学校编码：10384　　分类号_____密级_____
学号：23220071152866　　UDC _____

厦门大学

硕士学位论文

基于 Tiling Array 的植物基因结构分段与表达差异分析

Segmentation and Expression Analysis of Plant Tiling Array Data

武姗姗

指导教师姓名：吉国力 教授
专业名称：系统工程
论文提交时间：2010 年 5 月
论文答辩日期：2010 年　月
学位授予日期：2010 年　月

答辩委员会主席：_____
评　阅　人：_____

2010 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）：

2010 年 5 月 30 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）：

2010 年 5 月 30 日

摘 要

在对植物基因芯片数据分析的相关研究领域中，基于 Tiling Array 芯片技术的植物基因结构分段和基因表达差异分析是当前生物信息学的研究热点之一。作为迄今为止分辨率最高的基因芯片类型，Tiling Array 芯片技术已经广泛应用于生物学和医学的各个研究领域，成为目前研究人类基因组和其他各种模式生物基因组复杂性的最强有力的工具。

本文是基于实验室课题组与美国迈阿密大学李教授（Dr.Q. Quinn Li）实验室的项目合作对植物选择性多聚腺苷酸化相关问题的研究成果，应用计算机和数学算法对植物基因结构分段和基因表达差异问题进行分析和研究。论文首先探讨了 Tiling Array 芯片的设计模式及其独特的 PM-MM 探针选取方法以及在此基础上发展而来的一些 Tiling Array 经典信号识别算法。然后针对本课题的实际情况，经过合理的实验设计、试验流程，设计了一个完整的基于拟南芥 Tiling Array 芯片数据分析的数据处理流程，一步一步描述了从拿到芯片原始探针浓度数据后应该如何进行数据处理。应用了包括 R 语言、 Bioconductor 平台、matlab 生物信息工具箱等一系列适合分析 Tiling Array 芯片数据的计算机分析平台、软件和数学算法，对拟南芥 Tiling Array 芯片进行了数据处理，特别是在拟南芥数据预处理和标准化方面做了重点介绍，如为消除非特异性杂交干扰影响的 VSN 背景校正、为消除不同样本之间系统偏差所进行的 Quantile 芯片间归一化、DNA 参考标准化、为从生物学角度上更好的解释及使数据满足特定分布的数据转换、探针过滤等步骤，并给出了对应算法所进行的芯片数据处理结果。最后通过 R、SQL Server 数据库将构造后的探针数据导入到 Matlab 分段识别算法中进行分段识别并运用统计学分析方法进行差异表达基因的挑选。实验结果表明，本文基于 Tiling Array 芯片数据的拟南芥基因分段和基因表达差异分析方法是可行和有效的。

关键词： Tiling Array；拟南芥；基因分段；差异分析

Abstract

Segmentation and expression analysis of plant Tiling Array data is an interesting but challenging problem. As an important branch of gene chips, which has the highest resolution of all gene chips by far, Tiling Array chip technology has been widely used in all research areas of biology and medicine, and is the most powerful tool for the study of the human genome and the complexity of other genomes of various model organisms.

Cooperating with Dr.Q.Quinn Li at the Department of Botany, Miami University, and based on the research results of differentially expressed gene, this thesis is to research on the problem of segmentation and differentially expression gene using computer and mathematic algorithm. In this article, we describe a whole data processing framework based on Arabidopsis thaliana Tiling Array, introducing how to deal with the raw Tiling Array probe data. First we discussed the Tiling Array chip design patterns and unique PM-MM probe selection method, and some classical signal recognition algorithm of Tiling Array. Then according to our actual condition, we designed a analysis procedure for Arabidopsis thaliana Tiling Array data, including the application of a series of computer analysis softwares, such as R, Bioconductor platform, matlab bioinformatics toolbox, and applied them to the actual array data to do VSN background correction to eliminate the interference effects of non-specific hybridization, Quantile normalization to eliminate systematic bias between different samples, DNA reference normalization, Log transformation and probe filtering and gives the corresponding results of each algorithm. Finally, we conducted a segmentation analysis of the data, selecte the differentially expressed genes, using the sub-recognition algorithms and statistical analysis, and then compared different algorithms' results. Experimental results showed that the methods we used on Arabidopsis Tiling Array data are feasible and effective.

Key Words: Tiling Array; Arabidopsis Thaliana; Segmetation; Expression Analysis

目　录

Contents

第一章 绪论

1.1 引言

我们正处在一个激动人心的时代，科技的进步已使人类可以窥视生命的秘密。2000 年 6 月 26 日，美、英、日、法、德、中 6 个国家 16 个研究中心联合宣布完成覆盖人的大部分基因组、准确率超过 90%的 DNA 序列人类基因组“工作框架图”，这部由 30 亿个字符组成的人类遗传密码本已活生生的摆在了我们面前。于此同时，来自其它生物的基因组信息源源不断从自动测序仪中涌出，堆集如山，浩如烟海。这些海量的生物信息是用特殊的“遗传语言”——DNA 的四个碱基字符（A、T、G 和 C）和蛋白质的 20 个氨基酸字符一一写成。从这一时刻开始，人类历史进入了一个崭新的时代——后基因组（post-genome）时代[1]。

原始的生物信息资源挖掘出来后，生命科学工作者面临着严峻的挑战。数以亿计的 ACGT 序列中包涵着什么信息？基因组中的这些信息怎样控制有机体的发育？基因组本身又是怎样进化的？人们迫切希望可以研究出这些分子数据所具有的丰富内涵，以及其背后隐藏着的尚不为人所知道的生物学知识，这就形成了以基因组为主要研究对象的生物信息学[2][3]。《科学》（Science）在 2001 年 2 月 16 日人类基因组专刊上配发了一篇题为“生物信息学：努力在数据的海洋里畅游”的文章，文章写道：“我们身处急速上涨的数据海洋中…，我们如何避免生物信息的没顶之灾呢？一叶轻舟也许可以救命！生物信息学便是我们找到的这样一条“轻舟”。作为一门年青的学科，它充满挑战、机遇且引人入胜。”

一般意义上，生物信息学是研究生物信息的采集、处理、存储、传播、分析和解释等各方面的一门学科，它通过综合利用生物学、计算机科学和信息技术而揭示大量而复杂的生物数据所赋有的生物学奥秘[4]。具体而言，生物信息学作为一门新的学科领域，它是把基因组 DNA 序列信息分析作为源头，在获得蛋白质编码区的信息后进行蛋白质空间结构模拟和预测，然后依据特定蛋白质的功能进行必要的药物设计。基因组信息学、蛋白质空间结构模拟以及药物设计构成了生物信息学的 3 个重要组成部分。

最初生物信息学仅用于数据的贮存，关注于存储来自基因组测序计划完成的序列数据。随着人类基因组计划的完成，生物信息学也被赋予了更多的内容，人们开始感兴趣如何利用计算方法分析生物数据，如根据核酸序列预测基因结构、功能的算法等[5]。生物信息学的研究重点从以前的以 DNA 测序转移到了系统了解基因组内所有基因的生物学功能，即功能基因组（ functional genomics）。研究内容主要包括基因功能发现、基因表达分析及突变检测[6]。

传统研究功能基因的方法包括减法杂交，差示筛选，cDNA 代表差异分析以及 mRNA 差异显示等，但这些技术不能对基因进行全面系统的分析。对于人类基因组，曾经估计编码蛋白质的基因数大约为 2 万个。除此之外，还有许多不编码蛋白质的 RNA 基因，如 rRNA，tRNA，microRNA，snoRNA 等。全长 cDNA 测序（cDNA sequencing）方法识别了目前已知、高质量的编码蛋白基因，但几乎所有类似的克隆方法都偏向于探测在生物组织中被充分表达的基因。这类技术往往很难深入地探索基因组水平所有的表达信息，也不适用于不同的组织和不同条件下转录信息的提取。在刚刚过去的 5 年间，从微阵列（microarray）基因芯片技术发展而来的新技术 Tiling Array 使高通量、全基因组水平的表达探测在理论上得以开展。随着芯片探针设计密度的不断增加，Tiling Array 在高等真核生物全基因组水平的应用也得以逐渐实现。

1.2 本文研究背景及意义

微阵列（microarray）或芯片（chip）技术是近年发展起来的一种新的分子生物学研究工具[7]。它利用光导化学合成、照相平板印刷以及固相表面化学合成等技术，在固相表面合成成千上万个寡核苷酸探针，与放射性同位素或荧光物标记的 DNA 或 cDNA 杂交，用于分析 DNA 突变及多态性、DNA 测序、监测同一组织细胞在不同状态下或同一状态下多种组织细胞基因表达水平的差异、发现新的致病基因或疾病相关基因等多个研究领域。

Tiling Array 实验技术，中文译为“嵌合芯片”，是从传统的微阵列（microarray）技术发展而来，是迄今为止分辨率最高的基因芯片类型，Tiling Array 的探针设计几乎涵盖了目标 DNA 的全部序列[8]。