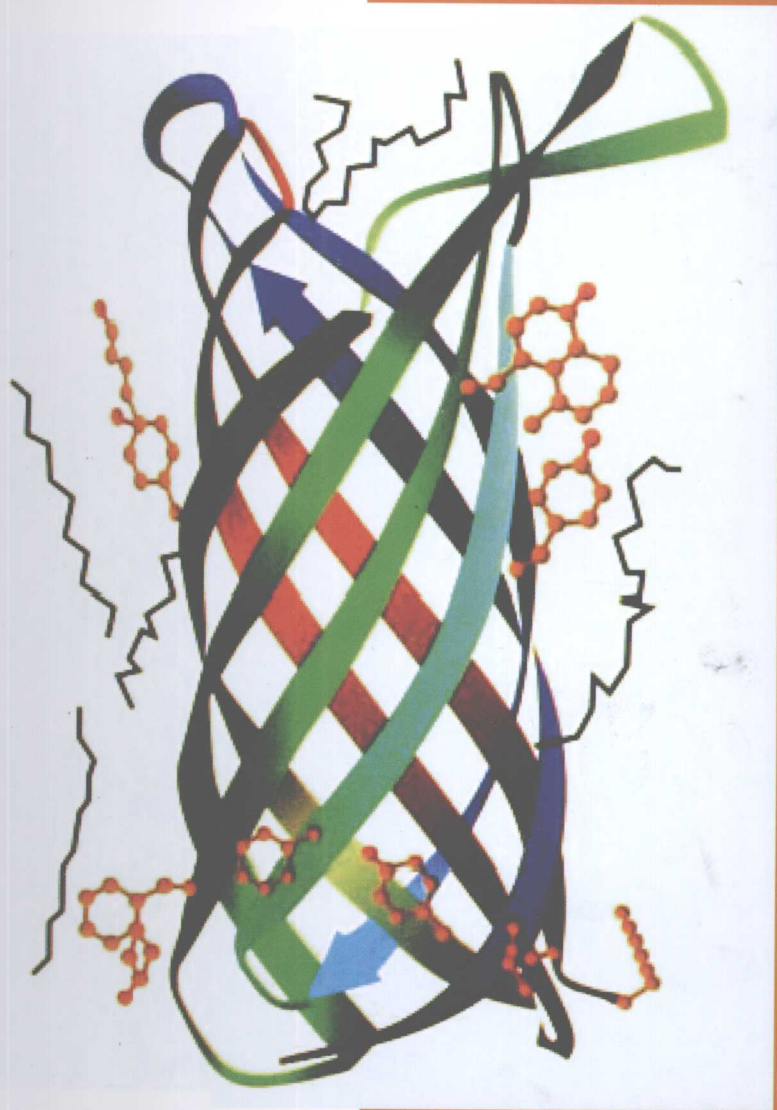


SHENGWU XINXIXUE

# 生物信息学

许忠能 主编

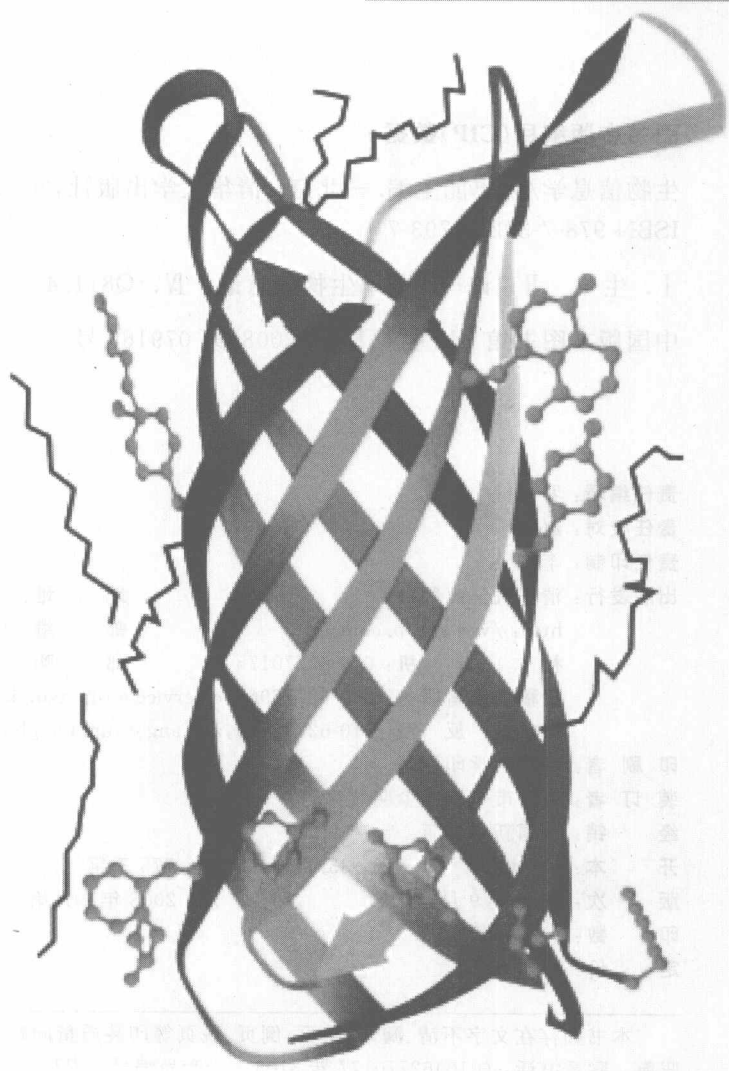


清华大学出版社

SHENGWU XINXIXUE

# 生物信息学

许忠能 主编



清华大学出版社  
北京

## 内 容 简 介

生物信息学是一门新兴的交叉科学。本书共分 16 章,详细介绍了生物信息学的定义、研究内容、生物学基础、数据库网络基础、算法与数学基础以及其在序列拼接、基因预测、引物设计、生物进化与分子发育分析、蛋白质结构预测、RNA 结构预测、生物芯片、计算机辅助药物设计、生物分子网络与生物系统仿真、DNA 计算中的应用与发展状况等内容。

本书结构清晰,系统完整,文笔流畅,既可作为高等院校相关专业师生的教材,也可作为该领域中研究、教学、软件开发等科研人员的参考用书。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

## 图书在版编目(CIP)数据

生物信息学/许忠能主编.—北京:清华大学出版社,2008.9  
ISBN 978-7-302-17793-7

I. 生… II. 许… III. 生物信息论 IV. Q811.4

中国版本图书馆 CIP 数据核字(2008)第 079169 号

责任编辑:罗 健

责任校对:赵丽敏

责任印制:李红英

出版发行:清华大学出版社

<http://www.tup.com.cn>

社 总 机:010-62770175

投稿与读者服务:010-62776969, [c-service@tup.tsinghua.edu.cn](mailto:c-service@tup.tsinghua.edu.cn)

质 量 反 馈:010-62772015, [zhiliang@tup.tsinghua.edu.cn](mailto:zhiliang@tup.tsinghua.edu.cn)

地 址:北京清华大学学研大厦 A 座

邮 编:100084

邮 购:010-62786544

印 刷 者:清华大学印刷厂

装 订 者:三河市李旗庄少明装订厂

经 销:全国新华书店

开 本:185×260 印 张:33.75 字 数:775 千字

版 次:2008 年 9 月第 1 版 印 次:2008 年 9 月第 1 次印刷

印 数:1~3000

定 价:59.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题,请与清华大学出版社出版部联系调换。联系电话:(010)62770177 转 3103 产品编号:022990-01

# 编 委 会

主 编 许忠能(暨南大学生命科学技术学院)

编 委 (按姓氏笔画排序)

Ken Chan (澳大利亚 University of the Sunshine Coast)

石 宏 (中国科学院昆明动物研究所)

许龙飞 (暨南大学信息科学技术学院)

刘吉平 (华南农业大学动物科学学院)

刘顺会 (广东药学院生命科学与生物制药学院)

陈 宣 (暨南大学信息科学技术学院)

周 杰 (中国科学院微生物研究所)

周 晖 (广东海洋大学水产学院)

黄蓓蓓 (武汉大学软件工程国家重点实验室)

蒙进芳 (西南林学院资源学院)

# 前 言

生物信息学融合了生命科学、信息科学等 21 世纪多个朝阳学科,越来越深入影响科学研究与社会生活的多个方面。开设生物信息学课程有利于提高学生科研理念及科技产业化意识。

许多学者已编写了不少优秀的生物信息学教材与专著,涵盖教学与科研的多个方面,但这些教材与专著往往仅适用于某一专业的学生。本书编写的目标是为不同专业背景的读者提供一本学习生物信息学的入门教科书。编者设计了大量的例子,便于读者理解相关原理。希望为推动国内生物信息学教学贡献绵薄之力。

2005 年,许忠能博士在暨南大学教务处的谢绍潮老师的建议下及清华大学出版社罗健编辑的帮助下,以试用多个学期的生物信息学课程教案为蓝本编写了提纲,向在生物信息学及相关领域教学与科研第一线的若干学者发出邀请,希望他们加入本书编委会。令人振奋的是, Ken Chan 博士、许龙飞教授、陈宣老师、黄蓓蓓博士生、刘顺会副教授、周杰博士、刘吉平副教授、周晖老师、石宏博士、蒙进芳老师欣然同意共同编写本书。 Ken Chan 博士参与编写了第 2、9 章的部分内容;许龙飞教授编写了第 3 章,同时参与编写了第 4、5 章的部分内容;刘顺会副教授编写了第 14 章;刘吉平副教授参与编写了第 8、15 章的部分内容;黄蓓蓓博士生编写了第 7 章,并参与编写了第 4、6 章的部分内容;陈宣老师参与编写了第 4 章的部分内容;周杰博士编写了第 10 章;周晖老师编写了第 12 章;蒙进芳老师参与编写了第 13 章的部分内容;石宏博士参与编写了第 13 章的部分内容;许忠能博士编写了第 1、11、16 章,并参与编写了第 2、5、6、8、9、15 章的部分内容,同时负责全书的统编、审校工作。暨南大学计算机系蔡利栋教授对第 5 章进行了审核。

本书得到暨南大学水生生物学国家重点学科相关基金的资助。编委会感谢暨南大学水生生物研究所韩博平教授对本书的大力支持,感谢中山大学邓日强副教授对本书编写过程中提供的帮助,感谢各编者所在单位的相关学生与老师、清华大学出版社、暨南大学教务处对本书的支持。同时,衷心感谢在此过程中,各编者的家人对编写工作的鼓励。

生物信息学涉及的领域众多,学科发展迅速,本书的内容难免出现错误,恳请各位读者批评指正。

本书编委会

2008 年 5 月于暨南大学

# 目 录

第 1 章 生物信息学概述 .....	1
1.1 背景与定义 .....	1
1.1.1 生物学原始数据量的急剧扩增 .....	1
1.1.2 名词“bioinformatics”的第一次出现 .....	2
1.1.3 定义 .....	3
1.2 研究内容 .....	4
1.2.1 生物信息的存储与获取 .....	4
1.2.2 序列比对 .....	4
1.2.3 测序与拼接 .....	5
1.2.4 基因预测 .....	5
1.2.5 生物进化与系统发育分析 .....	6
1.2.6 蛋白质结构预测 .....	6
1.2.7 RNA 结构预测 .....	6
1.2.8 分子设计及药物设计 .....	7
1.2.9 代谢网络分析 .....	7
1.2.10 生物芯片 .....	7
1.2.11 DNA 计算 .....	8
1.3 数据库、软件、科研教育机构 .....	8
1.3.1 数据库 .....	8
1.3.2 软件 .....	9
1.3.3 科研教育机构 .....	9
1.4 期刊与著作 .....	12
1.4.1 期刊 .....	12
1.4.2 著作 .....	13
1.5 生物学、计算机技术与数学基础 .....	13
1.5.1 生物学 .....	14
1.5.2 计算机技术 .....	14
1.5.3 数学 .....	15
1.6 展望 .....	15
1.6.1 研究内容的展望 .....	15
1.6.2 应用领域的拓展 .....	16
1.6.3 研究者的回报 .....	16
可免费登录的相关网站 .....	17

习题 .....	18
参考文献 .....	18
<b>第 2 章 生物信息学的生物学基础</b> .....	20
2.1 生物学研究的层次 .....	20
2.1.1 宇宙生命的研究 .....	20
2.1.2 生物与环境的关系 .....	22
2.1.3 生物种类 .....	25
2.1.4 生理 .....	29
2.1.5 细胞 .....	32
2.1.6 生物分子 .....	34
2.1.7 生物进化 .....	35
2.2 分子生物学基础 .....	37
2.2.1 核酸的结构 .....	38
2.2.2 蛋白质的结构 .....	42
2.2.3 DNA 的复制 .....	45
2.2.4 基因的转录 .....	47
2.2.5 蛋白质的生物合成 .....	48
2.3 人类基因组计划 .....	49
2.3.1 目标与意义 .....	50
2.3.2 资助 .....	50
2.3.3 研究机构 .....	51
2.3.4 研究方法 .....	51
2.3.5 目前结果 .....	55
可免费登录的相关网站 .....	57
习题 .....	57
参考文献 .....	57
<b>第 3 章 数据库与网络基础</b> .....	60
3.1 数据库技术基础 .....	60
3.1.1 数据库的基本概念 .....	60
3.1.2 数据库的体系结构和数据独立性 .....	60
3.1.3 关系数据库系统 .....	62
3.1.4 生物数据处理常用的数据库系统 .....	63
3.2 网络技术简介 .....	64
3.2.1 网络基础知识 .....	64
3.2.2 Internet 及其应用 .....	65
3.2.3 基于 Web 的数据库系统 .....	68

3.2.4 基于网络的搜索引擎 .....	70
可免费登录的相关网站 .....	73
习题 .....	73
参考文献 .....	73
<b>第4章 UNIX操作系统与计算机语言 .....</b>	<b>75</b>
4.1 UNIX操作系统 .....	75
4.1.1 UNIX历史 .....	75
4.1.2 UNIX系统的特点 .....	76
4.1.3 Redhat Linux 9的安装 .....	77
4.1.4 UNIX的基本使用 .....	81
4.1.5 大型应用软件 .....	85
4.2 计算机语言 .....	85
4.2.1 Perl语言简介 .....	86
4.2.2 Java语言简介 .....	89
可免费登录的相关网站 .....	89
习题 .....	90
参考文献 .....	90
<b>第5章 算法与数学基础 .....</b>	<b>91</b>
5.1 算法 .....	91
5.2 图论 .....	93
5.2.1 图 .....	93
5.2.2 寻找最短路 .....	95
5.2.3 欧拉图与哈密顿图 .....	98
5.2.4 树 .....	100
5.2.5 图论在生物信息学中的应用 .....	101
5.3 动态规划 .....	102
5.4 贝叶斯统计 .....	104
5.4.1 经典统计学的几个概念 .....	104
5.4.2 经典统计与贝叶斯统计的差异 .....	105
5.4.3 贝叶斯定理 .....	105
5.4.4 贝叶斯统计在生物信息学中的应用 .....	106
5.5 马尔可夫模型 .....	107
5.5.1 概念 .....	107
5.5.2 转移概率 .....	107
5.5.3 算法过程 .....	108
5.5.4 马尔可夫模型在生物信息学中的应用 .....	109



5.6	隐马尔可夫模型 .....	109
5.6.1	概念 .....	109
5.6.2	算法过程 .....	110
5.6.3	隐马尔可夫模型三个问题的研究 .....	113
5.6.4	隐马尔可夫模型在生物信息学中的应用 .....	113
5.7	神经网络模型 .....	114
5.7.1	神经网络的分类 .....	114
5.7.2	神经网络的学习方法 .....	115
5.7.3	神经网络模型在生物信息学中的应用 .....	123
5.8	遗传算法 .....	123
5.8.1	概念 .....	123
5.8.2	遗传算法运算过程 .....	124
5.8.3	遗传算法在生物信息学中的应用 .....	127
5.9	聚类分析 .....	128
5.9.1	相似性测度及聚类准则 .....	128
5.9.2	聚类算法 .....	128
5.9.3	聚类分析在生物信息学中的应用 .....	131
5.10	其他应用于生物信息学中的算法 .....	132
5.11	生物信息学中算法的发展 .....	132
	可免费登录的相关网站 .....	132
	习题 .....	132
	参考文献 .....	133
<b>第6章</b>	<b>序列比对 .....</b>	<b>134</b>
6.1	序列比对的概念 .....	134
6.2	序列比对的意义 .....	135
6.3	全局比对与局部比对 .....	136
6.3.1	全局比对 .....	136
6.3.2	局部比对 .....	136
6.4	计分方法 .....	137
6.4.1	匹配计分 .....	137
6.4.2	结构与性质的计分 .....	137
6.4.3	可观测变换计分 .....	137
6.4.4	空格罚分 .....	150
6.5	比对的算法过程 .....	150
6.5.1	两个序列比对 .....	150
6.5.2	多序列比对 .....	158
6.6	比对软件的使用 .....	165

6.6.1 用比对软件进行两序列比对 .....	165
6.6.2 用比对软件进行多序列比对 .....	168
6.7 计算机语言编写程序进行序列比对 .....	169
可免费登录的相关网站 .....	173
习题 .....	173
参考文献 .....	173
<b>第7章 序列拼接 .....</b>	<b>175</b>
7.1 霰弹法测序的 DNA 序列拼接 .....	175
7.1.1 霰弹法测序原理 .....	175
7.1.2 霰弹法测序拼接的计算模型 .....	176
7.2 杂交测序法的 DNA 序列拼接 .....	178
7.2.1 杂交测序法原理 .....	178
7.2.2 杂交法测序拼接的计算模型 .....	179
可免费登录的相关网站 .....	180
习题 .....	180
参考文献 .....	181
<b>第8章 生物信息数据库的查询与搜索 .....</b>	<b>182</b>
8.1 生物信息数据库 .....	182
8.1.1 核酸序列数据库 .....	182
8.1.2 蛋白质序列数据库 .....	183
8.1.3 结构数据库 .....	184
8.1.4 基因组数据库 .....	185
8.1.5 蛋白组数据库 .....	185
8.1.6 代谢组数据库 .....	185
8.1.7 疾病数据库 .....	185
8.1.8 药物与分子设计数据库 .....	186
8.1.9 分析与记录方式数据库 .....	186
8.2 生物信息数据库的字符匹配查询 .....	186
8.2.1 查询系统 SRS .....	186
8.2.2 查询系统 Entrez .....	193
8.3 生物信息数据库的相似性搜索 .....	199
8.3.1 BLAST .....	199
8.3.2 FASTA .....	207
可免费登录的相关网站 .....	210
习题 .....	210
参考文献 .....	211

<b>第 9 章 生物进化与分子系统发育分析</b> .....	213
9.1 生物进化 .....	213
9.1.1 进化理论的历史 .....	213
9.1.2 进化与自然选择的证据 .....	219
9.1.3 分子进化 .....	228
9.1.4 生物进化与生物信息学的关系 .....	234
9.2 分子系统发育分析 .....	234
9.2.1 分子系统发育分析的概念 .....	234
9.2.2 构建进化树的方法 .....	235
9.2.3 用网上软件构建进化树 .....	266
可免费登录的相关网站 .....	280
习题 .....	280
参考文献 .....	281
<b>第 10 章 基因预测与引物设计</b> .....	285
10.1 基因特征 .....	285
10.1.1 原核生物的基因特征 .....	285
10.1.2 真核生物的基因特征 .....	286
10.2 基于 EST 的基因鉴定 .....	288
10.2.1 EST 概念 .....	288
10.2.2 EST 的获得 .....	288
10.2.3 EST 与基因识别 .....	288
10.2.4 EST 的其他用途 .....	289
10.2.5 EST 数据的不足 .....	289
10.3 基因预测的算法 .....	289
10.3.1 相似性比较预测 .....	289
10.3.2 隐马尔可夫模型 .....	290
10.3.3 神经网络方法 .....	290
10.3.4 密码学方法 .....	290
10.3.5 Z-曲线法 .....	290
10.3.6 其他算法 .....	291
10.4 引物设计 .....	291
10.4.1 上、下游引物的 3'末端与 5'末端 .....	291
10.4.2 引物分子内不互补 .....	291
10.4.3 引物的长度、组分与解链温度 .....	291
10.5 网上的基因预测软件 .....	292
可免费登录的相关网站 .....	293
习题 .....	294

参考文献 .....	294
<b>第 11 章 蛋白质结构及其预测 .....</b>	<b>295</b>
11.1 蛋白质的结构及其实验测定方法 .....	295
11.1.1 蛋白质的结构概述 .....	295
11.1.2 维系蛋白质结构的作用力 .....	295
11.1.3 蛋白质结构的显示软件 .....	296
11.1.4 蛋白质结构的实验测定方法 .....	297
11.2 蛋白质分类 .....	304
11.2.1 按序列特征分类 .....	304
11.2.2 按在生物体中的位置分类 .....	304
11.2.3 按折叠类型分类 .....	304
11.3 蛋白质结构预测算法 .....	308
11.3.1 特殊序列预测 .....	308
11.3.2 蛋白质二级结构的预测 .....	309
11.3.3 蛋白质三级结构的预测 .....	314
11.4 蛋白质结构预测软件 .....	318
11.4.1 蛋白质二级结构预测软件 .....	318
11.4.2 蛋白质三级结构预测软件 .....	320
11.5 编写计算机程序进行蛋白质二级结构预测 .....	322
可免费登录的相关网站 .....	325
习题 .....	326
参考文献 .....	326
<b>第 12 章 RNA 结构与预测 .....</b>	<b>328</b>
12.1 RNA 的发现及其功能研究 .....	328
12.2 RNA 的结构特征及其与功能的关系 .....	330
12.2.1 RNA 的结构层次 .....	330
12.2.2 核糖体 RNA 的结构 .....	331
12.2.3 tRNA 的结构 .....	333
12.2.4 mRNA 的结构与功能 .....	336
12.2.5 核酶的结构与功能 .....	340
12.2.6 形成 RNA 特定结构的序列特征 .....	341
12.3 RNA 二级结构的预测算法 .....	343
12.3.1 比较序列分析方法 .....	343
12.3.2 动态规划算法 .....	344
12.3.3 对 RNA 结构预测算法的评价 .....	345
12.4 网上 RNA 二级结构分析软件 .....	346

可免费登录的相关网站 .....	347
习题 .....	347
参考文献 .....	347
<b>第 13 章 生物芯片 .....</b>	<b>349</b>
13.1 引言 .....	349
13.2 生物芯片的原理 .....	350
13.2.1 生物芯片的制备 .....	350
13.2.2 待检生物样品制备和标记 .....	355
13.2.3 生物分子之间的结合 .....	355
13.2.4 检测原理 .....	356
13.3 数据分析 .....	356
13.3.1 图像分析 .....	356
13.3.2 标准化处理(normalization) .....	357
13.3.3 Ratio 分析(ratio analysis) .....	358
13.3.4 聚类分析(clustering analysis) .....	358
13.3.5 基因表达数据库 .....	358
13.4 其他生物芯片技术 .....	359
13.4.1 微流路芯片 .....	359
13.4.2 活体化芯片 .....	359
13.4.3 芯片实验室(lab-on-a chip) .....	359
可免费登录的相关网站 .....	359
习题 .....	359
参考文献 .....	359
<b>第 14 章 计算机辅助药物设计 .....</b>	<b>361</b>
14.1 计算机辅助药物设计的概念 .....	361
14.2 药物设计的理论基础 .....	364
14.2.1 受体与配体 .....	364
14.2.2 理论计算方法 .....	370
14.3 结合自由能的计算 .....	375
14.3.1 自由能微扰/热力学积分方法 .....	375
14.3.2 线性相互作用能方法 .....	376
14.3.3 打分函数 .....	377
14.4 基于配体的药物设计 .....	377
14.4.1 定量构效关系方法 .....	378
14.4.2 药效团模型方法 .....	380
14.5 基于受体的药物设计 .....	383
14.5.1 重新配体设计 .....	384

14.5.2 分子对接虚拟筛选 .....	388
14.5.3 生物大分子建模和药物设计集成软件包——Insight II .....	401
14.6 药物发现集成平台 .....	404
可免费登录的相关网站 .....	406
习题 .....	407
参考文献 .....	407
<b>第 15 章 生物分子网络与生物系统仿真 .....</b>	<b>408</b>
15.1 生物分子网络 .....	408
15.1.1 生物分子网络的特征与研究方法 .....	408
15.1.2 代谢网络 .....	410
15.1.3 基因调控网络 .....	412
15.1.4 蛋白质相互作用网络 .....	414
15.2 生物系统仿真 .....	415
15.3 系统生物学概况 .....	417
可免费登录的相关网站 .....	418
习题 .....	419
参考文献 .....	419
<b>第 16 章 DNA 计算 .....</b>	<b>420</b>
16.1 DNA 计算的生物学基础 .....	420
16.1.1 DNA 的组成 .....	420
16.1.2 碱基配对 .....	420
16.1.3 DNA 分子的制备 .....	421
16.1.4 连接、合成 DNA 与 RNA 分子的酶类的作用 .....	422
16.1.5 切割 DNA 的酶类的作用 .....	422
16.1.6 DNA 序列的测定 .....	423
16.2 Adleman 开创 DNA 计算研究领域的实验 .....	423
16.3 DNA 计算的应用 .....	434
16.4 问题与展望 .....	435
可免费登录的相关网站 .....	436
习题 .....	436
参考文献 .....	436
<b>附表 1 生物信息数据库 .....</b>	<b>437</b>
<b>附表 2 中国、美国、英国、加拿大、澳大利亚科研教育机构开设生物信息学     专业的情况 .....</b>	<b>490</b>
汉英名词索引 .....	515
英汉名词索引 .....	519

# 第 1 章 生物信息学概述

**本章提要：**本章介绍了生物信息学的概况。生物信息学是一门交叉学科，它综合运用数学、计算机科学和生物学的各种工具，来阐明大量数据所包含的生物学意义。生物学、计算机技术、数学等学科是学习生物信息学基本理论及运用相关软件的基础。生物信息学研究的内容包括生物信息的存储与获取、序列比对、测序与拼接、基因预测、生物进化与系统发育分析、蛋白质结构预测、RNA 结构预测、分子设计与药物设计、代谢网络分析、基因芯片、DNA 计算等。生物数据库、相关软件是生物信息学研究与应用的重要资源。专门机构、期刊、著作在生物信息学的研究、教育、推广应用上起着重要的作用。生物信息学有非常好的发展前景及广泛的应用性，它给从事该专业的人有较高的回报。

生物信息学(bioinformatics)是一门新兴的交叉学科,生物学与医学、数学、计算机科学是其中三个主要组成部分。生物学家与医学家认为生物信息学主要是搞清生物学意义或医学意义,数学家认为算法与数学模型是其核心,计算机专业人员则觉得数据库及相关软件是生物信息学得以发展的基础。可见,这个学科涉及面广,其作用和价值在人类基因组计划的实现过程中得以充分体现。本章即介绍一下这个领域的概况。

## 1.1 背景与定义

### 1.1.1 生物学原始数据量的急剧扩增

在近代科学发展中,数学、计算机科学、生物学互相渗透、结合,用以解决许多重大的科学问题。生物学数据量剧增是近二十年来生物学出现的重大问题之一,它需要多学科结合去解决。

由于 20 世纪初遗传学、分子生物学、生物化学等生物学分支的研究异军突起,使众多学科纷纷在技术与理论上支持生物学研究。物理学领域的一位伟大的科学家 Delbrück 的著作《生命是什么》引领众多优秀的物理学家投入到生物学研究中去,并在其中产生了若干位诺贝尔生理医学奖得主。应用先进的物理学方法,科学家提出了生物大分子如蛋白质、核酸的结构模型,并最终得到验证。之后,对生物大分子的研究突飞猛进。在此基础上,人类对自身生物背景信息的探求欲望不断加深。1986 年,诺贝尔生理医学奖得主 Dulbecco 在 *Science* 杂志上提出“人类基因组计划”。经过几年的论证,最终由美国政府出资 30 亿美元开展此计划,这是美国历史上第三个国家计划。

美国国家科学院人类基因组制图与测序委员会在 1988 年确定了测序计划及所需资金,经美国国会审批,人类基因组计划从 1990 年开始,历时 15 年,由美国国立卫生研究

所、美国能源部资助 30 亿美元来完成。该计划的目的是测定包括人类基因组在内的若干基因组序列,开发相关基因组序列分析及储存技术,研究该计划对社会、法律及伦理的影响。实际看来,美国出的 30 亿美元是高收益的,从该计划实施带来的影响使美国稳居现代生物学界领头羊的位置,多国生物学精英汇聚美国,同时该计划为美国培养了一批生物学顶尖人才,仅这些方面的价值就不止 30 亿美元了。由于这项计划对人类的未来意义重大,同时学术界也不希望重要的生物学技术被垄断,因此先后有 5 个国家加盟这项计划,当然这些国家的相关机构也出钱资助。这次计划最初测定的是其他生物的基因组,以寻求合适的方案对人类基因组测序。这些过程推动了 DNA 测序技术的发展,使 DNA 序列量急剧增长。在 2001 年 2 月,人类基因组草图公开发表。2003 年,人类基因组项目成果已达到设定的目标,因此美国政府宣布该项目已完成,比原计划提前了两年。然而测序工作及对测序结果质量的验证工作将在此后多年继续进行。

人类基因组计划的完成,部分得益于数学与计算机技术的广泛应用。如在艰巨的测序过程中,多种序列分析软件起到关键的作用,保证了测序顺利进行。大型国际公用的生物大分子序列数据库不断接受各测序中心提交的序列信息,数据量以指数形式上升。图 1-1 是世界三大序列数据库之一的 EMBL 在人类基因组计划期间核酸序列数量的增长情况。

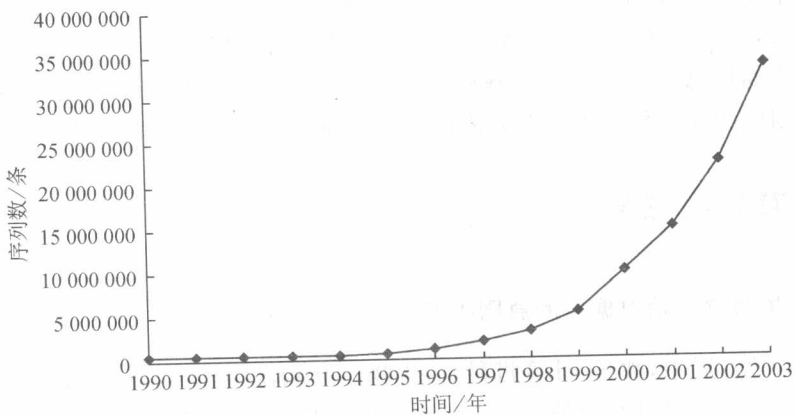


图 1-1 EMBL 数据库在人类基因组计划期间核酸序列数量增长情况

面对庞大的生物学数据,如何去分析、管理它们就变得非常重要。这时,传统的生物学研究手段已经束手无策,一门由生物学、数学、计算机科学紧密结合的技术——生物信息学应运而生,用以对付正在等待管理、分析与开发的海量生物学数据,就这样,生物信息学在人类基因组计划进行过程中产生不断发展壮大。

### 1.1.2 名词“bioinformatics”的第一次出现

生物信息学研究内容的源头较多,很难找到一个公认的标志性研究内容或研究论文作为该领域开创性的里程碑。有的学者觉得生物信息学的主要内容是计算分子生物学,分子生物学方面的事件左右着生物信息学,所以认为应将 1953 年 Watson 和 Crick 提出



DNA 双螺旋结构模型作为生物信息学的开端。而有的学者认为第一次公开提出将生物与信息结合的事件更有意义,所以 1956 年美国田纳西州的 Gatlinburg 召开的“生物学中的信息理论研究会”应被认为是揭开生物信息学序幕的事件。还有 20 世纪 60 年代 Dayhoff 等完成的《蛋白质序列与结构图册》、20 世纪 70 年代 Needleman 和 Wunsch 提出的序列比对算法、20 世纪 80 年代诞生的三大序列数据库等都是生物信息学的某些研究内容的开创性事件。然而这些事件中仍未提到目前生物信息学的专用名字——“bioinformatics”。

“bioinformatics”一词是由林华安(Hwa A. Lim)博士在 1987 年首创的。林华安博士出生于马来西亚,1981 年由英国伦敦大学帝国学院大学毕业,1986 年获美国 Rochester 大学博士学位。1987 年他到佛罗里达州立大学的超级计算机中心担任基因与生物物理组的主任,并于同年年底开始发表跨学科研究的论文。他认为信息学与生物学相结合是未来科学研究的一个潮流,并构思一个新的名词为这个新学科命名。1987 年,他最终敲定“bioinformatics”一词作为这个新学科的名字。1990 年,林华安博士组织了第一届生物信息学与基因组研究国际会议(Bioinformatics and Genome Research International Conference),并担任会议主席。此后 10 年,他连续担任该大会主席,于是“bioinformatics”一词也在学术界更加深入人心。

### 1.1.3 定义

生物信息学的定义随着其研究发展与实际需要而几经改动,一般认为,1995 年美国人类基因组计划第一个五年总结报告中生物信息学的定义较为完整。这个定义是:生物信息学是一门交叉学科,它包含了生物信息的获取、处理、存储、分发、分析和解释等在内的所有方面,它综合运用数学、计算机科学和生物学的各种工具,来阐明和理解大量数据所包含的生物学意义。

在实际的工作及相关文献中常出现一些词与生物信息学相关,如计算分子生物学、后基因组学、生物数学、生物学、信息学等。由于不少词所包括的内容有交叉或不同的作者对同一名词的解释有所偏重,所以不少初学者容易对这些词产生混淆。以下简单介绍一些相关的专业名词的定义,以示与生物信息学的区别。

**生物学**(田清涑 2000) 是研究生命的科学,是研究生命现象的本质及探讨生物发生、发展及其活动规律的科学,又称生命科学(bioscience)。生命有几个主要特征:生命体内的化学元素种类与各元素的比例相似,体内有相似的生物大分子,能进行新陈代谢、生长发育与繁殖,会产生遗传和变异,具有应激性。

**信息理论**(常迥 1993) 是研究信息的产生、获取、度量、变换、传输、处理、识别及其应用的一门科学。来自生物方面的信息源是自然信息源的一个组成部分。

**生物数学**(徐克学 1999) 是一门介于生物学与数学之间的边缘学科,它以数学方法研究和解决生物学问题,并对与生物学有关的数学方法进行理论研究。

**计算分子生物学**(João Setubal 和 João Meidanis 1997) 计算分子生物学是开发和使用数学与计算机技术,以帮助解决分子生物学中的问题的一门学科。

**生物统计学**(李春来,王志和,王文林 2004) 是用数理统计的原理和方法来分析和解释生物界各种现象和实验资料的科学。