学校编码：10384　　分类号_____密级_____

学号：19820091152551　　UDC_____

厦门大学

硕 士 学 位 论 文

核磁共振代谢组学中干扰因素的抑制方法

Study on Bias Factors Filtering for NMR-based Metabolomics

丁 俊

指导教师姓名：董继扬教授

专 业 名 称：无线电物理

论文提交日期：2012 年 5 月

论文答辩时间：2012 年 6 月

学位授予日期：2012 年 月

答辩委员会主席：_____

评阅人：_____

2012 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

部分缩写专业名词英汉对照表

NMR	Nuclear Magnetic Resonance	核磁共振
FID	Free Induction Decay	自由感应衰减
MS	Mass Spectrometry	质谱
CSN	Constant Sum Normalization	面归一化
PQN	Probabilistic Quotient Normalization	概率商归一化
GAN	Group Aggregating Normalization	组内聚合归一化
PCA	Principal Component Analysis	主成分分析
MCA	Minor Component Analysis	次成分分析
PLS-DA	Partial Least Squares-Discriminant Analysis	偏最小二乘判别分析
OSC	Orthogonal Signal Correction	正交信号校正
EISC	Extended Inverted Signal Correction	扩展反向信号校正
ISC	Inverted Signal Correction	反向信号校正
MSC	Multiplicative Signal Correction	多元散射校正
OPLS	Orthogonal Partial Least Squares	正交偏最小二乘
CPF	Climaco-Pinto Filter	Climaco-Pinto 滤波
MCF	MCA-based Filter	基于 MCA 的滤波
ANOVA	ANalysis Of VAriance	方差分析
Deviation	difference between an observed value and the expected value	测量值和期望值之差，偏差
Error	difference between an observed value and the true value	测量值和真实值之差，误差
Variation	the amount of difference between normal expected output and observed output	测量值和期望值总差异，总偏差
Variance	mean square of error	均方误差，方差
Standard Deviation	square root of variance	标准差
SS	Sum Square of error	总离差

目　　录

CONTENTS

摘　要

代谢组学借助高通量、高灵敏度与高精确度的现代分析技术考察生物体受内外界刺激或扰动后(如特定基因的变异或环境的变化)，其内源性代谢产物的组成及其随时间的变化规律。凭借核磁共振(NMR)技术的众多优势，基于NMR的代谢组学近年来得到了迅速发展，目前已被广泛应用于病理学、药理学、生物科学等许多领域。处于复杂环境中的生物体不可避免地受到各种内外界刺激因素的作用，当我们试图探究某种刺激因素(如疾病、饮食、药物干预等)对生物体代谢过程的作用时，其它刺激因素便成为了干扰因素。若这些干扰因素对生物体的作用过大，则会影响后续的分析结果的准确性，造成异常代谢通路和相关生物标志物的辨识错误。因此，如何减少干扰因素影响成为了代谢组学数据预处理中亟待解决的问题。对此，本文做了以下两方面的工作：

1. 阐述了方差分析(ANOVA)方法用于代谢组学已知偏向性因素干扰滤除的原理，并与偏最小二乘判别分析(PLS-DA)方法结合用于不同饮食人群的代谢分析，分别考察了把性别因素和饮食因素其中一种作为感兴趣因素，另一种因素作为干扰因素的滤除效果，并与未经干扰滤除的PLS-DA识别结果比较。结果表明，ANOVA方法能有效降低干扰因素的影响，获取与感兴趣因素相关的更准确的代谢信息。用7-折交叉验证法对ANOVA干扰滤除前后的PLS-DA模型进行验证，结果显示干扰滤除后的模型预测能力更强。最后，文中还讨论了基于ANOVA方法滤除干扰因素的适用范围和基本假充，以及当各因素水平下样本数据非平衡时，常规ANOVA方法可能引起的偏置问题。

2. 详细分析了偏向性因素方差与样本个体差异在特征空间中的分布差异，提出了基于次成分分析(MCA)的未知偏向性因素干扰抑制方法。采用素食数据及高脂数据进行验证，并与正交信号较正(OSC)、Climaco-Pinto等人的方法进行比较，对比干扰抑制后PLS-DA模型前两个主成分的预测能力、解释能力及得分图中样本的可分性，结果显示本文方法对未知干扰因素有更好的抑制效果。

本研究为代谢组学偏向性因素干扰抑制提供了新的方法，新方法能有效抑制偏向性因素的干扰，使后续的统计分析的鲁棒性更好，更具生物学意义。

关键词：代谢组学；干扰因素；次成分分析；方差分析

ABSTRACT

As a newly-developing technique, metabolomics is commonly defined as the study of metabolic profiles of organisms and the changes corresponding metabolism, either arising from natural fluctuations or induced by external perturbation. By employing modern analytical technique with high throughput, high sensitivity and high accuracy, it analyzes cells, tissues and endogenous metabolites in biological fluids. It also identifies and illustrates the specific pathological condition through inspecting the dynamic and complicated changes of the metabolisms. Due to the unique advantages of NMR technique, NMR-based metabolomics has gained rapid development in recent years and it has also been applied to many fields such as pathology, pharmacology and biological science. With the application of metabolomics in the various fields, improving the processing method of metabolomics' data has aroused great concern. How to reduce the effect of interference factors have become the most urgent issue. To solve the problem, this thesis focuses on the following two aspects.

1. ANOVA was applied to filter the bias factors in metabolomics analysis, in combination with PLS-DA. ANOVA-based model-identification technique was used to analyze vegetarian data. In which one factor in sex or diet was regarded as the interested factor and another one as the unknown bias factor and vice versa. Their effect was compared with the model of PLS-DA without filtering the bias factors. As a result, ANOVA can effectively reduce the influence of bias factors and achieve more accurate metabolic information associated with the interested factors. According to a 7-fold cross validation, the data post filtering bias factors demonstrated a more accurate and predictable model. It was also pointed out that ANOVA is inadaptable while the data is unbalance under various factors, which may affect subsequent model identification.

2. The distribution difference was analyzed between the variance of bias factors and individual difference in feature space. In addition, a novel reducing method for unknown bias factors based on MCA was posed. In which one factor in sex or diet was regarded as the interested factor and another one as the unknown bias factor and vice versa. Compared with OSC-based filtering and Climaco-Pinto's filtration (CPF), the new method MCF was proved to be the more effective one.

In this thesis, a novel method was proposed to filter the bias factors in metabolomics study, which can effectively reduce the complex in subsequent multivariable analysis and statistics, thus result in the more detailed and significant metabolic information.

Keywords: Metabolomics, Bias Factor, MCA, ANOVA