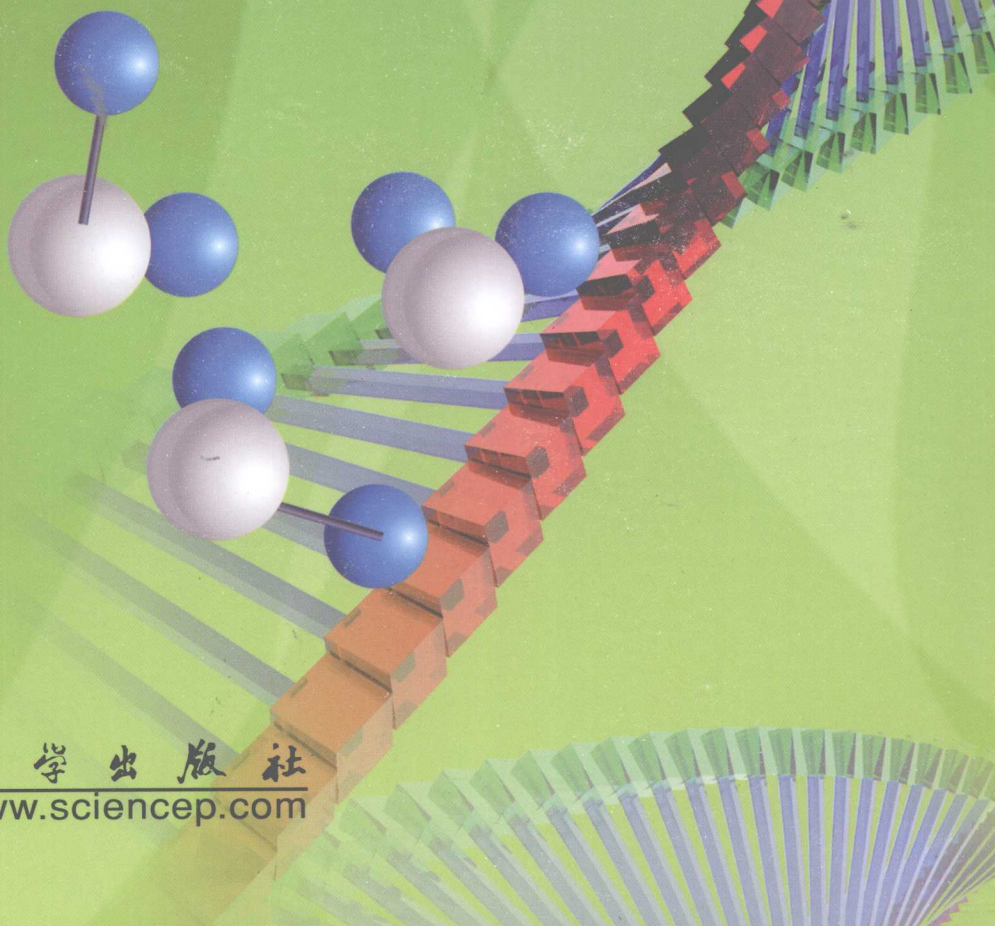
生物信息学分析实践

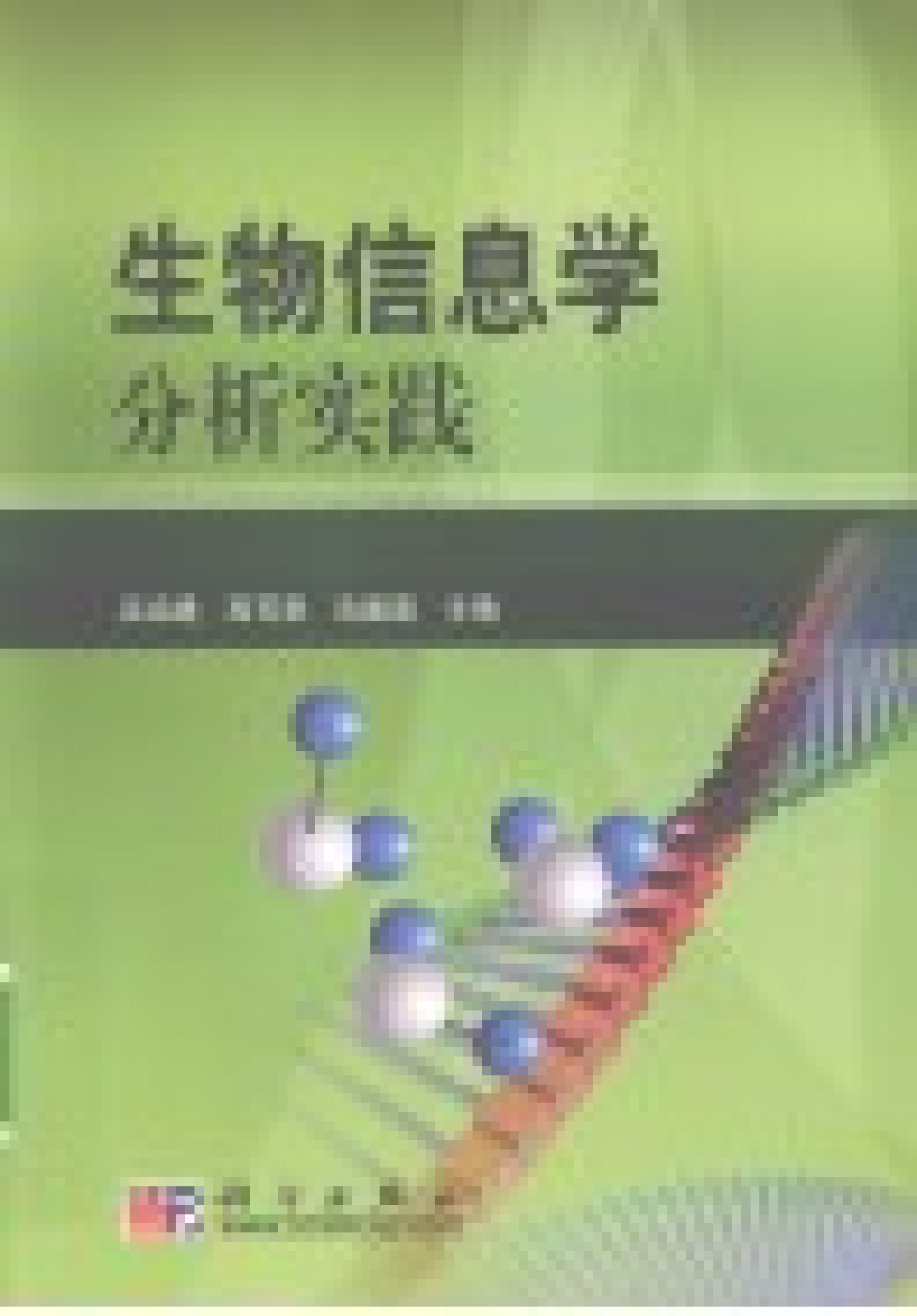
吴祖建　高芳銮　沈建国　主编



科学出版社
www.sciencep.com

生物信息学
分析实践

[redacted]



生物信息学分析实践

吴祖建　高芳銮　沈建国　主编

科学出版社
北京

内 容 简 介

本书内容主要包括生物信息学简介、三大数据库检索、引物设计及测序结果分析、核酸序列分析、蛋白质序列分析、蛋白质空间结构预测、系统发育分析、RNA 分析、参考文献管理。本书的一大特色在于丰富的实例和图表，使读者可以很直观地了解和掌握书中的内容。

本书取材新颖，实践性强，是一本实用的生物信息学分析手册与操作指南，适用于生命科学、农学、医学等相关专业学生使用，也可用于从事生物学相关的科研人员、教师参考使用。

图书在版编目(CIP)数据

生物信息学分析实践/吴祖建，高芳銮，沈建国主编. —北京：科学出版社，2010. 6

ISBN 978-7-03-027831-9

Ⅰ. ①生… Ⅱ. ①吴…②高…③沈… Ⅲ. ①生物信息论 Ⅳ. ①Q811. 4

中国版本图书馆 CIP 数据核字(2010)第 103496 号

责任编辑：丛 楠 甄文全 / 责任校对：赵桂芬
责任印制：张克忠 / 封面设计：耕者设计工作室

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

骏走印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2010 年 6 月第 一 版 开本：787×1092 1/16
2010 年 6 月第一次印刷 印张：14 3/4 插页：2
印数：1—3 000 字数：349 000

定价：30.00 元

（如有印装质量问题，我社负责调换）

序

1990年以来，随着各物种基因组测序计划的启动和分子结构测定技术的突破以及互联网的普及，数以百计的生物学数据库陆续建立。如何对这些海量数据进行收集、整理、解析和服务，从中发现隐藏在序列信息中的生物学规律是亟待解决的问题。于是一门新兴的交叉科学——生物信息学便应运而生。作为一门融合了数学、生命科学、计算机信息科学和生物统计学在内的交叉学科，生物信息学目前已成为生命科学发展一个不可或缺的重要工具。

近年来，有关生物信息学的著作甚多，其中不乏以算法理论为主调的专著，或以教学为目的的“综述性”教材，而有关技术性著作尚不多见。为方便广大生物信息学初学者快速掌握实验数据分析方法，一批活跃在科研一线的学者们聚集在一起，精心编撰了这本《生物信息学分析实践》。内容涵盖了生物信息学简介、数据库检索、引物设计及测序结果分析、核酸序列分析、氨基酸序列分析、蛋白质空间结构预测、系统发育分析、RNA分析等诸多技术问题。在系统阐述原理和方法基础上，突出实战技巧，循序渐进，全面介绍生物学软件及数据库的应用，具有较强的实用性和可操作性。

实验数据分析需借助恰当的工具。《生物信息学分析实践》一书图文并茂、通俗易懂，语言生动活泼，表述深入浅出，且大量实例均来自科研实践，具有经典的普适性，是一本实用的生物信息学分析手册与操作指南，可为广大科研工作者和大专院校师生提供参考。当然，生命科学和生物信息学及其相关技术，还在不断发展，这就必然要在今后的再版中不断修订和完善，而目前这本书会对我国相关科学工作者起到一定的借鉴作用。

中国科学院院士

2010年4月

前　言

随着计算机科学、信息科学等飞速发展，一门以数据分析处理为本质的新学科——生物信息学悄然兴起，它融合了数学、生命科学、计算机信息科学等多门学科，至今已经广泛地渗透到科学研究的各个方面，成为一门用途极为广泛的工具学科。核酸序列、蛋白质序列的分析，以及各类生物学软件和数据库的应用，已经成为科研工作者必备的基本技能。然而，由于擅长专业领域的不同，研究人员面对大量的生物学软件及数据库常无所适从。目前国内外已有不少优秀的生物信息学教材与专著，但大多数以理论为主，或侧重于算法研究，或倾向于网络数据库介绍，对于数据实例分析实践的专著尚不多见。为此，本书编写的目的就是为不同专业背景的读者提供一本实用的生物信息学分析入门书，让读者在实例分析的过程中，不仅学会工具软件的操作使用，更能学会思考与解决科研问题。

2008 年夏，在福建农林大学植物病毒研究所的倡议下，编著者通过基因酷等论坛，向在生物信息学及相关领域教学与科研第一线的人员发出邀请，希望他们加入《生物信息学分析实践》编委会，共同编写本书。很快，中国科学院遗传与发育生物学研究所乔楠和刘林川、中国科学院南京地质古生物研究所盖永华、南京师范大学生命科学学院李鹏、西北农林科技大学林文超、福建省出入境检验检疫局沈建国、浙江省农业科学院鹿连明、南方医科大学郭玲等欣然同意共同编写本书。本书由吴祖建、高芳銮、沈建国担任主编。

全书共包括七章。第一章为生物信息学简介，主要包括生物信息学基础及应用；第二章介绍了数据库检索，主要包括综合性数据库 NCBI、EMBL-EBI 和基因组信息数据库 UCSC；第三章介绍了引物设计及测序结果分析；第四章介绍了核酸序列分析，主要包括核酸序列常规分析、序列比对分析及基因结构识别；第五章介绍了蛋白质序列分析，主要包括蛋白质基本性质分析、结构域分析及蛋白质空间结构预测；第六章介绍了分子进化与系统发育分析，主要包括系统发育分析、检验及评估；第七章介绍了 RNA 分析，主要包括 siRNA 设计和 microRNA 分析。其中林文超、吴祖建编写第一章；乔楠、沈建国编写第二章；鹿连明、冯佩富、庄军编写第三章；高芳銮、朱萧、沈建国编写第四章；高芳銮、万祥辉编写第五章；李鹏、盖永华编写第六章；刘林川、郭玲编写第七章。吴祖建研究员参与编写了第一、四、五章的部分内容，同时负责本书的统编工作，高芳銮、沈建国共同负责本书的审校工作。

中国科学院院士谢联辉教授欣然作序，生物农药与化学生物学教育部重点实验室主任关雄教授、福建农林大学汪世华教授等对本书提出了非常有价值的建议和意见，科学出版社甄文全编辑、福建农林大学植物病毒研究所谢荔岩老师对本书的编写给予了热情的支持和帮助，华中科技大学出版社朱建丽、上海交通大学生命科学与技术学院曹又方、基因酷站长胡洋、中国水产科学研究院南海水产研究所吕俊霖、福建省漳州市林业

局吴子毅、中山大学张海丽等作为本书的第一批读者给本书提出了许多宝贵的意见，国家转基因生物新品种培育重大专项重点课题资助项目（2009ZX08009-044B）、公益性行业（农业）科研专项（nyhyzx07-051）、福建省自然科学基金（2006J0047，2009J01046）、福建出入境检验检疫局科技项目（FK2007-25）为本书提供了出版基金，基因酷（genecool. com）、生物秀（bbioo. com）、螺旋网（helixnet. cn）等论坛为本书编委会提供了交流平台。在此，一并致以衷心的感谢！

由于生物信息学发展日新月异，作者的水平有限，书中难免存在错误，恳请各位读者批评指正，以期再版修订，来函请发至邮箱 wuzujian@126. com 或 raindyok@126. com。

编著者

2010 年 4 月　福州

目　　录

序

前言

第一章　生物信息学简介 …… 1

1.1　生物信息学基础 …… 1

1.1.1　什么是生物信息学 …… 1

1.1.2　生物信息学的形成与发展 …… 2

1.1.3　生物信息学的研究内容 …… 4

1.2　生物信息学应用 …… 6

1.2.1　生物信息学的热点领域 …… 6

1.2.2　信息技术与生物信息学 …… 8

第二章　数据库检索 …… 11

2.1　综合性数据库 NCBI …… 11

2.1.1　NCBI 简介 …… 11

2.1.2　NCBI 数据库介绍 …… 12

2.1.3　Entrez 简介 …… 14

2.1.4　Entrez 检索实例 …… 15

2.2　综合性数据库 EMBL-EBI …… 19

2.2.1　EBI 简介 …… 19

2.2.2　EBI 数据库简介 …… 20

2.2.3　SRS 简介 …… 21

2.2.4　SRS 检索实例 …… 22

2.2.5　BioMart 简介 …… 23

2.2.6　BioMart 检索实例 …… 23

2.3　UCSC 基因组浏览器 …… 27

2.3.1　UCSC 基因组浏览器简介 …… 27

2.3.2　UCSC 基因组浏览器检索实例 …… 28

第三章　引物设计及测序结果分析 …… 30

3.1　引物设计 …… 30

3.1.1 概述 …… 30
3.1.2 常规 PCR 引物设计实例分析 …… 36
3.1.3 简并引物设计 …… 53
3.2 测序结果分析 …… 60
3.3 序列拼接实例分析 …… 61
第四章 核酸序列分析 …… 75
4.1 常规分析 …… 75
4.1.1 核酸序列的检索 …… 75
4.1.2 核酸序列组分分析 …… 75
4.1.3 序列变换 …… 77
4.1.4 限制性酶切分析 …… 78
4.2 比对分析 …… 82
4.2.1 BLAST 比对 …… 82
4.2.2 双序列比对 …… 87
4.2.3 多序列比对 …… 89
4.3 基因结构识别 …… 93
4.3.1 ORF 识别及其可靠性验证 …… 93
4.3.2 启动子及转录因子结合位点分析 …… 97
4.3.3 重复序列分析 …… 99
4.3.4 CpG island …… 101
4.3.5 3′UTR 区 …… 102
第五章 蛋白质序列分析 …… 105
5.1 蛋白质序列的基本性质分析 …… 105
5.1.1 理化性质分析 …… 105
5.1.2 疏水性分析 …… 107
5.1.3 跨膜区分析 …… 110
5.1.4 信号肽预测 …… 112
5.1.5 Coil 区分析 …… 116
5.1.6 亚细胞定位 …… 118
5.2 结构域分析及 motif 搜索 …… 120
5.2.1 结构域分析 …… 120
5.2.2 motif 搜索 …… 123
5.3 空间结构预测 …… 126
5.3.1 蛋白质二级结构预测 …… 126

5.3.2　蛋白质三级结构预测 …… 129
5.3.3　蛋白质结构预测方法评价 …… 138
5.4　抗原表位预测分析 …… 139
5.4.1　B淋巴细胞抗原表位预测分析 …… 140
5.4.2　T淋巴细胞抗原表位预测分析 …… 145
第六章　分子进化与系统发育分析 …… 149
6.1　进化的分子基础 …… 149
6.1.1　进化树与分子系统学 …… 149
6.1.2　相似性与同源性 …… 151
6.1.3　分子进化 …… 151
6.2　系统发育分析 …… 151
6.2.1　DNA和氨基酸序列的进化演变 …… 153
6.2.2　系统发育树的种类 …… 153
6.2.3　用于系统发育分析的分子标记选择 …… 154
6.2.4　常用的构树方法及其甄选 …… 156
6.2.5　系统发育分析常用软件 …… 159
6.3　系统发育分析的检验 …… 165
6.3.1　系统发育分析方法可靠性的评价 …… 165
6.3.2　系统树的误差分析及消除方法 …… 166
6.4　系统树的评估 …… 168
6.5　系统发育分析实例 …… 168
6.5.1　使用MEGA软件重建NJ树 …… 172
6.5.2　使用PAUP软件重建NJ树 …… 174
6.5.3　使用MEGA软件重建MP树 …… 175
6.5.4　使用PAUP软件重建MP树 …… 176
6.5.5　使用PAUP软件重建ML树 …… 177
6.5.6　贝叶斯树 …… 177
第七章　RNA分析 …… 181
7.1　RNA简介 …… 181
7.1.1　RNA的结构 …… 182
7.1.2　RNA的功能 …… 182
7.1.3　RNA研究进展与展望 …… 184
7.2　RNA二级结构 …… 184
7.2.1　RNA的二级结构概述 …… 184

7.2.2 RNA 二级结构的预测方法 …… 186
7.2.3 RNA 结构预测实例分析 …… 188
7.3 高效 siRNA 的设计 …… 191
7.3.1 RNAi 的作用机制 …… 192
7.3.2 siRNA 的设计原则 …… 193
7.3.3 影响 RNAi 的其他因素 …… 193
7.3.4 siRNA 的设计步骤 …… 194
7.3.5 siRNA 的合成 …… 194
7.3.6 siRNA 干涉效果的评判 …… 195
7.3.7 siRNA 相关数据库介绍 …… 195
7.3.8 siRNA 设计实例分析 …… 196
7.4 microRNA 分析 …… 198
7.4.1 microRNA 作用机制概述 …… 198
7.4.2 miRNA 功能与研究方法 …… 200
7.4.3 miRNA 生物信息学分析 …… 201
7.4.4 miRNA 及其靶基因预测的实例分析 …… 206
主要参考文献及网址 …… 211
附录 …… 218
附录 1 常用生物学数据库 …… 218
附录 2 各种主要的 RNA 二级结构预测软件比较 …… 219
附录 3 名词解释 …… 221
彩图

第一章　生物信息学简介

1.1　生物信息学基础

1.1.1　什么是生物信息学

“生物信息学”是英文单词“bioinformatics”的中文译名。从字面上可以看出，它是一门与生物学和信息学两大热门学科都密切相关的学科。什么是生物信息学呢？我们还是从这一词语的来源谈起。bioinformatics 一词是由美国学者林华安博士（H. A. Lim）在 1987 年首创的。他最初使用的是“compbio”这一单词，随后更改为“bioinformatique”。不久，他又进一步将其更改为 bio-informatics（或 bio/informatics），又因名称中的“-”或“/”符号容易导致当时的电子邮件系统问题，故将其去除，于是今天我们所看到的“bioinformatics”就正式诞生了，林华安博士也因此赢得了“生物信息学之父”的美誉（蒋彦等，2003）。

众所周知，生物信息学是一门崭新的、正如火如荼发展的交叉学科，它吸引了越来越多的来自生物学、医学、数学以及信息科学背景的专家学者投入到生物信息学的研究当中去。但是由于不同领域科学家兴趣的侧重点不同，对生物信息学的定义往往具有局限性，犹如盲人摸象（图 1-1）。关于生物信息学的定义，到目前为止，国内外众说纷纭，

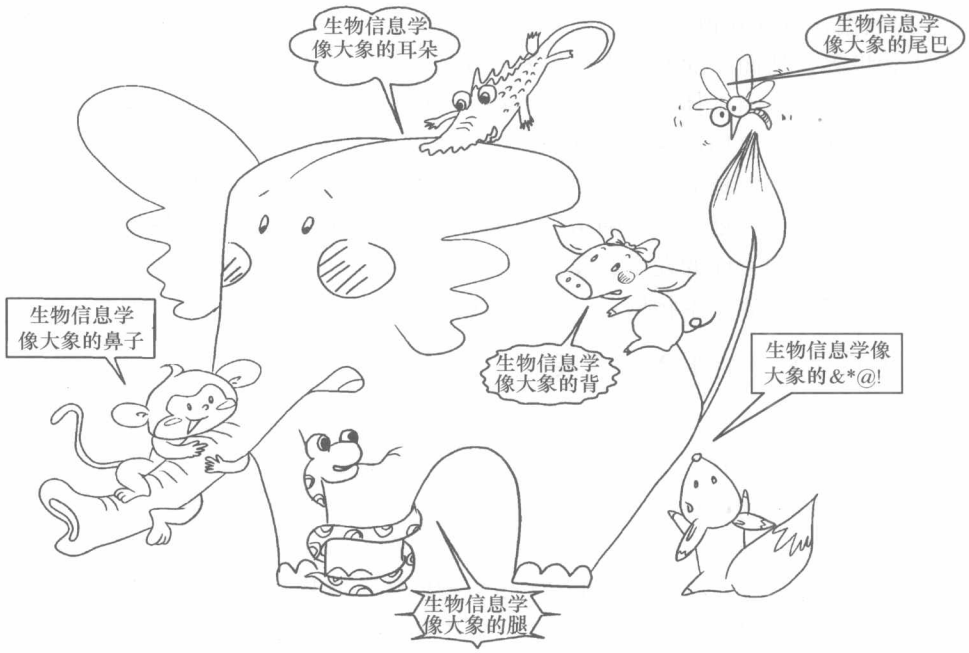


图 1-1　不同领域科学家对生物信息学的认识

没有形成统一的、普遍公认的标准。美国国家基因组研究中心认为，生物信息学是一门生物学、数学和计算机相互交叉融合而产生的新兴学科；美国佐治亚理工大学认为，生物信息学是采用数学、统计学和计算机等方法分析生物学、生物化学和生物物理学数据的一门综合学科；美国密苏里大学认为，生物信息学是获取、存储和处理生物学信息的一门科学与技术；美国加利福尼亚大学洛杉矶分校则认为，生物信息学是对生物学信息和生物学系统内在结构的研究，它运用数学和计算机科学的分析理论和实用工具将分散的生物学数据联系起来。笔者根据自己对生物信息学的认识，综合多方面内容，从广义和狭义两个方面对什么是生物信息学做了相关的解释。

1. 广义的生物信息学

广义的生物信息学是指从事对基因组研究相关的生物信息的获取、加工、储存、分配、分析和解释。它包括了两层含义，一个是对海量数据的收集、整理与服务，也就是管好这些数据；另一个是从中发现新的规律，也就是用好这些数据。具体地说，生物信息学是以 DNA 和蛋白质序列等生物信息分析作为源头，破译隐藏在序列信息中的语义规律，阐明海量生物信息的信息实质，在此基础上归纳、整理与遗传语义信息释放及调控相关的转录谱和蛋白质谱的数据，从而认识代谢、发育、分化和进化的规律。广义的生物信息学涉及生命的信息交换和传递的各个层次，如核酸、蛋白质、细胞、组织、器官、系统和生物体等。

2. 狭义的生物信息学

狭义的生物信息学是指综合应用信息科学、数学的理论、方法和技术，管理、分析和利用生物分子数据的科学，即分子生物信息学（molecular bioinformatics）。最初的生物信息学以基因组 DNA 序列信息挖掘为出发点，通过收集、组织、管理基因组 DNA 序列数据，以期获得隐藏在基因组序列之中丰富的生物学知识，从更深层次认识未知的生物世界。然而，随着技术的不断进步，生物分子数据已经不仅仅限于基因组序列数据，微阵列、基因本体（gene ontology，GO）注释、分子图谱、结构数据等其他数据的积累也在快速增加，这些数据同样具有丰富的内涵。科学地利用生物分子数据及其分析结果，可以大大提高研究和开发的科学性及效率。但是，如何发展应用数理统计、模式识别、动态规划、神经网络、遗传算法及隐马尔科夫模型等挖掘海量数据中可以阐明细胞、器官和个体的发生、发育、病变、衰亡的基本规律的方法，仍是生物学家所面临的一个严峻挑战，生物信息学也正是在这种挑战中不断发展起来的交叉学科。

1.1.2 生物信息学的形成与发展

生物信息学是建立在分子生物学基础上的。因此，要了解生物信息学的发展就必须先对分子生物学的发展有一个简单的了解。1866 年孟德尔（G. J. Mendel）根据实验结果提出了基因是以实物存在的假设。1871 年 Miescher 从死的白细胞核中分离出脱氧核糖核酸。1944 年阿佛莱（O. T. Avery）、麦克李沃（C. M. Mac Leod）和麦克卡（M. Mc Carty）三人通过试验证明 DNA 是生物的遗传物质以前，人们仍然认为染色体蛋白质携带基因，而 DNA 则被认为是一个小角色。1944 年 Chargaff 发现了著名的 Chargaff 规律，即 DNA 中鸟嘌呤的量与胞嘧啶的量总是相等，腺嘌呤与胸腺嘧啶的量

总是相等。与此同时，Wilkins 与 Franklin 用 X 射线衍射技术测定了 DNA 纤维的结构（图 1-2）。1953 年 James Watson 和 Francis Crick 通过大量研究后在 *Nature* 杂志上发表论文，提出了 DNA 的双螺旋三维结构。他们的理论奠定了分子生物学的基础。DNA 双螺旋模型已经预示出了 DNA 复制的规则。1956 年 Kornberg 从大肠杆菌中分离出了 DNA 聚合酶Ⅰ（DNA polymerase Ⅰ），这种酶能使 dNTP 连接成 DNA。Meselson 和 Stahl 于 1958 年证明了 DNA 半保留复制。Crick 于 1954 年提出了遗传信息传递的规律，根据当时有限的资料，他把中心法则（central dogma）的公式表述为“DNA→RNA→蛋白质”，虽然现在的中心法则已经得到新的修正，但 Click 提出的中心法则对后来分子生物学和生物信息学的发展都起到了极其重要的指导作用。1966 年，Nirenberg 和 Khorana 破译了全部遗传密码字典的 64 个密码子，限制性内切酶的发现和重组 DNA 的克隆（clone）奠定了基因工程的技术基础。正是由于分子生物学的研究对生命科学的发展有巨大的推动作用，生物信息学的出现也就成了一种必然。

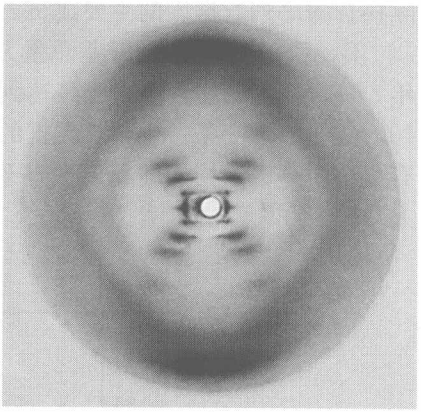


图 1-2　DNA 的 X 射线衍射照片
（引自 http://www.scifun.ed.ac.uk/card/flakes.html）

1. 生物信息学的萌生

生物信息学的产生与发展仅有十几年的时间，bioinformatics 这一名词在 1991 年左右才在文献中出现，而且大多出现在电子出版物中。事实上，生物信息学已经存在了几十年，可以说是一门有“较长历史”的学科，早在计算机初创的 1956 年就已经在美国田纳西州的盖特林堡召开过首次“生物学中的信息理论讨论会”，这应该算作生物信息学的雏形，只不过最初人们把生物信息学称为基因组信息学。

20 世纪 80 年代末，随着人类基因组计划的启动，生物实验和衍生数据的大量储存，促使这一新兴交叉学科形成。但由于当时技术的局限性，这一学科并没有得到重视和发展。在沉寂了 20 余年后，才依赖计算机科学、工程学和应用数学的迅猛发展得以飞速发展。随着互联网的发展和普及，基于 Internet 的庞大信息网络无疑对信息学的发展起了巨大的推动作用，也为生物信息学这一全新领域的萌芽和发展奠定了坚实的基础。当世界各地的科学家开始认识到计算机的重要性并着手尝试利用计算机来组织、储备和分析生物学的观测资料时，生物信息学就已经萌芽了。然而，那些科学家当时也许并没有意识到生物信息学已经在他们手里诞生了。

2. 生物信息学数据库

与生命科学其他学科的发展不同的是，生物信息学的高速发展并不是主要依赖于理论上的重大突破，从实验中获得的大量数据以及基于这些数据建立的生物信息数据库在一定程度上是生物信息学发展的动力。美国洛斯阿拉莫斯国家实验室于 1979 年建立起 GenBank 数据库；1982 年欧洲分子生物学实验室的核酸序列数据库（European Molecular Biology Laboratory，EMBL）开始提供服务；日本也于 1984 年着手建立国家级的

核酸序列数据库（DNA data bank of Japan，DDBJ），并于 1987 年开始提供服务。1999 年，在美国国立卫生研究院（NIH）建议下，美国又相继建立了 20 个生物计算中心，专门从事生物信息学相关研究和人才培养。目前，美国已经成为世界上生物信息学研究实力最强的国家之一。GenBank 由 NIH 提供维护，并与 DDBJ 和 EMBL 一起，成为国际核苷酸序列数据库的主要成员。

GenBank 数据库的碱基总量大约每 13～15 个月翻一番，如图 1-3 所示。截至 2009 年 8 月，登录的序列总量已经达到 108 431 692 条，DNA 碱基对（base pairs）已经达到 106 533 156 756 条。生物学数据的积累并不仅仅表现在 DNA 序列方面，与其同步的还有蛋白质一级结构，即氨基酸序列的增长。此外，还有其他大量数据库资源已经面向大众。

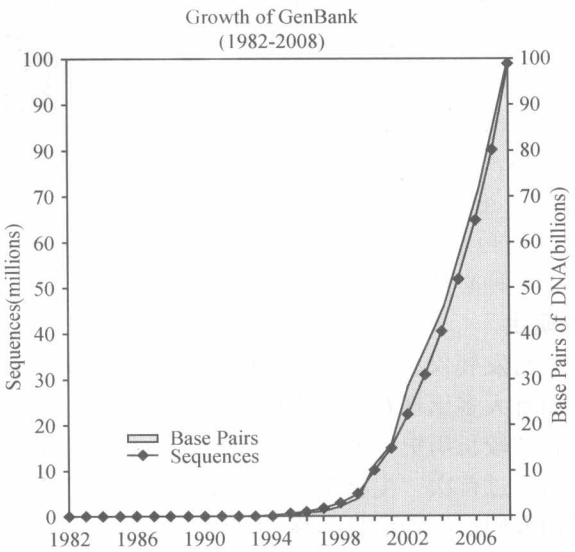


图 1-3　GenBank 数据增长

（引自 http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html）

Nucleic Acids Research 杂志在每年第一期中会详细介绍各种最新生物信息学数据库。在 2000 年 1 月 1 日出版的 28 卷第一期中共介绍了 115 种通用和专用的数据库。到 2009 年，这一数据已达到 1170，增长了 10 多倍。本书后面的章节将详细介绍有关数据库的使用方法，附录 1 中列出的是一些比较常用的在线生物学数据库，供读者参考。

1.1.3　生物信息学的研究内容

生物信息学在短短十几年间，已经发展成多个研究方向，国际上公认的生物信息学研究内容大致包括以下几个方面。

1. 生物信息的收集、存储、管理与提供

生物信息的收集、存储、管理与提供主要是指建立国际基本生物信息库及相关的评估与检测系统，从而提供与生物学研究相关的在线服务。目前，大部分的生物学数据库

是由一些国家支持的非赢利性机构和一些知名大学的研究院所专门提供，也有部分是商业性质的数据库（如 MDL）。分子生物信息数据库种类繁多，一般而言，这些数据库可以分为一级数据库和二级数据库。一级数据库的数据都直接来源于实验获得的原始数据，只经过简单的归类整理和注释。因此，一级数据库具有容量大、更新速度快、用户面广等特点，运行维护也需要具有高性能的计算机硬件和专门的数据库管理系统（Database Management System，DBMS）作支撑。例如，欧洲生物信息学研究所维护 EMBL 使用的是 Oracle，而基因组数据库 GDB 的管理、运行则基于 Sybase 数据库系统。较为流行的数据库管理软件有 Oracle、Sybase、Informix 和 Microsoft SQL Server 等。比较重要的一级核酸数据库有 GenBank、EMBL 和 DDBJ 等；蛋白质序列数据库有 Swiss-Prot 蛋白质序列数据库（Swiss-Prot Protein Sequence Database）、蛋白质信息资源（Protein Information Resouce，PIR）等；蛋白质结构库有 PDB（Protein Data Bank）等，序列数据库来自序列测定，结构数据库来自 X 射线衍射和核磁共振结构测定。二级数据库是在一级数据库、实验数据和理论分析的基础上针对特定目标衍生而来，是对生物学信息的进一步整理。二级数据库也称专门数据库、专业数据库与专用数据库。虽然二级数据库的种类繁多，但其容量小，可以不用大型商业数据库软件支撑。

2. 基因组序列信息的提取和分析

利用国际 EST 数据库（dbEST）和各实验室自己测定的 EST，通过电子克隆的方法发现新基因和新的单核苷酸多态性（single nucleotide polymorphism，SNP）位点以及各种功能位点；基因组中非编码区的信息结构分析，提出理论模型，阐明该区域的重要生物学功能；进行模式生物完整基因组的信息结构分析和比较研究；利用生物信息研究遗传密码起源、基因组结构的演化、基因组空间结构与 DNA 折叠的关系，以及基因组信息与生物进化关系等生物学的重大问题。

3. 功能基因组分析

功能基因组分析包括与大规模基因表达谱分析相关的算法、软件研究与开发、基因组比较、基因表达调控网络及相关通道的研究；与基因组信息相关的核酸、蛋白质空间结构的预测和模拟，以及蛋白质功能预测与蛋白质相互作用网络的研究。

4. 生物分子设计

生物分子设计包括 DNA 和 RNA 等核酸的结构建模和各种 RNA 分子的设计；发展生物大分子空间结构模拟、电子结构模拟和分子设计；具有不同模体（motif）的复合蛋白质以及连接肽（linker）的设计；生物活性分子的电子结构计算和设计；纳米生物材料的模拟与设计。

5. 药物设计

在已知蛋白质三级结构的基础上，可以利用分子对齐算法，采用分子模拟软件分析结合部位的结构性质，如静电场、疏水场、氢键作用位点分布等信息。可以在计算机上设计抑制剂分子，然后再运用数据库搜寻或者全新药物分子设计技术，获得分子形状和理化性质与受体作用位点相匹配的分子，合成并测试这些分子的生物活性，经过几轮循环，即可以发现新的先导化合物。因此，计算机辅助药物设计大致包括活性位点分析法、数据库搜寻及全新药物设计。

6. 生物信息分析的技术与方法研究

生物信息分析的技术与方法研究包括：发展有效的能支持大尺度作图与测序需要的软件、数据库以及若干数据库工具，如电子网络等远程通讯工具；改进现有的理论分析方法，如统计方法、模式识别方法、隐马尔科夫过程方法、分维方法、神经网络方法、复杂性分析方法、密码学方法和多序列比较方法等；创建一切适用于基因组信息分析的新方法、新技术，包括引入复杂系统分析技术、信息系统分析技术等；建立严格的多序列比较方法；发展与应用密码学方法及其他算法和分析技术，用于解释基因组的信息，探索 DNA 序列及其空间结构信息的新表征；发展研究基因组完整信息结构和信息网络的研究方法等。

7. 应用与发展研究

应用与发展研究汇集与疾病相关的人类基因信息，发展患者样品序列信息检测技术和基于序列信息选择表达载体、引物的技术，从而进行基因疾病诊断和基因药物开发等，建立与动植物良种繁育相关的数据库以及与大分子设计和药物设计相关的数据库。

8. 系统生物学研究

长期以来，生物学研究是在规模较小的实验室进行的，然而人类基因组计划、人类单体型图谱计划、人类蛋白质组学计划等的实施将生物学的学科交叉和国际研究合作扩展到了更大范围和更高层次，这也意味着我们有能力利用生物信息学研究方法对生物进行更全面、更系统的研究探索。通过计算生物学的方法来定量描述和预测生物功能、表型和行为。2008 年 10 月 6 日，第三届全国生物信息学与系统生物学学术大会在武汉召开，会议的主题是："生物信息学研究：从基因组信息学到系统生物学"。这也意味着生物信息学的研究内容已经触及系统生物学这一全新的领域，分子网络与通路、细胞层次系统生物学等也成为这次大会讨论的主题之一。

1.2 生物信息学应用

1.2.1 生物信息学的热点领域

目前，生物信息学的研究内容几乎涵盖了生命科学的各个领域，它的发展给生命科学研究带来重大的变革。生物信息学的发展也对生命科学本身的发展产生革命性的影响，其研究成果大大地促进了生命科学其他研究领域的进步。生物信息学的发展是目前基因组学、蛋白质组学、生物芯片等生命科学前沿研究领域发展的直接推动力。

1. 人类基因组计划

人类基因组计划（human genome project，HGP）由美国科学家于 1985 年率先提出，并于 1990 年正式启动。美国、英国、法国、德国、日本和中国科学家先后共同参与了人类基因组计划。这一计划旨在为 30 多亿个碱基对构成的人类基因组精确测序，发现所有人类基因并搞清其在染色体上的位置，破译人类全部遗传信息。我国于 1993 年加入该计划，并承担其中人类第 3 号染色体短臂上约 30Mb 的测序任务。2000 年 6 月 26 日，参加人类基因组工程项目的 6 国科学家共同宣布，人类基因组草图的绘制工作已经完成。最终完成图要求测序所用的克隆能忠实地代表常染色体的基因组结构，序