Developing Bioinformatics Computer Skills

生物信息学中的计算机技术

O'REILLY®
中国电力出版社

Cynthia Gibas & Per Jambeck 著
孙超 郭庆民 刘相国 吴斌 译

生物信息学中的计算机技术

Cynthia Gibas & Per Jambeck 著

孙超 郭庆民 刘相国 吴斌 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Paris • Sebastopol • Taipei • Tokyo

O'Reilly & Associates, Inc. 授权中国电力出版社出版

中国电力出版社

图书在版编目（CIP）数据

生物信息学中的计算机技术/（美）吉伯斯（Gibas, C.），詹姆贝克（Jambeck, P.）著；孙超等译．-北京：中国电力出版社，2002.6

书名原文：Developing Bioinformatics Computer Skills

ISBN 7-5083-1052-7

I. 生… II. ①吉… ②詹… ③孙… III. 计算机应用－生物信息论

IV. Q811.4-39

中国版本图书馆CIP数据核字（2002）第033523号

北京市版权局著作权合同登记

图字：01-2001-1209号

书　　名/ 生物信息学中的计算机技术

书　　号/ ISBN 7-5083-1052-7

责任编辑/ 关敏

封面设计/ Ellie Volckhausen，张健

出版发行/ 中国电力出版社（www.infopower.com.cn）

地　　址/ 北京三里河路6号（邮政编码100044）

经　　销/ 全国新华书店

印　　刷/ 北京市地矿印刷厂

开　　本/ 787毫米×1092毫米　16开本　28印张　406千字

版　　次/ 2002年7月第一版　2002年7月第一次印刷

印　　数/ 0001-5000册

定　　价/ 49.00元（册）

O'Reilly & Associates 公司介绍

为了满足读者对网络和软件技术知识的迫切需求，世界著名计算机图书出版机构O'Reilly & Associates公司授权中国电力出版社，翻译出版一批该公司久负盛名的英文经典技术专著。

O'Reilly & Associates 公司是世界上在 UNIX、X、Internet 和其他开放系统图书领域具有领导地位的出版公司，同时是联机出版的先锋。

从最畅销的《The Whole Internet User's Guide & Catalog》（被纽约公共图书馆评为二十世纪最重要的50本书之一）到GNN（最早的Internet门户和商业网站），再到WebSite（第一个桌面PC的Web服务器软件），O'Reilly & Associates一直处于Internet发展的最前沿。

许多书店的反馈表明，O'Reilly & Associates是最稳定的计算机图书出版商——每一本书都一版再版。与大多数计算机图书出版商相比，O'Reilly & Associates公司具有深厚的计算机专业背景，这使得O'Reilly & Associates形成了一个非常不同于其他出版商的出版方针。O'Reilly & Associates所有的编辑人员以前都是程序员，或者是顶尖级的技术专家。O'Reilly & Associates还有许多固定的作者群体——他们本身是相关领域的技术专家、咨询专家，而现在编写著作，O'Reilly & Associates依靠他们及时地推出图书。因为O'Reilly & Associates紧密地与计算机业界联系着，所以O'Reilly & Associates知道市场上真正需要什么图书。

作者简介

Cynthia Gibas，位于弗吉尼亚州Blackburg的弗吉尼亚理工学院的生物学副教授。她在计算生物学受到青睐之前就已经是计算生物学家了。目前她正在推动全新的家庭Linux集群。她的研究兴趣包括基因组结构和进化、蛋白质表面和界面特性，以及蛋白质结构预测。她为生物学家们讲授生物信息学方法的课程。她正期盼着下一个假期的来临，可能在2006年的某个时候。

Per Jambeck，加州大学圣迭戈分校生物工程系的博士研究生。他从1994年开始从事计算生物学的研究，兴趣主要集中在如何将机器学习应用于多维生物学数据的理解。在谈及空闲时间时，Per总是充满渴望地微笑着，他时常组织主持社区演出和学生广播站。

封面介绍

本书封面上的动物是一种很小的线虫，Caenorhabditis elegan，与它的那些肮脏的近亲不同，这种线虫生活在土壤中，靠微生物和细菌为生。它大概能长到1毫米长。

尽管是一种原始的生物，这种动物在本质上却和人类有着许多相同的生物学特征。C. elegan的生命是从单细胞开始的，经分裂长成多细胞的成熟个体。它有神经系统和脑（更确切地说是围咽环，circumpharyngeal ring）以及支持运动的肌肉系统。它有一定的行为，能进行简单的学习。与人类一样，它也分为两性，但它分为雄性和自体受精的雌雄同体。C. elegan在实验室条件下很容易大量生长，寿命只有短短的2~3周，可在复杂的实验中对它进行操作。这些特性使得它成为进行科学研究的理想生物。

雌雄同体的C. elegan有959个细胞，300个是神经元，81个是肌肉细胞。通过发育学研究，完整的细胞世系已经搞清楚了。成体头部有大量感受味觉、嗅觉、触觉和温度的感觉器官。虽然没有眼睛，但它对光有轻微的反应。C. elegan有大概17800个不同的基因，它的基因组已全部测序。与果蝇、小鼠、拟南芥菜一起，C. elegan已成为自Sydney Brenner几十年前首次注意这种生物以来、生物学中最常见的模式生物之一。

目录

第三部分 生物信息学工具

前言

计算机和万维网（World Wide Web）正在迅速而显著地改变着生物学研究的面貌。现在，“模式转变（paradigm shift）”这个词可以来描述一切事物——从新的经济趋势到新口味的可乐。但是，生物科学的模式转变还是传统意义上的。理论和计算生物学已经在生物科学的“边缘”中存在了几十年。但就在短短的几年间，基因组研究中大量新的生物学数据和分析这些基因组数据所必需的计算机应用开始影响到生物学的每一个方面。过去科学家要在实验室进行的研究，现在可以在计算机上进行，因为需要查找数据库信息，从中提出新的假说。

在过去的20年中，PC机和超级计算机已经走进了各学科科学家的生活。一开始，PC机是作为实际计算能力很小的昂贵的新生事物出现的，现在它已经具有10年前超级计算机的计算能力。就像计算机已经代替了写作者的打字机和会计的分类账目一样，在实验室中，它们在从实验室仪器中控制和收集资料方面发挥着作用，作为一种存储资料的方法，计算机已有能力完全代替实验室记录和文件。计算机数据库存储的数据比其他非电子形式的记录要容易访问。除了在数据的存储、分析和可视化方面的作用以外，在对任何可用数学方式进行描述的系统的理解上，计算机的功能也非常强大，因而出现了计算生物学这一学科，更新的叫法是生物信息学。

生物信息学（bioinformatics）是信息技术在生物数据处理上的应用，它是一门快速发展的学科。在过去的20年中，将生物数据存储在公共的数据库中变得越来越普遍了，而且这些数据库也在呈指数增长。生物学的文献也在呈指数增长。如果没有计

算机工具的帮助，即使是最热情的研究工作者，想要得到相关领域内最新的必要信息也是不可能的。通过Web，用户在任何地方都可以与任何站点上的数据和资料进行交互 —— 只要他们知道怎样建立正确的工具。

生物信息学首先是生物科学，但是它又不仅仅是生物科学。它往往不是致力于研究精致的算法，而是解决实际中的问题。生物信息学家是工具建立者，为了建立有用的工具，对于他们来说，对生物学问题的理解和对计算方法的理解同等重要。生物信息学算法需要涉及复杂的科学推测，这些推测以独特的方式使程序开发和数据建模变得更加复杂。

生物信息学和计算生物学的研究范围包罗万象：从将生物系统中提取出的特性转变为数学或物理模型，到实现数据分析的新算法，再到开发数据库和Web工具来访问它们。为了从事计算方面的研究，生物学家必须熟练应用在各种操作系统中运行的多种软件工具。本书介绍并解释了在生物信息学研究中最常用的工具。本书包括许多附加信息和背景资料，以帮助理解如何很好地利用这些工具，并解释它们的重要性。我们希望本书能帮助你迈出在研究工作中高效应用计算机的第一步。

本书的读者

大多数生物科学的学生和研究人员应用计算机的目的越来越不仅限于进行文字处理或数据收集和作图。但是，许多人没有计算机科学或计算理论的背景，对他们而言，计算生物学和生物信息学领域可能非常广泛而复杂。本书是受我们的学生和同事的启发而撰写的，它绝不是涵盖了生物信息学各方面的圣经。然而，本书对生物信息学一些最重要的主题的介绍却是非常深入的。为了找到生物序列、基因组和分子结构数据库的信息，我们介绍了标准的计算技术；我们讨论了如何鉴定基因并检测可辨识基因家族的特征模式；讨论了系统发育关系、分子结构和生物化学特性建模。我们还讨论了如何将计算机作为一个工具来组织数据，系统地思考数据分析过程，并开始思考数据处理的自动化。

生物信息学是相当前沿的，因此即使看懂一本像这样介绍性的书也需要有一定的背景知识。为了很好地理解本书，应该有分子生物学、化学和数学的理论知识或学习经历。在计算机处理方面修一门或两门大学课程也是很有帮助的。

本书的结构

我们精心地安排了书中的材料，既可以从头到尾阅读，也可以跳过前面的章节学习后面的内容。本书分为四部分：

第一部分，概述

第一章，计算机时代的生物学，对生物信息学这门学科进行定义，讲述一些有关的历史，简要介绍本书包含的内容以及为何包含这些内容。

第二章，生物学问题的计算方法，介绍生物信息学和分子生物学的中心概念，以及产生大量生物学数据的技术和研究动机，还包括每个生物学家都需要知道的不断增长的基本计算机程序列表。

第二部分，生物信息学工作站

第三章，建立工作站，介绍Unix，然后是在PC机上安装Linux的基础知识，安装并运行软件。

第四章，Unix中的文件和目录，全面讲述了Unix文件系统，包括文件层次、命名模式、常用的目录命令和多用户环境下的工作。

第五章，在Unix系统下工作，解释用户日常最可能遇到的Unix命令，包括从文件中阅读、编辑和提取信息；正则表达式；shell脚本；与其他计算机的信息交流。

第三部分，生物信息学工具

第六章，在Web上进行生物学研究，讲述如何在Web上寻找信息的艺术。本章包括搜索引擎和搜索过程，到哪里去寻找科学文献和软件，怎样应用在线信息资源以及公共生物学数据库。

第七章，序列分析、成对比对以及数据库搜索，本章一开始对分子进化进行了回顾，然后介绍了成对（pairwise）序列分析技术的基础，比如：预测基因的位置，整体和局部的比对（alignment），应用BLAST和FASTA以局部的比对为基础在数据库中进行查找。本章包括用于序列分析的多功能工具。

第八章，多序列比对、进化树和简图，开始学习一系列相关的基因和蛋白质。它包含用ClustalW和Jalview工具进行多序列比对的策略，然后讨论用于系统发育的分析、构建简图（profile）和基序的工具。

第九章，蛋白质结构的可视化和结构性质的计算，包括蛋白质的三维分析和用来计算蛋白质结构性质的工具。本章一开始回顾了蛋白质化学的历史，然后转到以Web为基础的蛋白质结构分析工具的讨论；结构分类、比对和分析；溶剂的可接近性以及溶剂的相互作用；计算蛋白质的物理化学性质。最后以结构优化及浏览蛋白质结构数据库结束本章内容。

第十章，根据序列预测蛋白质的结构和功能，包括根据蛋白质的序列确定蛋白质结构的工具。本章讨论蛋白质序列的特征检测，二级结构预测以及三维结构预测。最后举例说明蛋白质建模。

第十一章，基因组学和蛋白质组学工具，将它们放在一起进行分析。至今为止，我们已经讲述了分析单一序列或结构的工具和技术，以及单基因长度的多序列比较的工具和技术。本章讨论在研究基因组中所有基因的整合功能时可用到的一些数据类型和工具，包括两个完整的基因组的序列测定，在Web上访问基因组的信息，注释和分析整个基因组序列以及新出现的技术以及蛋白质组学。

第四部分，数据库和可视化

第十二章，用Perl进行数据自动化分析，告诉大家怎样应用Perl语言，从大量的数据中提取出想得到的信息。本章目的不是教大家如何用Perl语言进行编程，但是在本章对Perl语言进行了一些简单的介绍。并举例以帮助大家用自己的方式学习编程。

第十三章，构建生物学数据库，介绍数据库的概念。包括在生物学研究中的数据库的类型，构建生物学数据库的数据库软件，数据库语言（尤其是SQL语言）以及正在发展的与数据库交互的基于Web的软件。

第十四章，可视化和数据采集，包括分析实验结果的计算工具和技术。本章的第一部分介绍在生物信息学研究中发展起来用于使数据可视的程序。其范围从一般意义上的线图和用于大量数据的统计包，如Grace和gnuplot，到以可解释的形式来提交序列和结构信息的程序，如TEXshade。本章第二部分讲述在生物信息学应用中现有的数据采集工具——在大量数据中寻找、解释、求取模式的过程。

研究生物信息学的方法

我们承认，我们是结构生物学家（实际上是生物物理学家）。对我们来讲，不考虑基因的蛋白质产物而研究基因是非常困难的。DNA序列对我们来说不仅仅意味着序列本身。对结构生物学家来说，基因（只有少数例外）意味着三维结构、分子形状和构象的变化、活性位点、化学反应和细微的分子内相互作用。在本书中，我们着眼于结构生物学家和生化学家以理解生物学功能的化学基础可能使用的序列信息。也许本书忽略了对分子生物学家和遗传学家来说很重要的一些序列分析的应用，所以欢迎读者发表自己的意见。

本书引用的URL

要了解在本书中引用的URL和作为生物信息学附加材料的更多信息，可参看本书的网页，列在“建议与评论”部分。

本书排版约定

本书英文使用以下字体约定:

斜体（*Italic*）
: 用于命令、文件名、目录、变量以及URL。

等宽字体（`Constant width`）
: 用于代码实例以及显示命令的输出。

等宽斜体（*`Constant width italic`*）
: 用于“用法”短语表示变量。

建议与评论

本书的内容都经过测试，尽管我们做了最大的努力，但错误和疏忽仍然是在所难免的。如果你发现有什么错误，或者是对将来的版本有什么建议，请通过下面的地址告诉我们:

美国:

O'Reilly & Associates,Inc.
101 Morris Street
Sebastopol, CA 95472

中国:

100080 北京市海淀区知春路 49 号希格玛公寓 B 座 809 室
奥莱理软件（北京）有限公司

本书有一个网页，上面附有勘误表、实例和其他附加信息。可以访问以下网页:

http://www.oreilly com/catalog/bioskills/

如评论或探讨技术问题，可发 email 至:

bookquestions@oreilly.com
info@mail.oreilly.com.cn

要查找有关本书更多的内容、评论、软件、资源中心以及 O'Reilly 网络，请访问我们的网站:

http://www.oreilly.com
http://www.oreilly.com.cn

致谢

Cynthia的致谢: 首先感谢所有听到我说“这本书就要写完了”而没有嘲笑我的人，因为这句话我说了有一千次了。感谢我的家人和朋友，因为在最近几个月中，我很少给他们打电话，但是他们能容忍我; 感谢在 2000 年秋季修生物信息学课程的学生，因为他们在我第一次生物信息学实验中积极参与，帮我确定了哪些主题需要更系统地解释; 感谢我在弗吉尼亚理工学院的同事，因为在一年来，就有关什么是生物信息学以及进行生物信息学学习的学生应该学习什么这些问题，我与他们进行了有趣的讨论; 感谢我的朋友和同事 Jim Fenton，因为他在本书形成的早期做出了不少贡献; 感谢我的论文导师 Shankar Subramaniam。我也要感谢我们的技术审校，