, many types of human biological data have been discovered and published, which play a more and more important role in the prediction of pathogenic genes. There are three main strategies for predicting gene–disease associations: positioning candidate cloning strategy, positioning strategy and non-positioning strategy. With the completion of Human Genome Project; Human Genome Project and the development of the related biological data, positional candidate strategy has gradually become the main method to find the causative gene. With the advent of large-scale genetic and disease related biological data, we can use the link prediction algorithm to complete the prediction of pathogenic genes.

In this dissertation, we focus on predicting related genes for diseases on heterogeneous biological network with data mining approaches. Specifically, we do the following two tasks and innovations:

(1) Based on the data of other nonhuman homologous genes, this dissertation constructs a heterogeneous network and proposes a probability-based collaborative filtering model for predicting gene–disease associations. We assume that in the same feature space, if the Euclidean distance between two nodes is closer, they get more similar. Based on this hypothesis, the relationship between diseases and genes was changed into a two classification problem by probability distribution. In order to improve the accuracy of prediction, we also add a lot of prior information, and add two different constraints to form the two improved models. In order to check the

effectiveness of the proposed model, we make a number of experiments on real biological data, and compare with the existing prediction algorithms to analyze the performance of this model.

(2) In order to deal with the lack of known negative samples on the heterogeneous biological network, we propose an PU Learning-based matrix decomposition model. With this model, we consider predictions on heterogeneous biological network as a recommendation problem on recommender systems; we add the Learning PU method to the inductive matrix completion model. The proposed model solves the problem of the lack of the negative samples on biological network, and can be used to predict novel pathogenic genes which are not on the training sets.

Key words: Biological Heterogeneous Network; Prediction of Pathogenic Genes; Collaborative Filtering

目 录

Contents