学校编码：10384　　分类号_____ 密级_____

学号：19020141152607　　UDC_____

厦门大学

硕 士 学 位 论 文

基于带有连续延时的ODE模型的基因调控关系的识别

Detection of Gene Regulatory Relationships based on ODE Model with Continuous Time-Lag

秦慧慧

指导教师姓名：方明, 胡杰 讲 师

专 业 名 称：计 算 数 学

论文提交日期：2017 年 4 月

论文答辩日期：2017 年 5 月

学位授予日期：

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）：

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

中文摘要

多年来，基因之间调控关系的识别研究层出不穷。本文主要提出带有连续延时的常微分方程模型来刻画两两基因之间的调控关系，然后进行模拟数据试验，并用此模型判别真实的两两基因对之间是否有调控关系。

首先，我们建立了带有连续延时的常微分方程模型；其次，基于真实实验数据，我们模拟生成有调控关系以及没有调控关系的基因对数据；然后，我们将模拟数据代入我们提出的带有连续延时的ODE模型中并求解相关参数，再选取相应特征对判别基因对之间是否有调控关系的SVM分类器进行训练与测试，我们得到测试准确率高达89.5%；再次，我们比较了所提出的模型和线性模型应用在相同模拟数据上的结果的优劣，得到我们提出的模型能更好地刻画基因对之间的调控关系；最后，我们将训练好的SVM分类器运用到真实数据上来判别两两基因之间是否有调控关系。

关键词：基因调控关系；ODE模型；连续延时；SVM分类器；

Abstract

Over the past years, there are many studies on the detection of regulatory relationships between genes. This paper mainly proposed an Ordinary Differential Equation model with continuous time-lag to describe the relationships between genes, then did simulations and predicted whether two genes of real data have the regulatory relationship.

Firstly, we construct an Ordinary Differential Equation model with continuous time-lag. Secondly, we generate the synthetic data with or without the regulatory relationship based on the real data from biology experiments. Thirdly, we apply our proposed model to the synthetic data, solve the relevant parameters, and use SVM to predict whether two genes have the regulatory relationship or not. The result shows the accuracy of prediction is 89.5%. Then, we compare the performance of our model with that of linear model using the same synthetic data. Finally, we apply our model to real data to determine whether two genes have the regulatory relationship.

Key words: gene regulatory relationships; ODE model; continuous time-lag; SVM classifier

目 录

Contents

第 一 章　引言

1.1　研究背景

基因是具有遗传效应的DNA片段。基因支持着生命的基本结构和性能，生物体的生、长、衰、病、老、死等一切生命现象都与基因有关，它也是决定生命健康的内在因素。基因表达是指细胞在生命过程中，把存储在DNA片段中的遗传信息经过转录和翻译，转变成具有生物活性的蛋白质分子。在RNA聚合酶的催化下，以DNA 为模板合成mRNA的过程称为转录。以mRNA作为模板，tRNA作为运载工具，在有关酶、辅助因子和能量的作用下将活化的氨基酸在核糖体上装配为蛋白质多肽链的过程称为翻译。

从DNA到蛋白质的过程叫基因表达，基因表达的研究对研究生物的特征有重要的意义。有关基因表达的研究有很多，例如文献[1]中对基因表达的分析，文献[2]研究了通过基因表达控制的癌症分子间的分类，文献[3]研究了巴斯德毕赤酵母的基因表达系统，文献[4]研究了不同水稻的基因表达的差异，文献[5]研究了肝癌细胞的基因表达，文献[6]研究了人类乳腺癌的基因表达。由此，我们可以看出，基因表达的研究已经涉及各个方面，意义重大，影响深远。

基因表达的过程复杂而有序，对这个过程的调节即为基因表达调控，基因表达调控是现代分子生物研究的中心课题之一。相关研究也有很多，例如早期文献[7]研究了大脑海马回神经元的基因表达调控，文献[8]研究了植物激素对基因表达的调控，文献[9]分析了两种家蚕基因的启动子，文献[10]分析了某种蛋白对基因表达的调控。从中我们可以看出，对于基因表达的调控机制的研究从未间断，研究者希望更加清楚基因表达的调控机制从而更好地去研究基因表达。

在研究基因之间的调控关系时，线性模型已经被很多研究人员使用，例如文献[11]

用了有限状态下的线性模型去重建了基因调控网络，文献[12]用了线性的离散动态系统去刻画产乙醇的酵母的基因调控关系，文献[13]用了分段线性模型去研究基因调控网络，文献[14]用了线性时变模型去研究基因调控。但是线性模型由于其简单，往往不能很准确地刻画基因之间的调控关系，我们通过文献发现，研究人员更多地是建立ODE调控模型。例如文献[15][16][17]在研究基因的表达和转录因子的关系时，都用了ODE调控模型，再如文献[18] 比较了不同的ODE方法应用在基因调控网络上，文献[19]运用了高维ODE 研究动态基因调控网络。但是，由于基因表达所生成的产物的合成是需要时间的，在生物学背景下，我们认为调控关系是有一定的延时效应的，例如文献 [20][21]在研究基因调控关系时都加入了延时效应。本文中，我们建立了带有连续延时的ODE调控模型去刻画基因对之间的调控关系，此模型更加符合生物学背景。

裂殖酵母(Schizosaccharomyces pombe)被广泛地运用于细胞生物学的研究中，它的无性繁殖为分裂生殖，是至今细胞分裂的最典型生物模型。本文模拟试验采用的数据就是基于裂殖酵母基因表达产物的数据。

1.2　本文主要工作

本文主要研究两两基因之间的调控关系，针对两两基因，建立带有连续延时的常微分方程调控模型，并且运用SVM判定两两基因之间是否有调控关系。

第二章中，对于两两基因之间的调控关系进行建立模型，通过参考文献，我们首先考虑ODE 模型，又考虑生物学背景，基因表达的产物合成等都需要时间，我们加入延时效应。于是，我们提出带有连续延时的ODE模型来刻画两两基因之间的调控关系。建立模型，最终需要求解5个参数，将其转化为求解两个参数的非线性优化问题，目标函数为拟合优度，最后根据改进的控制随机搜索算法来求解这些参数。

第三章中，我们模拟数据来检验第二章中提出的带有连续延时的ODE调控模型。首先对实验数据进行预处理，得到有调控关系基因对数据与无调控关系基因对数据；其次将基因对数据代入第二章中的调控模型，求解相应参数；最后我们选取相应特征

值对SVM分类器进行训练与测试，得到可以判别基因对之间是否有调控关系的分类器，并看测试集分类效果如何，以此来说明我们提出的带有连续延时的ODE模型是否能很好地刻画两两基因之间的调控关系。

第四章中，对于两两基因之间的调控关系我们建立了最简单的线性模型，我们将我们的ODE 模型与线性模型进行对比，将相同的测试数据集应用到线性模型中，我们根据线性模型拟合优度来判断两两基因之间是否有调控关系，再与第三节章我们的分类结果作比较，看看我们提出的带有连续延时的ODE调控模型表现如何。

第五章中，我们将我们的分类模型应用到真实数据中，研究它们之间两两是否有调控关系。

本文提出了带有连续延时的ODE调控模型去研究两两基因对之间是否有调控关系，对于新进的实验数据，我们可以直接应用到我们的模型中，再用统计学分类的方法直接去判断两两基因之间是否有调控关系。

第 二 章　模型与方法

2.1　数据来源

本文的数据来源基于文献[22]，我们采用了文献中裂殖酵母(Schizosaccharomyces pombe)基因表达产物的实验观测值数据。实验中，是用荧光蛋白标记裂殖酵母基因的表达产物的，每隔10分钟记录下荧光蛋白的强度值，对于每个基因观测了51次，最后得到的是具有时间序列的51个荧光蛋白强度的实验观测值。其数据下载地址为：http://publications.redgreengene.com/oliva_plos_2005/。

2.2　B样条拟合模型

我们令$f_m(t_p)$表示基因m在t_p时刻的表达水平，通常，我们用荧光蛋白去标记基因表达的产物；$y_m(t_p)$表示基因m在时刻t_p的表达水平的观测值，即观测的荧光蛋白强度值。由实验观测，我们得到的$y_m(t_p)$是一系列带有时间序列的离散点数据。

首先，由于数据$y_i(t_p)$为实验观测所得，存在一定的实验观测噪声，我们对其进行平滑降噪处理，将离散的实验数据点拟合为连续曲线$\tilde{y}_i(t_p)$，我们采用B样条拟合。

其次，在下文求解我们的模型时，我们需要的是调控基因的连续数据，因此，我们对预处理后的数据再次采用B 样条拟合模型将离散的实验数据点拟合成连续曲线。

B样条拟合模型中，我们设$d(d = 3 + knots)$为拟合模型的自由度，范围为$d \in [d_1, d_2]$，$knots$为定义域内结点个数。对于B 样条拟合模型的选择，我们考虑到选取的模型最后拟合出来的连续曲线需要尽量真实地反应基因的表达水平，即我们的拟合模型的拟合优度r^2应该尽量大，但是又需考虑不能过拟合，拟合的连续曲线还应该有一定的光滑性。

最后通过试验，我们采取最小BIC准则来选取B样条拟合模型的自由度d，从而确定模型，将离散数据点$y_m(t_p)$拟合为连续曲线$\tilde{y}_m(t_p)$。

2.3　带有连续延时的ODE模型

同上文，我们用$f_m(t_p)$表示基因m在时间t_p时刻的基因表达水平，实验中我们通常用荧光蛋白去标记基因表达的产物；$y_m(t_p)$表示基因m在时刻t_p的表达水平的观测值，通常为实验中测量的荧光蛋白强度值，其中p表示等间隔的时间序列。

值得注意的是，观测值$y_m(t_p)$并不等于$f_m(t_p)$，基因表达产物的总量并不等于实验中测量的荧光蛋白强度值。我们假设基因m在t_p时刻测量的荧光蛋白强度值$y_m(t_p)$等于$f_m(t_p)$的线性表达，即

$$y_m(t_p) = a_m * f_m(t_p) + b_m \tag{2-1}$$

我们仅仅用相应荧光蛋白强度值的观测值去反应一个基因的真实表达水平。

接下来我们考虑两两基因的调控模型，我们假设基因i调控基因k，则我们提出的带有连续延时的ODE 调控模型如下：

$$f_k^{'}(t_p) = \beta_{ik} f_i(t_p - \tau_{ik}) - \mu_k f_k(t_p) \tag{2-2}$$

其中，τ_{ik}就表示基因i调控基因k的延时时间，$f_i(t_p - \tau_{ik})$表示基因i在$(t_p - \tau_{ik})$时刻的表达水平，$f_k(t_p)$表示基因k在t_p时刻的表达水平，$f_k^{'}(t_p)$ 表示基因k在t_p时刻表达水平的改变速度，β_{ik}为系数，μ_k表示基因k表达产物的降解速率，即实验中测得的荧光蛋白的降解速率。

式(2-2)即为一阶线性非齐次方程，我们求解得到：

$$f_k(t_p) = f_k(s_k)e^{-\mu_k(t_p - s_k)} + e^{-\mu_k(t_p - s_k)} \cdot \int_{s_k}^{t_p} \beta_{ik} f_i(\delta - \tau_{ik}) e^{\mu_k(\delta - s_k)} d\delta \tag{2-3}$$

$$(p = 1, 2, ..., n),\ \ t_p \geq s_k$$

其中s_k表示基因k开始表达的理想时刻，根据此模型考虑了基因i调控基因k的时间滞后性，我们设置$s_k = s_i + max(\tau_{ik})$，即基因$k$表达的理想时刻为基因$i$表达的理想时刻加上最长滞后时间，$s_i$即为基因$i$的理想表达时刻，本文中$s_i = t_1$。

至此，我们得到了描述基因i和基因k表达水平关系的式子(2-3)，我们将其转换为我们可以得到的实验观测数据，即用标记基因表达产物的荧光蛋白强度值去描述基因的表达水平。如同式(2-1)所示，$y_i(t_p)$ 和$y_k(t_p)$都可表示为如下：

$$y_k(t_p) = a_2 * f_k(t_p) + b_2 \tag{2-4}$$

$$y_k(s_k) = a_3 * f_k(s_k) + b_3 \tag{2-5}$$

$$y_i(t_p - \tau_{ik}) = a_1 * f_i(t_p - \tau_{ik}) + b_1 \tag{2-6}$$

其中$y_k(s_k)$表示基因k在s_k时刻时，即基因k开始表达的理想时刻，测得的相应的荧光蛋白强度值，我们认为$y_k(s_k) = 0$，其他参数$a_1, a_2, a_3, b_1, b_2, b_3$ 均为常数。

我们将式(2-4)(2-5)(2-6)代入式(2-3)中，整理后得到

$$y_k(t_p) = C_{ik} + D_{ik}e^{-\mu_k(t_p - s_k)} + \beta_{ik}e^{-\mu_k(t_p - s_k)} \int_{s_k}^{t_p} y_i(\delta - \tau_{ik})e^{\mu_k(\delta - s_k)}d\delta + \epsilon_{kp} \tag{2-7}$$

$$(p = 1, 2, ..., n),\ \ t_p \geq s_k$$

其中，C_{ik}, D_{ik}以及β_{ik}为常数系数，我们注意到，对于$y_i(\delta - \tau_{ik})$，实验中我们测量的是一系列具有时间序列的离散点，在这里，我们求积分时需要的是连续值，所以，我们采用我们2.2中提及的B样条拟合曲线上数值$\tilde{y}_i(\delta - \tau_{ik})$。

最终，我们得到我们提出的基因i调控基因k的带有连续延时的ODE调控模型为：

$$y_k(t_p) = C_{ik} + D_{ik}e^{-\mu_k(t_p - s_k)} + \beta_{ik}e^{-\mu_k(t_p - s_k)} \int_{s_k}^{t_p} \tilde{y}_i(\delta - \tau_{ik})e^{\mu_k(\delta - s_k)}d\delta + \epsilon_{kp} \tag{2-8}$$

$$(p = 1, 2, ..., n),\ \ t_p \geq s_k$$

如上文分析，$s_k = s_i + max(\tau_{ik}) = t_1 + max(\tau_{ik})$，所以，我们的模型中有5 个参数：$\mu_k$，$\tau_{ik}$，$C_{ik}$，$D_{ik}$和$\beta_{ik}$。