

北京大学生命科学译丛-1

生物信息学概论

T K Attwood & D J Parry-Smith 著

罗静初 等 译

北京大学出版社
北 京

图书在版编目(CIP)数据 图字: 01-2001-0055

生物信息学概论/(英)谈莲莎(Attwood, T. K.), (英)史密斯(Parry-Smith, D. J.) 著; 罗静初等译. - 北京: 北京大学出版社, 2002. 4

(北京大学生命科学译丛; 1)

ISBN 7-301-05468-8

I. 生… II. ①谈… ②史… ③罗… III. 生物信息学
IV. Q 811.4

中国版本图书馆 CIP 数据核字 (2002) 第 014224 号

书 名: 生物信息学概论

著作责任者: [英] T K Attwood & D J Parry-Smith 著

译作责任者: 罗静初 等

责任编辑: 赵学范

标准书号: ISBN 7-301-05468-8/Q·0089

出版者: 北京大学出版社

地址: 北京市海淀区中关村北京大学校内 100871

网址: <http://cbs.pku.edu.cn>

电话: 出版部 62754962 发行部 62754140 编辑部 62752021

电子信箱: zpup@pup.pku.edu.cn

排版者: 兴盛达打字服务社 62549189

印刷者: 北京大学印刷厂

发行者: 北京大学出版社

经销者: 新华书店

850 毫米×1168 毫米 32 开本 9.75 印张 280 千字

2002 年 4 月第 1 版 2002 年 4 月第 1 次印刷

定 价: 16.00 元

CELL AND MOLECULAR BIOLOGY IN ACTION SERIES

Introduction to Bioinformatics

Teresa K. Attwood and
David J. Parry-Smith



Longman

本书中文版由英国 Longman 授予北京大学出版社出版

Originally published in English under the title: "Introduction to Bioinformatics" by Teresa K. Attwood and David J. Parry-Smith

Addison Wesley Longman Limited

Edinburgh Gate, Harlow
Essex CM20 2JE
England

and Associated Companies throughout the World

© Addison Wesley Longman Limited 1999

The right of Teresa K. Attwood and David J. Parry-Smith to be identified as authors of this Work has been asserted by them in accordance with the Copyright, Design and Patents Act 1988.

All rights reserved; no part of this publication may be reproduced, stored in any retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise without either the prior written permission of the Publishers or a licence permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1P 9HE.

First published 1999

ISBN 0 582 327881

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

A catalog entry for this title is available from the Library of Congress.

Set by 30 in Concorde BE

Produced by

Printed in Great Britain by Henry Ling Ltd., at the Dorset Press, Dorchester, Dorset.

内 容 简 介

本书从生物信息学的研究对象、意义出发,介绍生物信息学研究的基本方法和常用工具;绪论中概要介绍了全书的主要内容;第1章介绍生物信息学的一些基本概念;第2章介绍互联网上大量的生物信息资源;第3~5章介绍蛋白质和核酸序列和结构数据库;第6~8章介绍序列比对和数据库搜索的常用方法;第9章用一个实例说明这些方法和资源的使用;第10章介绍常用的软件和工具。

本书深入浅出,实用性强,可以作为高等院校生物信息学教材,也可作为生命科学和生物技术各领域分子生物学研究和开发工作者生物信息学参考书。

译 序

众所周知,20 世纪 90 年代诞生的因特网及其随之而来的信息高速公路计划,意味着信息时代的到来。几乎与因特网同时诞生的人类基因组计划,是迄今为止耗资最为巨大、竞争最为激烈、意义最为深远的大型国际合作研究项目。完成人类基因组 30 亿个碱基对的全序列测定,只是人类基因组计划的第一步。搞清楚人类基因组全套遗传密码的全部含义,则是这一计划的目标。由此引发的结构基因组、功能基因组、药物基因组、蛋白组计划已经开始实施。DNA 序列测定开拓了一个测序仪制造和测序试剂生产的高新技术产业,以基因芯片技术为代表的一系列生物技术则是功能基因组、药物基因组研究和开发的基础。

基因组计划和因特网推动了生物信息学的诞生。它以核酸、蛋白质等生物大分子数据为主要对象,以数学、信息学、计算机科学为主要手段,以计算机硬件、软件和计算机网络为主要工具,对浩如烟海的原始数据进行存储、管理、注释、加工,使之成为具有明确生物意义的生物信息。并通过对生物信息的查询、搜索、比较、分析,从中获取基因编码、基因调控、核酸和蛋白质结构功能及其相互关系等理性知识。在大量信息和知识的基础上,探索生命起源、生物进化以及细胞、器官和个体的发生、发育、病变、衰亡等生命科学中重大问题,搞清它们的基本规律和时空联系,建立“生物学周期表”。

国内计算机网络的开通,特别是北京、上海等大城市网络基础设施的发展,为在我国实现生物信息共享提供了必要条件。我国人类基因组、水稻基因组等大型研究计划以及国家 863 计划、973 计划、攻关项目、自然科学基金项目的实施,对生物信息在生命科学和生物

技术领域的研究、开发和应用提出了迫切要求。国内一大批单位已经开始从事生物信息学研究。不少大专院校已经和正在准备开设生物信息学课程。为此,我们组织翻译了这本“生物信息学概论”,为国内读者,特别是生命科学和生物技术领域从事分子生物学和基因组研究的师生,提供一本入门教材和参考书。

需要特别说明的是,由于生物信息学和计算机网络正在以日新月异的速度发展,本书原作中有些素材已经过时,不少网址有所变更,无法一一注明,请读者见谅。此外,原书中加框的双色图(灰度及淡蓝色),译著中标以“图框”并排仿宋体字以示区分。限于译者水平,对原书的理解肯定有不少欠妥之处,译文中难免有不少错误,敬请读者批评指正。如能将错误疏漏之处用电子邮件方式发给译者(luojc@pku.edu.cn),将不胜感激。

本书初稿由以下译者执笔:罗洪:前言、第1章;曲红:第2~3章;方刚:第4~5章;李兵、刘翟:第6~7章;禹胃:第8章;黄弋:第9~10章;陈蕴佳:图表注释和校对;吴昕:词汇表。全书由罗静初统一修改、整理和审定。

本书翻译过程中得到作者 Teresa Attwood 和 David Parry-Smith 的帮助,特此致谢。郑伟谋、官山和北京大学选修“生物信息学概论”课的许多研究生在成稿过程中提出了许多宝贵意见,一并表示感谢。感谢北京大学出版社赵学范编审为本书出版所作的努力。

罗静初

2001年2月于北京大学燕北园

目 录

献 辞	(7)
致 谢	(9)
原 序	(11)
绪 论	(15)
第 1 章 概论	(1)
1.1 本章简介	(1)
1.2 序列测定	(1)
1.3 什么是生物信息学	(3)
1.4 序列和结构	(4)
1.5 基因组计划	(5)
1.6 人类基因组计划现状	(7)
1.7 生物信息学的重要性	(8)
1.8 模式识别和结构功能预测	(10)
1.9 蛋白质折叠问题	(12)
1.10 分子伴侣	(13)
1.11 序列分析	(14)
1.12 同源性和相似性	(15)
1.13 生物信息学研究中的挑战性问题	(17)
1.14 本章小结	(20)
1.15 进一步阅读指南	(21)

第 2 章 信息网络	(24)
2.1 本章简介	(24)
2.2 因特网	(24)
2.3 IP 地址	(25)
2.4 因特网上的实用工具	(26)
2.5 万维网	(27)
2.6 网络浏览器	(27)
2.7 HTTP, HTML 和 URL	(29)
2.8 欧洲分子生物学网络组织	(31)
2.9 美国国家生物技术信息中心—NCBI	(38)
2.10 网络导游	(39)
2.11 本章小结	(41)
2.12 进一步阅读指南	(41)
2.13 有关网址	(42)
第 3 章 蛋白质数据库	(43)
3.1 本章简介	(43)
3.2 生物信息数据库	(43)
3.3 序列数据库	(45)
3.4 蛋白质序列复合数据库	(52)
3.5 蛋白质序列二次数据库	(55)
3.6 蛋白质序列模式复合数据库	(75)
3.7 结构分类数据库	(75)
3.8 本章小结	(79)
3.9 进一步阅读指南	(80)
3.10 有关网址	(82)
第 4 章 基因组信息资源	(83)
4.1 本章简介	(83)

4.2	DNA 序列数据库	(83)
4.3	特定基因组资源	(91)
4.4	本章小结	(95)
4.5	进一步阅读指南	(95)
4.6	有关网址	(96)
第 5 章	DNA 序列分析	(97)
5.1	本章简介	(97)
5.2	DNA 序列分析的意义	(97)
5.3	基因结构与 DNA 序列	(99)
5.4	DNA 序列分析	(102)
5.5	EST 数据库搜索	(108)
5.6	寻找基因的两种途径	(112)
5.7	细胞表达谱	(113)
5.8	cDNA 文库和 EST	(114)
5.9	EST 序列数据库	(115)
5.10	EST 数据对 DNA 数据库的影响	(120)
5.11	EST 序列分析示例	(124)
5.12	本章小结	(129)
5.13	进一步阅读指南	(130)
第 6 章	双序列比对	(131)
6.1	本章简介	(131)
6.2	数据库检索和数据库搜索	(131)
6.3	字母表和复杂度	(133)
6.4	算法和程序	(133)
6.5	双序列比对简例	(133)
6.6	子序列	(134)
6.7	同一性和相似性	(135)

6.8	点阵图	(142)
6.9	局部相似性和整体相似性	(145)
6.10	整体比对算法	(146)
6.11	局部比对算法	(150)
6.12	动态规划算法	(153)
6.13	基于双序列比对的数据库搜索	(153)
6.14	本章小结	(159)
6.15	进一步阅读指南	(160)
第7章	多序列比对	(162)
7.1	本章简介	(162)
7.2	多序列比对的意义	(162)
7.3	多序列比对的定义	(163)
7.4	调和序列	(164)
7.5	算法复杂性	(165)
7.6	手工比对方法	(166)
7.7	同步法	(168)
7.8	步进法	(168)
7.9	多序列比对数据库	(170)
7.10	基于多序列比对的数据库搜索	(174)
7.11	本章小结	(175)
7.12	进一步阅读指南	(176)
第8章	二次数据库搜索	(177)
8.1	本章简介	(177)
8.2	二次数据库搜索的意义	(177)
8.3	二次数据库的内容	(178)
8.4	本章小结	(198)
8.5	进一步阅读指南	(199)

第 9 章 数据库搜索实例	(200)
9.1 本章简介	(200)
9.2 基本方法	(200)
9.3 判断搜索结果	(219)
9.4 结构和功能分析	(220)
9.5 本章小结	(226)
9.6 进一步阅读指南	(227)
 第 10 章 序列分析软件包	 (228)
10.1 本章简介	(228)
10.2 序列分析软件包	(228)
10.3 商业数据库	(232)
10.4 商业软件包	(232)
10.5 综合软件包	(233)
10.6 其他 DNA 序列分析软件	(237)
10.7 内联网软件包	(238)
10.8 因特网软件包	(240)
10.9 本章小结	(243)
10.10 进一步阅读指南	(243)
10.11 有关网址	(244)
 中译版参考书目	 (245)
词汇表(汉英部分)	(247)
词汇表(英汉部分)	(262)
译后记	(277)

谨将此书献给

John Utting 教授

和

Jean Utting 博士

致 谢

首先,我们谨向下表中列出的网站和有关作者深表谢意,感谢他们为本书提供了许多精美图表。

编号	出 处	作 者
图 1.2	http://www.biochem.ucl.ac.uk/bsm/pdbsum/	R. Laskowski
图 2.3	http://www.ebi.ac.uk/	R. Lopez
图 2.4	http://www.ebi.ac.uk/	T. Etzold
图 2.5	http://www.vtourist.com/webmap/maps.htm	B. Plewe
	http://www.angis.org.au/	T. Littlejohn
图 3.3	http://www.expasy.ch/	A. Bairoch
图 3.4	http://www.expasy.ch/	A. Bairoch
图 3.6	http://www.blocks.fhcrc.org/	J. Henikoff
图 3.7	http://www.blocks.fhcrc.org/	J. Henikoff
图 3.8	http://www.expasy.ch/	A. Bairoch
图 3.9	http://www.sanger.ac.uk/	R. Durbin
图 4.3	http://www.sanger.ac.uk/	R. Durbin
图 7.2	http://www.sanger.ac.uk/	R. Durbin
图 8.3	http://www.bioinf.man.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图 8.4	http://www.blocks.fhcrc.org/	J. Henikoff
图 8.5	http://www.expasy.ch/	A. Bairoch
图 9.10	http://www.bioinf.man.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图 9.11	http://www.bioinf.man.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图 9.12	http://www.blocks.fhcrc.org/	J. Henikoff
图 9.13	http://www.blocks.fhcrc.org/	J. Henikoff
图 9.14	http://www.blocks.fhcrc.org/	J. Henikoff
图 9.15	http://www.stanford.edu/identify/	C. Nevill-Mannine
图 9.17	http://www.biochem.ucl.ac.uk/bsm/cath/	C. Orengo
图 9.18	http://www.biochem.ucl.ac.uk/bsm/pdbsum/	R. Laskowski
图 9.19	http://www.biochem.ucl.ac.uk/bsm/pdbsum/	R. Laskowski
图 9.7	http://www.expasy.ch/	A. Bairoch
图 9.9	http://www.sanger.ac.uk/	R. Durbin
图 10.2	http://www.mrcImb.cam.ac.uk:80/pubseq/	R. Staden
图框 1.3	Rube Goldberg 公司和 United Media 公司	
图框 8.1	http://www.biochem.ucl.ac.uk/cgi-bin/fingerPRINTScan/	P. Scordis
图框 9.1	http://www.blocks.fhcrc.org/	J. Henikoff

感谢 Anne Parry-Smith 女士始终如一的支持,在写作本书的艰苦日子里,她对我们嘘寒问暖,关心备至。感谢英国剑桥 Drug Discovery 研究所 Jeremy Packer 为本书所做的十分细致的校对工作。感谢本书主要策划者 Addison Wesley Longman 图书公司的 Alex Seabrook 先生和编辑 Kate Henderson 女士。他们的耐心十分可贵,堪称楷模。最后,作者 Attwood 还要感谢伦敦大学 University College 生物化学系序列分析小组的 Phil Scordis、Julian Selly、Jane Mabey、Will Wright 和 Maria Karmirantzou。由于本书的写作,Attwood 博士无法顾及他们的研究课题,有时甚至因为写作不顺利而向他们发脾气,而他们却以微笑表示理解。每当在写作中遇到各种困难时,他们总是雪中送炭,帮助解决那些在关键时刻故意作弄人的难题。

真诚感谢所有为本书出版尽心尽力的同事、亲友和学生们!

原 序

过去的十年,对生命科学来说,是一个不同寻常的十年。“基于硅片的生物学”已经在向我们招手。它的出现,为我们开拓了生物学研究的新途径,使我们有可能对全基因组进行系统的研究和全面的比较。旨在阐明生物遗传密码和基因产物的基因组研究是当前生命科学的热点,其最终目的是要揭示生物进化的机制,搞清蛋白质折叠的机理,探索蛋白质结构功能关系等生命科学中一系列重大问题。

迄今为止,利用计算机模拟生物过程依然具有很大的局限性。究其根源,主要在于我们对生命科学中重大问题的了解仍然十分有限。必须承认,蛋白质折叠的机制尚未得到完全阐明;对某个特定的序列或折叠方式,究竟源于趋同进化还是趋异进化,尚难给出明确的结论;仅由单个蛋白质序列或结构,尚难孤立地推断其功能。只有真正懂得计算机并非魔术师手中的魔杖,才能更好地利用计算机,解决那些可以解决的问题。不懂得这一点,就会被某些计算机程序或软件所导致的错误结果引入歧途。

当今生物信息学领域的主要研究方向之一,是从研究蛋白质功能入手,进而阐明进化关系。常用方法可以分为两类:一类以序列为基础,用序列分析的手段以及所得结果推断生物大分子的功能;另一类则以结构为基础,其基本思路基于蛋白质分子的结构功能关系。第一类方法的基本出发点,是通过和数据库中已知功能的序列进行相似性比对,确定那些新测定的序列与已知序列之间的关系。也就是说,通过数据库搜索,找出可能只有若干相同残基的功能位点,由某个初看起来完全不同的蛋白质分子确定该未知蛋白的功能。此时,未知蛋白和已知蛋白序列整体相似性并不很高,之所以有相同功

能,很可能是趋同进化的结果。而那些序列、结构相似性较高的蛋白质分子,则可能是同一祖先趋异进化的结果。

应该肯定,利用计算机进行核酸和蛋白质序列分析的方法已经比较成熟。那么,由蛋白质序列识别其功能,似乎应该顺理成章。然而,实际情况并非如此乐观。由于生物学实验远远落后于基因组测序的进程,对于大量的序列数据无法及时用实验方法验证。截至1998年,所有已经测定的基因组序列中,约三分之一蛋白质序列的功能无法推断。序列数据库中很大一部分序列除了基因名称外,没有其他注释信息,或仅仅标上“假想蛋白质”(hypothetical proteins)。此外,有些具有相同结构的蛋白质,却具有不同的功能。典型的例子是溶菌酶(lysozyme)和 α -乳清蛋白(α -lactalbumin)。这两种蛋白质具有高达50%的相同残基,序列相似性分数达70%,其功能却完全不同。前者是催化分解细菌细胞壁多糖的水解酶,后者不具任何催化作用,而是一个起调控作用的蛋白质,其作用是将半乳糖(galactose)分子从UDP-半乳糖(UDP-galactose)转化为D-葡萄糖(D-glucose)。为什么两者的序列相似性程度如此之高,折叠方式又十分相似,它们的功能却大相径庭呢?这是因为,溶菌酶中起关键作用的两个催化残基谷氨酸(glutamic acid)和门冬氨酸(aspartic acid)在 α -乳清蛋白中却没有。相反, α -乳清蛋白中起关键作用的酸性钙结合序列模体只在个别溶菌酶分子中存在。这一实例并不意味着利用计算机进行序列分析和结构模拟毫无用处,而是强调这种分析和模拟必须与基本的生物学知识相结合,才能得出较为可靠的结论。

不言而喻,蛋白质分子的序列和结构是决定其功能的基础,两者相辅相成,缺一不可。如果用简单的数学公式表示,不妨写成:“功能 = 序列 + 结构”。可以将蛋白质的各种折叠单元看做各种不同的基本部件,用来整合各种不同序列模体,以实现各种不同功能。可以用房子作比喻,形象地说明这个问题。把一间空房比做基本单位,就像蛋白质的某种折叠单元。若在这间房子里放入一张桌子和一把椅