HANDBOOK of COMPARATIVE GENOMICS

PRINCIPLES AND METHODOLOGY

比较基因组学手册

——原理与方法

[意] C. 萨科内（Cecilia Saccone）
G. 佩索莱（Graziano Pesole） 著

王 进 严 明 等译

化学工业出版社

HANDBOOK of COMPARATIVE GENOMICS

PRINCIPLES AND METHODOLOGY

比较基因组学手册

——原理与方法

[意] C. 萨科内（Cecilia Saccone）
G. 佩索莱（Graziano Pesole） 著

王 进 严 明 等译

化学工业出版社

·北 京·

图书在版编目（CIP）数据

比较基因组学手册——原理与方法/［意］萨科内（Saccone，C.），［意］佩索莱（Pesole，G.）著；王进等译. —北京：化学工业出版社，2008. 2
书名原文：Handbook of Comparative Genomics:Principles and Methodology
ISBN 978-7-122-02085-7

Ⅰ. 比… Ⅱ. ①萨…②佩…③王… Ⅲ. 基因组-对比研究 Ⅳ. Q343. 1

中国版本图书馆 CIP 数据核字（2008）第 019452 号

Handbook of Comparative Genomics: Principles and Methodology, by Cecilia Saccone and Graziano Pesole
ISBN 0-471-39128-X

北京市版权局著作权合同登记号：01-2005-2927

责任编辑：傅四周　孟　嘉　　文字编辑：张春娥
责任校对：战河红　　装帧设计：关　飞

出版发行：化学工业出版社（北京市东城区青年湖南街 13 号　邮政编码 100011）
印　　刷：北京市振南印刷有限责任公司
装　　订：三河市宇新装订厂
720mm×1000mm　1/16　印张 25¼　字数 569 千字　　2008 年 7 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899
网　　址：http://www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：65.00 元

谨以此书献给 Ernesto Quagliariello!

前　言

基因组时代的序幕刚刚拉开，后基因组时代即迅速来临。大规模测序实施以来，人们一直在探寻新的方法和新的工具以用于从分子水平上研究生命体，特别是对推动生物学发展的新假说进行研究。全基因组测序所带来的前景被寄予了厚望，但目前仍然存在许多问题有待解决；基因组信息的开发利用还处于萌芽阶段，需要发展新的方法学来对基因组信息进行充分利用。

一些基本的问题包括：大规模测序在多大程度上增加了我们关于物种性质和功能（或广义地看，有关基因组结构与功能）的知识？测序技术广泛应用以后出现了什么更有价值的观点或想法？人们对生物过程分子基础的认识是如何得到推进的，如果有的话，是否是根本性的？我们在进行比较基因组研究方面有什么样的能力？可以得出哪些一般性的规律？

众所周知，比较生物学研究大多建立在类比和同源概念的基础上，由于这个原因，必须进行全基因组水平的比较，而不是仅仅对单个基因、基因家族或特定的基因组区域进行比较。与之前只能专注于局部不同，原核生物、真核生物和细胞器的全基因组测序已经有效地促进了将基因组作为整体进行结构与进化的研究工作。因而产生了进化基因组学这个新的领域，并将在揭示核酸、蛋白质线性结构、三维折叠、细胞遗传学、基因表达和调控途径方面产生革命性的影响。

我们相信比较进化基因组学是揭示隐藏于生命体中的信息的关键，我们使用的概念“基因组学”包括其表达和调控机制，这是所有生物过程的基础，也就是说，基因组学也包括转录组学、蛋白质组学和其他“组学”，这些概念相互衔接、相互交叉。“组学”研究需要广泛的数据库和支持包括生物学、化学、信息学、数学和物理学多学科间合作研究的动态技术。

本书旨在为具有生物学背景并试图接触现代生物学的新方向——基因组学的读者提供一本工具书，对技术开发人员和管理人员、工业界和基金机构也很有用。原则上，任何有兴趣了解这一新领域的激动人心的新应用的人都可一读。必须说明的

是，这本书只陈述个人的观点，由于涉及范围广，问题复杂，所以无法囊括所有的文献，尽管这样，我们已经尽力突出里程碑性的工作、新产生的概念、关键的方法技术和这一新领域最急需的知识。本书以描述完整测序的生物全基因组为起点，提炼从已获得的数据中涌现出来的新概念，同时也介绍用于研究完整基因组及其进化的最新方法。

总体来说，本书是作为学生和研究人员的一般性的指导书，读者不一定是专门研究基因组学的，任何人都可一读，不管是理论研究还是带有实际目的，只要有兴趣涉猎这一迷人的生物学新篇章——实际上已经成长为一个新的学科领域。我们将本书分为三个部分。

第一部分描述了利用现代生物技术所获得的主要生物学过程的最新分子知识。介绍了由基因组测序带来的新认识。特别地，我们总结了完整测序的原核生物、真核生物和细胞器基因组的主要特征。

第二部分是基因组学中最新的方法学。介绍了现有实验技术和生物信息学工具，特别强调的是分子生物学技术、生物学数据库和用于序列分析的计算方法。

第三部分介绍了从比较研究中获得的结果。我们在这一部分讨论了一些基础性的前沿问题，如基因组大小的进化、碱基组成的约束以及分子层次上的生物体结构与起源。最后，我们论述了分子系统发生学的新进展。

致谢

感谢我们怀念的亲爱的朋友和同事 Giuliano Preparata，一位著名的理论学家，他将我们引入了分子进化研究的多学科探索；也热忱地感谢为本书提供过指导和建议的好几位同事和朋友，他们是 Marcella Attimonelli、Giorgio Bernardi、Rita Casadio、Nicla Cataldo、Victor De Lorenzo、Annamaria D' Erchia、Ilenia D' Errico、Carmela Gissi、Alessandro Minelli、Aurelio Reyes、Teresa M. R. Regina、Elisabetta Sbisa和 Apollonia Tullo。最后，特别感谢 Alessandra Larizza 整理文献和组织资料工作，以及 Marilina Lonigro 在语言文字方面的帮助。

目　录

上篇　基因组特性

第 1 章　原核生物 …… 3
1.1　引言 …… 3
1.2　形态和分类 …… 9
1.3　基因组的形状和大小 …… 12
1.4　基因数目和组成方式 …… 15
1.5　碱基组成 …… 19
1.6　密码子的使用 …… 26
1.7　复制和表达 …… 28
第 2 章　真核生物 …… 34
2.1　引言 …… 34
2.2　真核生物的分类与时间尺度 …… 34
2.3　基因组的形状和大小 …… 37
2.4　碱基组分 …… 42
2.5　复制、修复和重组 …… 44
2.6　基因表达 …… 48
2.6.1　转录与转录后调控 …… 48
2.6.2　遗传密码和密码子的使用 …… 54
2.6.3　翻译与翻译后修饰 …… 55
2.7　完整测序的真核生物基因组 …… 58
2.7.1　酿酒酵母基因组 …… 60
2.7.2　粟酒裂殖酵母基因组 …… 63
2.7.3　线虫基因组 …… 63

2.7.4 黑腹果蝇基因组 …… 65
2.7.5 拟南芥基因组 …… 67
2.7.6 水稻基因组 …… 69
2.7.7 人基因组 …… 70
第3章 细胞器 …… 77
3.1 线粒体 …… 77
3.1.1 一般结构和功能 …… 77
3.1.2 DNA与遗传系统 …… 79
3.1.3 基因组的特征 …… 88
3.2 叶绿体和其他质体 …… 114

中篇 方法学

第4章 基因组学中的分子生物学技术 …… 121
4.1 基因组DNA测序 …… 121
4.1.1 DNA测序技术 …… 121
4.1.2 人类基因组计划 …… 124
4.2 转录组分析 …… 125
4.2.1 基因表达分析 …… 125
4.2.2 表达序列标签 …… 126
4.2.3 基因表达序列分析 …… 127
4.2.4 差异显示 …… 129
4.2.5 表现度示差分析 …… 130
4.2.6 DNA微阵列 …… 131
4.3 蛋白质组分析 …… 137
4.3.1 二维凝胶电泳 …… 137
4.3.2 蛋白质鉴定 …… 139
4.3.3 蛋白质-DNA和蛋白质-蛋白质相互作用的研究 …… 141
4.3.4 生物芯片法分析蛋白质组 …… 143
第5章 基因组时代的生物学数据库 …… 145
5.1 引言 …… 145
5.2 基础数据库和专业化的数据库 …… 145
5.3 数据库结构 …… 148
5.4 数据库链接与互通性 …… 150
5.5 数据库注释 …… 151
5.6 检索系统 …… 154
5.6.1 SRS …… 154
5.6.2 Entrez …… 155

5.6.3 其他检索系统 …… 155
5.7 核酸数据库 …… 155
5.8 蛋白质数据库 …… 156
5.9 其他蛋白质数据库 …… 156
5.10 基因组数据库和资源 …… 159
5.11 基因数据库与资源 …… 163
5.12 转录组数据库 …… 164
5.13 代谢组数据库 …… 165
5.14 突变数据库 …… 167
5.15 线粒体数据库和资源 …… 168
第6章 基因组序列分析的计算方法 …… 171
6.1 引言 …… 171
6.2 点阵图 …… 172
6.3 两序列比对 …… 175
6.3.1 Needleman-Wunsch 全局比对算法 …… 175
6.3.2 Smith-Waterman 算法 …… 176
6.3.3 cDNA 以及基因组 DNA 序列的比对 …… 179
6.3.4 基因组比对 …… 179
6.3.5 清除序列库中的冗余序列 …… 181
6.3.6 同源序列相似度的计算 …… 185
6.4 数据库搜索 …… 189
6.4.1 FASTA …… 190
6.4.2 BLAST …… 192
6.4.3 BLAST 和 FASTA 程序包 …… 198
6.4.4 过滤不需要的序列匹配块 …… 200
6.4.5 重复序列匹配的过滤 …… 201
6.4.6 比对分值的统计学意义 …… 201
6.5 多序列比对 …… 205
6.6 可识别远源相关蛋白或蛋白质模块的比对谱 …… 209
6.7 序列的组装方法 …… 213
6.7.1 序列的净化 …… 213
6.7.2 序列的聚类 …… 214
6.7.3 构建一致序列 …… 215
6.7.4 用电子 PCR 绘制序列图谱 …… 217
6.7.5 基因组和 EST 项目的序列组装 …… 217
6.7.6 构建用于基因索引的组装序列 …… 219
6.8 生物序列的语言学分析 …… 222
6.8.1 马尔可夫链式生物序列 …… 224

6.8.2 生物序列语言的复杂度 …… 225
6.8.3 基因组重复序列的识别 …… 228
6.8.4 生物序列中的模式搜索 …… 229
6.8.5 识别染色体序列中的启动子区域 …… 234
6.8.6 发掘识别基因调控元件和蛋白质模块组件的模式 …… 237
6.8.7 基因预测 …… 239
6.8.8 基因组序列中 CpG 岛的识别 …… 243
6.8.9 密码子使用策略分析 …… 244
6.9 RNA 二级结构预测 …… 245
6.10 蛋白质序列分析 …… 252
6.10.1 蛋白质一级序列的分析 …… 253
6.10.2 预测跨膜蛋白螺旋 …… 258
6.10.3 蛋白质信号肽的识别及亚细胞定位的预测 …… 259
6.10.4 蛋白质二级结构的预测 …… 265
6.10.5 预测卷曲螺旋和螺旋-转角-螺旋结构 …… 270
6.10.6 蛋白质三级结构预测 …… 270
6.10.7 蛋白质折叠的识别与分类 …… 273
6.10.8 蛋白质功能预测的基因组比较工具 …… 273
6.11 进化与系统发育分析 …… 274
6.11.1 同源序列间遗传距离的评价 …… 276
6.11.2 分子系统发育学 …… 278

下篇 比较基因组学

第 7 章 分子进化 …… 293
7.1 引言 …… 293
7.2 基因组尺寸的进化 …… 294
7.3 碱基组分在进化中的作用 …… 297
7.4 原核基因组的进化 …… 301
7.5 从原核生物到真核生物 …… 302
7.5.1 真核细胞的起源 …… 302
7.5.2 线粒体基因组的进化 …… 304
7.5.3 质体的起源与进化 …… 309
7.6 从单细胞到多细胞 …… 312
7.7 核基因组的进化 …… 312
7.7.1 内含子 …… 313
7.7.2 基因数目和蛋白质数目 …… 313
7.7.3 非编码元件 …… 314

7.7.4 基因家族的扩展 …… 315
7.7.5 基因组倍增 …… 320
7.7.6 结论 …… 321
第8章 分子系统发育学 …… 322
8.1 引言 …… 322
8.2 分子钟 …… 322
8.3 相似性测量：直系同源与旁系同源 …… 324
8.4 基因组学时代的分子系统发育学 …… 326
8.5 远源物种间的相互关系：进化树 …… 328
8.6 后生动物的系统发育 …… 329
8.6.1 细胞器分类学与核分类学 …… 329
8.6.2 哺乳动物系统发育 …… 330
附录 本书引用的URL …… 333
参考文献 …… 335
索引 …… 384
译后记 …… 390

上篇　基因组特性

上篇　基因组特性

第 1 章
原核生物

1.1 引言

就在我们编写本书的同时，人们正以每月两个细菌基因组的速度完成测序工作，所产生的原核生物基因组序列可以从美国国家生物技术信息中心（NCBI）的网站上获得。本章涉及的数据都是已完成测序的全基因组序列，见表 1.1。很显然，到本书面世的时候，肯定会有更多的基因组完成了测序，并且某些观点也可能随之而发生改变。可以想象，如果技术取得巨大进步的话，我们目前所拥有的知识体系将会发生革命性的变化。

表 1.1 列出了迄今为止已完成测序的原核生物基因组，包括它们的名称和属性，如 EMBL 数据库索引号、大小、形状、是否存在染色体外遗传因子以及参考文献。从表中可以看出选择测序物种的原因是多种多样的，主要取决于它在基础科学或应用科学研究中的地位，如：对系统发育、代谢机制（主要是古细菌）研究的重要性，作为人类或动物病原体以及用于工业化制酶的重要性。换句话说，优先对那些目前熟知或者应用前景广泛的物种进行测序，而从系统发育的角度看，这种选择是非常随意的。

我们当前正处于基因组时代的起步阶段，尽管完成测序的物种数目非常少，但它们已经显示出惊人的迹象。本章将介绍测序方面的重要成就，这些成就为我们提供了原核生物基因组的新知识，也促进了相关领域研究方法的发展。本章还包括原核生物的一些基本知识，如形态、分类，以及遗传物质的结构、复制和表达等主要性质。当然，这些描述较为简短，因为我们侧重的是全基因组测序方面的重要问题。希望读者进一步参考该领域更深入的研究工作以及各种综述和论文。

表 1.1 已完成测序的原核生物基因组

物种	染色体		染色体外遗传因子		文献
	索引号	大小/bp	索引号	大小/bp	
古细菌(Archaea)					
敏捷好氧炽热球菌(*Aeropyrum pernix*)	BA000002	1669695			Kawarabayasi, Hino 等(1999)
闪烁古球菌(*Archaeoglobus fulgidus*)	AE000782	2178400			Klenk, Clayton 等(1997)
盐杆菌 NRC-1(3 个染色体)(*Halobacterium* sp. NRC-1)	AE004437 AE004438 AF016485	2014239 365425 191346	AF016485 AE004438	191346 365425	Ng, Kennedy 等(2000)
热自养甲烷杆菌(*Methanobacterium thermoautotrophicum*)	AE000666	1751377			Amith, Doucette-Stamm 等(1997)
詹氏甲烷球菌(*Methanococcus jannaschii*)	L77117	1664970	L77118 L77119	58407 16550	Bult, White 等(1996)
坎氏甲烷球菌 AV19(*Methanococcus kandleri* AV19)	AE009439	1694969			Slesarev, Mezhevaya 等(2002)
嗜乙酸甲烷八叠球菌 C2A(*Methanosarcina acetivorans* str. C2A)	AE010299	5751492			Galagan, Nusbaum 等(2002)
马氏甲烷八叠球菌 Goe1(*Methanosarcina mazei* Goe1)	AE008384	4096345			Deppenmeier, Johann 等(2002)
嗜气热棒菌(*Pyrobaculum aerophilum*)	AE009441	2222430			Fitz-Gibbon, Ladner 等(1998)
深海热球菌(*Pyrococcus abyssi*)	AL096836	1765118			Lecompte, Ripp 等(2001)
强烈热球菌 DSM3638(*Pyrococcus furiosus* DSM3638)	AE009950	1908256			Robb, Maeder 等(2001)
掘越热球菌(*Pyrococcus horikoshii*)	AP000001- AP000007	1738505			Kawarabayasi, Sawada 等(1998)
硫黄矿硫化叶菌(*Sulfolobus solfataricus*)	AE006641	2992245			She, Singh 等(2001)
常田硫化叶菌(*Sulfolobus tokodaii*)	BA000023	2694765	AJ010405	41229	Kawarabayasi, Hino 等(2001)
嗜酸热原体(*Thermoplasma acidophilum*)	AL445063- AL445067	1564905			Ruepp, Graml 等(2000)
火山热原体(*Thermoplasma volcanium*)	AP000991- AP000996	1584799			Kawashima, Amano 等(2000)

续表

物　种	染色体		染色体外遗传因子		文　献
	索引号	大小/bp	索引号	大小/bp	
细菌(Bacteria)					
根癌土壤杆菌 C58 (Cereon) [*Agrobacterium tumefaciens* str. C58 (Cereon)]	AE007869	2841581			Goodner, Hinkle 等(2001)
根癌土壤杆菌 C58 (U. Washington)[*Agrobacterium tumefaciens* str. C58 (U. Washington)]	AE008688	2841490	AE008687	542780	Wood, Setubal 等(2001)
			AE008690	214234	
风产液菌(*Aquifex aeolicus*)	AE000657	1551335	AE000667	39456	Deckert, Warren 等(1998)
耐盐芽孢杆菌(*Bacillus halodurans*)	BA000004	4202353			Takami, Nakasone 等(2000)
枯草芽孢杆菌(*Bacillus subtilis*)	AL009126	4214814			Kunst, Ogasawara 等(1997)
伯氏疏螺旋体(*Borrelia burgdorferi*[①])	AE000783	910725	AE000791	9386	Fraser, Casjens 等(1997);
			AE000792	26498	Casjens, Palmer 等(2000)
			AE001575	30750	
			AE001576	30223	
			AE001577	30299	
			AE001578	29838	
			AE001579	30800	
			AE001580	30885	
			AE001581	30651	
			AE001583[①]	5228	
			AE000793[①]	16823	
			AE001582[①]	18753	
			AE000785[①]	24177	
			AE000794[①]	26921	
			AE000786[①]	29766	
			AE000784[①]	28601	
			AE000789[①]	27323	
			AE000788[①]	36849	
			AE000787[①]	38829	
			AE000790[①]	53561	
			AE001584[①]	52971	

续表

物　种	染色体		染色体外遗传因子		文　献
	索引号	大小/bp	索引号	大小/bp	
细菌(Bacteria)					
马尔他布鲁菌(*Brucella melitensis*)	AE008917	2117144			DelVecchio, Kapatral 等(2002)
巴克纳菌 APS (*Buchnera* sp. APS)	AP000398	640681	AP001070	7258	Shigenobu, Watanabe 等(2000)
			AP001071	7786	
空肠弯曲杆菌(*Campylobacter jejuni*)	AL111168	1641481			Parkhill, Wren 等(2000)
新月柄杆菌(*Caulobacter crescentus*)	AE005673	4016947			Nierman, Feldblyum 等(2001)
肺炎衣原体 AR39(*Chlamydia pneumoniae* AR39)	AE002161	1229853		4524	Read, Brunham 等(2000)
肺炎衣原体 CWL029(*Chlamydia pneumoniae* CWL029)	AE001363	1230230			Kalman, Mitchell 等(1999)
沙眼衣原体 MoPn(*Chlamydia trachomatis* MoPn)	AE002160	1069412	AE002162	7501	Read, Brunham 等(2000)
沙眼衣原体血清型 D(*Chlamydia trachomatis* serovar D)	AE001273	1042519			Stephens, Kalman 等(1998)
肺炎嗜性衣原体 J138(*Chlamydophila pneumoniae* J138)	BA000008	1226565			Shirai, Hirakawa 等(2000)
丙酮丁醇梭菌(*Clostridium acetobutylicum*)	AE001437	3940880	NC_001988	192000	Nolling, Breton 等(2001)
产气荚膜梭菌(*Clostridium perfringens*)	BA000016	3031430			Shimizu, Ohtani 等(2002)
谷氨酸棒杆菌(*Corynebacterium glutamicum*)	AX114121	3309400			Tauch, Homann 等(2002)
耐辐射异常球菌 R1(2 个染色体)[*Deinococcus radiodurans* R1 (2 chromosomes)]	AE000513	2648638	AE001826	177466	White, Eisen 等(1999)
	AE001825	412348	AE001827	45704	
大肠杆菌 K-12(*Escherichia coli* K-12)	U00096	4639221			Blattner, Plunkett 等(1997)
大肠杆菌 O157：H7 EDL933(*Escherichia coli* O157：H7 EDL933)	AE005174	5528970			Perna, Plunkett 等(2001)
大肠杆菌 O157：H7 Sakai(*Escherichia coli* O157：H7 Sakai)	BA000007	5498450	AB011549	92721	Hayashi, Makino 等(2001)
			AB011548	3306	
具核梭杆菌具核亚种 ATCC 25586(*Fusobacterium nucleatum* subsp. nucleatum ATCC 25586)	AE009951	2174500			Kapatral. Anderson 等(2002)
流感嗜血杆菌 Rd(*Haemophilus influenzae* Rd)	L42023	1830138			Fleischmann, Adams 等(1995)
幽门螺杆菌 26695(*Helicobacter pylori* 26695)	AE000511	1667867			Tomb, White 等(1997)
幽门螺杆菌 J99(*Helicobacter pylori* J99)	AE001439	1643831			Alm, Ling 等(1999)
乳酸乳球菌乳亚种(*Lactococcus lactis* subsp. *lactis*)	AE005176	2365589			Bolotin, Wincker 等(2001)
无害李斯特菌(*Listeria innocua*)	AL592022	3011208	AL592102	81900	Glaser, Frangeul 等(2001)
单核细胞增生李斯特菌 EGD-e(*Listeria monocytogenes* EGD-e)	NC_003210	2944528			Glaser, Frangeul 等(2001)
百脉根中慢生根瘤菌(*Mesorhizobium loti*)	BA000012	7036074	AP003015-16	351911	Kaneko, Nakamura 等(2000)
			AP003017	208315	