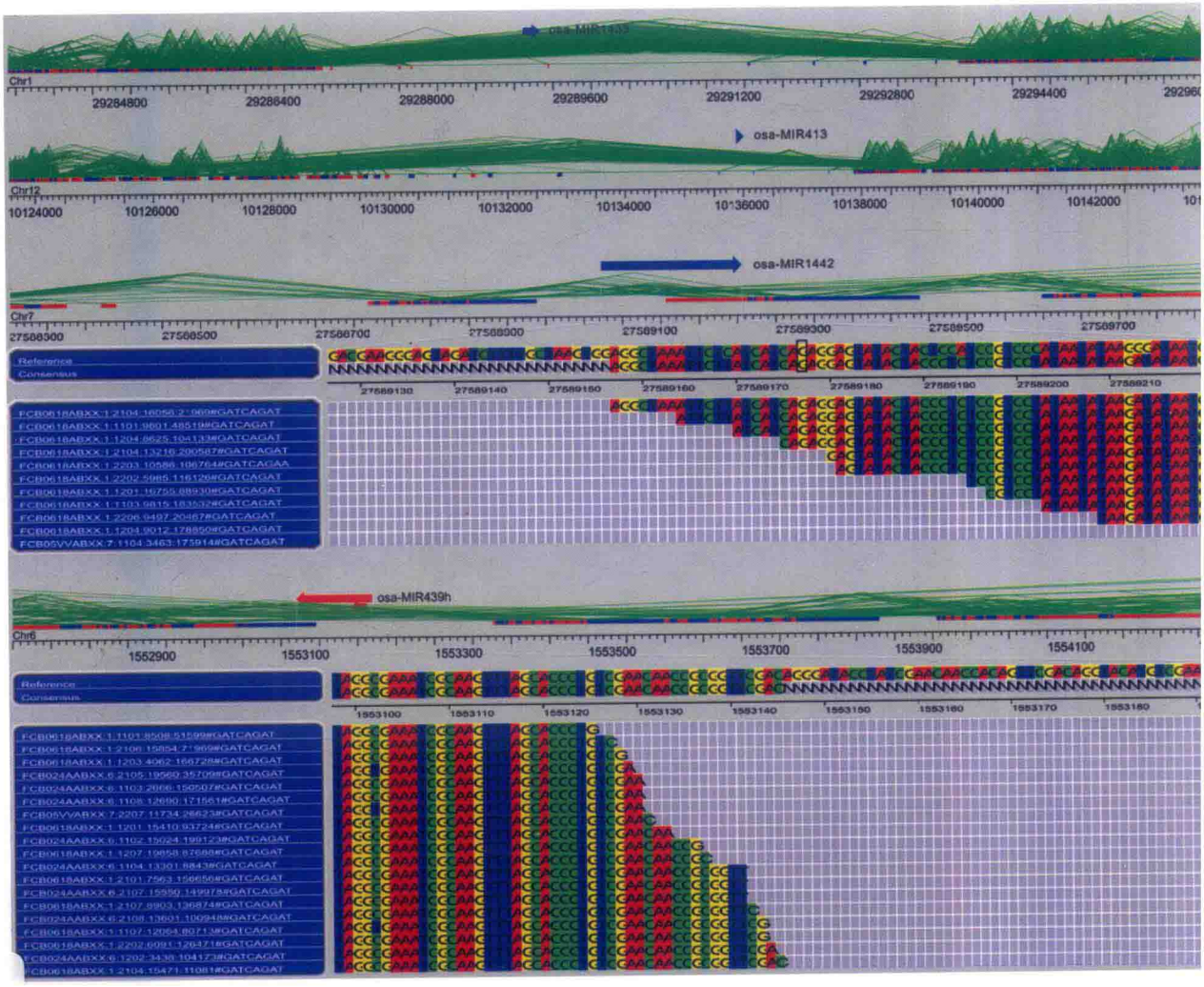
高等院校农学与生物技术专业规划教材

樊龙江　主编

生物信息学

Bioinformatics

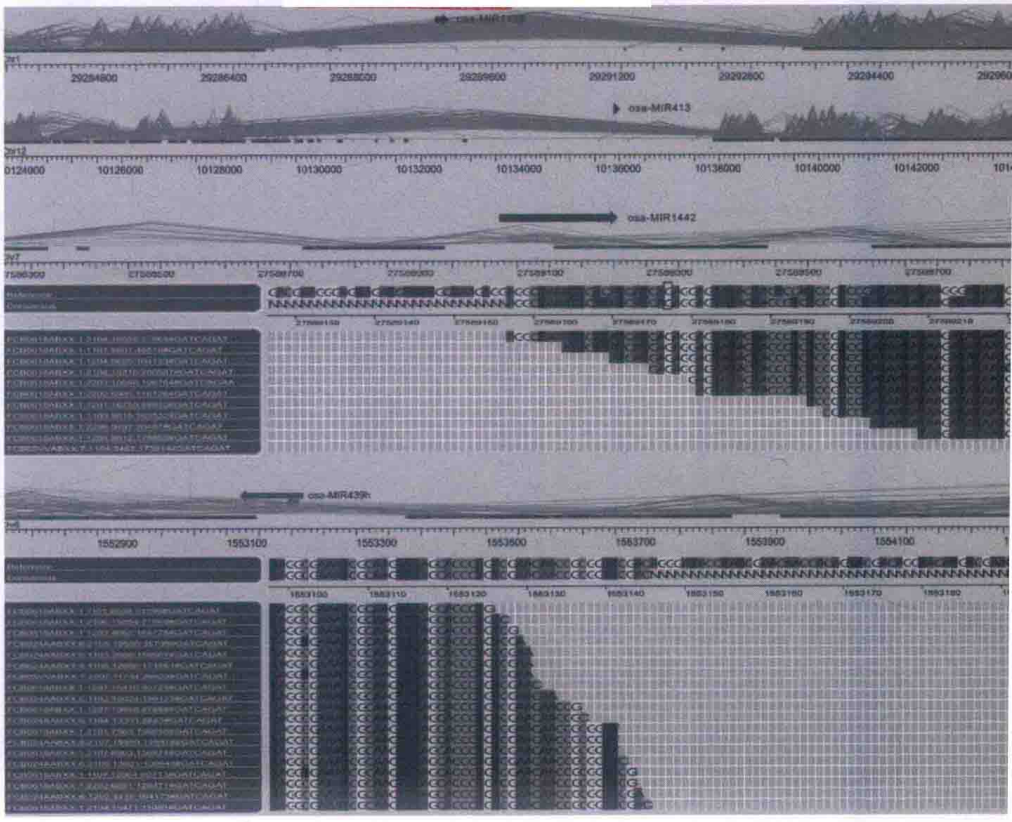


ZHEJIANG UNIVERSITY PRESS
浙江大学出版社
国家一级出版社
全国百佳图书出版单位

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

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

❋ 第二篇 高通量测序数据分析

❋ 第三篇 生物信息学外延与交叉

❋ 第四篇　生物信息学资源与实践

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）。

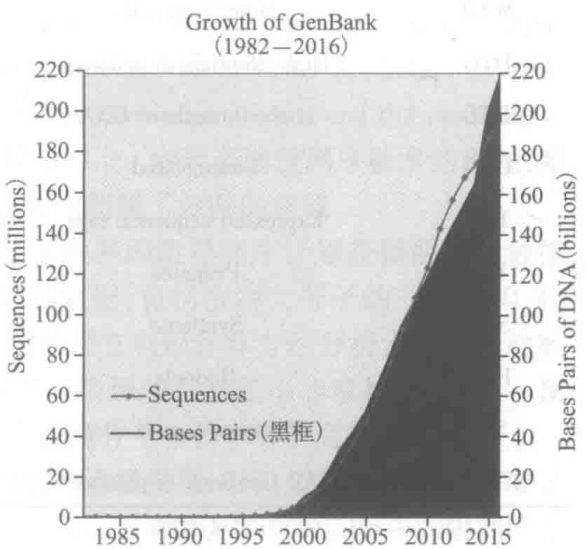


图 0-1.1 国际核酸序列数据库 GenBank 记录数量增长情况（截至 2016 年 8 月）（www.ncbi.nlm.nih.gov/genbank/statistics/）

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