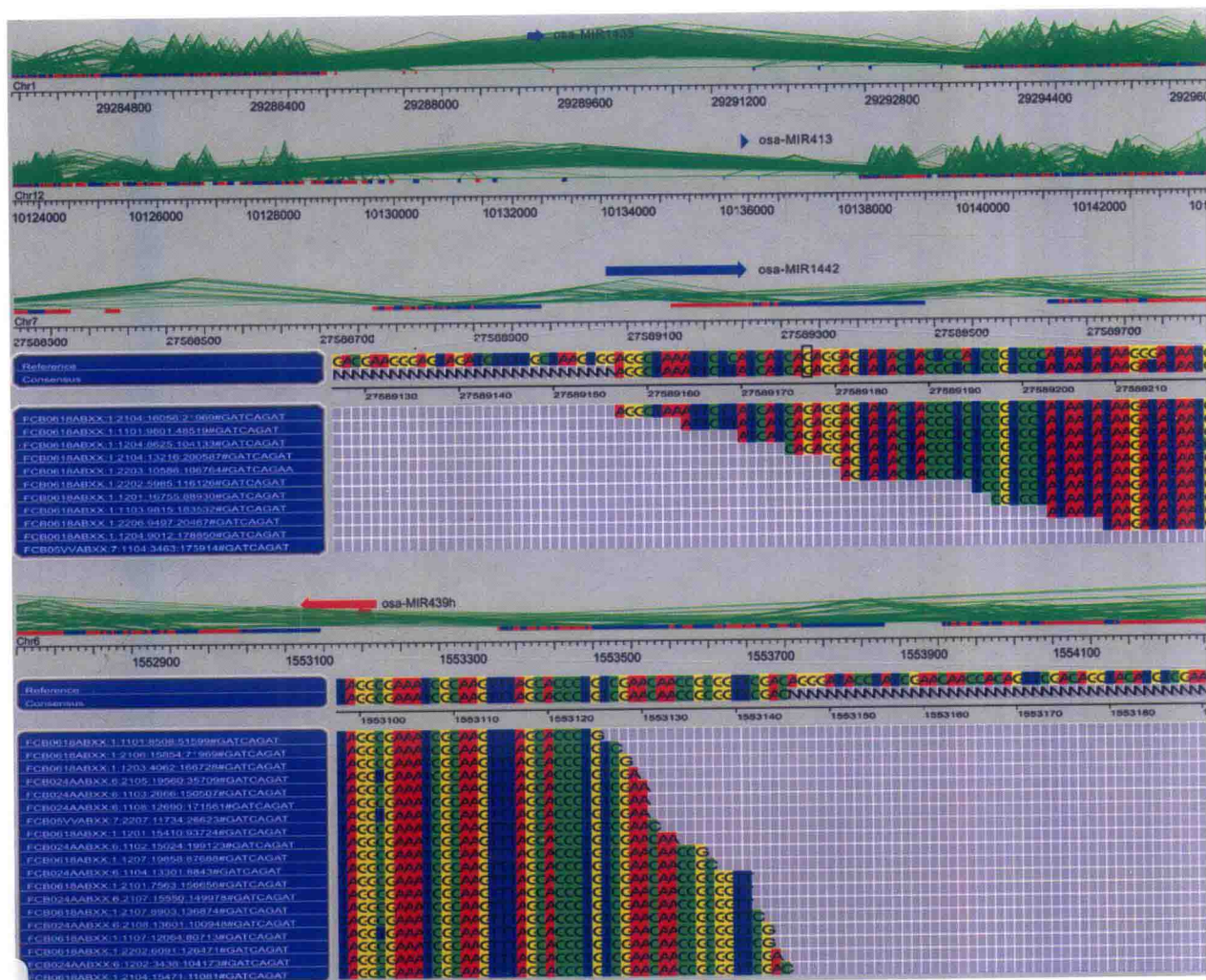




生物信息学

Bioinformatics





高等院校农学与生物技术专业规划教材

生物信息学

樊龙江 主编



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

国家一级出版社
全国百佳图书出版单位

图书在版编目 (CIP) 数据

生物信息学 / 樊龙江主编. — 杭州: 浙江大学出版社, 2017.9

ISBN 978-7-308-17147-2

I. ①生… II. ①樊… III. ①生物信息论—教材
IV. ①Q811.4

中国版本图书馆CIP数据核字(2017)第176137号

生物信息学

樊龙江 主编

策划编辑 阮海潮 (ruanhc@zju.edu.cn)

责任编辑 阮海潮

责任校对 陈静毅 王安安 吴水燕

封面设计 续设计

排 版 杭州兴邦电子印务有限公司

出版发行 浙江大学出版社

(杭州市天目山路148号 邮政编码310007)

(网址: <http://www.zjupress.com>)

印 刷 嘉兴华源印刷厂

开 本 787mm×1092mm 1/16

印 张 33.25

字 数 830千

版 印 次 2017年9月第1版 2017年9月第1次印刷

书 号 ISBN 978-7-308-17147-2

定 价 98.00元

版权所有 翻印必究 印装差错 负责调换

浙江大学出版社发行中心联系方式: (0571) 88925591; <http://zjdxcsb.tmall.com>

生物信息学

作者名单

主 编 樊龙江

参编人员 吴三玲 邱 杰 叶楚玉 徐海明

王 煜 阮松林 沈恩惠 沈一飞

褚琴洁 毛凌峰

序 言

1959年9月,我国自行研制的104真空管电子计算机通过国家鉴定。它每秒钟可以执行1万条浮点运算指令。2016年6月,在世界超级计算机500强名单中位居首位的是我国无锡超算中心的神威太湖之光计算机,其峰值运算速度达到每秒9亿亿次(93104.6 Tflops)。57年间,运算速度提高了9万亿倍。信息技术的如此发展速度是人类在所有其他科学技术领域不能比拟的,它注定要改变社会生产和生活的方方面面,生物学和医学的研究也不例外。

1953年,DNA双螺旋结构的发现,把生物学推进到了分子水平。生命活动的核心过程由核酸和蛋白质两大类高分子,以及它们与其他分子的相互作用决定。DNA和蛋白质序列的测定,特别是永无止境的基因组测序,导致生物大数据的迅猛增长。生物信息学应运而生。

1999年,我提出建立国家级生物医学信息中心的建议。虽然建立“中心”的计划由于科学管理体制问题而长期搁浅,但我国生物信息学的研究和教学在广大同行推动下仍然不断进步。2001年初,我和张淑誉在杭州参加华大基因的籼稻基因组测序任务。相当一部分测序工作在西湖边上曲苑风荷附近的杭州华大基因完成。西湖“西进”之后,现在那里只剩下金庸茶馆的一座亭子。

那时华大基因杨焕明教授等学者与浙江大学相关院系商议,着手建立生物信息学研究生点。我自始至终参与了筹划过程,并且承诺为2001—2003年的三届研究生讲授“生物信息学引论”大课。浙江大学请当时已经是副教授的农学院樊龙江博士做我的“助教”。这是一位极其称职的“助教”。他每课必在,认真地批改学生作业,同时还参加了水稻基因组的研究。

2004年以后,朱军教授和樊龙江等继续生物信息学的讲授和研究。我高兴地看到,十几年来浙江大学的生物信息学无论在学生培养,还是在科学研究方面都取得了明显成绩。现在樊龙江聚团队之力,主编了《生物信息学》一书,更是值得祝贺的好事。不过我自己只有同一两位合作者共同写书的经历,对于现在比较时兴的团队著述没有经验,也不大放心。好在樊龙江告诉我,他在统一全书文字和体例方面,下了很大功夫。我想,读者们是会对此有所评价的。



复旦大学理论生命科学研究中心

2017年7月26日

前 言

自开始接触生物信息学以来,一晃已近二十年了。我是在攻读博士学位期间开始注意并学习生物信息学的。我的博士生导师胡秉民为应用数学专业教授,主要从事生态系统模型模拟研究。虽然已具备一定数量统计和数量遗传学基础,但当时对于生物信息学,我还是非常陌生的,通过自学才开始一点点了解这门新兴学科。2001—2003年间,中国科学院理论物理研究所郝柏林院士在浙江大学首次开设“生物信息学”研究生课程,我作为他的助教,系统地学习了生物信息学;同时,在他的带领下从事水稻基因组分析。自那时起,浙江大学生物信息学学科和相应研究机构也逐步建立起来,如,2001年浙江大学成立生物信息学研究所,朱军和杨焕明任所长;2003年浙江大学建立IBM生物计算联合实验室等。2004年郝院士离开杭州加入复旦大学,生物信息学研究生课程就由朱军教授和我承担下来。现在该课程作为浙江大学全校研究生公共课程,已成为一门重点建设课程,每年选课人数都在150人左右。

20世纪末,我国生物信息学还处于起步阶段,学习资料很少。学生时常索要学习材料,于是我整理了备课笔记,取名《生物信息学札记》,于2001年6月上传到实验室主页上供学生参考。随着生物信息学的发展,分别于2005年3月和2010年1月更新札记两次。由于网络传播的作用,许多生物信息学初学者都读过该札记,在国内产生了一定的影响。本书是在该札记框架基础上,补充大量新材料编写而成的。

生物信息学学科内容涵盖广且发展很快。基于国内外生物信息学相关教材,以及自身对生物信息学的粗浅理解,我把生物信息学大致分为四部分(篇)内容:第一部分即基础篇,为生物信息学的基础知识。这部分内容总体变化不大(与10~15年前比较),它是生物信息学的核心知识,生物信息学教学最重要的部分,应为必讲内容。第二部分高通量测序数据分析篇,是最近十年才出现的生物信息学新内容。2005年高通量测序技术突破后,针对该技术产生的序列数据,出现大量生物信息学新算法和新工具。第三部分生物信息学外延与交叉,重点介绍与生物信息学密切相关的其他生物学学科。生物信息学引入了这些学科的部分核心技术(或反过来被引入),如数量遗传学、群体遗传学和新兴学科合成生物学。第四部分为生物信息学资源与实践篇。生物信息学数据库和软件工具对生物学学科至关重要,所以这部分也是生物信息学重要组成部分。同时,该篇中以

实践为目的的生物信息学教学资源是课堂教学的一个很好补充。

我重点编写了本书第一部分基础篇。我的学生参与撰写了有关章节,同时也邀请了相应领域研究者参与部分章节撰写(徐海明:数量遗传学;阮松林:蛋白质组学),最后由我统稿。朱军教授帮助审阅了部分内容。我们尽可能完整地列出参考书目,标注材料来源,但一定还会有所遗漏。本书受浙江大学本科专业核心课程教材建设专项经费资助出版。

每次拿起书稿总是能发现一些错误或不准确的地方,但由于时间关系,只好交稿了。如果你发现书中问题,望赐教指正(fanlj@zju.edu.cn),以便我们再版时更正。

樊龙江

2017年8月

目 录

绪 论	1
第一节 生物信息与生物信息学	1
一、迅速增长的生物信息	1
二、生物信息学的概念	2
第二节 生物信息学简史与展望	5
一、生物信息学发展简史	5
二、生物信息学技术的应用	8
三、生物信息学学科展望	10
第三节 本书的组织与使用	13

✧ 第一篇 生物信息学基础

第 1-1 章 生物信息类型及其产生途径	16
第一节 生物信息的类型	16
第二节 DNA 测序技术	18
一、第一代测序技术	18
二、第二代测序技术	21
三、第三代测序技术	29
第三节 高通量测序技术的应用	32
一、DNA/RNA 相关测序	32
二、蛋白质-DNA/RNA 互作测序	34
三、甲基化/宏基因组测序	34
第四节 蛋白质序列及其结构测定	35
一、蛋白质序列与蛋白质互作测定	35
二、蛋白质结构测定	37
第 1-2 章 分子数据库	38
第一节 分子数据库概述	38
一、分子数据库概念	38
二、数据库记录格式	38
三、数据库冗余、序列递交和检索	41

第二节 核苷酸及其相关数据库	45
一、DNA/RNA 序列数据库	45
二、基因组数据库	47
三、非编码 RNA 数据库	49
第三节 蛋白质及其相关数据库	51
第四节 代谢途径等专业数据库	54
一、代谢途径数据库	54
二、代谢组学数据库和表型数据库	55
第 1-3 章 两条序列联配算法及序列搜索	57
第一节 序列联配基本概念	57
第二节 计分矩阵	58
一、计分矩阵的一般原理	58
二、氨基酸替换矩阵	60
三、位置特异性计分矩阵(PSSM)	62
第三节 两条序列联配算法	64
一、Needleman-Wunsch 算法	64
二、Smith-Waterman 算法	67
第四节 BLAST 算法及数据库搜索	69
一、BLAST 算法	70
二、利用 BLAST 进行数据库序列搜索	71
三、序列相似性的统计推断	78
第 1-4 章 多条序列联配算法及功能域分析	81
第一节 多序列联配概念及其算法	81
一、多序列联配概念	81
二、多序列全局联配算法	81
三、多序列局部联配算法	83
第二节 蛋白质序列功能域分析与模型	86
一、功能域概念	86
二、功能域模型	88
第三节 熵与信息量	91
一、不确定性与信息量	91
二、信息熵的应用	92
第 1-5 章 基因预测与功能注释	94
第一节 基因组序列构成与基因预测	94
一、基因组序列的基本构成	94
二、基因预测及其基本方法	96

三、基因注释流程	99
第二节 从头预测——隐马尔可夫模型(HMM)方法	101
一、马尔可夫和隐马尔可夫模型	101
二、隐马尔可夫模型问题及其算法	103
三、HMM 基因预测模型及其应用	104
第三节 贝叶斯统计及其在基因预测方面的应用	107
一、贝叶斯统计与生物信息学	107
二、利用贝叶斯统计进行基因预测	110
第四节 基因功能注释	112
一、利用序列和结构域数据库进行注释	112
二、利用功能分类和代谢途径信息进行注释	114
第五节 基因序列构成分析	114
一、碱基构成与分布	114
二、DNA 行走与 Z 曲线	118
三、同向重复序列分析	119
四、蛋白质序列跨膜等特征分析	123
第 1-6 章 系统发生树构建	126
第一节 系统发生树与遗传模型	126
一、系统发生树概述	126
二、遗传模型	129
第二节 距离法	131
一、非加权平均连接聚类法(UPGMA 法)	132
二、Fitch-Margoliash 算法	134
三、邻接法(NJ 法)	137
第三节 简约法	139
第四节 似然法	141
一、DNA 序列的似然模型	141
二、两条序列系统发生树	142
三、三条及多条序列系统发生树	143
第五节 基因组组分矢量方法	145
一、组分矢量方法(CVTree 算法)	145
二、基因组关联“距离”与系统发生树构建	146
第 1-7 章 蛋白质结构预测与药物设计	148
第一节 蛋白质结构概述	148
一、蛋白质结构及其预测	148
二、蛋白质结构数据库	150

三、蛋白质结构主要预测工具	151
第二节 蛋白质二级结构预测	153
一、蛋白质二级结构预测方法	153
二、结构预测实例	154
第三节 蛋白质三级结构预测	156
一、同源建模法	157
二、折叠识别法	159
第四节 计算机辅助药物设计	160
一、间接药物设计	160
二、直接药物设计	161
第 1-8 章 生物信息学计算机基础	163
第一节 使用 Unix/Linux 操作系统	163
一、Unix/Linux 操作系统及其结构	163
二、Linux Shell 常用命令	165
第二节 掌握一门计算机编程语言	168
一、计算机编程语言	168
二、Python 语言	170
三、R 语言	183
四、MySQL 语言	188
第三节 并行与自动化	195
一、并行式计算	196
二、并行化模型及其实例	197
第四节 其他	202
一、算法	202
二、可视化与画图	204

* 第二篇 高通量测序数据分析

第 2-1 章 基因组拼接与分析	210
第一节 基因组序列拼接概念	210
一、基因组短序列拼接问题	210
二、基因组从头拼接主要方法	211
三、利用遗传图谱等进行基因组组装	211
第二节 基于图论的拼接算法	213
一、图论	213
二、基于德布鲁因图的拼接算法	215

第三节 第三代测序数据拼接方法	220
第四节 基于字符串(<i>K</i> -mer)的基因组调查与分析	223
一、基因组大小估计	223
二、基因组复杂度估计	224
三、基因组“肖像”及缺失字符串分析	225
第 2-2 章 基因组变异与分析	228
第一节 基因组变异类型与检测方法	228
一、基因组遗传变异类型	228
二、基因组变异检测方法	228
第二节 基因组重测序及其应用	232
一、基因组重测序应用领域	233
二、基因组重测序数据分析	235
第 2-3 章 转录组分析	241
第一节 转录组测序与拼接	241
一、转录组及其技术平台	241
二、转录组序列拼接	244
第二节 基因表达分析	245
一、差异表达基因的鉴定	246
二、差异表达基因富集分析	246
第三节 可变剪切和基因融合分析	251
一、基因可变剪切	251
二、融合基因	253
第 2-4 章 非编码 RNA 分析	257
第一节 非编码 RNA 简介	257
一、非编码 RNA 类型与功能	257
二、非编码 RNA 进化	259
三、样品采集及其测序方法	266
四、非编码 RNA 主要数据库	269
第二节 小 RNA 计算识别与靶基因预测	273
一、miRNA 主要特征及计算识别	273
二、siRNA 主要特征及计算识别	277
三、miRNA 和 siRNA 靶基因预测	280
第三节 长非编码 RNA 鉴定与功能分析	282
一、线性 lncRNA 鉴定	282
二、环化 RNA 鉴定	283
三、lncRNA 功能预测	288

第 2-5 章 甲基化与组蛋白修饰分析	291
第一节 表观遗传机制概述	291
第二节 甲基化测序与分析	292
一、甲基化测序原理	292
二、生物信息学分析方法	296
第三节 组蛋白修饰测定与分析	298
一、组蛋白样品的制备	298
二、组蛋白修饰分析方法	298
第 2-6 章 宏基因组分析	301
第一节 宏基因组及其分析方法	301
一、宏基因组概述	301
二、宏基因组学技术的应用	304
第二节 16S rDNA 序列分析	305
一、质控与分析流程	306
二、物种多样性分析	309
三、群落结构分析	313
第三节 全基因组序列数据分析	316
一、分析内容与流程	316
二、基因预测及功能注释	321
第 2-7 章 蛋白质组分析	324
第一节 蛋白质组学概述	324
一、蛋白质组及其分析	324
二、高通量分离和鉴定技术	325
第二节 双向电泳图像与质谱组合分析	332
一、胶图获取与分析	332
二、利用指纹图谱鉴定蛋白质	334
第三节 质谱数据采集与分析	335
一、质谱数据采集策略	335
二、肽段数据库搜索与质量控制	340
第四节 定量蛋白质组分析	345
一、同位素标记定量分析	345
二、非同位素标记定量分析	348

* 第三篇 生物信息学外延与交叉

第 3-1 章 系统生物学	352
第一节 系统生物学概述	352
一、系统生物学的兴起和基本概念	352
二、研究领域及其与生物信息学的交叉	353
第二节 网络与生物网络	353
一、无标度和阶层网络	353
二、生物网络模块及其算法工具	355
第三节 基因调控网络	356
一、布尔网络模型	357
二、贝叶斯网络模型	360
第 3-2 章 群体遗传学	363
第一节 群体遗传多态性与结构	363
一、遗传多态性及其估计	363
二、群体结构	366
第二节 正向选择的统计检验	368
一、自然选择与中性检验	368
二、基于种内多态性的检验方法	369
三、基于种内多态和种间分歧度的检验方法	373
第三节 群体进化的溯祖测验	374
一、溯祖理论	374
二、溯祖测验的应用	376
第四节 统计测验相关问题与策略	379
一、复合测验问题	379
二、全基因组检测的优势和假阳性问题	379
三、影响统计测验的因素	381
第 3-3 章 数量遗传学	383
第一节 数量性状遗传基本概念	383
第二节 连锁分析	387
一、连锁分析原理	387
二、试验群体的连锁分析	387
三、常用连锁分析软件	393
第三节 关联分析	397
一、关联分析基本原理	397
二、常用关联分析软件	402

第 3-4 章 合成生物学	409
第一节 合成生物学概述	409
一、合成生物学定义和研究内容	409
二、合成生物学引发的争议	411
第二节 基因线路:模块化工程化	412
一、基因线路的基本概念	412
二、几个经典基因线路设计	415
第三节 基因组人工合成与重构	418
一、噬菌体基因组人工合成与重构	418
二、细菌基因组人工合成与重构	420

✦ 第四篇 生物信息学资源与实践

第 4-1 章 生物信息学常用代码和关键词	424
第一节 核苷酸和氨基酸代码	424
第二节 遗传密码	426
第三节 核苷酸和蛋白质序列记录特征关键词	426
一、核苷酸序列记录关键词及其说明	426
二、蛋白质序列记录相关的关键词及其说明	430
第 4-2 章 生物信息学数据库和在线分析工具	433
第一节 重要门户网站与分子数据库	433
一、主要门户网站	433
二、主要分子数据库	433
第二节 主要在线分析工具	436
第三节 主要开源分析软件	438
第 4-3 章 生物信息学实验	440
实验 1 分子序列数据库记录格式与检索	440
实验 2 数据库搜索与未知序列功能预测	443
实验 3 多序列联配及其功能域预测	445
实验 4 蛋白质编码基因预测与功能注释	447
实验 5 非编码 miRNA 二级结构及其靶基因预测	450
实验 6 基因组浏览器 GBrowser 及其应用	453
实验 7 系统发生树构建	457
实验 8 蛋白质结构预测	460
第 4-4 章 生物信息学常用英文术语及释义	464
参考文献	499

绪论

我们处在一个激动人心的时代——基因组时代。科学的进步已使人类可以窥探生命的奥秘,甚至包括人类自身。人类基因组在世纪之交被人类自己破译了,这部由 30 亿个字符组成的人类遗传密码本已活生生地摆在了我们面前。与此同时,来自其他生物的基因组信息源源不断地从自动测序仪中涌出,堆积如山,浩如烟海。这些海量的生物信息是用特殊的“遗传语言”——DNA 的四个碱基字符(A、T、G 和 C)和蛋白质的 20 个氨基酸字符(A、R、N、D、C、Q、E、G、H、I、L、K、M、F、P、S、T、W、Y 和 V)——写成。

《科学》(*Science*)杂志在 2001 年 2 月 16 日人类基因组专刊上配发了一篇题为《生物信息学:努力在数据的海洋里畅游》(*Bioinformatics—Trying to swim in a sea of data*)的文章 (Roos, 2001)。文章写道:“我们身处急速上涨的数据海洋中……我们如何避免生物信息的没顶之灾呢?”近年来高通量测序技术的出现,使数据海洋更显排山倒海之势。生物信息学便是我们能找到的可以畅游数据海洋的一条“轻舟”。生物信息学是一门年轻的学科,它充满挑战和机遇,且引人入胜。

第一节 生物信息与生物信息学

一、迅速增长的生物信息

近 20 年来,分子生物学发展的一个显著特点是生物信息的剧烈膨胀,且迅速形成了巨量的生物信息库。这里所指的生物信息包括多种数据类型,如分子序列数据(核酸和蛋白质)、蛋白质二级结构和三维结构数据等(详见第 1-1 章)。由测序仪等产生的大量核酸序列和三维结构数据被存在各类数据库中,这些原始数据构成的数据库就是所谓的初级数据库(primary database);那些由原始数据分析而来的诸如功能区(domain)、二级结构、疏水位点等数据,则组成了所谓的二级数据库(secondary database)。

生物信息的增长是惊人的。近年来,特别是随着高通量测序技术的出现,核酸库的数据每 14 个月左右就要翻一番。2000 年底,数据库数据超过 100 亿个碱基对(GenBank Release 120, 2000)(图 0-1.1),而 2016 年已达到 2 200 亿个碱基对,如果



图 0-1.1 国际核酸序列数据库 GenBank 记录数量增长情况(截至 2016 年 8 月)(www.ncbi.nlm.nih.gov/genbank/statistics/)