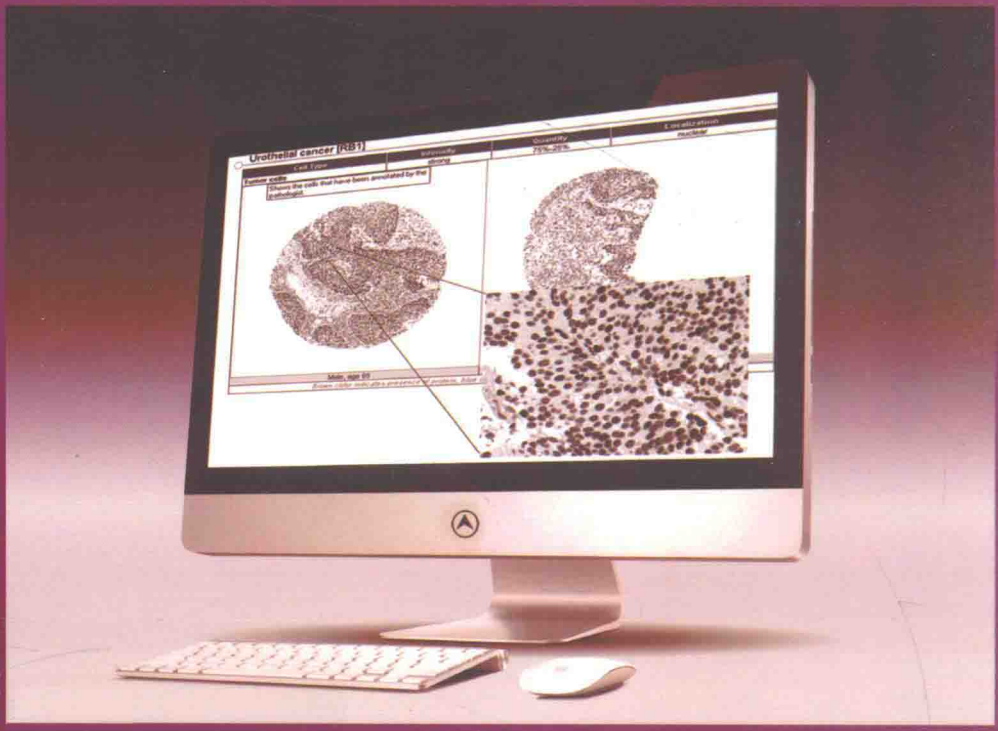
Humana Press

临床生物信息学

Clinical Bioinformatics

Ronald J.A. Trent / 著
卢学春 杨 波 张 峰 / 主译



军事医学科学出版社

临床生物信息学

Clinical Bioinformatics

Ronald J. A. Trent 著

卢学春 杨 波 张 峰 主译

军事医学科学出版社

·北京·

Ronald J. A. Trent
Clinical Bioinformatics
EISBN: 978-1-60327-148-6
Library of Congress Control Number: 2007933471

图书在版编目（CIP）数据
临床生物信息学 / 卢学春，杨波，张峰主译. —北京：军事医学科学出版社，2013.5
ISBN 978-7-5163-0235-4

Ⅰ.①临… Ⅱ.①卢… ②杨… ③张… Ⅲ.①生物信息论 Ⅳ.①Q811.4

中国版本图书馆CIP数据核字（2013）第097416号
责任编辑：孙 宇 于庆兰 吕连婷
出 版 人：孙 宇
出 版：军事医学科学出版社
地 址：北京市海淀区太平路27号
邮 编：100850
联系电话：发行部：（010）66931049
编辑部：（010）66931127，66931039，66931038
传 真：（010）63801284
网 址：http://www.mmsp.cn
印 装：北京宏伟双华印刷有限公司
发 行：新华书店
开 本：787mm × 1092mm 1/16
印 张：24.5
字 数：380千字
版 次：2014年1月第1版
印 次：2014年1月第1次
定 价：78.00元

《临床生物信息学》编委会

主　译：

卢学春（中国人民解放军总医院老年血液科 主任医师、副教授、副主任）

杨　波（中国人民解放军总医院老年血液科 副主任医师、讲师）

张　峰（中国科学院北京基因组研究所 助理研究员）

主　审：

韩　进（中国人民解放军总医院 主任医师、教授、副院长）

副主译：

迟小华（中国人民解放军第二炮兵总医院药剂科 副主任药师）

于睿莉（中国人民解放军总医院耳鼻咽喉头颈外科研究所 主治医师）

脱　帅（中国人民解放军第202医院医务处 主治医师）

责任编辑：

罗奇斌（中国科学院北京基因组研究所 助理研究员）

杨蕊菁（中国科学院北京基因组研究所 助理研究员）

译　者：（以姓氏汉语拼音排序）

蔡力力（中国人民解放军总医院老年临检科 副主任技师）

迟小华（中国人民解放军第二炮兵总医院药剂科 副主任药师）

董红宇（首都医科大学石景山医院风湿免疫科 主任医师、主任）

荆玉红（中国医学科学院协和医科大学药物研究所 助理研究员）

李惠子（中国人民解放军第二炮兵总医院营养科 主治医师）

李　新（中国科学院北京基因组研究所 助理研究员）

李蕴博（吉林大学第一医院神经外科 主治医师）

卢学春（中国人民解放军总医院老年血液科 主任医师、副教授、副主任）
罗奇斌（中国科学院北京基因组研究所 助理研究员）
脱朝伟（中国人民解放军第202医院电镜室 主任医师、教授、主任）
脱　帅（中国人民解放军第202医院医务处 主治医师）
王　飞（中国人民解放军第208医院创伤外科 主治医师）
王　杨（吉林大学中日联谊医院心内科 主治医师）
杨　波（中国人民解放军总医院老年血液科 副主任医师、讲师）
杨蕊菁（中国科学院北京基因组研究所 助理研究员）
于金海（吉林大学第一医院普通外科 主治医师）
于睿莉（中国人民解放军总医院耳鼻咽喉头颈外科研究所 主治医师）
张　帆（中国人民解放军总医院泌尿外科 主治医师）
张　峰（中国科学院北京基因组研究所 助理研究员）
张丽丽（中国科学院北京基因组研究所 助理研究员）
张永彪（中国科学院北京基因组研究所 助理研究员）

Contributors

JONATHAN W. ARTHUR • *Discipline of Medicine, Central Clinical School, University of Sydney and Sydney Bioinformatics, New South Wales, Australia*

JENNIFER H. BARRETT • *Section of Genetic Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, St James's University Hospital, Leeds, United Kingdom*

MICHAEL A. BLACK • *Department of Biochemistry, University of Otago, Dunedin, New Zealand*

PAUL C. BOUTROS • *Department of Medical Biophysics, University of Toronto, Ontario, Canada*

ALLEN K. L. CHEUNG • *Centre for Virus Research, Westmead Millennium Institute, Westmead, New South Wales, Australia*

ENRICO COIERA • *Centre for Health Informatics, University of New South Wales, New South Wales, Australia*

STUART J. CORDWELL • *School of Molecular and Microbial Biosciences, University of Sydney, New South Wales, Australia*

BEN CROSSETT • *School of Molecular and Microbial Biosciences, University of Sydney, New South Wales, Australia*

ANDREW DUBOWSKY • *Genetic Pathology, Flinders Medical Centre, Bedford Park South Adelaide, Australia*

ALISTAIR V. G. EDWARDS • *Discipline of Pathology, School of Medical Sciences, University of Sydney, New South Wales, Australia*

PIOTR G. FAJER • *Institute of Molecular Biophysics, Department of Biological Sciences, Florida State University, Tallahassee, Florida, USA*

DAVID C. Y. FUNG • *School of Information Technologies, Faculty of Science, University of Sydney, New South Wales, Australia*

SCOTT A. GRIST • *Genetic Pathology, Flinders Medical Centre, Bedford Park South Adelaide, Australia*

CHRISTOPHER A. HAIMAN • *Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA*

BRETT D. HAMBLY • *Pathology Discipline, Bosch Institute, School of Medical Science, University of Sydney, New South Wales, Australia*

MARCUS HINCHCLIFFE • *Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

NAN HU • *Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, NCI, Bethesda, Maryland, USA*

ANTHONY M. JOSHUA • *Department of Medical Oncology, Princess Margaret Hospital, Toronto, Canada*

HUONG LE • *Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

MAXWELL P. LEE • *Laboratory of Population Genetics, Center for Cancer Research, NCI, Bethesda, Maryland, USA*

SIAW-TENG LIAW • *School of Rural Health, University of Melbourne, Shepparton, Victoria, Australia*

DONALD R. LOVE • *School of Biological Sciences, University of Auckland, Auckland, New Zealand*

ALEXANDRE MENDES • *Newcastle Bioinformatics Initiative, University of Newcastle, New South Wales, Australia*

PABLO MOSCATO • *Newcastle Bioinformatics Initiative, University of Newcastle, New South Wales, Australia*

CECILY E. OAKLEY • *Institute of Molecular Biophysics, Department of Biological Sciences, Florida State University, Tallahassee, Florida, USA*

CHI N. I. PANG • *Biomolecular Sciences, Faculty of Science, University of New South Wales, Australia*

ANASSUYA RAMACHANDRAN • *Department of Obstetrics and Gynaecology, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand*

PETER SCHATTNER • *Department of General Practice, Monash University, Clayton, Victoria, Australia*

RODNEY J. SCOTT • *Discipline of Medical Genetics, University of Newcastle, New South Wales, Australia*

ANDREW N. SHELLING • *Department of Obstetrics and Gynaecology, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand*

VITALI SINTCHENKO • *Centre for Infectious Diseases and Microbiology-Public Health, Western Clinical School, The University of Sydney, New South Wales, Australia*

BARRY SLOBEDMAN • *Centre for Virus Research, Westmead Millennium Institute, Westmead, New South Wales, Australia*

DANIEL O. STRAM • *Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA*

GRAEME SUTHERS • *Familial Cancer Unit, Children's Youth and Women's Health Service, North Adelaide, Australia*

PHILIP R. TAYLOR • *Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, NCI, Bethesda, Maryland, USA*

RONALD J. A. TRENT • *Department of Molecular and Clinical Genetics, University of Sydney Central Clinical School, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

CARL VIRTANEN • *University Health Network, Microarray Centre, Toronto Medical Discovery Tower, Toronto, Ontario, Canada*

MELANIE Y. WHITE • *Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA*

MARC R. WILKINS • *Biomolecular Sciences, Faculty of Science, University of New South Wales, Australia*

JAMES WOODGETT • *Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada*

HOWARD H. YANG • *Laboratory of Population Genetics, Center for Cancer Research, NCI, Bethesda, Maryland, USA*

JEAN YEE HWA YANG • *School of Mathematics and Statistics, University of Sydney, New South Wales, Australia*

BING YU • *University of Sydney Central Clinical School, Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

序 言

生物信息学是20世纪80年代诞生和兴起的一门新的应用生物学学科。生物信息学不仅在过去的三十年里有了长足的发展，而且在基础生物学、医学、农学和法医学等领域都有了深入和广泛的应用。这个学科产生的原因很多，但一言以蔽之，就是科学发展对新领域、新方法和新技术的需求。科学前沿会不断拉动学科交叉、融合和细分，导致细分点上的纵深发展，生物信息学就是在这样的细分点上。

因此，生物信息学首先是学科交叉和融合的产物。一边是计算机科学与技术的高速发展和基础数学、统计学和工程学的坚实基础；另一边是基因组学所引领的高通量信息获取技术（比如各种“组学”技术）的迅猛发展和相关信息在生物各学科领域的高速度积累。两者的结合，产生了不可低估的力量，从而造就了一批从各个学科汇聚而来的科学家，推出了新的学科生长点。其次，生物信息学是一门以技术和方法为基础的学科，与分子生物学很相似，自身的发展依赖技术和方法上的突破，需要大科学项目作为催化剂。自“人类基因组计划”以来，大的科学计划和全球性的合作项目层出不穷，生物信息学便如鱼得水，技术和方法也因此而得到拓展、改革与创新，使用者和开发者通过紧密互动，来适应和满足科学前沿的发展。第三，生物信息学是一门应用生物学学科，随着需求的诞生而诞生，随着需求的消失而消失。被淘汰的技术甚至数据都会退出市场和使用终端。可见，这个领域的开发者和获利者都必须学会借鉴和舍弃，

善于发现新的开发点和应用点。同样，一个新技术在进入市场的早期也需要大家共同来开发新的方法和应用。第四，生物信息学的学科基础是数据。数据的管理需要构建各种各样的数据库，数据的分析需要流程和算法，数据还要不断地收集、整理和挖掘。数据的基本特征是积累。基本数据是不能随便丢掉的（被淘汰的技术产生的数据除外），因此，数据会越积越多，给数据的管理和挖掘带来很多压力。好在数据存储的价格不高，这得感谢计算机技术的不断发展。第五，数据并不是信息。要生物信息学家来精心“耕耘”才会使数据变成信息。而最终完成从信息到知识的转化则往往需要生物学来合作，生物学家对生物学的概念和命题要把握得更准些。信息库最终要成为知识库，内容更多、更实用。

Ronald J. A. Trent 主编的这本《临床生物信息学》收集了生物信息学在临床领域的主要应用，并为一些基本方法与技术的使用提供了入门性的实验指导。本书的各个章节都是由澳大利亚、美国和加拿大的生物信息学家分别撰写，内容简要、实用，是一本很好的启蒙手册。因为生物信息学发展得很快，我们不能只依赖这本书来全面了解临床生物信息学，尤其是这个领域的最新进展。但是本书的内容并没有过时，仍然代表了这个领域发展早期的基本研究方法和内容。希望译者们的心血和努力化作中国临床生物信息学发展的第一块基石。

于　军

中国科学院北京基因组研究所所长

译者序

如何利用已有的开放性数据库进行临床研究是本书的核心内容。

随着基础研究的进步，基因组学、表观基因组学和蛋白质组学的海量数据不断增多，全球范围内每天产生海量的DNA、RNA、蛋白质等生物信息数据。临床医生如何利用这些基础研究的数据，真正解决临床实践过程中所面对的诊断、治疗和预后评估的问题，并直接利用这些数据为患者服务？这是每一个基础研究和临床工作者所面临的难题。根据不同研究目的统计、分析这些数据，成为限制这些数据快速转化为临床信息的瓶颈。在当今转化医学逐渐引起学者们关注的大背景下，以开发、利用上述这些海量生物信息数据为临床服务为目的的一门新学科应运而生，即临床生物信息学。为此，国外在这方面研究得较早，相对成熟，国内才刚起步，多数学者对这门学问还很不了解。《临床生物信息学》一书的出版为临床医生和研究人员全面了解临床生物信息学这门科学提供了一个很好的契机，是临床医生开展临床研究的一个很好的切入点和敲门砖。

本书系统介绍了临床生物信息学所能解决的问题和最常应用的临床生物信息学数据库。尤其难能可贵的是，还有一部分针对具体肿瘤患者的具体基因检查结果如何进行相应临床治疗、预后评估和随访的案例。各章节作者均为生物信息领域的权威专家。读者将会发现本书为研究人员、临床学家和卫生专业人士解读不断增长的海量而复杂的医学信息提供最佳的指导。

我们将此书编译付梓，旨在为国内的生物医学研究人员及临床医生提供有价值的参考信息，推动生物信息学在临床上的转化和应用，希望最终能够帮助临床医生主动利用临床生物信息学方法去解决患者的诊断、治疗和预后判断中所遇到的种种问题。

由于译者水平有限，疏漏之处在所难免，恳请广大读者批评指正。

卢学春　杨波　张峰

中国人民解放军总医院老年血液科

中国科学院北京基因组研究所

北京 2012 年 11 月 10 日

前 言

临床医学所面临的一个挑战就是如何处理不断增长的生物信息数据，这在现今新兴的“组学”时代更是如此，并且对研究人员、卫生专业人士和广泛学术团体将会产生深刻影响。为了迎接和应对这项挑战，就需要以计算机为基础进行数据的贮存、处理和传播（即生物信息学）。在《分子医学研究方法》™这套丛书中，论述了许多如何利用临床生物信息学的策略。本书重点论述如何应用软件将各种生物信息转化成具有临床意义的成果，包括六个主题：①基因发现——第 1 ~ 4 章；②基因功能（芯片）——第 5 ~ 9 章；③ DNA 突变分析——第 10 ~ 12 章；④蛋白质组学——第 13 ~ 15 章；⑤互联网在线分析方法和资源——第 16、17 章；⑥信息学在临床实践中的应用——第 18、19 章。

最后，我还要感谢 Carol Yeung 为本书的编撰所提供的帮助。

Ronald J. A. Trent

悉尼，2006 年 12 月

目 录

Contents

第一章

致病基因发现的生物信息分析

Bing Yu

摘要

复杂疾病涉及多个基因及环境因素的相互影响。直接发现这些基因并非易事，基于生物信息的分析策略能显著提高发现这些基因的可能性。数据挖掘和自动化分析对基因定位起到了很大的推动作用。在基因发现阶段，利用生物信息技术优化关联和连锁分析来查找候选基因是一种快捷方法。应用生物信息也能将编码区突变分开，并可以主要分析预测非编码区突变（特别是启动子区域）的功能。

关键词：基因发现，复杂疾病，数据挖掘，预测，数据存取，生物信息，单体型推算，模拟。

缩写：cM——厘摩，EST——表达序列标签，LD——连锁不平衡，OMIM——人类孟德尔遗传在线数据库，SNP——单核苷酸多态性。

1 简介

人类基因组计划的完成是医学科学史上的一个里程碑。获得了约30亿个碱基对的序列，但是这对于破译整个人类基因组来说仅仅是一个开始而已。人类基因组计划的最终目的是要解密人类健康和疾病的生物学及内在的生理学机制理。现今，常见公共卫生问题，包括心血管疾病，中风，癌症，糖尿病，肥胖及心理健康疾病已经成为社会群体的主要健康隐患[1-3]。不同于单基因疾病和符合孟德尔遗传的简单疾病，我们研究的对象是复杂疾病。它们在病理生理学上包含了多个基因和环境风险因素，同时基因和基因之间，基因和环境之间的相互影响也很复杂[4,5]。但与疾病相关基因的发现还是大大促进了我们对疾病的病因学、发病机理的认识，并推进了常见疾病在诊断和治疗方法上的进步。

致病基因的定位开始于20世纪70年代后期，是用基因产物和基因功能来确定基因与疾病相关性的一种方法[4,5]。运用此方法成功地发现了例如地中海贫血症中编码 α 珠蛋白的基因。但是在缺少目标基因或者家系缺失的情况下去确定致病的基因蛋白却不是总能行得通的。因此，从二十世纪80年代的后期开始，克隆定位的方法成为了致病基因发现的重要手段[5]。这种方法绕过了蛋白质的部分，只针对染色体特定位置上的基因进行克隆。在过去近20年里，这种方法已经在简单的单基因疾病，如囊性纤维化、亨廷顿症及其他一些罕见疾病的基因定位中取得了很高的成功率。大多数致病基因的发现都基于家系连锁分析的支持。但家系连锁分析受外显率高低和多基因效应的影响。这两点在常见复杂疾病的预测上来说却是非常重要的特征指标。因此，关联分析作为一种替代策略，在复杂疾病的基因发现方面渐受追捧。关联分析将复杂疾病相关的基因突变进行检测，可能发现：相对于正常人群，疾病等位基因在病例组中或多或少存在普遍性。关联性分析比较连锁分析来说，在捕捉多基因效应上更有效一些（见第二章中关于孟德尔及复杂疾病的讨论）。

生物信息分析技术作为一种有效的补充方法，大大地提高了基因发现的可能性。它将生物学、计算机科学、数学知识结合在一起，为基因定位、基因注释和确定基因变异开拓了新思路（图1）[5]。“生物信息”不仅仅是对