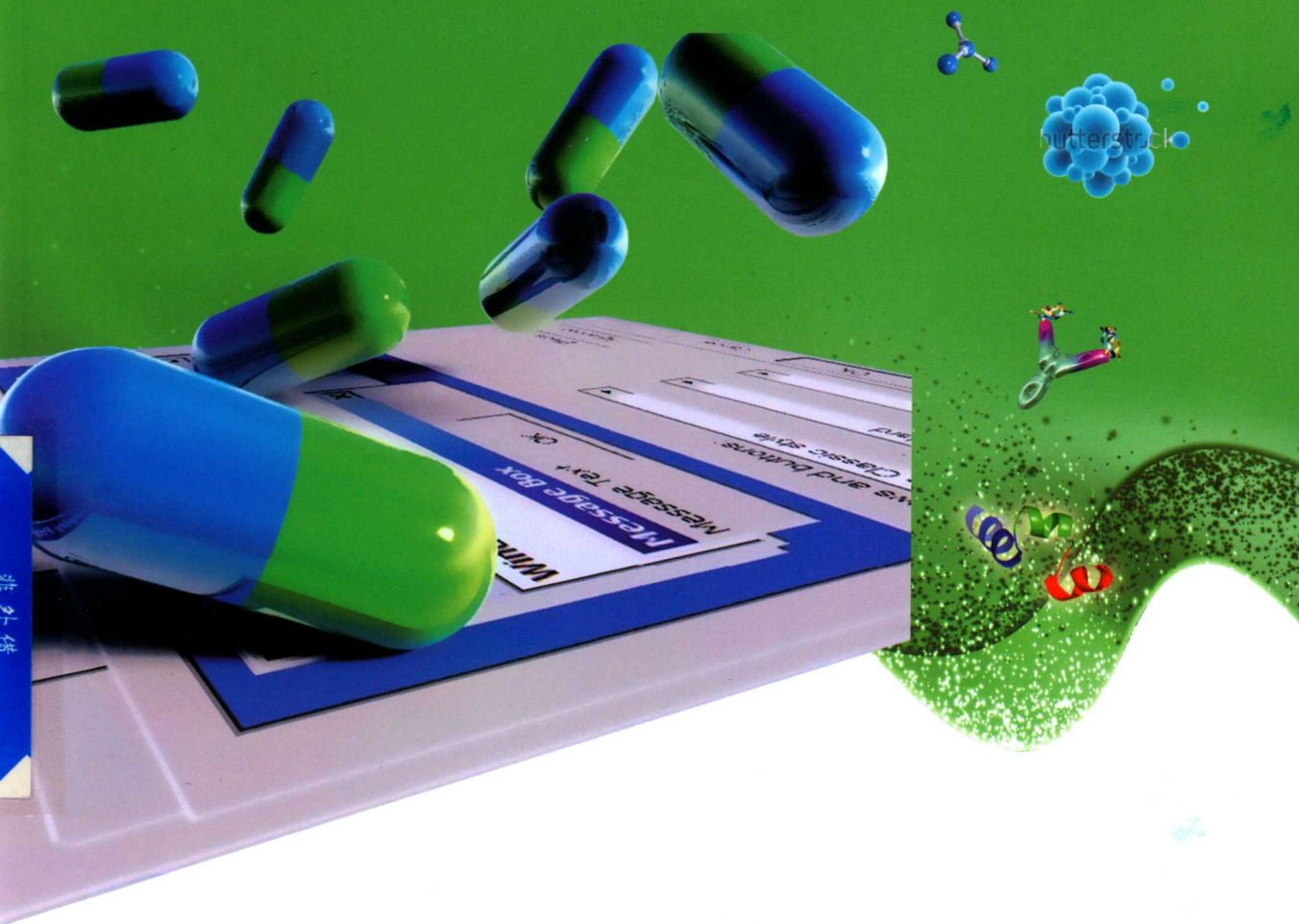




生命科学与信息技术丛书

实用生物信息学

冯世鹏 汤华 周犀 周智 著



 中国工信出版集团

 电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

生命科学与信息技术丛书

实用生物信息学

冯世鹏 汤华 周犀 周智 著



电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

生物信息学是生命科学领域中一门新兴的前沿科学，其发展在很大程度上改变了人们研究利用生物的方式方法，成为生命科学研究最重要的工具之一。本书由 Windows 篇和 Linux 篇两部分组成，重点阐述生物信息学的基本概念与相关技术。Windows 篇讨论了在 Windows 操作系统下进行数据库检索、引物设计、核酸蛋白质序列转换、物种进化树构建、蛋白质高级结构预测分析、非编码 miRNA 研究应用等理论与方法，以及给出相关软件的使用介绍。Linux 篇讨论了在 Linux 操作系统下进行基因组测序，RNA-seq、miRNA-seq 等测序数据质控，基因组组装，转录组分析，基因预测、注释，基因表达分析等操作训练，并且给出对应软件的使用介绍。

本书的特色是紧贴科研实践、图文并茂、内容深浅合适、可操作性强、实用性高，读者可根据书中的步骤轻松实现相应分析，可供从事生物学、农学、医学的科技工作者、教师、学生作为生物信息学入门的参考书籍。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

图书在版编目(CIP)数据

实用生物信息学 / 冯世鹏等著. —北京: 电子工业出版社, 2017.8

(生命科学与信息技术丛书)

ISBN 978-7-121-30224-4

I. ①实… II. ①冯… III. ①生物信息论 IV. ①Q811.4

中国版本图书馆 CIP 数据核字 (2016) 第 258348 号

策划编辑: 冯小贝

责任编辑: 周宏敏

印 刷: 三河市鑫金马印装有限公司

装 订: 三河市鑫金马印装有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本: 787×1092 1/16 印张: 21.25 字数: 550 千字

版 次: 2017 年 8 月第 1 版

印 次: 2017 年 8 月第 1 次印刷

定 价: 69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zltts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式: fengxiaobei@phei.com.cn。

本书得到以下单位的支持，特此感谢。

国家自然科学基金地区基金项目（31460178）

海南大学 2015 年度自编教材资助项目

海南省中西部高校提升综合实力工作基金项目

海南省热带资源可持续利用重点实验室

海南省自然科学基金项目（314042）

前 言

关于本书的成因：希望通过本书让读者了解生物信息学，并能利用生物信息学工具进行常规的分析；对于学有余力或者对生物信息学有浓厚兴趣的读者，则读完本书后可进行二代测序数据的初步深度分析。本书主要针对生物科学相关专业本科生、研究生或者其他有志于学习生物信息学的初学者，希望本书能起到抛砖引玉的作用，带领他们进入生物信息学领域。

关于本书的内容：全书分为两篇，Windows 篇属于生物信息学基础，相关生物信息学软件在装有 Windows 系统的计算机上即可运行，这部分内容要求每个生物科学专业的本科生或读者必须了解掌握，主要包括生物信息相关数据库、序列比对、引物设计、序列分析、进化分析等；Linux 篇属于生物信息学的深度应用，主要软件及其应用需要在安装 Linux 系统的计算机上才能最有效地运行，这部分的内容供学有余力或者有志于进行生物信息学研究应用的学生或工作人员学习，主要包括基因组、转录组的测序、组装、注释等分析内容。

关于学习生物信息学的态度：不贪多、不畏多、自学为主、教学为辅。所谓“不贪多”，就是生物信息学涉及多个学科门类，一个人几乎不可能精通所有相关门类，因此最好根据个人兴趣选择其中一个方向刻苦钻研，勤以练习，融会贯通，同时兼顾其他方面。所谓“不畏多”，就是不要被生物信息学所需要学习的知识吓到，有的知识够用即可，遇到需要进一步学习的时候再去学习新的知识，循序渐进，学得也快。所谓“自学为主、教学为辅”，就是强调学习的主动性，带着强烈的兴趣学习，学习效果要远好于被迫学习。自学过程中不可避免地会遇到一些问题，此时力求通过查阅资料自行解决问题，因此会自然而然地产生自豪感；如果自己查阅资料无法解决的时候最好能有人给以辅助，否则会卡在那里、无法进行后续的学习，这就是要有教学为辅的作用。生物信息学注重实际分析，由于软硬件的差异，对于同样的数据，不同的人处理得到的结果可能不一致，这就要勤加练习，积累经验，分析导致不同结果产生的原因，并能对结果进行取舍，或者改变条件重新分析。

生物信息学，你可以爱它，因为它帮你解决了很多生物学的问题；你也可以恨它，因为有时候你的问题它无法解决。但不管你是爱还是恨，它就在那里，如果你的工作或者学习跟生物有关，你就必须要了解它！

冯世鹏

2017年6月12日于海大

写作任务分工

篇	章	标题	任务分工
	第 0 章	绪论	冯世鹏、汤华
Windows 篇	第 1 章	文献信息检索	冯世鹏、汤华
	第 2 章	生物信息数据资源	冯世鹏、汤华
	第 3 章	序列比对	周犀
	第 4 章	核酸序列分析	周犀
	第 5 章	蛋白质序列分析	周犀
	第 6 章	基因表达分析	周犀
	第 7 章	进化分析	周智
	第 8 章	非编码 miRNA 分析	周智
Linux 篇	第 9 章	Linux 系统	冯世鹏
	第 10 章	Perl 语言	冯世鹏
	第 11 章	测序方法及数据处理	冯世鹏
	第 12 章	基因组组装	冯世鹏
	第 13 章	小 RNA 测序数据分析	周犀
	第 14 章	RNA-seq 数据分析	周智
	第 15 章	基因预测	冯世鹏
	第 16 章	基因注释及功能分析	冯世鹏
	附录 A	生物信息学文件格式	冯世鹏

作者简介

冯世鹏：博士，讲师，海南大学热带农林学院

汤 华：博士，教授，海南大学热带农林学院

周 犀：博士，讲师，海南大学热带农林学院

周 智：博士，副教授，海南大学海洋学院

2.5	代谢通路数据库	52	习题	93
2.5.1	在 KEGG 数据库搜索	53	参考文献	93
2.5.2	主页快速链接	54		
2.5.3	KEGG 通路图及其元素意义	55		
2.6	基因组浏览器	57	第 5 章 蛋白质序列分析	94
2.6.1	基因组数据展示内容	58	5.1 蛋白质理化性质和一级结构分析	94
2.6.2	BLAT 搜索	61	5.1.1 蛋白质理化性质分析	94
2.7	非编码 RNA 数据库	62	5.1.2 蛋白质理化性质分布图	95
2.7.1	miRNA 数据库	62	5.1.3 蛋白质信号肽预测	97
2.7.2	NONCODE 数据库	63	5.2 蛋白质二级结构分析	99
习题		66	5.2.1 蛋白质跨膜结构区分析	99
参考文献		66	5.2.2 蛋白质卷曲螺旋分析	101
			5.2.3 蛋白质二级结构预测分析	103
第 3 章 序列比对		68	5.3 蛋白质三维结构预测分析	104
3.1	比对程序介绍	68	习题	105
3.2	比对序列相似性的统计特性	69	参考文献	105
3.3	在线 BLAST 序列比对	72		
3.4	本地运行 BLAST	75	第 6 章 基因表达分析	106
3.4.1	BLAST 程序的下载和安装	75	6.1 qPCR 数据分析	106
3.4.2	搜索数据库的索引格式化	75	6.1.1 绝对定量分析方法	107
3.4.3	运行 BLAST 程序, 搜索本地序列数据库	76	6.1.2 相对定量方法分析	108
3.5	多序列比对	77	6.2 基因芯片数据分析	111
3.5.1	ClustalX 的使用	77	6.2.1 从 GEO 上下载基因芯片表达谱数据	111
习题		80	6.2.2 将表达谱数据导入 MATLAB 软件	112
参考文献		80	6.2.3 对 soft 格式文件的标准化	113
			6.2.4 差异表达基因筛选	114
第 4 章 核酸序列分析		81	习题	114
4.1	基因阅读框的识别	81	参考文献	115
4.2	基因其他结构区预测	82		
4.2.1	CpG 岛的预测	82	第 7 章 进化分析	116
4.2.2	转录终止信号预测	84	7.1 进化理论介绍	116
4.2.3	启动子区域的预测	84	7.1.1 种群是生物进化的基本单位	116
4.2.4	密码子偏好性计算	86	7.1.2 可遗传的变异是生物进化的原始材料	116
4.3	引物设计	88	7.1.3 分子进化中性学说	117
4.3.1	引物设计的基本原则	88	7.2 进化分析(以 MEGA 为例)	117
4.3.2	Primer 5 引物设计	88	7.2.1 序列准备	118
4.3.3	利用 Primer 5 进行酶切位点分析	91	7.2.2 序列比对	119
4.4	核酸序列的其他转换	92		

7.2.3 建树计算	119
7.2.4 进化树的调整	121
习题	121
参考文献	122
第 8 章 非编码 miRNA 分析	123
8.1 miRNA 简介	123
8.1.1 miRNA 的生物合成	123
8.1.2 miRNA 调控基因表达的机理	124
8.1.3 miRNA 的生理调节作用	125
8.2 miRNA 靶基因预测	125
8.2.1 miRNA 靶基因的预测原理	125
8.2.2 miRNA 靶基因的预测软件	126
8.2.3 miRNA 靶基因的预测步骤	127
8.3 调控靶基因的 miRNA 预测	130
8.4 miRBase 数据库的使用	131
8.4.1 miRBase 数据库的搜索	131
8.4.2 miRBase 数据库批量下载	132
8.4.3 miRNA 记录信息	133
习题	134
参考文献	134

Linux 篇

第 9 章 Linux 系统	138
9.1 Linux 简介	138
9.1.1 什么是 Linux 系统	138
9.1.2 为什么要学习 Linux 系统	139
9.1.3 如何学习 Linux 系统	140
9.2 Linux 系统安装	140
9.2.1 Linux 系统下载	140
9.2.2 系统安装盘制作	142
9.2.3 CentOS 6.5 操作系统安装	144
9.2.4 更新 yum 源	154
9.3 Linux 命令行模式——终端	155
9.4 Linux 系统开机	156
9.5 Linux 系统文件	157
9.5.1 Linux 文件夹及其主要作用 (以 CentOS 6.5 为例)	157
9.5.2 Linux 的文件信息的意义	158

9.5.3 Linux 命令帮助文件	159
9.6 几个重要的快捷键	161
9.7 Linux 系统的命令	161
9.7.1 Linux 系统命令的输入格式	161
9.7.2 常用命令及其常用选项介绍	161
9.7.3 数据流重定向	167
9.7.4 管道命令	168
9.7.5 vim 编辑器工具	168
9.7.6 其他命令	170
习题	177
参考文献	177

第 10 章 Perl 语言	178
10.1 Perl 版本	178
10.2 Perl 标量数据	179
10.2.1 Perl 运算符	180
10.2.2 标量变量	180
10.2.3 数字及字符串的比较 运算符	181
10.3 列表与数组	182
10.3.1 数组及其赋值操作	182
10.3.2 数组元素的引用	182
10.3.3 数组相关的几个命令	183
10.4 哈希	183
10.4.1 哈希赋值	184
10.4.2 哈希的相关函数	184
10.5 判断式及循环控制结构	185
10.5.1 if 条件判断式	185
10.5.2 while 循环结构	185
10.5.3 until 循环结构	186
10.5.4 foreach 循环结构	186
10.5.5 each 控制结构	186
10.6 正则表达式	187
10.6.1 正则表达式相关符号	187
10.6.2 捕获变量	188
10.6.3 正则表达式中特殊字符 的意义	188
10.7 Perl 的排序	189
10.7.1 sort 命令	189

10.7.2	sort 与比较运算符及默认函数的连用	189	习题	231
10.8	Perl 默认的函数的总结	189	参考文献	231
10.9	程序精解	190	第 12 章 基因组组装	232
10.9.1	实例一：从 fasta 文件中寻找特定的序列	190	12.1 Velvet 拼装软件	233
10.9.2	实例二：文本内容分类统计功能	193	12.1.1 Velvet 软件安装	234
10.9.3	实例三：统计文件内容是否有重复	195	12.1.2 Velvet 参数介绍	234
10.9.4	实例四：Scaffolds 序列的排序	196	12.1.3 Velvet 命令运行	237
	习题	196	12.1.4 Velvet 运行结果解读	237
	参考文献	197	12.2 SOAPdenovo 软件拼装	238
第 11 章 测序方法及数据处理		198	12.2.1 软件的安装	239
11.1 测序技术的发展		198	12.2.2 参数介绍	239
11.1.1 第一代测序方法		198	12.2.3 SOAPdenovo 命令运行	241
11.1.2 二代测序方法		201	12.2.4 SOAPdenovo 运行结果解读	242
11.1.3 测序文库插入片段大小选择		205	12.3 ABySS 软件拼装	242
11.1.4 测序类型		205	12.3.1 ABySS 的安装	242
11.1.5 测序方法的搭配		206	12.3.2 ABySS 主要参数介绍	243
11.1.6 测序质量值		206	12.3.3 ABySS 命令运行	245
11.2 测序数据处理		207	12.3.4 ABySS 运行命令结果解读	245
11.3 测序数据质量分析		208	12.4 ALLPATH-LG 软件拼装	245
11.3.1 用 FastQC 软件对测序数据进行评估		208	12.4.1 ALLPATH-LG 的安装	246
11.3.2 NGSQCToolKit 对测序 Reads 的处理		213	12.4.2 ALLPATH-LG 的主要参数	246
11.3.3 FASTX_Toolkit 对测序 Reads 的处理		216	12.4.3 ALLPATH-LG 测试数据运行过程解读	249
11.4 深度测序数据上传 SRA 数据库		218	12.4.4 运行结果解读	252
11.4.1 材料准备		220	12.5 Gaps 修补	252
11.4.2 注册项目信息		221	12.5.1 GapFiller 软件安装	252
11.4.3 提供技术信息		224	12.5.2 相关参数介绍	253
11.4.4 上传数据		227	12.5.3 程序运行命令	254
11.4.5 数据传输完毕状态		230	12.5.4 运行结果解读	254
			12.6 基因组组装效果评估	254
			习题	254
			参考文献	255
			第 13 章 小 RNA 测序数据分析	256
			13.1 小 RNA 测序简介	256
			13.2 小 RNA 测序数据质控	257
			13.3 miRNA 的识别	259
			习题	263

参考文献	263	15.2.4 结果解读	286
第 14 章 RNA-seq 数据分析	264	15.3 AUGUSTUS	286
14.1 转录组序列比对	265	15.3.1 AUGUSTUS 软件安装	286
14.1.1 数据准备	265	15.3.2 相关参数介绍	286
14.1.2 比对数据库	265	15.3.3 训练 AUGUSTUS	287
14.1.3 TopHat 软件下载及安装	266	15.4 PASA	291
14.1.4 Bowtie 软件和 SAMtools 软件下载及安装	266	15.4.1 PASA 软件安装	291
14.1.5 常用 TopHat 参数介绍	266	15.4.2 相关命令参数介绍	293
14.1.6 基因组数据库序列索引	267	15.4.3 命令运行	294
14.1.7 TopHat 使用实例	267	15.4.4 运行结果解读	296
14.1.8 输出文件说明	267	15.5 EVM(EvidenceModeler)	296
14.2 转录本组的组装	268	15.5.1 EVM 软件下载安装	296
14.2.1 cufflinks 的安装	268	15.5.2 相关参数介绍	297
14.2.2 cufflinks 的参数	269	15.5.3 EVM 软件的运行	298
14.2.3 cufflinks 的输出结果	269	习题	300
14.3 合并转录组	269	参考文献	300
14.3.1 用 cuffmerge 合并转录本 的命令	270	第 16 章 基因注释及功能分析	302
14.4 基因表达差异分析	270	16.1 BLAST 软件介绍	302
14.4.1 用 cuffquant 计算表达谱	270	16.1.1 BLAST 软件安装	302
14.4.2 用 cuffdiff 计算不同样本 表达谱的差异	271	16.1.2 相关命令参数介绍	303
14.5 差异表达结果的热图表示	272	16.2 NR 注释	308
习题	273	16.2.1 NR 数据库制备过程	308
参考文献	273	16.2.2 NR 注释过程	309
第 15 章 基因预测	275	16.3 COG 注释	310
15.1 GeneMark 软件序列	275	16.3.1 COG 数据库准备过程	310
15.1.1 GeneMarkS 的安装	275	16.3.2 COG 命令注释过程	311
15.1.2 相关参数介绍	276	16.4 Swiss-Prot 注释	311
15.1.3 GeneMarkS 命令运行	279	16.4.1 数据库准备	312
15.1.4 GeneMarkS 运行结果解释	280	16.4.2 Swiss-Prot 注释过程	312
15.2 Glimmer 软件	280	16.4.3 InterPro 注释	312
15.2.1 Glimmer 软件安装	280	16.5 KEGG 注释	314
15.2.2 相关命令参数介绍	281	16.6 GO 注释	317
15.2.3 程序运行	284	习题	320
		参考文献	321
		附录 A 生物信息学文件格式	322

第0章 绪 论

0.1 生物信息学的发展历史

以人类基因组计划实施为界，生物信息学的发展大致经历3个阶段，包括前基因组时代、基因组时代和后基因组时代。前基因组时代，有部分计算生物学家进行算法开发及核酸与蛋白质大分子数据收集及数据库构建。基因组时代，由人类基因组计划的实施开始，先后有6个国家的科学家直接参与人类基因组计划项目开发，同时也有像 Celera 公司为代表的其他科学家进行基因组测序及相应数据分析软件开发。后基因组时代，虽然进行了更广泛的生物物种的序列测定，但是基因组序列研究已经不是重点，更多的生物信息学研究人员转向研究蛋白质组、转录组、代谢组、比较基因组、结构基因组、功能基因组等研究领域。

0.1.1 Bioinformatics 的来源

Bioinformatics (生物信息学) 一词最早是荷兰科学家 Paulien Hogeweg 与 Ben Hesper 于 1978 年使用的，但其实他们更早在 1970 年用荷兰文字 (Bioinformatica) 发表文献，但是该文献关注的人较少。该词当时代表生物过程中的信息流，与现在生物信息学的定义有很大的差别。

华裔科学家林华安 (Hwa A. Lim) 博士 1987 年提出 Bioinformatics，并对生物信息学定义进行了探讨，并在 20 世纪 90 年代多次主持召开生物信息学相关会议，对生物信息学的推广做出了一定的贡献。

0.1.2 生物信息学的定义

生物信息学还没有统一的定义，不同的生物信息学家对其定义均有所差异，1995 年人类基因组计划第一个五年报告中给出了一个定义，在李霞教授主编的《生物信息学》一书中采用了该定义：“生物信息学是一门交叉学科，它包含了生物信息的获取、加工、储存、分配、分析、解释在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具来阐明和理解大量数据所包含的生物学意义。”

生物信息学包括广义生物信息学和狭义生物信息学，广义生物信息学是研究整个生命过程的相关信息；狭义生物信息学是研究生物大分子 (主要是核酸和蛋白质) 所包含的生物信息，有时候也称为分子生物信息学。目前的生物信息学研究主要集中在狭义生物信息学方面，因此本书的内容也主要集中在狭义生物信息学。

0.1.3 人类基因组计划

人类基因组计划 (Human Genome Project, HGP) 是由美国科学家发起，由美国 NIH 及能源部支持，先后有英国、法国、德国、日本和中国共 6 国科学家参与的大型国际合作项目，1990 年 10 月启动，总投资 30 亿美元，计划用 15 年时间完成人类基因组 30 亿个碱基对的测

序。人类基因组计划(1990—2003年)与阿波罗登月计划(1961—1969年)、曼哈顿原子弹计划(1942—1946年)并称为20世纪人类自然科学史上三大科学计划。人类基因组计划在生物信息学学科发展过程中的作用再怎么强调都不为过。国际人类基因组计划由美国1988年成立的国际人类基因组研究中心(现国际人类基因组研究所)的第一任主任 Watson Jamas 领导,后 Watson 于1992年辞职,改由 Francis Collins 任国际人类基因组研究所所长,并继续领导国际人类基因组计划。

国际人类基因组计划团队采用逐步克隆(clone by clone)的方法,并使用第一代测序仪进行测序。人类基因组包含22条常染色体和x、y两条性染色体,6国科学家分别承担其中一部分染色体测序任务,以杨焕明为首的我国科学家承担了第三号染色体短臂上约30Mb区域的测序任务,占人类基因组测序任务的1%。1998年以 Craig Venter 为首成立了 Celera 公司,号称要用3亿美金,以300台最新毛细管自动测序仪(ABI 3700)及全球第三的超大型计算机在3年内完成人类基因组的测序, Venter 采用的测序策略是全基因组鸟枪法(whole genome shotgun)。随着 Celera 公司人类基因组计划的启动,形成了与国际协作组的竞争,迫使国际协作组加快了测序步伐,最终于2000年6月26日由美国总统克林顿宣布人类基因组草图的完成,当时 Collins 和 Venter 并肩站在克林顿后面。国际协作组及 Venter 的人类基因组测序结果文章分别在2001年的 *Nature* 及 *Science* 杂志上发表。2003年4月宣布人类基因组序列图绘制成功,人类基因组计划的所有目标全部实现,人类基因组计划圆满结束,主要完成了人类基因组的4张图:遗传图谱、物理图谱、序列图谱、基因图谱;基因组测序覆盖度达到99%,测序准确度达到99.99%。2004年10月公布人类基因组完成图,后续科学家不断对人类基因组图谱进行补充完善,目前 UCSC 数据库人类基因组图谱已更新至第38版(GRCh 38/hg38,于2013年12月释放)。

0.1.4 生物信息学发展重要人物及大事件

1. 生物信息概念的建立

1956年在美国田纳西州盖特林堡首次召开生物学中的信息理论研讨会,产生了生物信息的概念。

1970年, Hogeweg 使用了 bioinformatics 一词。

1987年, Hwa A. Lim 再次提出 bioinformatics 一词,并陆续召开了一系列生物信息相关会议。

2. 算法的建立及发展

1962年, Zucherkan dl(Emile Zucherkan dl)和 Pauling(Linus Pauling)开创了分子进化这个全新的研究领域,主要通过序列分析研究序列变化与进化之间的关系,后来被称为分子钟(molecular clock)。

1966年 Dayhoff(Margaret Belle (Oakley) Dayhoff)运用计算机及 maximum parsimony 方法重建蛋白进化树。

1970年 Needleman(Saul B. Needleman)和 Wunsch(Christian D. Wunsch)一起开发了序列比对算法(Needleman-Wunsch 序列比对算法),是首个动态规划算法,用于序列间的全局比对,为序列间的比较及数据库的搜索提供了可能,可以说是生物信息学发展史上的里程碑事件。

1978年, Dayhoff 建立蛋白质序列比较的 PAM (Point Accepted Mutation) 替换矩阵, 大大提高了序列比较算法的性能。

1981年, Smith (Temple F. Smith) 和 Waterman (Michael S. Waterman) 提出了著名的局部对位排列算法 (Smith-Waterman 算法), 提高了序列比对的精确度。

1988年, Pearson (William R. Pearson) 和 Lipman (David J. Lipman) 发表了著名的 FASTA 序列运算法则, 其产生的 fasta 序列格式目前被广泛使用。

1990年, BLAST 算法及软件发表, 目前仍然被广泛使用, 并在其基础上有新的改进软件出现。

1992年, BLOSUM 打分矩阵发布, 是目前进行蛋白质序列比较的应用最广泛的打分方法。

3. 大型数据库的建立

1965年, Dayhoff 收集蛋白质序列及机构, 并以图书形式陆续出版, 1984年由该数据建立了 PIR 数据库。

1982年, 欧洲分子生物学实验室 EMBL 诞生, 提供核算序列数据库服务。

1982年, 美国国立卫生研究院下属的国立生物技术信息中心建立了 GeneBank 数据库。

1986年, 日本核酸序列数据库 DDBJ 诞生。

1988年, 三大数据库达成协议: 采用共同的数据库记录格式收集直接提交的数据, 并定期进行数据交换。

1986年, 创建专家注释非冗余蛋白质数据库 Swiss-Prot, 1996年创建蛋白翻译 TrEMBL 数据库, 2002年两个数据库整合为 UniProtKB 数据库。

1999年, 启动 Ensemble 计划, 目标在于开发工具及软件进行基因组的自动注释, 并将相关注释结果在网页呈现并进行共享。

2000年, 创建了 UCSC Genome Browse 基因组浏览器, 进行基因组注释信息的可视化操作。

2003年, 创建 miRBase 数据库, 用于存储非编码 miRNA 数据。

2003年, 启动 ENCODE (Encyclopedia of DNA Elements) 项目, 目标是解析人类基因组的功能区域, 2011年进行了第一次数据的大型释放, 目前该项目仍在继续。

4. 物种测序

1990年, 人类基因组计划启动, 2000年完成人类基因组草图, 2003年完成精细图。

1995年, 第一个细菌全基因组测序完成 (流感嗜血杆菌, *Haemophilus influenzae* Rd), 这是人类拥有的第一个能自由活动生物的全基因组序列, 使用全基因组鸟枪法, 基因组大小为 1.8Mb。

1996年, 第一个真核生物基因组测序完成 (面包酵母, *saccharomyces cerevisiae*), 基因组大小为 12.1Mb, 16条染色体。

1997年, 第一个模式生物完成测序 (大肠杆菌, *escherichia coli*), 基因组大小为 5.1Mb。

1998年, 第一个多细胞生物测序完成 (秀丽线虫, *caenorhabditis elegans*), 基因组大小为 99.2Mb, 6条染色体。

2000年, 第一个植物基因组测序完成 (拟南芥, *arabidopsis thaliana*), 基因组大小为 125Mb, 10条染色体。果蝇 (*drosophila melanogaster*) 的基因组测序完成, 基因组大小 148.5Mb。

2002年, 水稻 (*oryza sativa*) 和小鼠 (*mus musculus*) 基因组草图完成。

2002年,国际人类基因组单体型图计划(The International HapMap Project)启动,来自美国、日本、英国、加拿大、中国、尼日利亚的科学家协作,旨在确定和编目人类遗传的相似性和差异性,该计划于2005年完成。

2006年,癌症基因组图集(The Cancer Genome Atlas, TCGA)计划启动,计划测定10 000个肿瘤基因组,于2014年宣告结束。

2006年,深圳华大启动炎黄计划,计划测定100个黄种人基因组,旨在为黄种人疾病防治提供参考,2007年炎黄一号发布,其他项目仍在进行。

2009年,万种脊椎动物基因组计划(the Genome 10K project, G10K)启动,计划测定1万种脊椎动物基因组序列,目前未完成。

2010年,深圳华大启动千种动植物基因组计划,目前尚未完成。

5. 测序技术发展

1970年,吴瑞(Ray, Wu)开发了第一种DNA测序方法。

1977年,Sanger(Frederick Sanger)开发双脱氧测序技术,也称为Sanger测序法。

1986年,第一台商业化测序仪推出(370A, ABI),也称为第一代测序仪,1998年推出了3700测序仪,目前已经发展至3730x1全自动四色荧光聚焦测序仪,基于双脱氧测序技术,是目前测序通量最高的第一代测序仪。

2005年,454公司推出基于焦磷酸测序的454测序仪,是第一台二代测序仪。

2006年,Illumina推出基于Solexa技术的二代测序仪。

2007年,ABI推出基于连接法测序的Solid二代测序仪。

2010年,Life公司推出基于PCR反应释放 H^+ 导致PH变化进行检测的Ion Torrent二代测序仪。

2010年,PacBio公司推出第三代纳米孔单分子测序仪。

0.2 生物信息学的研究内容

0.2.1 生物分子数据的收集与管理

生物分子的类型:目前生物信息学研究的生物分子主要集中在核酸和蛋白质,因此生物分子的类型包括:蛋白质、DNA(包括基因组DNA、线粒体DNA、叶绿体DNA)、RNA(包括mRNA、tRNA、rRNA、miRNA、lncRNA等不同类型的RNA)。

生物分子数据类型:包括序列信息、结构信息、表达信息、定位信息、相互作用信息等。对于DNA分子来说,主要研究其序列信息、结构信息、片段注释等信息。对于RNA来说,主要研究其表达信息、结构信息、定位信息等。对于蛋白质来说,主要研究其结构信息、定位信息、相互作用信息等。

分子数据收集方法:序列信息的收集包括一代测序、二代测序、蛋白质序列测定等方法。结构信息的收集包括X-ray、核磁共振(NMR)、高效液相色谱技术(HPLC)等方法。表达信息的收集包括二代测序、定量PCR、芯片杂交、Northern Blot、Western Blot等。定位信息的收集主要包括原位杂交、融合蛋白标记、荧光共振能量转移(FRET)等方法。相互作用信息收集主要包括酵母杂交技术、凝胶阻滞实验(EMSA)等。

数据管理: 包括所收集的生物信息等数据提交给已有的数据库, 或者构建本地数据库等。

0.2.2 数据库搜索及序列比较

数据库的分类: 按照数据来源分为一级数据库, 主要收集实验获得的原始数据; 二级数据库, 在一级数据库的基础上加工而成的数据库, 目前的数据库大多都同时收录原始数据及在原始数据基础上的注释信息, 因此均兼具一级、二级数据库的特征。按照收录的数据类型分为核酸数据库, 如 GenBank、ENA、DDBJ 等; 蛋白质序列数据库, 如 Uniprot; 蛋白质结构数据库, 如 PDB 等。还有一些专有数据库, 如线虫基因组数据库 AceDB、拟南芥数据库 tair; 非编码 RNA 数据库, 如 ncRNAdb、miRBase; 蛋白质序列二级结构数据库, 如 Prosite。数据库的类型很多, 我们应该了解大型数据库及其所收录的数据类型, 以便于有针对性地查找相关数据库, 比如要找基因的序列, 可在三大核酸数据库进行搜索; 要找 miRNA 相关信息, 最好在 miRBase 数据库搜索等。

数据库搜索方法: 一般数据库搜索的设计比较人性化, 易于掌握, 而且各数据库网站会提供较详尽的帮助文件介绍数据库的搜索方法以供自学。当然, 本书也会展示重要数据库的搜索方法, 供读者举一反三。

序列比较: 序列比较在生物信息分析过程中应用广泛, 通过比较寻找序列的插入缺失等异同; 寻找同源蛋白并推测未知基因的功能; 通过序列比较可进行进化分析、寻找特定结构域等多方面的应用。序列比较可以使用数据库在线比较, 比如三大核酸数据库 GenBank、ENA、DDBJ 均提供序列的 blast 搜索比较方法; 也可以使用软件本地比较, 比如 Cluster 软件、DNASTAR 软件等。

0.2.3 基因组序列分析

随着二代测序技术的发展成熟, 目前进行基因组的测序费用已经大大降低, 一般的实验室都能承受, 测序完毕之后的基因组序列分析包括如下内容。

基因组组装: 通过不同软件将测序所得的短 Reads 组装成 Contigs 或者 Scaffolds; 对 Gaps 进行生物信息学修补。

基因预测: 包括基因(gene)、启动子(promoter)、多聚 A 位点(polyA)、mRNA、tRNA、rRNA、snoRNA、miRNA、重复序列等序列模块的预测。

基因注释: 对基因进行 NR、Swissprot、KEGG、InterProScan、COG 注释, 预测这些基因潜在的功能。对 tRNA、rRNA、miRNA 等非编码 RNA, 需通过比对, 找出已知的非编码 RNA 及可能新的非编码 RNA。对于重复序列, 要能区分这些重复序列的类型, 比如 Alu 序列、微卫星序列等。对于 miRNA, 还要预测其调控的潜在的靶基因。

功能基因分析: 对感兴趣的基因进行基因家族分析, 分析基因共有或特有结构域, 对不同物种的相似基因做进化分析等。

比较基因分析: 比较基因组里的插入、缺失(InDel)、转位、移位、SNP 位点、基因数量的变化、共有基因、特有基因分析等。

0.2.4 基因表达数据的分析与处理

基因表达数据分析, 依据数据测定的方法不同, 分析方法也会有很大的差别。

Northern Blot: 抽提 RNA 后电泳、杂交、显影/拍照, 通过软件对条带取灰度值进行分析

来获得基因表达差异数据,该方法灵敏度低,可能要对放射性同位素进行操作等,其应用受到一定的限制。

定量 PCR 数据: 抽提 RNA 定量 PCR 操作完成后,先通过定量扩增曲线确定各样本的 Ct 值,然后通过经典的 $2^{-\Delta\Delta Ct}$ 方法进行基因差异表达计算,该方法灵敏度高,操作相对容易,是使用最多的基因定量检测方法之一。

芯片数据: 抽提 RNA 后,通过标记、芯片杂交、扫描、取灰度值获得原始芯片数据,对原始数据进行质控分析及归一化处理,再使用 ArrayTools、R 软件等对芯片数据进行基因差异表达分析。芯片可同时检测成千上万的基因,因此该方法对于转录组研究使用较多,尤其适合基因组信息比较明确、并有商业化芯片的物种,如人、小鼠、拟南芥等模式生物。

二代测序数据: 抽提 RNA 后建库并进行二代测序 (RNA-seq),将所获得的原始 Reads 信息转换为各基因的 RPKM 值(此过程相当于对数据的归一化处理),再运用 R 软件、cufflinks 等软件对基因进行差异表达分析。该方法既可研究已知基因,亦可研究未知基因,因此对于基因组信息有限的物种的转录组研究,应用非常广泛。

0.2.5 蛋白质结构预测

蛋白质结构可通过实验方法直接测定,如利用 X 光(X-ray)晶体学、核磁共振(NMR)、荧光光谱、紫外光光谱等方法直接测定。

X 光晶体学: X 光射入晶体后会衍射,衍射线的分布和强度与晶体结构密切相关,类似晶体指纹,因此该方法是目前使用最多的蛋白晶体结构研究方法。实验过程需先对蛋白质过表达并创造条件使其形成结晶,再通过 x 光衍射测定衍射数据、对衍射数据进行分析并构建结构模型。该方法测定结果比较准确,但是获得良好的蛋白结晶相对困难。

核磁共振: 其原理是原子核在强磁场作用下,吸收外来电磁辐射产生核能级的跃迁,从而产生核磁共振(NMR)现象。物质的分子结构与所处的化学环境对 NMR 信号有影响,因此在化学环境一定时,即可研究物质的分子结构,尤其适合进行有机化合物结构研究。实验过程将蛋白质溶于一定的介质中,并置于强磁场条件下获得 NMR 图谱,将所得图谱与对照图谱进行比较分析,即可推断蛋白质的结构。该方法在溶液状态下对蛋白质结构进行测定,所测定的结构更符合生理状态真实结构,其应用越来越受到重视。

蛋白质结构预测的简单分类大概有两类方法。第一是通过与现有已知结构的蛋白质进行序列比较,通过序列的相似性来推断待测序列的结构,常用软件如 interProScan、SWISS-MODEL 等。第二是重新预测,即将已知的蛋白质结构分拆来构建模型构件数据库,然后通过对待测蛋白质序列进行构件分析,并最终预测其可能的结构。对蛋白质的预测涉及不同的算法,有兴趣的读者可自行查阅相关资料,如果对算法不了解也没有关系,因为很多算法已经开发出完整的软件工具,还有部分提供网页版的,大多数人仅需要知道如何使用这些工具即可,对于其具体的运算过程可以不必考虑。

0.2.6 非编码 RNA 研究

非编码 RNA 种类包括最早发现的 tRNA、rRNA;当前研究热点的非编码小 RNA(包括 miRNA、piRNA、snoRNA);长非编码 RNA(lncRNA)等。

tRNA 与 rRNA 是最早发现的非编码 RNA,其功能研究得比较清楚,对于新测序物种,