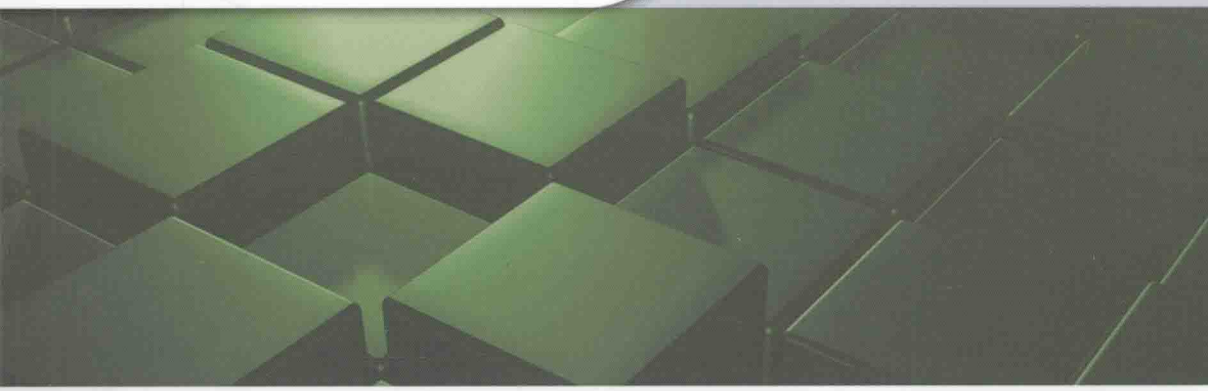
生物数据分析和生物系统模型中的参数估计

田立平 著



机械工业出版社
CHINA MACHINE PRESS

生物数据分析和生物系统模型中的参数估计

田立平　著

机械工业出版社

本书是在作者对时间序列基因表达数据和非线性动态生物系统参数估计领域的研究论文基础上完成的，它包括5个部分：第1部分给出了本书的研究背景和结构纲要；第2部分包括5章，每章介绍一个时间过程的基因表达数据的方法；第3部分也包括5章，每章描述一种用于非线性动态分子生物系统的参数估计方法；第4部分介绍了有关基因调控网络的建模及参数估计研究现状与进展情况；第5部分为附录。

本书在该领域具有一定前沿性和创新性。本书的第2～11章主要来源于作者近几年发表在著名的国际会议或期刊的研究论文。本书可以作为大学教师、研究生以及研究机构的专家、学者和工程师的参考用书。

图书在版编目（CIP）数据

生物数据分析和生物系统模型中的参数估计/田立平著. —北京：机械工业出版社，2016.3

ISBN 978-7-111-52459-5

Ⅰ. ①生… Ⅱ. ①田… Ⅲ. ①生物信息论—数据处理—研究 Ⅳ. ①Q811.4

中国版本图书馆CIP数据核字（2015）第301269号

机械工业出版社（北京市百万庄大街22号 邮政编码 100037）

策划编辑：张俊红 责任编辑：闫洪庆

责任校对：佟瑞鑫 封面设计：路恩中

责任印制：乔 宇

北京铭成印刷有限公司印刷

2016年3月第1版第1次印刷

169mm×239mm · 10印张 · 194千字

0001—3000册

标准书号：ISBN 978-7-111-52459-5

定价：30.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务 网络服务

服务咨询热线：010-88361066 机工官网：www.cmpbook.com

读者购书热线：010-68326294 机工官博：weibo.com/cmp1952

读者购书热线：010-88379203 金 书 网：www.golden-book.com

封面无防伪标均为盗版 教育服务网：www.cmpedu.com

前　言

许多疾病（如艾滋病、肥胖病、H1N1、H7N7禽流感等）起源于在分子水平上不能正常工作的生物系统。随着DNA微阵列和串联质谱等科学技术和实验技术的发展，产生了大量来自于动态分子生物系统的时间序列数据。这些生物数据包含的信息在由基因调控网络导致的疾病诊断和药物合成方面具有非常重要的作用。许多小的非线性基因调控网络通过试验来研究。然而对于大型的基因调控网络，特别是复杂疾病网络，通过研究小型网络的试验方法是无法完成的。即使我们有足够的能力来观测生物网络之间的功能状态和相互作用，但计算上的限制也会阻碍我们理解分子生物系统的行为，因为它们的复杂度会伴随着网络组成的相互作用的数量呈几何级数增长。

本书是在作者对时间序列基因表达数据和非线性动态生物系统参数估计领域的研究论文基础上完成的。它包括5个部分。第1部分给出了本书的研究背景和结构纲要。第2部分包括5章。每章介绍一个时间过程的基因表达数据的方法。第3部分也包括5章。每章描述一种用于非线性动态分子生物系统的参数估计方法。这些方法不仅提供了时间基因表达数据的分析方法，也提供了高质量的数据用于建立动态的基因调控网络。只要建立了一个动态的分子生物系统模型，就可以对生物系统的性能进行研究。本书为深入了解动态的基因表达数据和动态的分子生物系统（包括基因调控网络）的性能提供了一些新的方法。因此，希望能对许多疾病的治疗和药物的设计提供科学的理论依据。第4部分介绍了有关基因调控网络的建模及参数估计研究现状及进展情况。第5部分为附录。

本书有如下特点：1）与传统时间过程的基因表达数据的

分析方法不同，本书充分采用了基因表达序列的动态和周期性，并且采用非线性模型分析时间过程的基因表达数据。一般情况下，非线性模型的参数估计是研究的主要问题。然而，我们在第 2～6 章设计了很有效的算法来估计非线性模型的参数。2）基于生物化学反应原理，动态的分子生物系统的模型是较为复杂的非线性微分方程。因而估计这些模型中的参数非常具有挑战性。虽然传统的非线性优化算法，如牛顿法以及改进的牛顿法可以应用于解决这个问题，但这种传统的算法都具有对初始值敏感的不足，而且容易陷入局部最优。利用这些非线性微分方程的特殊结构，我们设计了几种有效的算法用于动态的分子生物系统的参数估计。3）本书的内容具有前沿性和较大的创新性，本书的主要内容来源于作者 2009～2013 年发表在著名的国际会议或期刊的研究论文。

我要感谢许多人。在他们的帮助下，本书才得以顺利出版。首先，我要感谢加拿大 Saskatchewan 大学的吴方向教授以及他的研究团队，衷心地感谢他们给予的鼎力合作、帮助和鼓励。其次，我要感谢我的家人，在我写作过程中，他们一直给予我理解和支持；也感谢我的研究生孙群、贾鹏飞、曾俊、颜斌斌 4 位同学的帮助。最后，我也非常感谢出版社以及编辑，感谢他们在编辑本书时所付出的辛勤努力。

本书可能会有很多不足之处和错误，恳请读者朋友们能发现错误，提出更正、建议，作者将不胜感激。

本书由国家自然科学基金面上项目“基于动态非线性的大型复杂基因调控网络的建模与分析”（61571052）和“北京市高层次人才创新创业计划（G01040011）”以及北京物资学院高级别科研项目培育基金项目（GJB20141004）支持。

北京物资学院信息学院　田立平

目　录

前言

第 1 章　导论……1

1.1　背景……1

1.2　本书的框架……2

第 2 章　周期性基因鉴定的参数估计方法……6

2.1　引言……6

2.2　方法……8

2.2.1　基因的周期性表达模型……8

2.2.2　假设检验……9

2.3　实验结果与讨论……10

2.4　结论和展望……13

参考文献……14

第 3 章　从微阵列时间过程表达探测近似周期性表达的基因……15

3.1　引言……15

3.2　方法……17

3.2.1　基因近似周期性的表达模型…17

3.2.2　假设检验……19

3.3　实验结果与讨论……20

3.4　结论和展望……25

参考文献……25

第 4 章　伪周期性基因表达谱鉴定……28

4.1　引言……28

4.2　方法……30

4.2.1　伪周期性基因表达序列的鉴别模型……30

4.2.2　假设检验……32

4.3　实验结果与讨论……33

4.4　结论和展望……35

参考文献……36

第 5 章　基于非线性模型的周期性表达基因数据的聚类分析法……38

5.1　背景……38

5.2　方法……40

5.2.1　周期性表达基因模型……40

5.2.2　基于聚类分析的非线性模型……42

5.2.3　验证……43

5.3　实验结果与讨论……44

5.4　总结……48

参考文献……48

第 6 章　基于非线性模型的时间序列基因表达数据分析……51

6.1　背景……51

6.2　方法……53

6.2.1　时间序列基因数据的非线性模型……53

6.2.2　非线性模型的显著性分析……55

6.2.3　基于非线性模型的聚类分析…56

6.2.4　数值计算……57

6.3　基因表达数据在现实生活中的应用……61

6.4　结论……62

参考文献……63

第 7 章 有约束的交互最小二乘法对 S 系统生物网络模型的参数估计……66
7.1 引言……66
7.2 算法描述……68
7.3 数值算例……70
7.4 结论……73
参考文献……73
第 8 章 线性分式模型中参数估计的迭代最小二乘法……75
8.1 引言……75
8.2 算法描述……76
8.3 说明性的例子……79
8.4 结论……82
参考文献……82
第 9 章 一种基于幂律的细胞凋亡模型及其参数估计……84
9.1 引言……84
9.2 模型与参数估计……85
9.3 仿真结果……89
9.4 结论和未来研究方向……91
参考文献……91
第 10 章 复杂度分析与动态代谢系统的参数估计……93
10.1 引言……93
10.2 参数估计的模型复杂性分析……94
10.3 参数估计算法……99
10.4 应用……102
10.5 结论和未来的研究方向……104
参考文献……104
第 11 章 基于逻辑和的基因调控网络中的参数估计……107
11.1 引言……107
11.2 基于逻辑和的基因调控网络……108
11.3 参数估计方法……110
11.4 说明性的例子……112
11.5 结论……114
参考文献……114
第 12 章 关于基因调控网络的建模及参数估计的研究现状及进展……117
12.1 研究意义……117
12.2 国内外研究现状分析……118
12.3 未来研究的主要问题……123
12.4 主要研究内容、研究目标，以及拟解决的关键科学问题……124
12.5 拟采取的研究方法、技术路线、实验手段、关键技术、特色及创新点……127
参考文献……129
附录……132
附录 A 一元线性回归的分析及最小二乘估计……132
附录 B 多元统计分析中的聚类分析……135
附录 C 数据挖掘中的聚类分析……137
附录 D F 分布定义及性质……141

第1章 导 论

1.1 背景

在过去的十年中，随着生物科学技术和实验技术的发展，产生了大量的分子和细胞水平上的生物数据和信息。这些成果的一个例子就是利用 DNA 微阵列技术产生了大量的基因表达数据。如今，如何将这些现有的生物数据和信息集成和利用，以及在系统层面上定量了解生物过程的动态行为，是研究人员面临的富有挑战性的新课题。在这种背景下，一个新兴的研究领域——系统生物学诞生了，它着重对生物系统的原理进行数学建模和定量分析。

系统生物学是一个较新的研究领域，其研究重点是在生物系统中组成成分之间的相互作用，以及这些相互作用怎样影响系统的具体功能与系统行为（例如，酶和代谢产物的代谢途径）。而不是像传统的生物学研究方式那样，只对单个组成部分或有机体方面做分析。系统生物学是关于作为一个系统的所有组成部分和相互作用的研究。从系统的角度来看，有可能发现新的性质并进一步解释生物系统的运行规律，从而帮助我们更好地理解整个生物过程的运行机制。所有这些问题的答案，如果从实验中去获取的话，不仅会浪费很多时间，而且也是无法完成的。

系统生物学研究的最终目标是通过分析和理解复杂的生物系统的机制和特点，设计和控制细胞的功能。为了达到这个目标，我们首先需要构建一个用来表示系统内的动力学和相互作用的模型。通过高科技的实验手段，例如，DNA 微阵列，串联质谱，时间序列的数据（如基因表达分子水平，蛋白酶，或代谢产物），可以从动态的生物系统中收集时间数据。利用这样的时间序列数据，生物系统模型的构建主要包括三个主要阶段：① 确定包含在相应的生物学过程的（典型的生物分子）组成成分；② 识别生物系统模型的结构；③ 估计模型中的参数值。

生物系统模型的构建可以由两种方法确定。一种方法是基于通过曲线拟合的方法描述所获得的数据，采用回归模型，比如，自回归（AR）模型、自回归移动平均（ARMA）模型等。这种模型可以很好地拟合数据。然而，这些模型的一个主要问题是模型中的参数通常不具有生物学意义，因此在实际中很难解释和使用。另一种方法是基于生物学中的分子反应原理（如分子热力学定律）。作为使用分子反应原理的一个结果，所构建的动态模型通常是非线性微分方程。虽然模型中的所有参数都有明确的生物学意义，但在这种模型中的参数估计一般来说是一个难以解决的非线性问题。虽然生物系统模型的构建与分析是非常重要的，但它不是本书的重点。在本书中，我们将主要集中于介绍在相应的生物过程中的成分识别和生物系统模型中的参数估计。

1.2 本书的框架

本书其余章节的内容可以分为4个部分。第1部分包括第2～6章。在这些章节中，我们基于基因的时间序列表达数据的显著差异性分析和聚类分析，给出了识别成分（基因）的各种方法。时间序列基因表达数据来源于现实生活中一些典型的非线性的生物过程，因而我们的分析方法也都建立在一些非线性模型的基础之上。

在第2章中，我们依照三角正弦和余弦函数的线性组合加上一个高斯噪声项来构建了周期性的基因表达数据的模型。在该模型中，我们给出了一个两阶段参数估计方法来估计模型中的参数。另一方面，非周期性的时间序列基因表达数据是由一个常数加高斯噪声项的模型来表示的。本书用统计学的F-假设检验方法来确定基因表达数据是否是周期性的，同时使用了一个合成数据和两个生物数据来检测所提出的方法的合理性。结果表明，该方法可以有效地识别周期性表达基因。

在第3章中，我们提出了一种从时间序列的表达数据来检测周期表达基因的新方法。在该方法中，一个近似的周期性的基因表达数据由一个近似的周期函数的模型表达，该模型是由一个三角正弦函数和余弦函数的线性组合、一个关于时间变量的线性函数和一个高斯噪声项组成。由于该模型中的参数和时间变量都是非线性的，我们分两个阶段估计模型的参数。另一方面，非周期性的基因表达序列是由一个常数加高斯噪声项构成的。本书采用统计学的F-假设检验方法来检测

基因是否周期性表达，同时也使用了一个合成数据和两个生物数据来检测所提出的方法的合理性。结果表明，该方法可以有效地识别近似周期表达基因。

在4章中，我们提出了一个用于识别伪周期的基因表达序列的方法。在该方法中，一个伪周期的基因表达序列是由三角函数和指数函数加上高斯噪声项的线性组合模型刻画。仍用两阶段参数估计方法来估计模型的参数。另一方面，非伪周期的基因表达序列是由一个常数加高斯噪声的模型刻画。本书采用统计学的F-假设检验方法来检测基因是否周期性表达，同时使用三个生物数据集来检测所提出的方法的可行性。结果表明，该方法可以有效地识别伪周期表达基因。

在第5章中，因为作为伪周期表达的基因总与一个周期生物过程相关，我们基于聚类方法给出了一种伪周期表达基因序列的非线性模型。该方法假定一个周期表达基因数据集是在数个周期过程中产生的。每个周期过程所建的模型都是由三角正弦和余弦函数的线性组合再加上一个高斯噪声项构成。能够估计模型参数和迭代算法的两阶段法被用于评估聚类质量。拔靴法和均值校正的兰德指数（AARI）来衡量聚类质量。一个合成数据集和两个生物数据来评估所提出的方法的性能。结果表明，对于周期性基因表达数据，我们的方法比其他聚类方法（例如 k-均值）聚类结果的质量更好，因此它是周期性基因表达数据的一种有效的聚类分析方法。

在第6章中，我们采用了广义非线性模型分析基于时间过程的基因表达数据。我们提出了非线性模型参数估计的一种有效方法。然后，利用这个模型来对显著差异表达基因进行分析和对一组基因的表达谱进行聚类分析。两个合成的数据集的验证表明，我们提出的显著差异性分析方法和聚类分析方法优于现有的一些方法。一个实际的生物数据集的应用表明，我们的方法的分析结果与已有结果一致。

第2部分包括第7～12章。在这些章节中，我们提出估计生物过程模型的参数估计方法。该模型构建依据质量作用定律（Mass Action Law）。因而，刻画生物过程的模型往往为非线性微分方程组。与传统的非线性参数估计的方法不同，我们针对不同非线性模型的不同特性，给出不同的估计参数的新方法。

可以观察到S模型中的参数可以分为两组：一组参数是线性的，而另一组参数是非线性的。基于这一观察，在第7章中，利用参数特殊的结构和参数的生物学意义，我们给出了一个交互的最小二乘法来

估计 S 系统模型中的参数的方法。为了验证该方法的可靠性，将交互最小二乘法应用于生物系统，并与其他的参数估计方法进行比较。模拟仿真结果表明，我们所提出的参数估计方法具有更好的优越性。

基于统计热力学原理和 Michaelis-Menten 动力学方程，生物系统模型包含有理分式反应速率，而参数和状态在有理分式反应速率中都是非线性的。但是在有理分式中，模型参数在分子和分母中是线性的。基于这一观察，在第 8 章中，我们采用一个迭代最小二乘法估计根据生物系统所建的有理分式函数模型中的参数。其基本思想是将非线性最小二乘目标函数优化转化为迭代求解线性最小二乘问题的一个序列。将该方法应用于有理分式函数和基因调控网络。模拟仿真结果表明，我们所提出的方法比其他已有的算法更优越。

基于米凯利斯的 Michaelis-Menten 动力学模型，用于描述细胞死亡过程的动力学模型是一组带有分式（非线性）反应速率的非线性微分方程。虽然目前已经有几种方法来估计非线性反应速率这一参数。然而，模型中总有一些参数无法估计，或者存在较大的误差，因而得不到令人满意的结果。另一方面，幂律函数已被用于描述动力学反应速率。在第 9 章中，我们使用幂律函数来描述通过一群蛋白酶激活的细胞凋亡模型，我们给出了一个估计在基于幂律反应速率中的参数的参数估计方法。结论和分析表明我们所给的幂律模型不仅与用基于 Michaelis-Menten 动力学模型描述的蛋白酶激活细胞凋亡的动态模型具有相同的性质，而且我们所给的方法更容易估计模型中的参数。

代谢系统是一类复杂的生物系统。基于生化反应原理，动态代谢系统可以由一组由参数、状态（包括分子浓度）和反应速率构成的模型来表示。通过质量作用定律，在一个状态中，反应速率既是含有常数参数的多项式函数，又是含有常数参数的有理函数。因此，描述动态代谢系统的微分方程关于参数和状态都是非线性的。在第 10 章中，我们提出了一个分析复杂动态代谢系统中参数估计的方法，也就是将动态代谢系统中的参数估计简化成一个有理函数加多项式（我们称之为假有理函数）或多项式的参数估计。进而，依据假有理函数结构的特殊性，我们给出了一个估计假有理函数中参数的一个很有效的算法，并将所提出的方法应用到动态代谢系统中的参数估计。模拟仿真结果表明，我们所提出的方法具有更好的优越性。

一般而言，基于逻辑和的基因调控网络常常可以由非线性微分方

程来描述。基于这种模型，很多学者对于基因调控网络的稳定性问题进行了广泛的研究。然而由于在这样一个非线性模型中进行参数估计是十分困难的，因而属于基因调控网络的非线性模型的参数估计问题。在第 11 章中，我们提出了一个具有逻辑和的基因调控网络模型中的参数估计方法。在这种基因调控网络中，每个基因对应的正则函数中关于参数和状态都是非线性的 Hill 函数的线性组合。我们用两个基因相互切换的一个例子来说明该方法的可行性。结果表明我们所给的方法可以精确地估计基于逻辑和的基因调控网模型中的参数。在第 3 部分即第 12 章中，我们介绍了有关基因调控网络建模和参数估计问题的研究现状、进展及未来有待解决的问题。最后第 4 部分即附录给出了有关一元线性回归的分析及最小二乘估计、假设检验和聚类分析。

第 2 章　周期性基因鉴定的参数估计方法

摘要：在分子水平上，诸如细胞分裂等周期性的生物过程已经有了大量的研究。由于这些生物过程的周期性，与其相联系的时间序列基因表达谱也会表现出周期性。鉴于此，通过检测基因表达的时间序列数据中的周期性，将有助于我们理解分子水平上生物过程的运作机制。在本章中，一方面运用由正弦和余弦三角函数的线性组合加上相应的高斯噪声项来模拟一个具有周期性的基因表达谱，并采用两步参数估计法对模型中的参数进行了估计；另一方面，运用一个常数加上高斯噪声项来模拟一个不具有周期性的基因表达谱。然后，通过 F 统计检验来判断某个基因是否会周期性表达。最后，利用一个模拟数据集和两个生物过程数据集来评估该方法的性能。结构表明，本章所提出的方法能够有效地识别出具有周期性表达的基因。

2.1　引言

诸如细胞分裂等许多生物过程表现为周期性行为。为了理解这些生物过程的运作机制，运用 DNA 微阵列实验来检测某一时间段内所产生的一系列基因表达量，进而产生相应的时间序列数据[1,2]。这些时间序列数据所呈现出的基因动态变化将大致反映出与之相关的生物发展过程。鉴于此，通过检测时间序列数据中周期性表达的基因，将有助于我们理解分子水平上生物过程的运作机制。

在过去的十年中，研究者已经提出了一些方法来检测基因的周期性表达。离散的傅里叶变换方法是检测基因周期性表达最早的方法。但是，微阵列实验通常仅能产生短期的时间序列数据。如参考文献[3]所述，基于短期时间序列数据，运用离散傅里叶变换所得到的频率分辨率不足以得到周期性的频率。作者在参考文献[4]中提出了一种检测周期性表达基因的方法，简称为 CORRCOS。由于该模型会产生 1000 个不同频率的正弦模型，且每个频率下的模型会产生 101 个不同相位

的模型，因此 CORRCOS 可以产生 101000 个周期性的模拟模型。为此，每一个基因表达谱都要与这 101000 个模型进行比较，通过交叉相关度检测来衡量该模型与基因表达谱之间的相似度，将同一个基因表达序列最相似的模型作为该基因的周期性表达模型。尽管 CORRCOS 算法可以检测出周期性表达的基因，但该方法相当费时，且其判断条件——交叉相关性，不是客观的衡量标准。参考文献[3]的作者提出了另一种检验周期性表达基因的算法——RAGE 方法。类似于 CORRCOS 算法，RAGE 算法也是一种基于模拟模型的方法。首先，其运用模拟模型和基因数据之间的自相关性来估计基因表达序列的频率。然后，RAGE 算法会因估计出的不同频率及不同相位而产生大量的模型。在此基础上，利用真实的标准——Hausdorff 距离来衡量所建模型与基因表达谱之间的相似度。与 CORRCOS 算法相比较，RAGE 算法更省时。

以上提到的所有方法都没有进行统计分析。Wichert 等在参考文献[5]中提出了一种可以从时间序列基因表达谱中识别出周期性表达基因的统计方法。该方法类似于用正弦函数模拟基因表达谱并用 Fisher G—检验进行统计分析。此研究的基础上，Chen 在参考文献[6]中提出了在时间序列基因表达谱中检测出周期性表达基因的步骤。然而，Fisher G—检验仅能有效检验等间距的基因表达谱；对于非等间距的基因表达谱，Chen 等提出了运用 Lomb-Scargle 周期图来统计识别具有显著周期性的基因表达[7,8]。但是，最新研究显示：在时间序列数据较少或数据缺失的条件下所作出的 Fisher G—检验并不可信。相关学者在参考文献[7,8]中指出，该数据时间序列长度应不小于 40。基于此标准，大多数的基因表达谱都不适合进行 Fisher G—统计检验。另外，由于在收集数据前研究所需的时间段是未知的，这就使得我们在实验过程中很难获得这个时间段内完整的基因表达谱。

此外，其他学者也提出了很多描述周期性基因表达谱的模型，例如，曲线 B-spline 模型[10]和自相关方程[11]组成的线性组合。然而，学者们普遍认为正弦函数能更好地描述周期性的基因表达谱。在本章中，一方面运用由正弦和余弦三角函数的线性组合加上相应的高斯噪声项来模拟一个具有周期性的基因表达谱。通过对参考文献[1-9]研究发现，该模型的难点在于频率在非线性变化下对模型参数的估计。为此，本章采用两步参数估计方法来估计模型中的参数。另一方面，运用一

个常数加上高斯噪声项来模拟一个不具有周期性的基因表达谱。然后，通过 F 统计检验而非 Fisher G—检验，判断某个基因是否会周期性表达。最后，利用一个模拟数据集和两个生物过程数据集来评估该方法的性能。

2.2 方法

2.2.1 基因的周期性表达模型

令 $x(t)$ (t=1,2,…, m)表示生物过程中所产生的时间序列基因表达谱，其中 m 表示我们研究的基因表达时间点的数量。在本研究中，我们始终将基因表达的时间序列值的均值转换为 0。为了模拟这些基因表达时间序列，我们采用正弦和余弦三角函数的线性组合加上高斯噪声项构建了以下模型：

$$x(t)=a\cos(\omega t)+b\sin(\omega t)+\varepsilon(t) \tag{2-1}$$

式中，a 和 b 分别是余弦和正弦函数的系数，ω 是周期性表达数据的频率，$\varepsilon(t)$代表随机误差。在本研究中，我们假设随机误差服从于均值为 0，方差为 σ^2 的正态分布。此时模型就等价于参考文献[1,2,5,9]中的正弦函数模型：

$$x(t)=A\sin(\omega t+\Phi)+\varepsilon(t) \tag{2-2}$$

该模型被广泛用于模拟基因周期性表达时间序列。

给定一个时间序列基因表达谱 $x(t)$ (t=1,2,…, m)，式（2-1）中的估计参数 a 和 b 及 ω 都是非线性的。此时，在我们所研究的无噪声模型式（2-1）中：

$$x(t)=a\cos(\omega t)+b\sin(\omega t) \tag{2-3}$$

可以看作是以下二阶常微分方程的通解：

$$\ddot{x}(t)+\omega^2 x(t)=0 \tag{2-4}$$

在式（2-4）中 ω^2 是线性的，且与估计参数 a 和 b 相独立。因此，我们提出了两阶段参数估计方法来估计式(2-1)中的参数 a 和 b 及 ω。

第一步：根据式（2-4），使用线性最小二乘法估计参数 ω^2，具体来说，令 $\boldsymbol{X}_2=[\ddot{x}(1),\cdots,\ddot{x}(l)]$，$\boldsymbol{X}_1=[x(1),\cdots,x(l)]$。

然后通过最小二乘法进行参数估计得到

$$\hat{\omega}^2=\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_2/\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1 \tag{2-5}$$

因此　　$$\hat{\omega}=\sqrt{\hat{\omega}^2} \tag{2-6}$$

由于时间序列的基因表达数据是离散的，可以通过中心差分公式估算出的二阶导数如下：

$$\ddot{x}(t)=\frac{x(t+1)+x(t-1)-2x(t)}{\Delta^2},\ t=2,\cdots,m-1 \tag{2-7}$$

式中，Δ表示两个连续基因表达数据点之间的时间差。从式（2-7）中可以得到，向量 $\boldsymbol{X}_2$ 和向量 $\boldsymbol{X}_1$ 的长度为 $m-2$。需要注意的是，式（2-7）仅适用于等时间差的时间序列数据。对于非等时间差的时间序列数据，我们需要对方程进行修正，具体的修正方法见参考文献[21]。如果根据式（2-5）计算出的数字是负值，我们就认定这个基因为非周期性表达。

第二步：将式（2-6）计算出 ω 的估计值代入式（2-1），并运用最小二乘法对式（2-1）中的参数 a 和 b 进行估计，具体来说，令

$$\boldsymbol{X}=[x(1),\cdots,x(m)]\text{ 和 }\boldsymbol{A}=\begin{bmatrix}\cos(\Delta\hat{w}),\cdots,\cos(m\Delta\hat{w})\\ \sin(\Delta\hat{w}),\cdots,\sin(m\Delta\hat{w})\end{bmatrix}$$

利用最小二乘法进行估计，得到

$$\alpha=\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix}=(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}(\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}) \tag{2-8}$$

2.2.2　假设检验

为了检测某个基因是否为周期性表达，我们检验的零假设为

$$\mathrm{H}_0:\quad x(t)=\varepsilon(t) \tag{2-9}$$

交换假设为

$$\mathrm{H}_a:\quad x(t)=a\cos(\omega t)+b\sin(\omega t)+\varepsilon(t)$$

依据以下 F 统计量：

$$F=\hat{\sigma}_0^2/\hat{\sigma}_1^2 \tag{2-10}$$

式中，$\hat{\sigma}_0^2$ 为式（2-1）中正常白噪声的估计方差，计算得

$$\hat{\sigma}_0^2=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}/(m-1) \tag{2-11}$$

$\hat{\sigma}_1^2$ 为式（2-9）中正常白噪声的估计方差，计算得

$$\hat{\sigma}_1^2=\left[\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\alpha\right]^{\mathrm{T}}\left[\boldsymbol{X}-\boldsymbol{A}^{\mathrm{T}}\alpha\right]/(m-1) \tag{2-12}$$

根据统计学理论[22]，由于式（2-1）和式（2-9）中的噪声为正常的白噪声，所以 F 统计量服从于自由度为（m-1，m-1）的 F 分布。当 F 统计量的取值大于某一值时，就拒绝式（2-9），则该基因表达出现周期性行为，否则认为该基因数据仅仅是白噪声。根据由用户指定的自由度（即时间序列数据的长度）和显著性水平（通常为 0.01、0.05、0.1、0.2 等）来确定其取值，具体数值可以通过 F 分布表或通过使用标准化后的 MATLAB 函数 ICDF（'f'，1-α，M-1，m-1）得到，其中 α 是显著性水平。如果一个相关基因显著水平小于设定的显著水平，则认为该基因是周期性表达的，否则认定为非周期性表达基因。

2.3 实验结果与讨论

本章通过对一个模拟数据集和两个生物过程数据集的研究，从不同角度来评估所建模型的性能。

模拟数据集

该模拟数据集是根据参考文献[17-19]中的正弦函数模拟基因表达的周期性行为所产生的。令 x_{ij} 表示在 j 时点上基因 i 的模拟表达序列的值，则有

$$x_{ij} = c_i * \sin(2\pi(j-1)/12 - w_i) + d * \varepsilon_{ij} \qquad (2\text{-}13)$$

对于不同基因，常数 c_i 取值不同，w_i 表示正态分布上区间 [0, 2π] 上选取的基因 i 的相位变换；ε_{ij} 代表标准正态分布下基因 i 在 j 时点上的观测噪声；d 为常数。

令 d=0.5，且从[1,2]中均匀地对 c_i 进行取样，根据式（2-13）得到在 14 个等间距的时间点上，由 300 个基因序列组成的周期性基因表达数据集（PEGED）。当 c=0 且 d=1 的条件下，运用式（2-13）同样可以得到在 14 个等间距的时间点上，由 300 个基因序列组成的非周期性基因表达数据集（nPEGED）。

本章所提出的方法是运用 PEGED 和 nPEGED 来探讨该模型在检测具有周期性表达基因的准确性。当显著性水平为 0.1 时，计算得到的临界值为 2.0802。根据这个临界值，我们可以判断 PEGED 中所有的模拟表达序列都具有周期性，这证明了本章所提出的模型在 PEGED 中的正确率为 100%。另一方面，该方法在 nPEGED 中检测出了 8 个