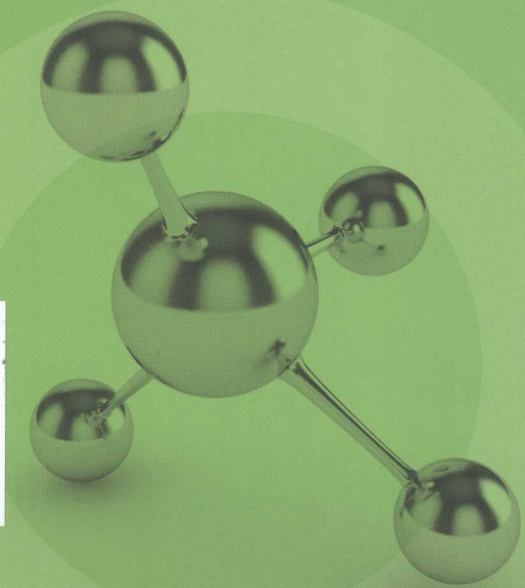
上海交通大学学术出版基金资助项目

生命科学数据处理与MATLAB应用

张雪洪 胡洪波 编著



上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

上海交通大学学术出版基金资助项目

生命科学数据处理与MATLAB应用

张雪洪　胡洪波　编著

上海交通大学出版社

内 容 提 要

本书以 MATLAB 语言为工具，以应用为目的，全面、系统而简洁地介绍了生命科学中常用的数据处理方法。全书共分十六章，内容主要包括生物信息检索与利用、生命科学中的数值方法、生物统计学、生命科学实验数据处理、生命科学中的数学模型及其求解、生命科学实验设计、生命科学中的常用软件等，具体实例涉及生命科学中的各个领域，重点相对突出。

本书可以作为从事生命科学领域工作科技人员的参考书，也可以作为相关专业高年级本科生、研究生的教学参考书。

图书在版编目(CIP)数据

生命科学数据处理与 MATLAB 应用/张雪洪，胡洪波编著. —上海：上海交通大学出版社，2013

ISBN 978-7-313-09439-1

Ⅰ. 生… Ⅱ. ①张…②胡… Ⅲ. 生命科学—数据处理—Matlab 软件 Ⅳ. Q1-0

中国版本图书馆 CIP 数据核字(2013)第 016623 号

生命科学数据处理与 MATLAB 应用

张雪洪 胡洪波 **编著**

上海交通大学出版社出版发行

(上海市番禺路 951 号 邮政编码 200030)

电话：64071208 出版人：韩建民

上海交大印务有限公司 印刷 全国新华书店经销

开本：787mm×960mm 1/16 印张：20.25 字数：383 千字

2013 年 4 月第 1 版 2013 年 4 月第 1 次印刷

ISBN 978-7-313-09439-1/Q 定价：43.00 元

前　言

现代生命学科的发展如果没有计算机的辅助是不可想象的，没有计算机处理海量的基因信息，就不可能有人类基因组学、蛋白质组学等的成就。随着和数学、物理、计算机等学科的交叉，生命科学得到了进一步的发展。目前我国生物学科人才在应用计算机进行数据处理并解决实际问题上显得过于薄弱，因此相关科技人员必须提高计算机应用和数据处理水平，利用计算机软件进行生命科学实验数据处理和实验设计，加深对生命过程的定量描述和本质的了解。

编写本书的出发点基于以下三个方面：

1. 现代生命科学发展的一个重要标志是定量化和模型化

现代生命科学的发展愈来愈多地要求用数学的方法进行定量研究，建立数学模型，以揭示生命现象的本质，加深对生命过程的了解，如种群生态学模型、数量植物生理学模型、数量分类学、数学遗传学等。生物工程更是如此，其主要任务之一是过程数学模型的建立，以有利于过程的参数分析、优化操作与计算机控制，如微生物生长动力学模型、酶催化反应动力学模型等。

为了建立定量的生命科学的数学模型，除了需要掌握相关学科如生物科学与技术、生物工程、生态学等专业基础理论外，因生命系统的复杂性，亟需学习模型化方法中的一些共性问题，利用计算机知识定量探索复杂生物体的发展规律，以及生物系统的发展趋势，以揭示生命现象的本质。

2. 许多生命科学的数学模型无法用经典的数学解析法求解

生命科学及生物工程提出了相当多的复杂数学模型与数学问题，对于求解生命科学领域的复杂数学模型，经典的数学解析法是无能为力的，必须借助计算机算法求解。因此，数值计算在生命科学领域占有极其重要的地位，是现代生命科学发展的促进因素。

读者通过本书的学习，应能掌握各种算法的原理及 MATLAB 软件在生命科学数据处理中的应用，按实际情况选择合理的算法，编写适用的计算机程序，针对实际问题在计算机上算出正确结果，从而根据结果说明数学模型的物理意义，如种群生长动力学的物理意义等。

3. 计算机软件工程方兴未艾

生命科学的发展，使有关的数据、信息极其丰富。为便于数据处理和实际问题的解决，目前已涌现大量与生命科学有关的计算机软件，这为解决实际问题提供了相当方便的条件。

为了科研工作的顺利开展，科技人员应该了解与生命科学相关的软件工程的发展。读者通过本书的学习，应能非常方便迅速地利用相关的软件，特别是目前国际通用的数学软件 MATLAB，解决实际问题，如生物信息检索与利用、生物统计学、生命科学数学模型等。

本书以 MATLAB 语言为工具，以应用为目的，阐述生命科学数据处理中的共性问题，内容包括生命科学中的数值方法、生物统计学、生命科学实验数据处理、生命科学中的数学模型及其求解、生命科学实验设计，以及生命科学中的常用软件和生物信息检索与利用等，具体实例涉及生命科学中的各个领域，如微生物学、遗传学、生物工程、分子生物学、生态学等。

目前虽然数学建模、算法、实验设计等内容在国内外已有许多专著和译著，但大多数侧重于计算机和数学知识，过于数学化，而在生命科学中的应用实例较少，体现出系统性、专业性不够。从事生命科学研究的专业人士阅读这些书也较困难，一般读者即使学过了，也不知如何在生命科学研究中进行应用。

近几年迅速发展起来的计算生物学这一新兴交叉学科，其所要求的数学基础和计算机基础超出了生物学科本科生甚至研究生的要求。对于生命科学技术人员来说，学习本书的目的在于了解实验数据、实验信息的基本处理方法，从众多的实验数据中得到对自己有用的数据，从中探索数据变化的规律，提高分析数据和处理数据的能力。本书通过综合生命科学方面的实例、数据和数学模型，了解计算机在生命科学的数据处理与分析、实验数据模型化中的应用。

本书由张雪洪、胡洪波编著，其中张雪洪编写了第一、第二、第十至第十六章，胡洪波编写了第三至第九章，并由胡洪波负责对本书的 MATLAB 程序作了调试。书中引用的实例来自各种公开文献，在此一并向原作者表示感谢。由于作者水平有限，本书存在的不足之处，恳请广大读者批评指正。

编　者

2013 年 1 月于上海

目　录

第一章　绪论 …… 1
第一节　生命科学数据处理与计算机应用 …… 1
第二节　生命科学中常用的计算机软件概述 …… 6

第二章　生物信息检索与利用 …… 10
第一节　生物信息学概述 …… 10
第二节　常用的生物信息数据库 …… 13
第三节　数据库的检索和应用 …… 19
第四节　蛋白质和核酸的结构与功能的预测分析 …… 23

第三章　数值方法中的误差 …… 32
第一节　学习生命科学中数值方法的意义 …… 32
第二节　近似值和舍入误差 …… 33
第三节　截断误差和泰勒级数 …… 35

第四章　MATLAB 软件与数值计算功能 …… 41
第一节　引言 …… 41
第二节　MATLAB 的语言结构 …… 42
第三节　矩阵、变量、运算和表达式 …… 43
第四节　绘图和控制语句 …… 47
第五节　MATLAB 在线帮助 …… 51

第五章　非线性方程的数值解法 …… 56
第一节　引言 …… 56
第二节　初值估计 …… 57
第三节　简单迭代法 …… 59
第四节　埃特金迭代法 …… 62
第五节　牛顿法 …… 65

第六节　插值法 …… 67
第七节　用 MATLAB 求解非线性方程 …… 70

第六章　线性方程组的数值解法 …… 74
第一节　引言 …… 74
第二节　解线性方程组的直接法 …… 75
第三节　解线性方程组的迭代法 …… 81
第四节　应用 MATLAB 求解线性方程组 …… 90

第七章　插值法和数值微分 …… 92
第一节　引言 …… 92
第二节　拉格朗日插值多项式 …… 93
第三节　二元插值 …… 98
第四节　三次样条插值 …… 99
第五节　数值微分 …… 104
第六节　应用 MATLAB 进行插值和微分计算 …… 106

第八章　数值积分 …… 113
第一节　引言 …… 113
第二节　牛顿-柯特斯公式 …… 115
第三节　变步长梯形求积法 …… 118
第四节　龙贝格求积法 …… 120
第五节　高斯求积法 …… 123
第六节　应用 MATLAB 计算积分 …… 127

第九章　常微分方程初值问题的数值解法 …… 130
第一节　引言 …… 130
第二节　数值解法的基本思想 …… 130
第三节　欧拉方法 …… 131
第四节　龙格-库塔法 …… 135
第五节　用 MATLAB 求常微分方程的数值解法 …… 141

第十章　生物统计学基础 …… 145
第一节　随机变量的分布 …… 145

第二节　随机变量的数字特征——数学期望和方差 …… 149
第三节　样本的特征值和常见分布 …… 150
第四节　参数估计 …… 154
第五节　假设检验 …… 157
第六节　MATLAB 统计工具箱应用简介 …… 161

第十一章　生命科学实验数据的误差分析 …… 167
第一节　实验数据的测量误差 …… 167
第二节　随机误差 …… 168
第三节　随机误差的传递 …… 170
第四节　实验数据的预处理 …… 176
第五节　系统误差 …… 180

第十二章　生命科学中的数学模型建立 …… 186
第一节　实验数据处理和数学模型的建立 …… 186
第二节　数学模型的建立方法 …… 187
第三节　数学模型的选择 …… 190
第四节　生命科学中的数学模型特征 …… 193

第十三章　生命科学中常见的数学模型 …… 196
第一节　生物传递模型 …… 196
第二节　生物种群的指数增长模型 …… 198
第三节　生物种群相互作用模型 …… 200
第四节　生态数学模型 …… 202
第五节　药物动力学模型 …… 204
第六节　群体遗传学模型 …… 207
第七节　生命科学的其他典型数学模型 …… 209

第十四章　数学模型的求解与线性回归 …… 212
第一节　数学模型的求解和最小二乘法原理 …… 212
第二节　实验数据的一元线性回归 …… 213
第三节　多元线性回归 …… 219
第四节　逐步回归法 …… 236
第五节　回归方程的预测和控制 …… 239

第六节　线性回归在 MATLAB 的实现 …… 241

第十五章　非线性模型的求解 …… 247
第一节　非线性模型的线性化 …… 247
第二节　非线性模型的拟合 …… 251
第三节　非线性回归模型的检验 …… 254
第四节　非线性回归在 MATLAB 的实现 …… 257
第五节　最优化方法及 MATLAB 优化求解 …… 259

第十六章　生命科学实验设计 …… 264
第一节　实验设计概述 …… 264
第二节　单因素设计和双因素设计 …… 266
第三节　正交实验设计 …… 271
第四节　回归正交设计 …… 278
第五节　序贯实验设计 …… 290

附录 …… 294
附录 1　标准正态分布表 …… 294
附录 2　t 分布表 …… 296
附录 3　χ^2 分布表 …… 297
附录 4　F 分布表 …… 299
附录 5　常用正交表 …… 303
附录 6　回归正交设计表 …… 306
附录 7　正交拉丁方表 …… 309
附录 8　MATLAB 常用操作符与函数 …… 310

主要参考文献 …… 315

第一章 绪 论

第一节 生命科学数据处理与计算机应用

随着生命科学和计算机技术的发展，计算机在生命科学领域中的应用越来越普遍，计算机已广泛应用于微生物学、遗传学、生态学、医学、人口学、药物动力学、生理学、分子生物学等领域。同时，生物信息学、数值方法、数据模型化、最优化实验设计等在生命科学中越来越显示出强有力的作用。

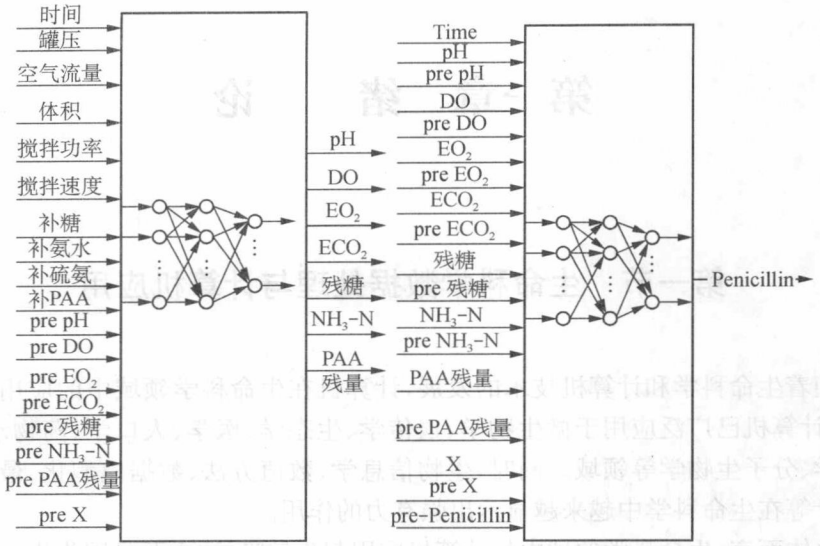
总体而言，生命科学领域中的计算机应用起步较晚，这主要是因为生命过程非常复杂，影响因素众多，内在机理研究难以深入，其具体的定量方法远远不能满足要求，需要人们对其进一步研究和探索。同时，生命科学也逐步从经验科学向理论科学发展，这对生命科学工作者提出了较高的数学和计算机应用的要求。如以计算机数据分析应用较多的微生物发酵领域为例，影响微生物发酵过程的参数很多，综述如下：

物理参数：温度、压力、搅拌速度、输入功率、气体流量、液体流量、黏度、泡沫程度、发酵液体积和发酵液密度等。

生化参数：溶氧和溶解的二氧化碳、排出的氧浓度、排出的二氧化碳浓度、基质浓度（如蛋白质、糖）、产物浓度、菌数、代谢物浓度、酸度和细胞内组分含量（DNA、RNA、ATP、ADP）等。

还有为了达到以上各个参数要求的控制参数，如：加入的酸、碱量和浓度、消泡剂量、溶液稀释率、前体、诱导剂等。

同时微生物发酵系统还是一个非线性、非静态、非稳态的多变量输入输出系统，因此系统的计算机应用相当复杂。如计算机在青霉素发酵的仿真中的应用（见图 1-1），以便对发酵工艺进行自动控制，仿真系统的左网络指各种操作变量与发酵的状态变量及有关计算变量之间的关系，右网络指各种状态变量与青霉素产出之间的关系，要建立整个系统的数学模型，了解各个参数之间的相互作用及应答关系非常困难。另外，计算机在微生物发酵过程的应用是多方面的，包括生物反应和反应器设计、产物的分离纯化、参数的测量控制、过程分析评价、过程的设计放大等。在其他生命领域的应用更是这样。



注：带 pre 前缀的为相同级网络的输出反馈

图 1-1　计算机在青霉素发酵的仿真中的应用

20 世纪 80 年代起，现代生物技术和分析测试方法的迅速发展，极大地丰富了生命科学的数据资源，而且数据的质量也大为提高。大量多样化的生命科学数据资源中蕴含着大量重要的生物学规律，这些规律是我们解决许多生命之谜的关键所在。面对如此繁杂数据的分析处理，采用传统的一些手段已力不从心，只有借助计算机的应用，才能逐步实现。

计算机在生命科学中的应用，总体上可分为以下五种类型：

(1) 计算机在生命科学领域的数据采集。实验数据的在线检测，如常规的温度、压力、pH 值、溶氧浓度；生物医学中的葡萄糖浓度、脑电流等生物电信号；实验数据的离线检测，如蛋白质浓度、酶活；DNA、RNA 等核酸浓度的测定，代谢中间产物等生物特性物质的检测，生物种群数目的统计等。由于生物数据的量大面广，依靠传统的人工采集数据的方法已不能适应需要。

以往，计算机在线检测生命科学的参数是相当困难的。生物数据采集的需求促进了新型生物传感器的设计和研究的快速发展，利用生物物质和酶等生物分子之间作用产生的光、电、热、质量等可以定量的物质，进行数学定量建模和计算，研究其相互之间的关系，通过计算机自动信号处理，测定氨基酸、胆固醇、糖、AMP、维生素等的浓度。利用这一原理，制成各种酶电极、细胞电极、生物分子电极及其检测系统等。对生物传感器，要求其测量误差小，灵敏度高，响应快，信号转换快，

因此生物传感器及生化测量仪器必然要应用计算机技术，特别是进行大量数据采集处理的仪器，如色谱仪、质谱仪、核磁共振仪等。

(2) 计算机对生命科学实验数据的处理。包括生命科学中各种实验数据的处理，生命科学数学模型的建立和求解，利用数学模型对实验的控制和实验监测，实验跟踪生物量、生物参数，以及生命科学和生物工程的实验设计包括最优化实验设计。如将所测定的与 DNA 序列对应的光谱数据进行整理和处理后确定核苷酸的位置；放射性示踪物在生物分子中的研究应用；利用计算机按分子量大小或其他特性自动分离生物物质；利用计算机对生物工厂进行工艺优化设计，对实验测量值的误差进行自动分析处理等。

基因芯片技术是基因研究领域中一项非常重要和关键的实验技术，对该技术所产生的大量实验数据，也必须采用计算机进行高效分析，从中获得基因研究的众多信息。

在所有的数据处理和数据分析中，应用计算机建立和求解生命科学领域的数学模型意义非常重大，而且生命科学数学模型化的研究正逐步由静态向动态发展。

计算生物学(computational biology)实际上是综合性、系统性的处理复杂生命科学实验数据的新兴学科。运用大规模高效的数据分析，统计学、信息论、集合论等数学与计算机理论方法，用数学建模、计算机仿真(或人工智能)等技术对生物数据进行处理，包括代谢网络的优化、药物结构分析、基因组序列识别等。计算生物学利用其卓越的数值计算能力进行生物学的研究。它处理的是数量巨大的生物学资料、数据及对其进行复杂的计算。要达到上述计算生物学的目标，不仅需要先进的计算机硬件，合适而且高效的软件或设计优良的演算方法也是必须的。

(3) 计算机在生物信息学中的应用。计算机对生物信息的处理是数据处理中的一个特殊部分，由于生物信息学的迅速发展，这已成为一个单独的应用领域。生物信息学是以计算机为工具对生物信息进行储存、检索、传输和分析的学科，涉及范围很广，其研究重点一般为两个方面，即基因组学(genomics)和蛋白质组学(proteomics)。它们涉及对核酸和蛋白质序列信息的获取、分析和存储，数据的查询和校对等，包括对大量基因组数据、蛋白质组数据信息，如 GenBank、生物分子结构数据库 MMDB，以及生物类文献，如 MEDLINE 数据库和 BA(biological abstract)数据库的检索等。

在蛋白质结构的分析和功能的预测方面，蛋白质的折叠类型与其氨基酸序列具有相关性，这样就有可能直接从蛋白质的氨基酸序列通过计算机辅助方法预测出蛋白质的三维结构。而由于蛋白质以及一些核酸、多糖的三维结构获得精确测定，基于生物大分子结构知识的计算机辅助药物设计也成为了当前的热点。

人类基因组计划需要对测定的 30 亿个碱基、3 万个基因进行分析和定位，进

而弄清楚其中所有功能单位的组织结构形式以及调节机制，没有计算机的帮助是难以想象的。近几年基因组学、蛋白质组学、代谢组学等的飞速发展，提供了海量的实验数据，迫切需要发展新的数据分析技术和计算机软件，以快速获取所希望的数据和信息。除了基因、蛋白质等信息外，生物信息学的研究可提供生物大分子及其空间结构的信息，还能提供电子结构包括能级、表面电荷分布、分子轨道相互作用以及动力学行为等的信息，如生物化学反应中的能量变化、电荷转移、构象变化等。生物信息学的理论模拟还可研究包括生物分子及其周围环境的复杂体系和生物分子的量子效应等。

(4) 计算机数值方法在生命科学中的应用。事实上，实验数据的处理，包括生物信息学的数据处理，都是以数值方法和统计学的知识为基础的。现代生命科学提出相当多的数学问题及复杂的数学模型，涉及许多非线性的代数或微分方程，这些方程常常是大量耦合的。对于这类复杂的数学模型的研究，经典的数学解析法是无能为力的，必须借助数值方法应用计算机求解。因此，计算机数值法在生命科学领域内占有极其重要的地位，是现代生命科学技术发展的促进因素。

数值方法又称非直接解法，它应用算术运算求解实际数学问题，其求解结果是近似且离散的。在现代科学研究和工程计算中，计算机数值方法已成为研究人员和工程技术人员必不可少的手段。

绝大多数实际问题的求解采用经典的解析方法并不适宜，甚至很难得到结果。例如五次以上的多项式方程就没有公式解法，所有的超越方程更没有公式解；有的问题虽有解析解，但由于函数关系太复杂，其实用价值也不大。尽管图解法通常可用来解决复杂计算问题，但只限于有限的能应用三维或更低维图形求解的实际数学问题，且结果很不准确，另外，无计算机帮助的图解法非常费时甚至很难实现。因此，采用计算机数值方法求解实际问题已成为一种重要的处理方法。只要掌握数值方法，合理地选择、使用或编写计算机程序，就能够利用计算机解决实际数学及计算问题。

由于数值方法的发展和许多实际问题的提出，计算机数值计算软件已大量涌现。子程序库的存量逐年迅速增加，为生命科学技术人员解决科学与工程问题提供了便利的条件。但是对于缺乏数值方法知识和应用能力的人而言，绝不可能有效地应用这些子程序解决实际计算问题。因为在使用任何精密而完善的子程序去解决具体问题时，很可能遇到种种难题，这些困难可能由如下某些原因所引起：数学模型没有准确地反映实际生命现象和过程，选用的数值方法不恰当，方法的误差超过实际问题允许的误差，选用子程序的实际使用条件不恰当，在解决具体工程问题时未能对子程序做相应的修改或调整，等等。实际上，在应用程序或使用任何子程序时，都需要用户根据实际问题进行二次开发。至于在众多的子程序中选择适

合于解决具体计算问题的最优子程序，则需要更为坚实的数值方法基础。因此，掌握数值方法对于生命科学技术人员是相当重要的。

(5) 计算机用于生物工程和生命活动的过程控制和过程监督。如在发酵工程中，控制方式有人工控制和计算机控制两种，但目前大部分还是以人工控制或半人工控制为主，包括经典的自动控制、顺序控制、模拟控制等。计算机全自动控制能直接实现人机对话，利用系统的数学模型实现过程优化，如医学上人工血液输送系统(人工心脏等)的控制。

过程控制应用的范围较广，如生产过程的最优化，包括原料的成分配比、传递过程的最优化、生产动力成本优化、降低生产劳动强度等。计算机在食品加工、发酵工程、生物制药、生物环境保护等生命科学领域中实现了大量的成功控制，提高了生产和管理的经济效益。

计算机在上述领域的应用，还包括与以上领域相关的计算机软件的开发。计算机在生命科学上的广泛应用，大大促进了生命科学的发展，促进了我们对生命现象和人类自身的了解，并对相关产业起了很大的推动作用。现在生命科学已不再是仅仅基于试验观察的科学，仅靠传统的研究手段是无济于事的，理论和计算将发挥越来越巨大的作用，数学、物理、计算机科学将日益深入地渗透到生物学研究中，大量数据必须经过计算机收集、分析和整理后才能成为有用的信息和知识，为人类所使用。

本书以计算机应用为目的，主要阐述计算机在生命科学数据处理中的共性问题，内容包括生物信息检索与利用、生命科学中的数值方法、生物统计学、生命科学实验数据处理、生命科学中的数学模型及其求解、生命科学实验设计、生命科学中的常用软件等部分，具体实例涉及生命科学中的各个领域。重点介绍生命科学中的数值方法、生命科学实验数据处理和数学模型。本书所涉及的数学和计算机基础内容仅是基本的高等数学、线性代数、概率论与数理统计及计算机编程知识。

对于研究生命科学的技术人员来说，学习和应用纯粹的数值方法比较困难。本书主要以 MATLAB 软件为工具，介绍如何掌握各种数值计算技巧，合理利用计算机解决实际计算问题，学习的重点放在各种算法的应用上，而对某些数学问题及其证明仅作一般性了解即可。学习数值方法的关键不仅在于能从原理上理解各种算法，更重要的是合理选择和应用这些算法去解题。检查自己对某个算法是否掌握，要看能否应用这种算法编写适用的计算机程序，并在计算机上对实际问题做出正确处理。也就是说，学习数值方法不仅要在理论上学懂，而且要注重计算机上的实践。本书中这一部分就是为使生命科学领域技术人员掌握数值方法的基本知识而编写的，其内容包括非线性方程求根、代数方程组求解、插值法、数值积分、常微分方程及其方程组求解等，同时列举了与生命科学领域有关的实例，力图让读者能

学以致用。

在生命科学中，传统的研究方法如经验归纳法等已不能满足学科发展的需要，在生物工程上数学模型已成为一种重要的研究方法。现代生命科学的发展越来越多地要求用数学的方法对生命过程进行定量研究，建立数学模型，以揭示生命现象的本质。

数据处理和实验设计是一门综合性的学科，目前它和各门具体学科相结合，已成为各门学科的重要组成部分，并随着各门具体学科的发展一起发展。它和生命科学相结合，被具体用于生命科学实验数据处理、生命科学的建模和生命科学中的实验设计，对生命科学的发展发挥了积极作用。本书中这一部分的具体内容包括生命科学实验数据的误差及其分布、实验数据常用的处理方法、生命科学中数学模型的建立方法以及生命科学中常见的数学模型。以最小二乘法为基础，介绍实验数据的回归分析及其检验，还介绍了实验数据常用的设计方法、回归正交设计和序贯实验设计。对于生命科学技术人员来说，学习这部分内容的目的在于了解实验数据的基本处理方法，从众多的实验数据中得到有用的数据，从中探索数据变化的规律，提高分析数据和处理数据的能力，综合生命科学方面的实例、数据和数学模型，了解和掌握计算机在生命科学数据处理与分析、实验数据模型化中应用的基本思想和方法。

第二节　生命科学中常用的计算机软件概述

生命科学中的各个领域以及和生命科学相交叉的各学科，均开发了多种应用软件，如用于数值计算的 MATLAB 软件、药物设计的分子构型软件、微生物发酵工程中的控制软件、生物医学的仿真软件、计算机辅助设计 AUTOCAD 等。从网络资源来看，国外互联网上的生物信息学、生物软件网点非常多，大到代表国家级的研究机构，小到代表专业实验室的网点都有，大型机构的网点一般提供相关新闻、数据库服务、应用软件和软件在线服务，小型科研机构一般还提供自己设计的算法、应用软件的在线服务。下面简单介绍几种在生命科学领域应用较广的一些软件。

1. GCG 软件

生物信息学中使用较广的 GCG (Genetics Computer Group)软件主要提供一种计算机集成环境，它将大量序列分析和数据库搜索程序集成在一起，可以访问各种不同来源的序列数据库。它提供的集成环境 SeqLab(图形用户界面)是 Wisconsin Package 的一部分。Wisconsin Package 则是一种综合性的序列分析程

序，由 130 个独立的程序组成，用户为适应不同要求，可对其程序进行组合使用。

GCG 支持 5 种数据库供 Wisconsin Package 使用，分别是 2 种核酸数据库和 3 种蛋白质数据库。2 种核酸数据库是 GenBank 数据库和 EMBL 核酸序列数据库，这 2 种数据库也可被组合成为一个库，称为 GenEMBLPlus。3 种蛋白质数据库是 Protein Information Resource(PIR)国际蛋白质序列数据库、SWISS－PROT 蛋白质序列数据库、SP－TrEMBL 数据库(由欧洲生物信息学研究所等开发)。GCG 可以用于核酸和蛋白质序列的编辑、搜索、比较、分析等，还可用于进化分析、引物选择等。但目前原开发商 Accelrys 公司已不再提供 GCG 的程序维护与更新工作，SeqLab 已被 Accelrys Pipeline Pilot 软件替代。

2. SAS 软件

SAS(Statistics Analysis System)软件是目前国际上较流行、较权威的一种统计分析软件，由美国 SAS 公司研制，其版本不断更新，它可作为统计计算和绘图的工具。由于在生命科学的实验数据处理上常需要应用回归的方法和统计的方法，因此 SAS 软件被广泛用于生物学、医学、药学等的研究中。

SAS 软件是模块式结构，具有约 30 个模块，其中常用的有 SAS/BASE(基础)模块、SAS/STAT(统计)模块、SAS/GRAPH(图形)模块、SAS/ETS(预测)模块、SAS/IML(矩阵运算)模块、SAS/QC(质量控制)模块等。这些 SAS 模块可以独立使用，也可以相互结合使用。SAS 软件能够解决统计分析和实验设计的一些问题。

SAS 软件的使用是建立在 SAS 数据库之上的，而实现 SAS 程序的成功应用必须由用户编制 SAS 引导程序。SAS 引导程序由一系列符合 SAS 语言的语法规则的语句组成。尽管使用前必须对 SAS 语言有一定的了解，但总体来说，使用还是较为方便的。

3. 数值计算用的 Excel 和 MATLAB

Excel 为 Microsoft 公司的产品，Excel 扩展图表是一种特制的数学软件，用户可在数据行或列中输入数据并运算，可完成大量的报表的运算和输出。图表中任何一个数据发生改变时，软件都会更新计算结果。

Excel 内部也具有部分数值计算的功能，如方程求解、曲线拟合和最优化等，同时 Excel 以 Visual BASIC 作为用户开发应用的编程语言，可用于数值计算，另外 Excel 也具有可视化工具(如三维立体作图)，将数值计算和作图结合起来使用，可以进行具有相当难度的数值分析。

MATLAB 为 Mathworks 公司的主要产品，MATLAB 出自 MATrix LABoratory，原意为矩阵实验室，最开始是专门用于矩阵计算的软件。随着 MATLAB 推向市场，MATLAB 不仅具有了数值运算功能、符号运算功能，而且还具有了数据图示功能。在目前的最新版本 MATLAB 2012a 和 2012b 中，

MATLAB不仅在数值、符号和图形等功能上有了进一步增强，而且又增加了一些工具箱，以方便不同专业技术人员使用。

MATLAB中的函数和运算器有助于本书中多种数值方法的实现，这将在以后章节中作详细的介绍。另外，MATLAB作为一种高级语言，它不仅可以以一种人机交互式的命令行指令操作方式工作，而且还可以如BASIC、FORTRAN、PASCAL、C等高级语言一样进行程序设计，编制一种以m为扩展名的文件，即M文件。由于MATLAB本身的特点，M文件的编制与BASIC、FORTRAN、PASCAL、C等比较起来，有许多无法比拟的优点，如语言简单、可读性强、调试容易、调用方便等，因此可以通过简单编程方便地实现数值计算。

4. 其他的应用软件

(1) DNA分析软件：Gene Construct Kit与大多数DNA分析的软件不同，它管理并显示克隆策略中的分子构建过程，包括分子构建、电泳条带。另外，还可以质粒作图(有序列没序列均可)。该软件附有详细的在线帮助，可供从事分子生物实验人员和克隆策略人员备用。

(2) 蛋白质分析软件：MACAW是一种多序列构建与分析软件。通过实施基因组计划，得到了大量的蛋白序列与DNA序列数据，但在如此多的数据中，了解其相互关系，查找有用的片段是非常困难的工作，这样一些片段常常显示类似的分子结构与生物特性。用人工的方法不可能完成如此大量的比较工作。可应用MACAW程序，借助统计学方法和一定的运算规则，来查找这些片段。

蛋白序列分析软件包ANTHEPROT 5.0，包括了蛋白质研究领域所包括的大多数内容，功能非常强大。应用此软件包，使用个人计算机便能进行各种蛋白序列分析与特性预测，包括以下的多种功能，如进行蛋白序列二级结构预测；在蛋白序列中查找符合PROSITES数据库的特征序列；绘制蛋白序列的所有理化特性曲线；在Internet或本地蛋白序列数据库中查找类似序列；计算蛋白序列相对分子质量、密度与各蛋白残基百分组成；计算蛋白序列滴定曲线与等电点；选定一个片段后，绘制Helical Wheel图；进行点阵图(Dot Plot)分析；计算信号肽潜在的断裂位点等。

(3) 搜索查看软件：Vector NTI Viewer虽然只是一个载体查看软件，但其功能方便易用：①输入文件格式广泛，除了molecule documents(.gb)是该公司本身文件格式外，还能识别各种数据库应用格式软件：EMBL，GenBank，FASTA，Sequence files。②全部或部分序列可以拷贝到剪贴板，提供给其他程序使用。③载体图像可以拷贝到剪贴板上供给其他程序使用，或直接成.wmf格式图像文件存盘。④可以查找特定序列、ORF(可以设置相关参数)、描述载体、限制酶位点、一些功能序列和附注。⑤整个界面由文本、图形和序列三部分构成，而且点击任意的