

生物信息学札记（第4版）

樊龙江

浙江大学作物科学研究所
浙江大学生物信息学研究所
浙江大学 IBM 生物计算实验室

2017年9月

本材料已由浙江大学出版社出版：《生物信息学》，樊龙江主编，2017

部分内容可通过下列网址获得：
<http://ibi.zju.edu.cn/bioinplant/>

札记前言

第一版

这份材料是我学习和讲授《生物信息学》课程时的备课笔记，材料大多是根据当时收集的一些外文资料翻译编辑而成。学生在学习过程中经常要求我给他们提供一些中文的讲义或材料，这促使我把我的这份笔记整理并放到网上，供大家参考。要提醒使用者的是，这份材料仅是根据我对生物信息学的一些浮浅的认识整理而成，其中的错误和偏颇只能请读者自鉴了。

2001年6月

第二版

自1999年开始接触生物信息学以来，一晃已近六年，而本札记也近四岁了。2001和2002年中国科学院理论物理所的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，借着我国水稻基因组测序计划的机遇，在他的带领下从2001年开始从事水稻基因组分析，从此自己便完全投入到这一崭新、引人入胜的领域中来。

不断有来信向我索要本札记的电子版文件，同时，在不少网站上看到推荐该札记的内容。生物信息学、基因组学等发展很快，现在再回头审看该札记，有些部分已惨不忍睹，这促使我下决心更新它。但因时间和学识问题，还是有不少部分自己不甚满意，就只有待日后再努力了。欢迎告诉我札记中的BUG，我的信箱 fanlj@zju.edu.cn 或 bioinplant@zju.edu.cn。

2005年3月30日

第三版

近年来高通量测序技术产生的序列数据大量出现（如小RNA和大规模群体SNP数据），本次更新根据这一进展增加了两章内容，分别是第七章有关小RNA的分析和第八章遗传多态性及正向选择检测。两章内容由我的博士生王煜为主编写，李泽峰和刘云参与了文献整理。另外还更新了第四章有关水稻基因组分析一节。

2010年1月

第四版

2014年浙江大学开展本科生教材建设工作，我当时作为系主任要带头，就承诺编写我主讲的《生物信息学》教材。编写教材的确不是一件容易的事，经过几番挣扎和多方努力，总算完成了编写，算是了却了一桩心思。该教材内容比较完整，也跟踪了生物信息学领域的最新进展。我就权且把该教材内容作为札记的第四版，也算给该札记一个完美的结尾。

2017年9月



高等院校农学与生物技术专业规划教材

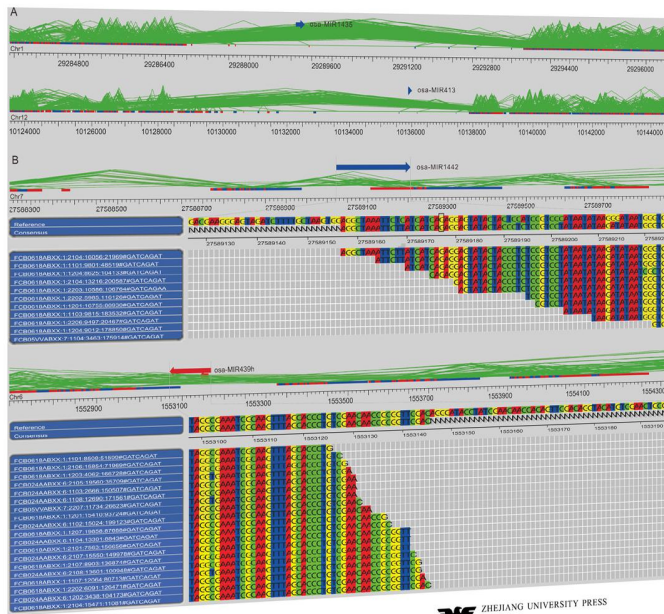
樊龙江 主编

生物信息学

生物信息学 Bioinformatics

樊龙江 主编

浙江
ZHEJIANG UNIVERSITY PRESS
浙江大学出版社



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

《生物信息学》

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。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我和朱军教授承担下来。现在该课程作为浙江大学全校性研究生公共课程，已成为一门重点建设课程，每年选课人数都在150人左右。

上个世纪末，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月挂到实验室主页上供学生参考。随着生物信息学发展，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，在国内形成一定的影响。本书是在该札记框架基础上，补充大量新材料编写而成。

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部分，为应为必讲内容；第二部分高通量测序数据分析篇，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。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，但由于出版计划一再拖延，只好交稿付印了。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
2016年9月

《生物信息学》简要目录及 PDF 下载（二校稿，以出版为准）

绪论		PDF
	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	PDF
第 1-2 章	分子数据库和常见记录格式	
第 1-3 章	两条序列联配及其算法	PDF
第 1-4 章	多条序列联配及功能域分析	PDF
第 1-5 章	基因预测与功能注释	PDF
第 1-6 章	系统发生树构建	
第 1-7 章	蛋白质结构预测与药物设计	
第 1-8 章	生物信息学计算机基础	
	第二篇：高通量测序数据分析	
第 2-1 章	基因组拼接与分析	PDF
第 2-2 章	基因组变异与分析	
第 2-3 章	转录组分析	
第 2-4 章	非编码 RNA 分析	PDF
第 2-5 章	甲基化与组蛋白修饰	
第 2-6 章	宏基因组分析	
第 2-7 章	蛋白质组分析	
	第三篇：生物信息学外延与交叉	
第 3-1 章	系统生物学	
第 3-2 章	群体遗传学	
第 3-3 章	数量遗传学	
第 3-4 章	合成生物学	
	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	
第 4-2 章	生物信息学常用英语术语及释义	
第 4-3 章	生物信息学主要数据库与工具	PDF
第 4-4 章	生物信息学实验	PDF
参考文献		

详细目录

序	郝柏林院士	
前言		
绪论		
第一节	生物信息与生物信息学	
第二节	生物信息学简史与展望	
第三节	本书的组织和使用	
	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	
第一节	生物信息的类型	
第二节	DNA 测序技术	
	一、第一代测序技术 二、第二代测序技术 三、第三代测序技术	
第三节	高通量测序技术的应用	
	一、DNA/RNA 相关测序 二、蛋白质-DNA/RNA 互作 三、甲基化/宏基因组	
第四节	蛋白质序列及其结构测定	
	一、蛋白质序列与蛋白质互作测定 二、蛋白质结构测定	
第 1-2 章	分子数据库和常见记录格式	
第一节	分子序列数据库概述	
	一、分子数据库概念 二、数据库记录格式 三、数据库冗余、序列递交和检索	
第二节	核苷酸及其相关数据库	
	一、DNA/RNA 序列数据库 二、基因组数据库 三、非编码 RNA 数据库	
第三节	蛋白质及其相关数据库	
第四节	代谢途径等专业数据库	
	一、代谢途径数据库 二、代谢组学等数据库	
第 1-3 章	两条序列联配及其算法	
第一节	序列联配基本概念	
第二节	计分矩阵	
	一、计分矩阵的一般原理 二、氨基酸替换矩阵 四、位置特异性计分矩阵 (PSSM)	
第三节	两条序列联配算法	
	一、Needleman-Wunsch 算法 二、Smith-Waterman 算法	
第四节	BLAST 算法及数据库搜索	
	一、BLAST 算法 二、利用 BLAST 进行数据库序列搜索 三、序列相似性的统计推断	
第 1-4 章	多条序列联配及功能域分析	
第一节	多序列联配概念及其算法	
	一、多序列联配概念 二、多序列全局联配算法 三、多序列局部联配算法	

第二节	蛋白质序列功能域分析与模型	
	一、功能域概念 二、功能域模型	
第三节	熵及矩阵信息量	
	一、不确定性与信息量 二、信息熵的应用	
第 1-5 章	基因预测与功能注释	
第一节	基因组序列构成与基因预测	
	一、基因组序列的基本构成 二、基因预测及其基本方法 三、基因注释流程	
第二节	从头预测——隐马尔可夫模型 (HMM) 方法	
	一、马尔可夫和隐马尔可夫模型 二、隐马尔可夫模型问题及其算法 三、HMM 基因预测模型及其应用	
第三节	贝叶斯统计及其基因预测应用	
	一、贝叶斯统计与生物信息学 二、利用贝叶斯统计进行基因预测	
第四节	基因功能注释	
	一、利用序列和结构域数据库进行注释	
	二、利用功能分类和代谢途径信息进行注释	
第五节	基因序列构成分析	
	一、碱基构成与分布 二、DNA 行走与 Z 曲线 三、同向重复序列分析 四、蛋白质序列跨膜等特征分析	
第 1-6 章	系统发生树构建	
第一节	系统发生树与遗传模型	
	一、系统发生树概述 二、遗传模型	
第二节	距离法	
	一、非加权平均连接聚类法 (UPGMA 法) 二、Fitch-Margoliash 算法 三、邻接法	
第三节	简约法	
第四节	似然法	
	一、DNA 序列的似然模型 二、两条序列系统发生树 三、三条及多条序列系统发生树	
第五节	基因组组分矢量方法	

	一、组分矢量方法 (CVTree 算法) 二、基因组关联“距离”与系统发生树构建	
第 1-7 章	蛋白质结构预测与药物设计	
第一节	蛋白质结构概述	
	一、蛋白质结构及其预测 二、蛋白质结构数据库 三、蛋白质结构主要预测工具	
第二节	蛋白质二级结构预测	
	一、二级结构预测方法 二、结构预测实例	
第三节	蛋白质三级结构预测	
	一、同源建模法 二、折叠识别法	
第四节	计算机辅助药物设计	
	一、间接药物设计 二、直接药物设计	
第 1-8 章	生物信息学计算机基础	
第一节	使用 Unix/Linux 操作平台	
	一、Unix/Linux 操作系统及其结构 二、Linux Shell 常用命令	
第二节	掌握一门计算机编程语言	
	一、计算机编程语言 二、Python 语言简介 三、R 语言 四、MySQL 语言	
第三节	并行与自动化	
	一、并行式计算 二、并行化模型及其实例	
第四节	其他	
	一、算法	
	二、可视化与画图	
	第二篇：高通量测序数据分析	
第 2-1 章	基因组拼接与分析	
第一节	基因组序列拼接概念	
	一、基因组短序列拼接问题 二、基因组从头拼接主要方法 三、利用遗传图谱等进行基因组组装	
第二节	图论及基于德布鲁因图拼接算法	
	一、图论	

	二、基于德布鲁因图的拼接算法	
第三节	第三代测序数据拼接方法	
第四节	基于字符串 (<i>K</i> -mer) 的基因组调查与分析	
	一、基因组大小估计 二、基因组复杂度估计 三、基因组“肖像”及缺失字符串分析	
第 2-2 章	基因组变异与分析	
第一节	基因组变异类型与检测方法	
	一、基因组变异类型 二、基因组变异检测方法	
第二节	基因组重测序及其应用	
	一、基因组重测序应用领域 二、基因组重测序数据分析	
第 2-3 章	转录组分析	
第一节	转录组测序与拼接	
	一、转录组及其技术平台 二、转录组序列拼接	
第二节	基因表达分析	
	一、差异表达基因的鉴定 二、差异表达基因富集分析	
第三节	可变剪切和基因融合分析	
	一、基因可变剪切 二、融合基因	
第 2-4 章	非编码 RNA 分析	
第一节	非编码 RNA 简介	
	一、非编码 RNA 类型与功能 二、非编码 RNA 进化 三、样品采集及其测序方法 四、非编码 RNA 主要数据库	
第二节	小 RNA 计算识别与靶基因预测	
	一、miRNA 主要特征及计算识别 二、siRNA 主要特征及计算识别 三、miRNA 和 siRNA 靶基因预测	
第三节	长非编码 RNA 鉴定与功能分析	
	一、线性 lncRNA 鉴定 二、环化 RNA 鉴定 三、lncRNA 功能预测	
第 2-5 章	甲基化与组蛋白修饰	
第一节	表观遗传机制	
第二节	甲基化测序与分析	

	一、甲基化测序原理 二、生物信息学分析方法	
第三节	组蛋白修饰测定与分析	
	一、组蛋白的样品制备 二、组蛋白修饰分析方法	
第 2-6 章	宏基因组分析	
第一节	宏基因组及其分析方法	
	一、宏基因组概述 二、宏基因组学技术应用	
第二节	16S rDNA 序列分析	
	一、技术方法与分析流程 二、物种多样性分析 三、物种丰富度估计 四、群落结构分析	
第三节	全基因组序列数据分析	
	一、分析流程与内容 二、基因预测及功能注释	
第 2-7 章	蛋白质组分析	
第一节	蛋白质组学概述	
	一、蛋白质组及其分析 二、高通量分离和鉴定技术	
第二节	双向电泳图像与质谱组合分析	
	一、胶图获取与分析 二、利用指纹图谱鉴定蛋白质	
第三节	质谱数据采集与分析	
	一、质谱数据采集策略 二、肽段数据库搜索与质量控制	
第四节	定量蛋白质组分析	
	一、同位素标记定量分析 二、非同位素标记定量分析	
	第三篇：生物信息学外延与交叉	
第 3-1 章	系统生物学	
第一节	系统生物学概述	
第二节	网络与生物网络	
	一、无标度和阶层网络 二、生物网络模块及其算法工具	
第三节	基因调控网络	
	一、布尔网络模型 二、贝叶斯网络模型	
第 3-2 章	群体遗传学	

第一节	群体遗传多态性与结构	
	一、遗传多态性及其估计 二、群体结构	
第二节	正向选择的统计检验	
	一、自然选择与中性检验 二、基于种内多态性的检验方法 三、基于种内多态和种间分歧度的检测方法	
第三节	群体进化的溯祖测验	
	一、溯祖理论 二、溯祖测验应用	
第四节	统计测验分析问题与策略	
第 3-3 章	数量遗传学	
第一节	数量性状遗传基本概念	
第二节	连锁分析	
	一、连锁分析原理 二、试验群体的连锁分析 三、常用连锁分析软件	
第三节	关联分析	
	一、关联分析基本原理 二、常用关联分析软件	
第 3-4 章	合成生物学	
第一节	什么是合成生物学？	
	一、合成生物学定义和研究内容 二、合成生物学引发的争议	
第二节	从“基因线路”开始：模块化工程化	
	一、基因线路的基本概念 二、几个经典基因线路设计	
第三节	从最小基因组开始：基因组人工合成	
	一、基因组的人工合成和重构 二、噬菌体基因组人工合成与重构 三、细菌基因组人工合成与重构	
	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	
第一节	核苷酸和氨基酸代码	
第二节	遗传密码	
第三节	核苷酸和蛋白质序列记录特征关键词	
第 4-2 章	生物信息学常用英语术语及释义	
第 4-3 章	生物信息学主要数据库与工具	
第一节	重要门户网站和分子数据库	
第二节	主要在线分析工具	
第三节	主要开放分析软件	
第 4-4 章	生物信息学实验	
实验 1	生物序列数据库记录格式与检索	
实验 2	数据库搜索与未知序列功能预测	

实验 3	抗性基因多序列联配及其功能域预测	
实验 4	蛋白质编码基因预测与功能注释	
实验 5	非编码 miRNA 二级结构及其靶基因预测	
实验 6	基因组浏览器 GBrowser 及其应用	
实验 7	系统发生树构建	
实验 8	蛋白质结构与功能预测	
参考文献		