学校编码：10384　　分类号___密级___
学号：23120090153690　　UDC___

厦门大学

博 士 学 位 论 文

基因芯片表达数据分析及基因调控网络模型研究

Data Analysis of Expression with Gene Microarray and Investigation for Gene Regulatory Networks

郭顺

指导教师姓名： 王守觉院士
郭东辉教授
专 业 名 称： 电路与系统
论文提交日期： 2017 年 月
论文答辩时间： 2017 年 月
学位授予日期： 2017 年 月

答辩委员会主席：_____
评　　阅　　人：_____

2017 年

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

摘 要

基因芯片表达数据可以揭示在各种不同条件下基因的活动情况及基因之间相互作用关系等，因此，对该数据的分析有着非常广泛的应用前景。近年来，随着基因芯片表达数据快速增长，传统的分析方法已经不能适应现代生物迅速发展的要求，迫切需要更加高效的方法来分析处理这些数据。

本文首先简述了基因芯片表达数据分析的关键技术及其研究进展，指出了现有方法存在处理冗余和噪声数据能力不足、结果缺乏生物学解释以及调控网络识别率有限等问题。针对这些问题，我们研究了基因芯片表达数据分析三种新方法，包括：基因芯片表达数据的特征选择方法、维度约简方法以及重构基因调控网络方法，为建立分类模型辅助诊断以及识别疾病相关基因等提供了解决方案，为生物进程分析、发现潜在药物靶点以及基因功能预测等提供了有效路径。本文的主要工作和创新点如下：

1. 提出了基于核函数为类分布中心的特征选择方法，即以类的分离性准则来定义求解的目标函数，给出具有较好的处理噪声和冗余数据的数据分析方法，提高特征选择的分类识别率，并且证明其在初始值不为零的条件下可获得全局最优解的结论；
2. 提出了基于 l_1 正则化的两阶段的局部维度约简方法，该方法不仅可以提高基因的分别识别率，且所获得的满足生物学解释的结果，其优点是所使用特征选择方法可去除冗余和不相关特征，并可实现对所筛选的特征进行具有生物学解释的特征提取；
3. 提出基于偏最小二乘(PLS)的基因调控网络重构方法，该方法将基因调控网络的模型重构问题分解成多个特征选择的子问题，每个子问题可通过基于 PLS 的特征选择方法来求解，并结合误差统计来进一步优化重构的调控网络，提高调控网络重构的准确率。

最后，基于本论文所提的技术与方法，我们设计一个基因芯片数据分析系统软件，该系统主要包含特征选择、维度约简以及重构调控网络这三种功能模块，可以直观的反映出本文工作成果，体现了本课题研究的实用性。

关键词：基因芯片；数据分析；特征选择；特征提取；调控网络

Abstract

The gene microarray data is able to reveal the gene activity in various conditions and the interactions between different genes, therefore, it has a very broad application prospects for analyzing gene microarray data. Recently, the amount of gene microarray data is growing faster than the rate at which it can be analyzed and more effective techniques and methods are needed to analysis these data.

In this dissertation, we introduce the key techniques and research progress of analysis of gene microarray data, and point out that many existed methods have some weakness including low capability of dealing with redundant and noisy data, unexplained mining results and limited accuracy of inferred regulatory gene networks. In light of these problems, we investigate three new gene microarray data mining methods, including feature selection and dimension reduction methods for gene microarray data, and the methods for the reconstruction of gene regulatory networks, which provide solutions for establishing classification model and identifying disease genes, and provide an effective way for biological process analysis, identifying the potential drug targets and predicting the gene function. The main work and innovation of this dissertation are as follows:

1. Proposing a centroid-based feature selection method. Theoretical analysis indicates that the global optimal solution of the proposed method can be reached with a non-zero initial point. In the proposed method, a kernel-based approach is used to estimate the class centroid to define the class separability criterion, based on which the objective function is formulated;
2. Proposing a two-stage l_1-regularized local dimension reduction method, which firstly uses a feature selection method to remove the redundant and irrelevant features, and then implements feature extraction on the selected features. The proposed method not only improves the classification accuracy, but also can generate biologically interpretable results;
3. Proposing a PLS-based method for the reconstruction of gene regulatory networks (GRNs). The proposed method decomposes the GRNs inference problem into multiple feature selection subproblems, each of which is solved by a PLS-based

feature selection method. Then, a statistical technique is used to refine the predictions of the inference network.

Finally, Based on the technologies and methods proposed in this dissertation, a Gene Microarray Data Analysis System is designed and implemented. The system contains three main functional modules, including Feature Selection, Dimension Reduction and Gene Regulatory Network Inference, which reflect the results of our work and the practicality of this research.

Key Words: Gene Microarray Data; Data Analysis; Feature Selection; Feature Extraction; Gene Regulatory Networks

目 录

Table of Contents