

生物医学类基础参考书

生物信息学 理论与医学实践

BIOINFORMATICS THEORY AND MEDICAL PRACTICE

主 编 李 霞 副主编 张 岩 陈秀杰 陈丽娜



 人民卫生出版社

BIOINFORMATICS
THEORY AND MEDICAL PRACTICE

生物信息学 理论与医学实践

主 编 李 霞

副 主 编 张 岩 陈秀杰 陈丽娜

编 委 (以姓氏笔画排序)

王艳秋 刘 磊 李 晋 李 霞

李春权 肖 云 汪强虎 张 岩

张绍军 陈 曦 陈丽娜 陈秀杰

陈晓文 周 猛 姜 伟 宫滨生

徐良德 徐建凯 智 慧

学术秘书 王 宏



人民卫生出版社

图书在版编目 (CIP) 数据

生物信息学理论与医学实践/李霞主编.—北京:
人民卫生出版社,2013.1

ISBN 978-7-117-16277-7

I. ①生… II. ①李… III. ①生物信息论 IV.
①Q811.4

中国版本图书馆CIP数据核字 (2012) 第283419号

人卫社官网	www.pmph.com	出版物查询, 在线购书
人卫医学网	www.ipmph.com	医学考试辅导, 医学数 据库服务, 医学教育资 源, 大众健康资讯

版权所有, 侵权必究!

生物信息学理论与医学实践

主 编: 李 霞

出版发行: 人民卫生出版社 (中继线 010-59780011)

地 址: 北京市朝阳区潘家园南里 19 号

邮 编: 100021

E - mail: pmph @ pmph.com

购书热线: 010-67605754 010-65264830

010-59787586 010-59787592

印 刷: 三河市宏达印刷有限公司

经 销: 新华书店

开 本: 787×1092 1/16 印张: 35

字 数: 852千字

版 次: 2013年1月第1版 2013年1月第1版第1次印刷

标准书号: ISBN 978-7-117-16277-7/R · 16278

定 价: 158.00 元

打击盗版举报电话: 010-59787491 E-mail: WQ @ pmph.com

(凡属印装质量问题请与本社销售中心联系退换)

序 言

20世纪90年代初,人类基因组计划(human genome project, HGP)的启动推动了生物学、医学、药学与信息科学之间的紧密联系,海量数据的收集、存储、分析及解释促使全世界科学家思考生物学、医学和药学发展的新思路,生物信息学就是在这样的背景下产生并蓬勃发展起来的。生物信息学(bioinformatics)是以数理科学为理论基础、以计算机技术为工具,进行深层次生物学海量数据挖掘与分析的多学科交叉的新兴学科。伴随着现代生物技术的发展,生物信息学在现代生物学、医学和药学的发展中发挥着重要作用。

随着新一代测序技术、生物芯片技术、药物筛选技术等快速发展,现代生物学、医学、药学研究已经由单一因素、单个分子层面进入到高通量、大规模的组学研究时代。面对信息含量大、数理逻辑强的生物学、医学、药学资源,传统的实验方法遇到巨大挑战,已经难以独立解决众多复杂的生物学、医学、药学问题。在此基础上,以海量数据分析为研究内容的生物信息学逐渐成为生物学、医学、药学研究领域不可或缺的组成部分。生物信息学理论能够广泛地应用于生物学、医学、药学等研究领域,如人类疾病病因学研究、临床诊断标志物识别、疾病分型和预后预测、遗传调控机制和分子通路建立、药物靶标识别与药物设计、新兴生物大分子发现与功能分析、生理模拟与病理推断、动植物育种与分子进化等方面,并能够极大地促进个性化医疗的发展。由此,我们总结多年积累的学术思想、研究心得及结果,编著了《生物信息学理论与医学实践》一书。本书旨在为生物学、医学、药学研究领域的科学工作者及生物信息学领域的同行、学生等人员介绍生物信息学基础理论、数据分析方法及其在生物学、医学、药学领域中的应用成果。

本书不仅对生物信息学研究领域的基础知识及基本理论进行了详细介绍,

如网络数据资源、序列比对、分子进化、基因芯片及蛋白质结构等;而且介绍了目前国内外生命科学研究应用的热门技术及热点领域,如新一代测序技术、富集分析技术、表观遗传学分析技术、microRNA与疾病研究及药物靶点筛查技术,并对书中涉及的各种分析技术给出详细的应用实例。我们希望能与感兴趣的读者交流,有机会完善本书。限于作者目前的水平,加之时间仓促,书中必有许多不足之处,希望能够得到读者的指正。

本书每一章的编者都有相关领域多年并丰富的研究经历,每一章都凝聚了他们的学术思想及科研成果。他们在百忙之中精心组织材料、字斟句酌编写本书,在此我们对全体编者的无私奉献表示衷心的感谢!多年来,我们的工作得到了哈尔滨医科大学各方面的大力支持与热情鼓励,同时也得到了国家自然科学基金的支持,谨在此一并表示诚挚的谢意!

李 霞
2012年12月

目 录

绪 论

INTRODUCTION TO BIOINFORMATICS MOLECULAR

第一节 生物信息学的产生及发展 / 02

- 一、生物信息学的产生 / 02
- 二、生物信息学的发展历史 / 03
- 三、生物信息学在未来生命科学研究中的作用 / 06

第二节 生物信息学的主要研究内容 / 07

- 一、基因组信息学 / 07
- 二、转录组信息学 / 09
- 三、蛋白质组信息学 / 11
- 四、代谢组信息学 / 12

第三节 当前生物信息学研究的热点 / 14

- 一、新一代测序数据的生物信息学分析 / 14
- 二、非编码区序列分析与功能识别 / 15
- 三、整合信息组学 / 15

四、转化医学和临床生物信息学 / 16

五、生物信息学与新药研究 / 16

第一章 序列比对与序列特征分析

CHAPTER 1 SEQUENCE ALIGNMENT AND ANALYSIS OF SEQUENCE CHARACTERISTICS

第一节 获取DNA、RNA和蛋白质序列 / 20

一、DNA序列的获取 / 20

二、RNA序列的获取 / 28

三、蛋白质序列的获取 / 29

第二节 双序列比对 / 32

一、序列比对的相关概念 / 32

二、双序列比对算法 / 37

第三节 多序列比对 / 43

一、多序列比对简介 / 43

二、多序列比对的方法 / 44

三、多序列比对常用工具和数据库 / 48

第四节 核酸序列特征分析 / 56

一、基因开放读码框的识别 / 56

二、内含子/外显子剪切位点的识别 / 58

三、序列模体的查找和可视化工具 / 58

四、密码子使用模式的分析 / 59

五、限制性核酸内切酶位点分析 / 60

六、重复序列的查找 / 61

第五节 蛋白质序列特征分析 / 62

一、蛋白质的理化性质分析 / 62

二、蛋白质的跨膜结构分析 / 64

三、蛋白质信号肽的预测和识别 / 67

四、蛋白质的卷曲螺旋预测 / 67

五、糖基化位点的预测与识别 / 68

六、磷酸化位点的预测与识别 / 68

第六节 应用实例 / 69

一、利用Spidey工具识别mRNA/cDNA的外显子组成 / 69

二、利用Spidey工具进行可变性剪切的分析 / 71

第二章 基因芯片数据分析

CHAPTER 2 MICROARRAY DATA ANALYSIS

第一节 基因芯片平台简介 / 76

一、cDNA芯片 / 76

二、寡核苷酸芯片 / 77

三、原位合成芯片 / 77

四、光纤微珠芯片 / 79

第二节 基因芯片数据的预处理 / 81

一、基因芯片数据的提取 / 81

二、数据过滤 / 81

- 三、补缺失值 / 82
- 四、数据对数化处理 / 82
- 五、数据标准化 / 83
- 六、应用举例 / 85

第三节 特征基因挖掘 / 89

- 一、特征选择的过程 / 89
- 二、特征选择方法的分类 / 92
- 三、应用举例 / 95

第四节 基因芯片数据的聚类分析 / 98

- 一、聚类分析中的距离(相似性)尺度函数 / 98
- 二、常用的聚类方法 / 100
- 三、应用实例 / 104

第五节 基因芯片数据的分类分析 / 108

- 一、 k 近邻分类法 / 108
- 二、决策树 / 108
- 三、支持向量机 / 110
- 四、分类器的分类效能评价 / 111

第六节 基因芯片数据库及常用分析软件 / 113

- 一、基因表达数据库 (gene expression omnibus, GEO) / 113
- 二、基因芯片显著性分析
(significance analysis of microarray, SAM) / 114
- 三、Cluster and TreeView / 114
- 四、BRB-ArrayTools / 115
- 五、R语言和Bioconductor / 115
- 六、Matlab: Bioinformatics Toolbox / 115

第三章 基因注释与功能分类

CHAPTER 3 GENE ANNOTATION AND FUNCTIONAL CLASSIFICATION

第一节 引言 / 118

第二节 基因注释数据库 / 119

- 一、GO(gene ontology) 数据库 / 119
- 二、KEGG数据库 / 125
- 三、其他常见生物学通路数据库 / 130

第三节 基因功能预测 / 132

- 一、基因功能预测的目的和意义 / 132
- 二、基因功能预测的基本原理 / 133
- 三、基因功能预测的常用工具 / 136

第四节 基因集合富集分析 / 139

- 一、富集分析的目的和意义 / 139
- 二、富集分析的基本原理 / 139
- 三、富集分析常用工具 / 140
- 四、富集分析应用实例 / 142

第五节 基因功能比较 / 145

- 一、基因功能比较的意义 / 145
- 二、语义相似性原理与算法 / 146
- 三、生物学术语相似性 / 146
- 四、基因(基因产物) 功能比较 / 148
- 五、基因集合功能比较 / 149

第四章 蛋白质结构分析

CHAPTER 4 PROTEIN STRUCTURE ANALYSIS

第一节 蛋白质高级结构 / 154

- 一、蛋白质的高级结构特征 / 154
- 二、蛋白质结构域与家族分类 / 157
- 三、蛋白质结构可视化软件 / 158

第二节 蛋白质的结构解析与预测 / 162

- 一、蛋白质结构实验检测技术与结构解析 / 162
- 二、蛋白质结构数据库 / 164

第三节 蛋白质高级结构预测方法 / 175

- 一、蛋白质二级结构预测方法及软件 / 175
- 二、蛋白质三级结构的预测方法及软件 / 180
- 三、对结构预测结果的评价 / 186

第四节 基于蛋白质结构的功能分析 / 187

- 一、蛋白质结构与功能基础 / 187
- 二、蛋白质结构与功能关系数据库 / 188
- 三、从蛋白质结构推断其功能的方法与分析软件 / 193
- 四、蛋白质相互作用与蛋白质功能 / 197

第五节 蛋白质的结构异常与疾病 / 202

- 一、蛋白质序列变化引发疾病 / 202

- 二、蛋白质折叠错误引发疾病 / 202
- 三、疾病过程中蛋白质的相互作用 / 203

第六节 应用实例 / 205

- 一、蛋白质结构信息在类风湿性关节炎致病基因挖掘中的应用 / 205
- 二、蛋白质结构转换几率与疾病的发生 / 208

第五章 分子进化分析

CHAPTER 5 MOLECULAR EVOLUTION ANALYSIS

第一节 系统发生分析与重建 / 214

- 一、核苷酸置换模型与氨基酸置换模型 / 214
- 二、系统发育树重建方法 / 219
- 三、分子钟假说 / 224

第二节 核苷酸和蛋白质的适应性进化 / 226

- 一、中性与近中性理论 / 226
- 二、微观适应性进化的检验方法 / 227
- 三、宏观适应性进化的检验方法 / 229
- 四、适应性进化基因 / 232

第三节 分子进化与生物信息学 / 233

- 一、基因组进化概述 / 233
- 二、病毒基因组分析 / 234
- 三、原核生物基因组比较 / 234
- 四、蛋白质互作网络进化 / 236
- 五、代谢网络进化分析 / 238

第六章 新一代测序数据分析

CHAPTER 6 NEXT-GENERATION SEQUENCING DATA ANALYSIS

第一节 全基因组测序与重测序 / 246

第二节 新一代测序技术和工作流程 / 248

- 一、新一代测序法和常见的测序仪 / 249
- 二、样品准备 / 250
- 三、合成测序法 / 251
- 四、第三代测序技术 / 252

第三节 新一代测序数据存储、处理与分析 / 255

- 一、新一代测序数据格式与质量编码 / 255
- 二、新一代测序数据库与数据格式转化 / 256
- 三、测序短片段在参考基因组中的定位 / 256
- 四、短片段作图软件 / 257
- 五、基因表达水平估计 / 259
- 六、可变剪切作图软件包 / 259

第四节 DNA和RNA测序 / 261

- 一、DNA重测序与个体变异发现 / 261
- 二、细菌基因组测序与致病性位点发现 / 262
- 三、宏基因组测序与感染性疾病分析 / 262

四、古生物基因组和进化研究 / 263

五、外显子组测序 / 263

六、非编码RNA测序 / 264

七、核糖体印记与深度测序技术 / 264

第五节 研究实例：基于新一代测序技术的癌症组学研究 / 266

一、短序列数据准备 / 266

二、短序列数据格式转换 / 270

三、短序列在参考基因组上定位 / 272

四、短序列在参考基因组上定位和可变剪切的识别 / 275

五、数字基因表达谱提取 / 279

第七章 转录调控的信息学分析

CHAPTER 7 BIOINFORMATICS ANALYSIS ON TRANSCRIPTION REGULATION

第一节 基因的转录调控 / 284

一、转录 / 284

二、RNA聚合酶 / 285

三、转录调控元件 / 286

四、真核生物转录机制 / 290

第二节 启动子的信息学分析 / 291

一、启动子识别问题 / 291

二、启动子数据资源 / 291

三、真核生物启动子在线分析工具 / 296

四、启动子识别的信息学研究方法 / 297

第三节 转录因子结合位点的信息学研究 / 300

- 一、转录因子及其转录调控机制 / 300
- 二、转录因子结合位点的高通量试验技术 / 303
- 三、转录因子结合位点相关数据库:
TRANSFAC、JASPAR、SELEX_DB / 306
- 四、转录因子结合位点模型的建立及分析 / 307
- 五、利用从头预测算法识别转录因子结合位点 / 308
- 六、结合ChIP-seq等高通量实验数据的
转录因子结合位点预测方法 / 309

第四节 可变剪接的生物信息学研究 / 312

- 一、可变剪接的调控机制 / 312
- 二、可变剪接数据资源: ASD、ASTD / 315
- 三、利用基因芯片技术进行可变剪接研究 / 315
- 四、RNA-seq与可变剪接异构体 / 316

第五节 转录调控与人类疾病 / 318

- 一、顺式调控元件 / 318
- 二、反式作用因子 / 319
- 三、顺式作用元件与疾病 (心脏病、肾病、Alzheimer、cancer) / 320
- 四、反式作用因子与疾病 / 321

第六节 应用实例 / 324

- 一、结合序列和表达谱数据识别组合式
转录调控模式和调控网络 / 324
- 二、基于启动子结构元件的组合模式识别调控网络 / 324

第八章 生物分子网络与通路

CHAPTER 8 BIOLOGY MOLECULAR NETWORK AND PATHWAY

第一节 网络的基本概念 / 332

- 一、网络的定义 / 332
- 二、有向与无向网络 / 332
- 三、加权与无权网络 / 333
- 四、二分网络 / 333

第二节 生物分子网络种类 / 334

- 一、蛋白质相互作用网络 / 334
- 二、基因转录调控网络 / 336
- 三、代谢和信号传导网络 / 337
- 四、衍生网络 / 339

第三节 生物分子网络分析方法 / 342

- 一、网络的拓扑属性分析 / 342
- 二、网络的无尺度特性分析 / 345
- 三、网络的模序搜索 / 346
- 四、网络的功能模块识别 / 347
- 五、网络分析软件 / 350

第四节 通路的网络分析方法 / 353

- 一、影响分析方法 / 353
- 二、潜能通路识别分析 / 354
- 三、通路分析软件 / 356

第五节 应用实例：疾病代谢子通路识别、网络构建和分析 / 358

一、利用通路结构信息识别疾病风险子通路 / 358

二、疾病代谢网络构建和分析 / 359

第九章 复杂疾病的系统遗传学分析

CHAPTER 9 SYSTEM GENETICS ANALYSIS OF COMPLEX DISEASES

第一节 复杂疾病的分子遗传学特征 / 368

一、孟德尔遗传病与复杂疾病 / 368

二、复杂疾病的分子系统特征 / 368

三、重要的复杂疾病数据库 / 369

第二节 变异组信息学与复杂疾病遗传定位 / 375

一、变异组学与人类疾病 / 375

二、SNP与人类复杂疾病定位 / 378

三、变异组学研究资源 / 382

第三节 复杂疾病的分子系统分析 / 389

一、面向通路的基因组范围关联研究 / 389

二、表型性状的分子遗传学 / 391

三、复杂疾病的系统遗传学 / 392

第四节 系统遗传学集成软件工具 / 395

一、Plink软件包与基因互作 / 395

二、基因组范围关联研究软件包SNPtest / 397