张雪洪 杭晓敏 胡洪波 编著

计算机在生命科学中的应用

JISUANJI ZAI SHENGMING KEXUE ZHONG DE YINGYONG

上海交通大学出版社

计算机在生命科学中的应用

张雪洪　杭晓敏　胡洪波　编著

上海交通大学出版社

内 容 提 要

本书主要讲述计算机在生命科学中应用的共性问题，以计算机应用为目的，全面、系统而简洁地介绍了计算机在生命科学中的应用。理论联系实际，通俗易懂，有助于读者较快的掌握。内容主要包括生物信息学、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几个部分，具体实例涉及到生命科学中的各个领域，相对突出重点，并应用 MATLAB 软件为主要计算工具。

本书可以作为从事生命科学领域工作的科技人员的参考书，也可以作为相关专业本科生、研究生的参考书籍。

图书在版编目(CIP)数据

计算机在生命科学中的应用/张雪洪，杭晓敏，胡洪波编著. —上海：上海交通大学出版社，2003

ISBN7－313－03293－5

Ⅰ. 计... Ⅱ. ①张...②杭...③胡... Ⅲ. 电子计算机—应用—生命科学 Ⅳ. Q1－0

中国版本图书馆 CIP 数据核字(2003)第 004489 号

本书出版由上海科技专著出版资金资助

计算机在生命科学中的应用

张雪洪 杭晓敏 胡洪波 编著

上海交通大学出版社出版发行

（上海市番禺路 877 号 邮政编码 200030）

电话：64071208 出版人：张天蔚

常熟市文化印刷有限公司 印刷 全国新华书店经销

开本：787mm×960mm 1/16 印张：18.25 字数：343 千字

2003 年 1 月第 1 版 2003 年 1 月第 1 次印刷

印数：1－2050

ISBN7－313－03293－5/Q・010 定价：29.00 元

出版说明

科学技术是第一生产力。21世纪，科学技术和生产力必将发生新的革命性突破。

为贯彻落实“科教兴国”和“科教兴市”战略，上海市科学技术委员会和上海市新闻出版局于2000年设立“上海科技专著出版资金”，资助优秀科技著作在上海出版。

本书出版受“上海科技专著出版资金”资助。

上海科技专著出版资金管理委员会

推动科技出版事业
提高学术研究水平

为「上海科技专著出版资金」题

徐匡迪

二〇〇〇年十一月十一日

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。没有计算机的应用，就不可能有人类基因组学、蛋白质组学的成就。随着生命科学和数学、物理、计算机等学科的交叉，生命科学得到了进一步的发展。但是目前我国生物学科人才在应用计算机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的应用水平，利用计算机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，对过程的本质了解。

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等。

为了建立定量的生命科学的数学模型，我们除了需要掌握相关学科如生物技术、生物工程、生态工程等的专业基础理论外，因生命系统的复杂性，极需学习模型化方法中一些共性问题，利用计算机知识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的数学模型，经典的数学解析法是无能为力的，而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，能够掌握各种算法的原理及其在生命科学中的应用，根据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。

3. 计算机软件工程的方兴未艾

生命科学的发展，使有关的数据、信息变得极其丰富。为便于数据处理和实际问题的解决，目前已涌现出大量和生命科学有关的计算机软件，而且子程序库

的存量也在逐年迅速增加，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此相关的软件工程的发展。读者通过本书的学习，能够非常方便迅速地利用相关的软件解决实际问题，如生物信息学、生物统计学、生命科学数学模型等。

根据以上意图，本书主要讲述计算机在生命科学中应用的共性问题，以计算机应用为目的，内容主要包括以上介绍的生物信息学、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几个部分，具体实例涉及到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，但大多数均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较少，体现出系统性不够。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中进行应用。

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。

参加本书编写的有张雪洪、杭晓敏、胡洪波，其中张雪洪编写了第 1，2 和 10~16 章，杭晓敏、胡洪波编写了第 3~9 章，并由胡洪波负责对本书的 MATLAB 程序作了调试。书中引用的例子来自各种公开文献，在此一并向原作者表示感谢。特别感谢徐秋栋老师为本书修订所做的大量工作，最后感谢提供出版经费的上海市科技专著出版资金管委会。

由于作者水平有限，本书难免有错误和不足之处，恳请广大读者批评指正。

编 著 者

2002 年 7 月于上海

目　录

1 绪 论

1.1 计算机在生命科学中的应用

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。如在计算机应用较多的微生物发酵领域，影响微生物发酵过程的参数很多，其中有以下一些。

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

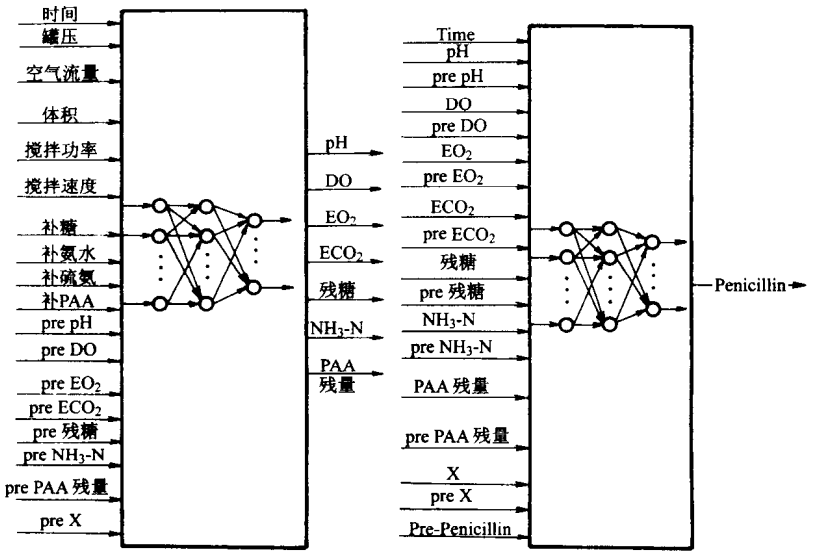
生化参数：溶氧、溶解的二氧化碳、排出的氧浓度、排出的二氧化碳浓度、基质浓度(如蛋白质、糖)、产物浓度、菌数、代谢物浓度、酸度和细胞内组份含量(DNA、RNA、ATP、ADP)等；

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、消泡剂量和溶液稀释率等。

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中的仿真应用[见图1-1，张嗣良等(生物化工年会 1996)]，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的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。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。

20 世纪 80 年代起，现代生物技术的迅速发展，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。对如此繁杂的数据的分析处理，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。



注:带Pre前缀的为相同级网络的输出反馈

图1-1 青霉素发酵仿真系统结构

计算机在生命科学中的应用总体上可分为下面五种类型。

1) 计算机在生命科学领域的数据采集

这包括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，如蛋白质浓度和 DNA、RNA 等核酸浓度的测定，生物特性物质的检测，生物种群数目的统计等等。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，研究其相互之间的关系，通过计算机自动信号处理，来测定氨基酸、胆固醇、糖、AMP、维生素等的浓度。利用这一原理，制成了各种酶电极、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需要应用计算机技术，特别是在进行大量数据的采集处理仪器中，如色谱仪、质谱仪中尤为重要。

2) 计算机对生命科学实验数据的处理

这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(含最优化实验设计)。如将所测定的 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，其意义非常重大，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。

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、传输和分析的科学，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。

人类基因组计划是要求测出人类基因组的全部脱氧核糖核苷酸序列(原估计其中编码有约十万多个蛋白质基因，现估计为四万个基因)，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，并绘制成直观图谱，该计划实现之后更深入的工作就是要弄清楚基因组所编码的所有蛋白质的表达情况。显然，没有计算机的帮助，人类基因组计划是无法完成的。

近几年人类基因组学和蛋白质组学上的飞速发展，为我们提供了大量的实验数据。生物信息学的研究不仅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。现代生命科学提出了相当多的数学问题及复杂的数学模型，涉及到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，它是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现代科学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可解；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实际工程问题提供了便利的条件。但是对于缺乏数值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，并对相关产业起到了很大的推动作用。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来，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，主要讲述计算机在生命科学中应用的共性问题，内容主要包括上述涉及的生物信息学、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几个部分，讲述的实例涉及到生命科学中的各个领域。重点讲述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。

对于生命科学技术人员来说，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，而且更重要的是在于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，而且注重计算机上的实践。本书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，并和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，介绍了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，使他们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。

1.2 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，大到代表国家级研究机构的，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

这是生物信息学中使用较广的软件。GCG(Genetics Computer Group)主要是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 120 多个独立的程序组成，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。