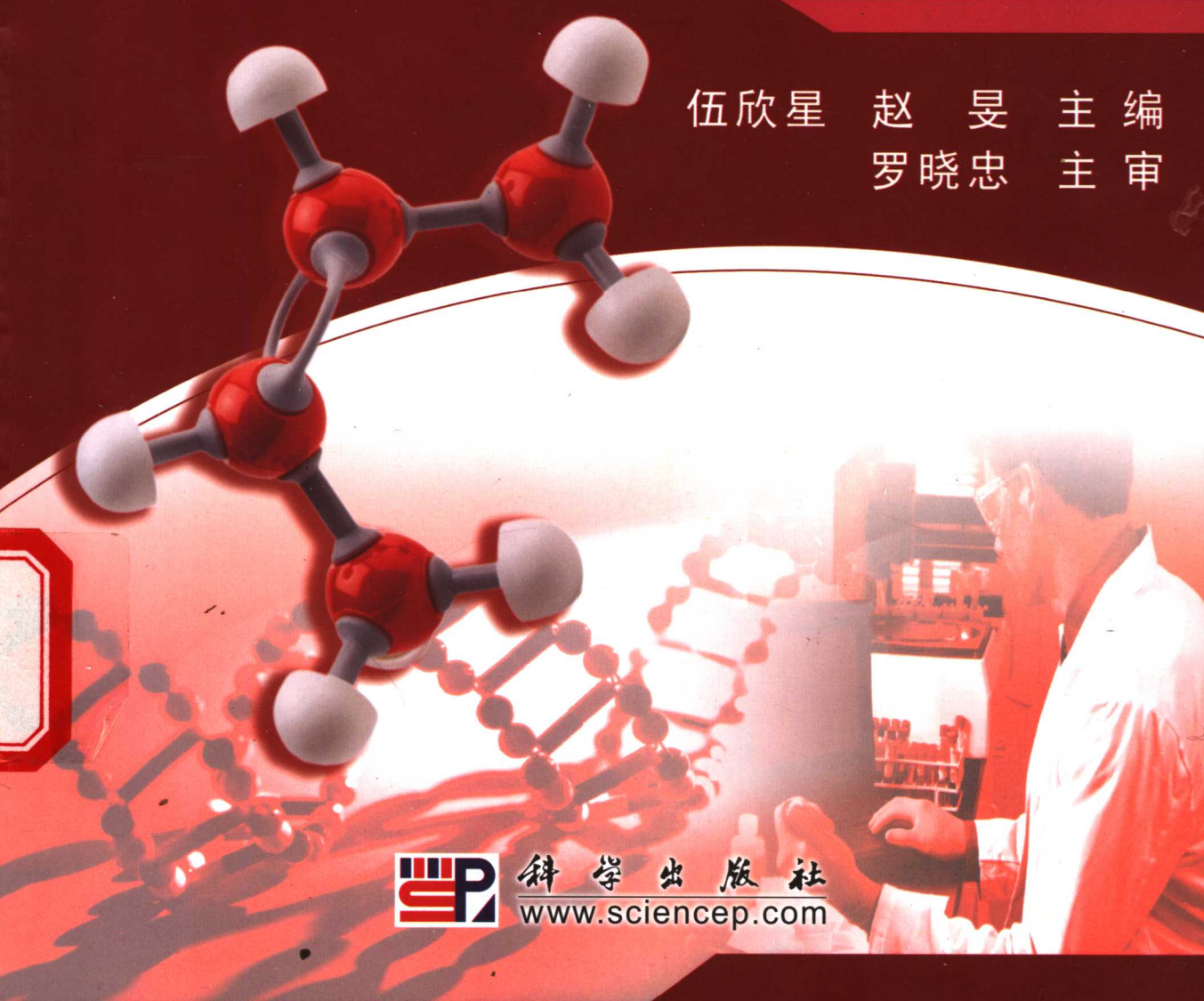
生物信息学
基础与临床医学应用指南

Bioinformatics:
A Practical Guide for Biomedical Research

伍欣星 赵 旻 主编
罗晓忠 主审



科学出版社
www.sciencep.com

生物信息学

——基础与临床医学应用指南

伍欣星　赵　旻　主　编

罗晓忠　主　审

国家自然科学基金资助项目(30171042)

科学出版社

北京

内 容 简 介

本书较为详尽地介绍了生物信息学在医学科研和临床应用中的最新信息及资料。全书分为上下两篇，共十四章，通过大量的实例，系统介绍了生物信息学的一些基本知识，以及生物信息学在功能基因组学研究中的应用，这些内容对于医学科研的设计和实施将极具指导意义。

本书对分子生物学以及生物信息学的一些名词给出了中英文对照和必要的解释，列出了一些常用的生物信息学相关网站，更加方便了读者的使用。

本书既可作为生物信息学课程的教材，也是一本实用性很强的生物信息学参考书。

图书在版编目（CIP）数据

生物信息学——基础与临床医学应用指南/伍欣星等主编. - 北京：科学出版社，2005

ISBN 7-03-015127-5

Ⅰ.生… Ⅱ.伍… Ⅲ.医学—生物信息论 Ⅳ.R318.04

中国版本图书馆 CIP 数据核字（2005）第 017075 号

责任编辑：王雨舸

责任印制：高 嵘／封面设计：李梦佳

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

武汉大学出版社印刷总厂印刷

科学出版社发行 各地新华书店经销

*

2005 年 3 月第 一 版 开本：787×1092 1/16

2005 年 3 月第一次印刷 印张：16 3/4

印数：1~4 500 字数：406 000

定价：30.00 元

（如有印装质量问题，我社负责调换）

前　言

随着人类基因组计划的初步完成，生命科学进入到了基因组学、蛋白质组学的时代，充满奥秘的生命科学研究也变得更为深广，更为精彩。与此同时，加工并处理各种已经和将要获得的诸如分子生物学、基因组学、蛋白质组学、基因芯片、蛋白质三维结构、分子进化学、生物化学、电生理学、系统生物学以及各生物分支学科大量的信息，成为生命科学研究的主要工作之一。于是，一门崭新的学科——生物信息学(bioinformatics)应运而生并快速发展起来。10 多年来，通过与计算机技术、网络技术密切联系，以各种计算理论与算法为理论基础，由高速发达的网络提供便捷的技术平台，生物信息学已经成为一门新兴的朝阳学科。这门学科所包含的强大信息分析技术手段和新颖的理论思路，为相关学科的研究提供了关键的支持。21 世纪是信息科学的时代，及时了解并掌握必要的生物信息学知识，不仅是生命科学研究者必备的技能之一，也是开阔研究思路和科研创新的基础，在科学研究中将起到事半功倍的作用。

当前的医学研究，早已深入到基因及基因组等分子水平，传统的医疗实践活动已经大量融入现代的基因诊断和治疗的理念与方法。很多以前停留在理论中的新技术和新理论，诸如 SNP 分析、生物芯片、EST 筛选等分子技术，已经成为医学研究者必不可少的工具。众多优秀的核酸、蛋白质数据库均可以通过互联网访问并加以利用，各种初级原始的实验数据和资料可以通过生物信息学的手段进一步加以分析和处理。为使医学研究者能系统地学习生物信息学的基本知识，并尽快掌握运用生物信息学的技术方法，我们从医学研究的角度出发，编写了本书。

本书分为上下两篇，共十四章。其中上篇包括第一章至第十章，主要介绍了生物信息学的基本知识，常用的核酸、蛋白质数据库，以及核酸、蛋白质的序列分析技术、微阵列及数据分析。其中包含有医学生物信息学的相关内容，如医学文献资料的查询检索、疾病相关数据库、药物及药物筛选数据库的应用、生物信息学常见软件的应用及常见的与医学相关的生物信息学网站等。各章均附有详细的实例分析，便于读者分析研究和自学。下篇则主要讨论了生物信息学在功能基因组学研究中的应用，详细介绍了通过生物信息学分析核酸及蛋白质基本信息、实验方案的设计和技术路线、实验方法的选择以及各种实验的基本原理，这些内容对于医学研究的设计和实施将极具指导意义。附录介绍了部分分子生物学、生物信息学术语，便于读者查询和运用。本书遵循简洁易懂的原则，立足于基础知识，着重于实践应用。全书每个章节以应用案例形式介绍生物信息学的基本原理和常用工具，通过视窗和图片等版面向读者展示生物信息学和功能基因组的精髓，使读者轻松高效地利用生物信息资源。本书也力求将生物信息学与功能基因组学相结合，为基础与临床医学工作者从事功能基因组研究提供一个互动平台。读者可以利用书本解决自己的问题，也能通过书本提出自己的问题。这种分析、假设和推理的过程，正是培养科学思维和科学精神的必由之路。

本书在编写过程中，武汉大学医学院病毒学研究所所分子病毒学研究室师生、深圳东湖医院聂广教授承担了大量的组织、撰写和编辑、校对工作，他们长期以来一直从事医学分子生物学、分子病毒学和肿瘤学的科研与教学工作，作为工作在医学科研第一线的研究人员，他们对于生物信息学在医学中的运用的重要性有独到而深刻的认识。此外，非常荣幸的是我们聘请了美国 SAIC 资深生物信息学家罗晓忠博士担任本书的主审，孙宏为博士、肖念清博士为编委。

罗晓忠博士现任 SAIC 资深生物信息学家(Sr Bioinfomatics Scientist)和学术负责人(Scientific and Technology Leadership)，先后在美国国立医学科学院(NIH)国立癌症研究所生物信息中心(NCICB)及其他生物信息、临床研究机构从事生物信息数据库和软件的开发、系统生物学及临床基因组研究。他多次主持生物信息科研和开发项目，并担任多个生物信息系统软件的总设计师，所研究的主要项目有 NIH 国立癌症研究所生物信息中心大型癌症芯片(cancer microarray)的数据库和数据分析项目、癌症基因组和综合生物信息(cancer genomic and integrated bioinformatics)、癌症临床基因组项目(cancer clinical genomics integration)、临床数据管理系统。孙宏为博士长期从事以计算机为主要手段的大规模生物数据分析，近五年来主要从事基因芯片微阵列的数据分析，先后担任德国拜耳公司(Bayer)及美国基因逻辑公司(Gene Logic)生物信息部的研究员及高级研究员，现任美国 NIH 国立关节炎、肌肉及皮肤病研究院生物信息部主任。肖念清博士现任美国科学应用公司(SAIC)资深研究员，作为项目负责人和学术带头人，多次参与美国国立医学科学院大型基因芯片数据库和分析平台的构建。

本书的出版得到了国家自然科学基金及微生物学国家重点学科经费的资助。

生物信息学是一门发展极快、实践性极强的学科，也是一门交叉性学科。学好并能熟练运用这门学科，还需要了解相关专业的知识、大量的实践应用，这是学好生物信息学的保证。当然，生物信息学发展很快，由于作者涉猎的范围及知识水平有限，书中一定存在很多的纰漏和错误之处，敬请各位同行专家和广大的读者批评指正。

编　者

2004 年 11 月

《生物信息学——基础与临床医学应用指南》编委会

主　编　伍欣星　赵　旻
副主编　左泽华　高桂芳
主　审　罗晓忠
编　委　(以姓氏笔画排序)

万　静　华中科技大学同济医院
王同祥　武汉大学医学院病毒学研究所
左泽华　武汉大学医学院病毒学研究所
伍欣星　武汉大学医学院病毒学研究所
刘　娟　中国科学院武汉病毒研究所
孙宏为　美国 NIH 国立关节炎、肌肉及皮肤病研究院生物信息部
张　力　武汉大学医学院病毒学研究所
张　涵　武汉大学中南医院
李　辉　武汉大学医学院病毒学研究所
肖念清　美国科学应用公司(SAIC)
欧　璇　武汉大学医学院病毒学研究所
罗晓忠　美国科学应用公司(SAIC)
荣　媛　武汉大学中南医院
赵　旻　武汉大学医学院病毒学研究所
高桂芳　武汉大学医学院病毒学研究所
熊金虎　武汉大学中南医院

目　录

上篇　生物信息学基础

第一章　生物信息学概述 …… 3
1.1　生物信息学的定义和研究范畴 …… 3
1.1.1　生物信息学的定义 …… 4
1.1.2　生物信息学中的数据库与网络 …… 4
1.1.3　生物信息学的主要研究范畴 …… 5
1.2　生物信息学的建立与发展 …… 7
1.3　医学生物信息学的发展与展望 …… 10
1.3.1　医学生物信息学的主要研究内容 …… 11
1.3.2　生物信息学的发展和展望 …… 11
第二章　医学生物信息数据库 …… 13
2.1　医学生物信息数据库简介 …… 13
2.2　国外常用医学文献数据库 …… 17
2.2.1　PubMed 文献数据库 …… 17
2.2.2　HighWire Press 电子期刊数据库 …… 20
2.3　国内常用生物医学文献检索数据库 …… 23
2.3.1　万方数据资源系统 …… 23
2.3.2　中国期刊网 …… 25
第三章　核酸数据库的应用 …… 29
3.1　常用的 DNA 数据库及软件 …… 29
3.1.1　GenBank——NCBI 核酸序列数据库 …… 29
3.1.2　EMBL——欧洲核酸序列数据库 …… 35
3.1.3　DDBJ——日本 DNA 数据库 …… 36
3.2　常用的 RNA 数据库及软件 …… 36
3.2.1　Transterm——mRNA 序列和翻译调控元件数据库 …… 37
3.2.2　RDP-II——核糖体数据库 …… 38
3.2.3　RNA 二级结构预测 …… 38
3.3　核酸同源性序列比对的策略和方法 …… 40
3.3.1　数据库中的相似性搜索 …… 40
3.3.2　BLAST 简介 …… 40
3.3.3　BLAST 应用举例 …… 43
3.4　新序列的提交 …… 46
第四章　人类基因组变异数据库 …… 48
4.1　SNP 数据库 …… 48

4.1.1 dbSNP 数据库……49
4.1.2 人类基因组变异数据库……51
4.2 突变数据库 …… 52
4.3 基因标记物与微卫星数据库 ……54
4.4 观察 SNP 和突变的工具……55
4.4.1 在基因组水平上观察 SNP 和突变的工具 …… 55
4.4.2 在基因水平上观察 SNP 和突变的工具 …… 55
第五章 蛋白质资源数据库 ……59
5.1 SWISS-PORT 蛋白序列数据库 ……59
5.1.1 SWISS-PORT 蛋白序列数据库区别于其他蛋白序列数据库的优点……59
5.1.2 SWISS-PROT 数据库的结构与级别……60
5.1.3 序列条目的结构……60
5.1.4 不同的行类型……63
5.1.5 数据库的检索……66
5.2 ASTRAL——蛋白质结构和序列分析体系 …… 67
第六章 生物芯片…… 70
6.1 概述…… 70
6.1.1 生物芯片简介……70
6.1.2 生物芯片分类……70
6.1.3 几种常见的生物芯片……71
6.2 基因芯片基本原理和基本流程 ……72
6.2.1 基因芯片的基本原理……72
6.2.2 基因芯片的基本流程……72
6.3 几种新型的芯片技术…… 76
6.4 生物芯片的应用 ……78
6.5 生物信息学中的新技术……80
附 基因芯片进行基因差异表达实际操作举例……82
第七章 疾病相关数据库 ……87
7.1 综合临床数据库 ……87
7.2 肿瘤相关数据库 ……91
7.2.1 Cancer.gov——肿瘤网……91
7.2.2 Oncolink ……94
7.2.3 癌症基因组剖析计划(CGAP) …… 95
7.2.4 中国癌症网 …… 97
7.3 心血管疾病相关数据库……97
7.3.1 心血管疾病相关医学数据库(Cardio) …… 97
7.3.2 中华心血管医学网……99
7.4 遗传性疾病数据库 …… 100
7.5 感染性疾病数据库 …… 102
第八章 生物信息学与药物设计 ……105

8.1 概述 …… 105
8.2 生物信息学在药物设计中的优势 …… 106
8.3 生物信息学在药物设计环节中的应用 …… 108
8.3.1 初始阶段:事半功倍的效果 …… 108
8.3.2 生物活性筛选阶段:提高筛选命中率 …… 109
8.3.3 药物开发阶段:联系遗传信息与药物疗效的桥梁 …… 110
8.4 药物设计过程中生物信息学应用流程 …… 110
8.4.1 综合分子生物学方法 …… 110
8.4.2 EST 数据库搜寻 …… 111
8.4.3 结构生物学方法 …… 111
8.5 生物信息学在药物设计中的其他应用 …… 112
8.5.1 药物作用的机制 …… 112
8.5.2 药物的药代动力学及毒理性质的研究 …… 112
8.5.3 计算机辅助药物设计 …… 113
8.6 后基因组时代药物研究的新进展和新趋势 …… 113
附 药物设计实例 …… 114
第九章 常用软件介绍 …… 115
9.1 Omiga 介绍 …… 115
9.2 Antheprot 介绍 …… 118
9.3 MACAW 介绍 …… 125
9.4 Primer Premier 介绍 …… 128
9.5 Reference Manager 介绍 …… 130
9.6 常用限制酶分析与质粒作图软件 …… 135
9.6.1 Gene Construction Kit 2.5 …… 135
9.6.2 Clone Manager 7 …… 139
9.7 RNA 二级结构预测及分析软件 …… 139
9.7.1 RNAdraw 1.1b …… 139
9.7.2 RNA Structure 3.2 …… 143
9.8 序列综合分析软件 …… 145
第十章 基因芯片微阵列数据分析 …… 148
10.1 常用基因芯片及其数据简介 …… 148
10.2 基因芯片数据处理与分析 …… 150
10.3 基因芯片数据分析的基本策略与方法 …… 151
10.4 基因微阵列数据分析中的常用软件 …… 155
10.4.1 Excel …… 155
10.4.2 SAM …… 159
10.4.3 R 及其在基因表达数据分析中的应用 …… 159

下篇 生物信息学与功能基因组学互动平台

第十一章 生物信息学与基因组学技术 …… 165

11.1 新基因分析的生物信息学策略 ……165
11.2 新基因的分离——cDNA 末端快速扩增技术 …… 168
11.3 基因突变检测(分析)技术……171
11.3.1 单链构象多态性技术……171
11.3.2 变性梯度凝胶电泳技术……173
11.3.3 测序与直接测序……175
11.3.4 单碱基延伸标签阵列技术(SBE-TAGS) …… 175
11.3.5 异源双链分析(HA)技术 …… 176
11.3.6 连接酶链反应(LCR) ……176
11.3.7 等位基因特异性寡核苷酸杂交(ASOH) …… 176
11.3.8 等位基因特异性扩增法(ASA) ……177
11.3.9 RNA 酶 A 切割法(RNase A cleavage) …… 177
11.3.10 基于PCR、酶切的技术……177
11.3.11 高通量检测技术 ……178
11.4 mRNA 差异显示技术……179
11.4.1 DD-PCR 基本原理……180
11.4.2 DDRT-PCR 技术路线 …… 180
11.4.3 DD-PCR 的优越性……182
11.5 比较基因组杂交技术……182
11.6 微阵列-比较基因组杂交技术…… 184
11.7 基因表达分析技术…… 186
11.7.1 基因表达系列分析技术……186
11.7.2 RNase 保护试验……189
11.7.3 RNA 印记杂交技术 …… 191
11.7.4 实时荧光定量 PCR 技术 …… 193
11.8 SNPs、ESTs 在研究新(未知)基因中的应用 …… 195
11.8.1 单核苷酸多态性(SNP) …… 195
11.8.2 表达序列标签(EST) …… 196
第十二章 RNA 组学及常用研究技术……199
12.1 反义核酸技术 ……199
12.2 核酶技术……202
12.3 RNA 错折叠技术…… 204
12.4 RNA 干扰技术…… 205
12.4.1 RNAi 作用的机制 ……206
12.4.2 RNAi 实验方案 ……207
12.4.3 RNAi 在基因功能分析中的应用 ……209
第十三章 模式生物体研究……210
13.1 转基因动物 …… 210
13.1.1 转基因动物概念……210
13.1.2 基本原理……211

13.1.3 嵌合体动物……212
13.1.4 转基因动物模型在医学研究中的应用……212
13.2 基因打靶技术 ……215
13.2.1 基因打靶技术的原理……216
13.2.2 基因打靶的操作要点……216
13.2.3 提高基因打靶效率的途径……218
13.2.4 基因打靶技术的应用……219
13.3 时空可调节性基因打靶技术与基因陷阱 ……219
13.3.1 时空可调节性基因打靶……219
13.3.2 基因陷阱……221
13.3.3 诱变技术在功能基因组学中的应用……223
第十四章 蛋白质组学技术……224
14.1 蛋白质组分离技术……225
14.1.1 二维聚丙烯酰胺凝胶电泳……225
14.1.2 高效液相色谱(HPLC) ……228
14.1.3 毛细管电泳及电色谱(CE/CEC) ……228
14.2 鉴定技术……229
14.2.1 质谱技术……229
14.2.2 图像分析技术……233
14.2.3 高流通量筛选(HTS) ……234
14.3 蛋白质芯片技术 ……235
14.4 酵母双杂交系统 ……237
附录 生物信息学及分子生物学术语……241

上　　篇

生物信息学基础

第一章　生物信息学概述

生物是一种能储存并加工信息的复杂系统。数量庞杂且种类繁多的生物信息在细胞之间、生物个体之间、生物种群之间相互交流并得以保存。人类对于生命信息的研究主要体现在用生物学和遗传学的理论与方法分析生命现象与规律上。但这种研究在人类历史中长期处于较低的水平，这主要是因为对生命本质的研究与理解没有取得令人满意的成果。直到 19 世纪，Gregor Mendel 发现并提出基因是遗传的基本单位，而且遗传信息在纵向传递过程中基因保持相对独立等现象及规律，使人类终于认识到生物信息在生命活动中的重要作用和相应的运行规律。人类对遗传及生命现象的研究开始迅猛发展，细胞核、染色体、基因组等成为 20 世纪人类科学研究的主要战场，大量的生物信息不断涌现，大量生物信息的积累和科学的进步已使人类有能力窥探生命，包括人类自身的秘密。人类基因组计划的初步完成，使人类基因组被初步破译，将人类推进到了一个激动人心的时代——基因组时代。随着人类基因组测序工作的初步完成，由 30 亿个字符组成的人类遗传密码本显现出来。来自其他模式生物的基因组信息也在不断地涌现和破译。所有这些海量的生物信息只是用简单的遗传语言，即 DNA 的四种碱基和蛋白质的 20 种氨基酸写成，但是其数量巨大，浩如烟海，其中蕴藏的规律仍是深奥难解。

生物科学正在经历从试验分析和数据积累到数据分析及其指导下的试验验证的转变，即从分析还原思维到系统整合思维的转变。急剧膨胀的数据资源，以及大量多样化的生物学数据资源中所蕴含的重要生物学规律，迫使人们寻求一种强有力的工具来有效地组织并研究这些生物信息，协助人脑完成庞杂的分析工作，以利于对这些生物学知识的储存和进一步加工利用。近年来以数据处理分析为本质的计算机科学技术和网络技术已经获得突飞猛进的发展，计算机科学技术和网络技术已经日益渗透到生物科学的方方面面。正是在这些学科的有力支持和相互作用下，一门崭新的拥有巨大发展潜力的生物信息学由此悄然而坚定地发展和成熟起来。人类基因组测序工作的初步完成在很大程度上受益于生物信息学的发展。生物信息学作为生物科学与计算科学融合体的诞生，是历史的偶然但更是历史的必然，可以说生物信息学是现代基因组、蛋白质组及以此为基础的现代生物学的孪生姐妹。生物信息学是一门年轻的学科，更是一门充满挑战和机遇、有广阔发展前景的新兴学科。

1.1　生物信息学的定义和研究范畴

加工并处理大量生物信息，是相关生命科学研究的主要工作。面对大量的分子生物学、基因组学、蛋白质组学、基因芯片、蛋白质三维结构、分子进化学、生物化学、电生理学、系统生物学以及各生物分支学科信息，在计算机科学、网络技术以及生物分析技术的相互作用和渗透下，诞生了一门崭新的学科——生物信息学(bioinformatics)。生物信息学是一门交叉学科，它包括了生物信息的获取、处理、存储、分发、分析和解释等各个方面，在综合运用数学算法、计算机科学及其他实验生物学的基础上，阐明和解释庞杂的数据中所包含的生物学信息和意义，并且辅助生物学的研究，模拟复杂的生物系统。

1.1.1 生物信息学的定义

生物信息学是在数学、计算机科学和生命科学的基础上形成的一门新型交叉学科，即为理解各种数据的生物学意义，运用数学、计算机科学与生物学手段进行生物信息的收集、加工、储存、传播、分析与解析、模拟的科学。随着人类对生命科学研究的不断深入，尤其是人类基因组计划的实施和初步研究结果的获得，现阶段的生物信息学也可以理解为是利用基因组 DNA 序列信息分析作为源头，在获得蛋白质编码区的信息后进行蛋白质空间结构模拟和预测，然后依据特定蛋白质的功能进行必要的药物设计的科学。基因芯片的高通量、并行分析，蛋白质组学以及研究蛋白质在不同时空情况下的表达及蛋白质结构和功能的预测，将为生物和医药的研究提供崭新的空间。

生物信息学还有另一个经常被使用的名字：计算生物学(computational biology)，此外计算分子生物学(computational molecular biology)和生物分子信息学(biomolecular informatics)等也被使用过。

1.1.2 生物信息学中的数据库与网络

生物信息学所处理的生物信息，数量巨大而且增长迅速。核酸库的数据每 10 个月左右就要翻一番。至 2000 年底，数据库数据达到了创记录的 100 亿个记录。特别是 2000 年 6 月 26 日人类基因组“工作框架图”绘制完成后，已经明确人类基因组共有 32 亿个碱基对，包含了大约 3 万多个蛋白编码基因。这些以及其他数量惊人的数据仍在不断呈几何级数地增长。

学习生物信息学，必须了解数据、数据库及计算机网络等相关的知识和理论。数据库技术是计算机学科的一个分支，将众多的信息数据储存并进行有效地管理，以实现信息资源的有效利用和共享，是一个优秀的数据库所应该具备的条件。同样，作为相关学科专业的科研人员也应该懂得将数据库应用到实际工作中，这样才能在未来的科学研究中事半功倍。

1. 数据及数据库

数据是指描述事物的符号，通常接触到的数据主要有文字、图像和声音等，它是数据库中存储的基本对象；数据库则是数据的集合，数据按一定的规律和模型进行组织和储存并可共享，这就组成了数据库；数据库系统则是指具体的数据库管理系统软件和相关的数据库的集合体，通常由软件、数据库和数据库管理员组成。生物信息学研究的一个核心问题是数据库的开发：如何整合和最有效地查询来自基因组 DNA 序列、mRNA 表达的空间和时间模式、蛋白质结构、免疫反应、文献记录等数据。另外核酸或蛋白质序列中识别模式的算法、相似性比较或系统发育构建、线性序列或高维结构的模序识别和基因表达的共有模式等，均是数据库研究的领域。

在现阶段的生物信息学研究中，面临着数据公开前的使用和已公开数据的保存限制问题。这些问题昭示了生物信息学理论中一个非常重要的理念，即数据的共享性和应用性。众所周知，数据的尽早公开发表对许多研究具有重要意义，例如人类基因组计划，其数据正式公布即上网公开发表，而且明确指出相关数据为全人类所共享，这样就极大地促进了相关领域的研究活动。当然这种利他主义的数据释放政策需要相应的知识产权保护。而数据的应用性限制问题，则主要是指来源于众多私营(人)公司的信息数据，由于受到限制而无法进一步加以重组和利用。只有在数据共享的情况下，生物信息学才能在人类基因组、蛋白质组等系统工程的研究中发挥最大的作用。

目前，国际核酸序列数据库(GenBank/EMBL/DDBJ)是国际上最大的核酸序列数据库，它由美国国家生物技术信息中心 NCBI、欧洲分子生物学实验室、日本国立遗传学研究所共同制作。截至 2004 年 8 月 15 日，该数据库收纳了包括人类、动物、植物、微生物及人工合成在内的 3734 万个序列记录、418 亿个碱基和 3734 个基因位点。GenBank/EMBL/DDBJ 数据库收录了国际上几乎所有的已知核酸序列数据，为研究者提供了不可或缺的技术和信息支持。这三个数据库每天都互相更新内容以保持一致。还有数以千计的数据库已经或正在建立之中，它们涵盖了核酸、蛋白质、糖类等生物大分子的生物信息，还包括大量的生物文献资料(PubMed)。

2. 网络

无论是与结构基因组学对应的“基因时代”，还是与功能基因组学、蛋白质组学相应的“后基因时代”；无论称 21 世纪是“信息时代”、“后信息时代”，还是“生物时代”，这些都说明 BT(生物技术)与 IT(信息技术)联系紧密，而网络在其中起着关键性的纽带桥梁作用。人类基因组计划及生物信息研究过程中产生的海量数据，需要大容量、高性能的超级计算机的支持。从序列拼接到基因预测，从蛋白结构预测到功能基因的分析，都离不开高性能服务器和网络的支持。因此，网络不但在日常生活中起着越来越重要的作用，而且从事生命科学研究的专业人士，了解熟悉网络，从中获取信息并能进一步加以研究，也已经成为其必备的基本技能。

1.1.3 生物信息学的主要研究范畴

生物信息学的研究领域主要涉及核酸序列和结构比对、蛋白质结构预测、基因获取与识别、分子进化、比较基因组学、生物芯片、基因表达的分析以及药物的设计等诸多领域。基因组信息学、蛋白质空间结构模拟、基因表达谱(gene expression profile)以及药物设计构成了生物信息学的重要组成部分和研究方向。在实际应用中生物信息学研究的具体内容应包括这几个主要部分：①新算法和统计学方法研究；②各类数据的分析和解释；③研制有效利用和管理数据的新工具；④运用生物信息进行生物医药研究和系统模拟。

1. 基因组相关信息的收集、储存、管理与提供

随着大量重要的生物学数据库及相关服务器的出现，有关基因组相关数据库的发展也相应受到研究者的广泛关注：建立基因组信息的评估与检测系统，数据标准化，将基因组信息以图表的方式展示，专家系统研究，次级及专业数据库的发展，以因特网为基础的基因组信息学传输网络。

2. 序列比对和结构比对

主要是指比较两个或两个以上符号序列的相似性或不相似性。序列比对(alignment)是生物信息学的基础内容之一，非常重要。两个序列的比对有较成熟的动态规划算法，在此基础上编写的比对软件包——BLAST 和 FASTA，可以免费下载使用，它们在数据库查询和搜索中有重要的应用。两个序列总体有时并不相似，但某些局部片断相似性很高。Smith-Waterman 算法是解决局部比对的好算法，缺点是速度较慢。两个以上序列的多重序列比对目前还缺乏快速而又十分有效的算法，目前基本问题是比较两个或两个以上蛋白质分子空间结构的相似性或不相似性。随着比较基因组学的发展，一个新的软件 Blat 已逐步应用于两个或多个生物体间整个基因组的比较。

3. 新基因的结构功能研究与鉴定(计算机辅助基因识别)

在给定基因组序列后，正确识别基因的范围及其在基因组序列中的精确位置成为重要的

课题之一。经过 20 余年的努力，已出现了数十种算法，有 10 种左右重要的算法和相应软件在网上提供免费服务。计算机辅助基因识别对原核生物的基因相对容易，而从具有较多内含子的真核生物基因组序列中正确识别起始密码子、剪切位点和终止密码子，是个相当困难的问题，研究现状不能令人满意，因此仍有大量的工作要做。

目前进行新基因的发现及功能研究，均利用 EST(expressed sequence tags，表达序列标签)的鉴定来进行，也就是通常所说的电子克隆。EST 是表达的 mRNA 序列，能与 EST 相配的 DNA 序列即是可以表达成 mRNA 的基因。常用的原理与方法包括：根据编码区具有的独特序列特征、根据编码区与非编码区在碱基组成上的差异、根据高维分布的统计方法(high dimensional distribution statistic)、根据神经网络方法(neural network)、根据分形方法(fractal)和根据密码学方法(cryptography)等计算并分析核酸数据。

4. 基因组非编码区分析及 DNA 语言研究

基因组非编码区约占人类基因组总量的95%以上，虽然目前对于其生物学意义尚不清楚，但其中必然蕴含着重要的生物学功能信息。据推测它们的生物学功能应体现在对基因表达的时空调控上。多数的 SNP(single nucleotide polymorphism)和重复序列(repetitive DNA)都在非编码区。寻找非编码区的编码特征及表达规律将是基因组研究的又一热点课题。对于基因组非编码区尚无成熟且成体系的理论和方法，通常根据已有的实验所证实的功能性 DNA 元件的序列特征，来预测非蛋白编码区中可能具有的功能性结构，并进而预测其可能的生物学功能，最后通过实验进行验证；还有的则是通过数理理论直接探索非蛋白编码区新的未知的序列特征，从理论上直接预测其可能的信息意义，最后同样通过实验验证。

5. 生物进化和比较基因组的研究

过去人们经常采用分析某些保守基因在不同物种之间的进化差异，进行物种进化的研究，取得了巨大的成就。但是现在已经注意到，是基因组整体组织方式而不是个别基因在物种演化历史中起着重要作用。例如，人类与黑猩猩之间仅有不足 3%的基因和蛋白质不相同，然而表型上却具有巨大的差异。由于基因组是物种所有遗传信息的储藏库，从基因组整体结构组织和整体时空调节网络方面，结合相应的生理机能表征，进行基因组整体的演化研究，将是揭示物种真实演化历史的最佳途径。

近年来由于较多模式生物基因组测序任务的完成，为从整个基因组的角度来研究分子进化提供了条件。可以设想，比较两个或多个完整基因组这一工作需要新的思路和方法，随着生物信息学时代的到来，使得分子进化的研究具备了更好的时机。

6. 基因组的比较研究与基因功能表达谱的分析

随着基因组研究计划的不断深入，众多的模式生物的基因组测序得以完成，使不同物种间的基因组比较研究成为可能。这种比较研究将为更好地分析理解生命科学中的众多问题提供契机。同时，基因组的研究正从结构基因组逐渐过渡到功能基因组的研究上，基因的功能表达谱的研究将真正实现静态的人类基因组研究，向时间、空间等多维多层次的新阶段迈进。大量的新技术如 DNA 芯片、二维凝胶电泳和测序质谱技术等，为基因功能表达谱的获得提供了技术上的支持。

7. 蛋白质分子空间结构预测、模拟和分子设计

蛋白质的研究长期以来是生物大分子研究中的难点，这是因为蛋白质的空间结构与功能密切相关。研究蛋白质仅仅知道其一级结构是不够的，因为蛋白质的功能是通过其三维高级结构来执行的，而且蛋白质三维结构也不是静态的，在行使功能的过程中其结构会相应地改