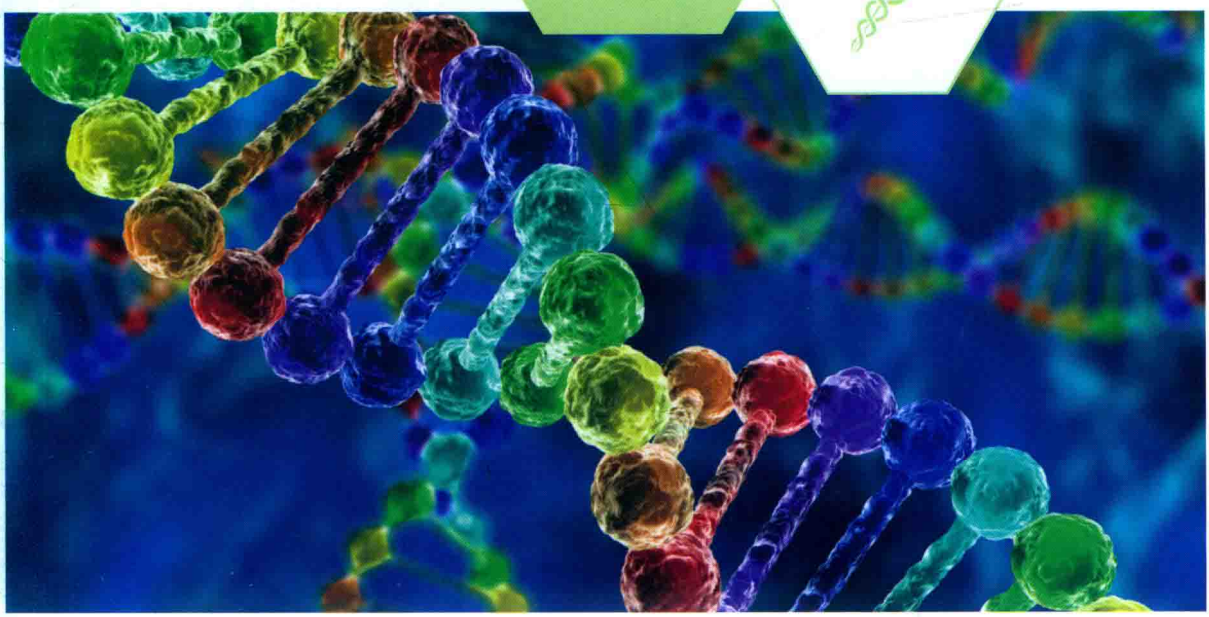
医学生物信息学
案例与实践

华琳　李林
主编

夏翊　郑卫英　安立
潘华　张骞
副主编



清华大学出版社

医学生物信息学案例与实践

华琳　李林
主编

夏翊　郑卫英　安立
潘华　张骞
副主编

清华大学出版社
北京

内 容 简 介

随着各种基因测序技术的兴起以及互联网的普及，医学科学已经进入基因组学、蛋白质组学、转化医学和精准医学的新时代，生物信息学在此背景下得到了快速发展。本书从医学和分子生物学角度出发，通过案例分析详细介绍常用的生物信息学数据库、基因芯片数据、RNA 测序数据、单核苷酸多态 SNP 数据、DNA 甲基化数据等的处理分析，还包括基因功能与通路分析技术、疾病风险通路的筛选、生物分子网络的构建等。本书还详细介绍了 R 软件及 Bioconductor 生物信息学软件包的使用，重点突出实用性和可操作性，以帮助读者对医学生物信息学方法的理解和掌握。

本书主要取材于编者近年来从事医学生物信息学的研究与教学工作内容，很多案例来自于编者近年来的科研实践。本书既可以作为基础医学、临床医学、预防医学等高年级本科生和研究生的“医学生物信息学”课程教材，也可供相关的生物科技人员阅读和参考。

图书在版编目(CIP)数据

医学生物信息学案例与实践/华琳，李林主编. —北京：清华大学出版社，2018
ISBN 978-7-302-48694-7

Ⅰ. ①医…　Ⅱ. ①华… ②李…　Ⅲ. ①医学—生物信息论　Ⅳ. ①R318.04

中国版本图书馆 CIP 数据核字(2017)第 270946 号

责任编辑：龙启铭　张爱华
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×230mm　　**印　　张**：16.75　　**字　　数**：349 千字
版　　次：2018 年 1 月第 1 版　　**印　　次**：2018 年 1 月第 1 次印刷
印　　数：1～1500
定　　价：39.00 元

产品编号：073432-01

编写人员名单

主编：

华　琳　　首都医科大学

李　林　　首都医科大学

副主编（按拼音先后顺序）：

安　立　　北京朝阳医院

潘　华　　北京天坛医院

夏　翊　　首都医科大学

张　骞　　北京大学第一医院

郑卫英　　首都医科大学

序

近年来，随着各种新兴基因测序技术的广泛开展以及互联网的普及，医学科学已经进入基因组学、蛋白质组学、转化医学和精准医学的新时代，生物信息学作为融合了现代生物学、数学、统计学和计算机科学等的前沿学科在此背景下得到了快速发展，并在许多方面影响着医学的发展。

包括人类基因组计划在内的生物基因组测序工程的里程碑式的进展、由此产生的包括生物体生老病死的生物数据以前所未有的速度递增，同时分子生物技术的快速更新和互联网的普及，极大丰富了各类型的生物医学数据，生物信息学作为处理这些数据的有效工具，将在未来的生物医学研究中发挥至关重要的作用。

本书主编华琳副教授多年从事生物医学统计与生物信息学的研究与教学工作，具有丰富的理论知识和实践经验，以第一作者和通讯作者在生物信息学领域发表 SCI 论文二十余篇。这本书取材于她近年来从事医学生物信息学的研究与教学工作内容，从医学和分子生物学角度出发，以案例的形式介绍了常用的生物信息学数据库、基因芯片、RNA 测序、单核苷酸多态 SNP、DNA 甲基化、表观遗传学等数据的分析方法，并详细讲解了生物信息学软件包的使用，突出了实用性和可操作性，便于读者对生物信息学的深入理解，给读者带来切实的帮助。

北京朝阳医院副院长 邢念增

前言

当前，随着各种基因测序技术的兴起以及互联网的普及，生命科学已经进入基因组学、蛋白质组学、转化医学和精准医学的时代。现代生物学技术促进了越来越多的临床医学研究在分子水平上开展，与此同时，加工并处理各种生物学分支产生的大量信息成为生命科学的主要研究工作。生物信息学这门崭新的学科由此应运而生并快速发展，通过与计算机技术、网络技术、统计理论和计算算法等相结合，为生物医学研究提供了关键的技术支持。各种临床实验数据和分子生物学数据都可以通过生物信息学手段进一步加工和处理。

为了使生物医学研究人员快速了解生物信息学的基本概念，初步掌握应用医学生物信息学方法对数据进行加工、处理、分析和结果阐释，本书从医学和分子生物学角度出发，通过案例分析详细介绍了医学生物信息学数据的处理方法。全书分为基础篇和提高篇。基础篇包括常用的生物信息学数据库、基因芯片数据、RNA 测序数据、单核苷酸多态 SNP 数据、DNA 甲基化数据等的处理分析，还包括基因功能与通路分析技术、疾病风险通路的筛选、生物分子网络的构建等。在提高篇中，以案例的形式翔实地介绍了生物信息学的综合分析方法和数据分析结果的可视化，具有较强的实用性和可操作性，便于读者研究和学习。此外，本书图文并茂，通俗易懂，避免使用大量烦琐的公式。书中很多案例来自编者近年来的科研实践。本书可以作为基础医学、临床医学、预防医学等高年级本科生和研究生的“医学或生物信息学”课程教材，也可供相关的生物科技人员阅读和参考。

在这里感谢首都医科大学生物医学工程学院领导的大力支持，感谢曾经与我们分享研究过程的研究人员，感谢参与讨论的研究生，还要感谢北京市自然科学基金项目（No. 7142015）的支持。

由于生物技术发展日新月异，生物信息学也发展很快，加之编者水平和所涉猎范围有限，书中不足和缺陷在所难免，希望得到专家、同行和读者的批评指正，以使本书不断完善。

编　者

2018 年 1 月

目 录

基 础 篇

提 高 篇

基　础　篇

第1章

生物信息学绪论

1.1 生物信息学概述

近年来，随着生物科学技术的迅猛发展，数据资源急剧膨胀，大量多样化的生物学数据资料产生，迫使人们寻求一种强有力的工具去组织这些数据，以利于存储、加工、分析和进一步使用，从而发现其中所蕴含的重要生物学规律。生命科学正在经历从实验分析和数据积累到数据分析及其指导下的实验验证的转变。近年来，以数据分析为本质的计算机科学技术和互联网技术正突飞猛进地发展，已经日益渗透到生物医学科学的方方面面。在数学、生物学、统计学、计算机科学等学科的有力支持下，一门崭新的、拥有巨大发展潜力的新学科——生物信息学(Bioinformatics)——悄然兴起。生物信息学是以数理科学为支柱，将数理科学、计算机与信息科学技术运用到生命科学尤其是分子生物学研究中的重大交叉学科前沿研究领域，它包含了生物信息的获取、处理、存储、分析和解释等所有方面，通过综合运用数学、统计学、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。

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综合体，充满着挑战和机遇，有广阔的发展前景。生物信息学不仅是一门新学科，更是一种重要的研究开发工具。只有通过生物信息学的计算处理，人们才能从众多分散的生物学观测数据中获得对生命运行机制的系统理解。只有根据生物信息学对大量数据资料进行分析，人们才能选择该领域正确的研发方向。从生物信息学研究的具体内容上看，生物信息学主要包括3部分：

(1) 生物信息学新算法和统计学方法。

(2) 生物技术产生的各类数据的分析和解释。

(3) 研制有效利用和管理数据的工具。

随着包括人类基因组计划在内的生物基因组测序工程的里程碑式的进展，包括生物体生老病死的生物数据以前所未有的速度递增，已达到每14个月翻一番的速度。同时随着互

联网的普及，大量的生物学数据库如雨后春笋般迅速出现和成长。因此，生物信息学将在未来的生物医学研究中起着至关重要的作用。

1.2 医学生物信息学的主要研究内容

生物信息学作为融合了现代生物学和计算机科学的前沿学科，自身也在许多方面影响着医学的发展。伴随着人类基因组计划（Human Genome Project，HGP）的初步完成，人类基因组逐渐被破译，组成人体约 3 万个基因的 30 亿个碱基对的秘密将被揭开，从而人类医学的发展将进入一个新的里程碑阶段。很多疾病的病因将被发现，从而可以更合理地指导药物研发，新的基因治疗技术和基因药物会越来越多地用于肿瘤和遗传性疾病的治疗中，人类的健康状况也将会大大改善。此外，21 世纪医学模式将发生革命性的变化，以细胞病理学为基础的医学模式，开始向分子医学模式转变。人类基因组计划正在建立起人类基因与生理、病理之间关系的知识视图。生物领域的新技术和新的研究方法在临床中逐步得到应用，更新了医学科学的基础。伴随着后基因组时代高通量组学技术涌现与生物信息学学科的飞速发展，出现了大量潜在的生物标记以及这些标记的模式，这些生物标记信息将会对临床研究带来巨大的应用潜力。因此，生物信息学也将为 21 世纪的医学奠定新的发展基础。医学生物信息学的主要研究内容包括以下几项。

1. 重大疾病的关键性基因鉴定

研究发现，很多疾病的发生与基因突变或基因多态性有关。目前，约有 6000 种以上的人类疾病与各种人类基因的变化相关联。很多复杂疾病（如肿瘤和一些免疫系统疾病）是基因与环境相互作用的结果。随着人类基因组计划的深入研究，当明确了人类全部基因的染色体位置、序列特征以及表达规律和产物特征后，人们就可以有效地了解疾病发展的分子机制，开发适宜的诊断和治疗手段。因此，发展有效的生物信息学算法进行疾病易感基因的定位，整合不同类型的数据研究基因表达规律及与疾病诊断治疗的关系，可以加快以个性化医疗为基础的精准医学的发展。

2. 流行病学研究应用

将流行病学的遗传和非遗传性研究与生物信息学方法结合，对疾病机理、个体对某种疾病的易感性和疾病在群体中的分布将会有更为明确的认识，从而对疾病的预防和治疗有极大的指导意义。

3. 药物设计

目前新药的研发主要有两个难点：一个是疾病相关靶点生物大分子的发现及确认；另

一个是具有生物活性的生物小分子的设计和发现。传统新药发现方式缺乏理论指导，主要依据大量的随机筛选，时间长，耗资大。生物信息学的兴起，为新药设计提供了新的理论和思路。通过开发生物信息学算法及相关的软件平台，可以指导药物作用靶点的选定和药物分子的设计，包括大分子结构功能模拟和预报、药物分子与大分子结合的模拟、关键性基因的致病机制及生物分子同源性分析等。

1.3　医学生物信息学面临的挑战

当前，精准医学和转化医学已成为 21 世纪生物医学研究的重点。应用现代遗传技术、分子影像技术、生物信息技术，结合患者生活环境和临床数据，实现精准的疾病分类及诊断，将有利于制定个性化的疾病预防和治疗方案。实施精准医学计划可以帮助提高疾病诊治水平，推动医学科技前沿发展，促进并带动大健康产业发展。尽管生物信息学在医学领域的研究中已经取得了不少的研究成果，从海量的数据中确定了数千个基因的功能，但是随着分子生物技术的日新月异，生物信息学的发展仍然面临着巨大的挑战，迫切需要新的研究手段和方法。

此外，生物信息学并不是一个足以乐观的领域，目前对于大规模数据内在的生成机制并没有完全明了，生物信息学在短期内很难有突破性的成果。但有理由相信，生物医学数据的巨大积累必将导致重大生物医学规律的发现，从而生物信息学这门学科也将更为成熟和完善，在生物医学研究中也将发挥越来越大的作用。

第2章

生物信息学数据库

2.1 生物信息学数据库简介

数据库是生物信息学工作的出发点。近年来，随着大量分子生物学实验数据的产生与积累，形成了数以千计的生物信息学数据库。当前，国际上已经建立的各类生物信息学数据库几乎覆盖了生命科学的各个领域，如核酸序列数据库、蛋白质序列数据库、蛋白质互作数据库、生物大分子结构数据库、人类疾病数据库以及基因组图谱数据库等。这些数据库由专门的机构建立和维护，负责收集、组织、管理和发布数据，并提供相关的数据查询、数据处理和数据分析工具，为生物医学研究人员服务。生物信息的基本数据资源主要来源于以下3类数据库：基因组数据库、核酸和蛋白质一级结构序列数据库以及蛋白质三维空间结构数据库。这些数据库主要来源于实验获得的原始数据，因此通常也称一级数据库。根据生命科学不同研究领域的需求，在一级数据库的基础上，针对某种特定目标对数据进行整理和分析而建立的专业数据库称为二级数据库。一级数据库数据量大，更新速度快，用户面广；二级数据库虽然容量小，但避免了过多的冗余数据，对于某些特定的研究则更为实用。以下各节将介绍几个常用的生物信息学数据库。

2.2 基因数据库

2.2.1 GenBank-NCBI 核酸序列数据库

GenBank-NCBI 核酸序列数据库（http：//www. ncbi. nlm. nih. gov/genbank/，见图2.1）包含了所有已知的核酸序列和根据DNA翻译的蛋白质序列，以及与它们相关的文献著作和生物学注释，由美国国立生物技术信息中心（National Center for Biotechnology Information，NCBI）建立和维护。该数据库中包含了已经公开的30万余种不同物种生物的