现代生物技术前沿

COMPARATIVE GENOMICS

〔美〕 M.克拉克 主编

邱幼祥 高翔 等 译

比较基因组学

科学出版社
www.sciencep.com

现代生物技术前沿

COMPARATIVE GENOMICS

〔美〕 M.克拉克 主编

邱幼祥 高 翔 等 译

比较基因组学

科学出版社

北 京

图字:01-2003-5019 号

内 容 简 介

本书讲述了各种模式生物基因组在人类基因组研究与人类基因鉴定中发挥的重要作用。介绍了研究模式生物基因组的不同方法,如突变研究、cDNA 表达图的构建、原位杂交和比较基因组(在基因和核苷酸水平)分析。这一切都将有助于我们对人类基因组的认识。在探索所有生物基因组进化的连续性和变异性、生物之间的亲缘关系、阐明基因结构和演化、调控序列的结构和功能、内含子的进化以及基因功能方面,比较基因组学更有其特殊的价值。

本书适合高等院校遗传学、分子生物学、生物化学等相关专业的学生、教师及科研人员参考。

Melody S. Clark
Comparative Genomics

图书在版编目(CIP)数据

比较基因组学/(美)M. 克拉克等主编;邱幼祥,高翔等译. —北京:科学出版社,2007

ISBN 978-7-03-019430-5

Ⅰ. 比… Ⅱ. ①克…②邱…③高… Ⅲ. 动物-基因组-对比研究 Ⅳ. Q 953

中国版本图书馆 CIP 数据核字(2004)第 033495 号

责任编辑:王海光 王 静 陈欣然 马学海/责任校对:李奕萱
责任印制:钱玉芬/封面设计:陈 敬

科学出版社出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2007 年 7 月第 一 版 开本:B5(720×1000)
2007 年 7 月第一次印刷 印张:14
印数:1—3 000 字数:273 000

定价:38.00 元

(如有印装质量问题,我社负责调换〈环伟〉)

参 编 者

Clark, Melody S. Fugu Genomics HGMP Resource Centre, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SB.
email : mclark@hgmp.mrc.ac.uk.

Elgar, Greg. Fugu Genomics HGMP Resource Centre, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SB.
email: gelgar@hgmp.mrc.ac.uk

Fröniecke, Lutz. National Institute of Health, National Cancer Institute, Basic Science Laboratory, Frederick, MD 21702. USA.
email: froenickel@ncifcrf.gov

Hertzog, Paul. Centre for Functional Genomics and Human Disease. Institute of Reproduction and Development, Monash University. 27-31 Wright Street, Clayton, Vic. 3168. Australia.
email: paul.hertzog@med.monash.edu.au

Jäckle, Herbert. Institut für biophysikalische Chemie, Abteilung Molekulare Entwicklungsbiologie, Am Fassberg 11, D-37077 Göttingen, Germany.
email: hjaeckl@gwdg.de

Jeffrey, William R. Department of Biology, University of Maryland, College Park, MD 20742-4415. USA.
email: wj33@umail.umd.edu

Kola, Ismail. 7245-24-110, Pharmacia and Upjohn, 301 Henrietta Street, Kalamazoo, MI 49007. USA
email: Ismail.kola@am.pnu.com

Lazner, Francesca. Centre for Functional Genomics and Human Disease. Institute of Reproduction and Development, Monash University. 27-31 Wright Street, Clayton, Vic. 3168. Australia
email: Francesca.cristiano@med.monash.edu.au

Lipkin, Ehud. Department of Genetics, The Hebrew University of Jerusalem, 91904 Jerusalem. Israel.
email: lipkin@vms.huji.ac.il

Marshall Graves, Jennifer. A. Department of Genetics and Evolution, La Troube University, Melbourne, Victoria 3083, Australia.
email: genjmg@plasmid.gen.latrobe.edu.au

Schäfer, Ulrich. Institut für biophysikalische Chemie, Abteilung Molekulare Entwicklungsbiologie, Am Fassberg 11, D-37077 Göttingen, Germany.
email: uschaef@gwdg.de

Shetty, Swathi. Department of Genetics and Evolution, La Troube University, Melbourne, Victoria 3083. Australia.
email: swathi@gen.latroube.edu.au

Soller, Morris. Department of Genetics, The Hebrew University of Jerusalem, 91904 Jerusalem. Israel.
email: soller@vms.huji.ac.il

Stanyon, Roscoe. National Institute of Health, National Cancer Institute, Basic Science Laboratory, Frederick, MD 21702. USA.
email: stanyonr@ncifcrf.gov

Wienberg, Johannes. National Institute of Health, National Cancer Institute, Basic Science Laboratory, Frederick, MD 21702. USA.
email: wienbergj@ncifcrf.gov

Wilson, Trevor. Centre for Functional Genomics and Human Disease. Institute of Reproduction and Development. Monash University. 27-31 Wright Street, Clayton, Vic. 3168. Australia.
email: Trevor.Wilson@med.monash.edu.au

目　录

1 比较基因组学介绍：测序计划和模式生物

Melody S. Clark，Fugu Genomics，HGMP Resource Centre，Wellcome Genome Campus，Hinxton，Cambridge，CB10 1SB，UK.

1.1 介绍：什么是比较基因组学

从字面上看，一言以蔽之，所谓比较基因组学就是将基因组进行比较的学科。人们马上会想到 DNA 和蛋白质序列，也会不可避免地想到要与人类基因组进行比较。然而比较基因组学远远不止这些。它还用于各种生物在任何层次上的比较：DNA 或蛋白质序列、基因在染色体上的定位和染色体图、功能和进化。其目的是破译基因有什么样的功能并且提出对于基因型和表型之间的联系的认识。它对于一整套遗传特征或疾病常常具有一些特殊的参考价值，因为这对于争取研究基金资助具有相当大的吸引力（当人们的研究数据进入到试验的方程式以求验证时，更是如此）。诸如牛、羊、猪、鱼等这些家畜，对于任何一个国家来说都具有巨大的经济利益，能够认识它们的优良性状甚至疾病的遗传模式都将是显而易见的商业需要。然而，只有通过学术上的或是“基础”研究的广泛试验，商业上的应用才能得到巩固加强。

当人们在研究计划上百花齐放、各抒己见时，并不总是能够决定究竟哪一个生物以及在哪个生物内有哪一套基因或遗传特征应该被研究。并不是所有的生物都经得起试验的检验，人类就是一个典型例子！不能什么事情都拿人来做试验。这样就有了“模式生物”这一概念。这个术语对自身含义做了自我解释，而且有越来越多的模式生物用来研究基因如何起作用，诸如控制序列的复杂因子之间如何相互作用、相邻基因环境、非编码序列（重复序列、反转录因子）和生物周围大环境的重要性。例如，在小鼠身上进行转基因，在酵母、线虫、果蝇、斑马鱼中进行突变研究，在家畜中进行数量性状的分析，在河豚鱼中识别进化上保守的控制因子，对许多物种间的基因组重排进行总的比较，这样的例子不胜枚举。随着世界范围测序能力的持续增加和高通量功能分析日新月异的发展，这一切都将证明，比较这一做法越来越重要，既表现在测序比较也表现在功能的生物学模型的应用方面。

这一章将集中对比较基因组学的内容做一介绍。其目的是对此给予一个简短的总揽全局的概括，主要集中在一些项目，诸如基因组测序计划和对于模式生物各式各样的利用，这些能够帮助解读基因的功能和进化。在本书的其他章节所涉及的这个新兴学科的研究范围、研究目的和研究手段会变得多样化，这反映了对比较基因组学这个主题进行研究的多种多样的途径。比较基因组学不但会告诉我们许许多多人类基因作用机制，而且还将告诉我们在许多其他生物中基因型—表型的联系以及进化的过程。

到了 2003 年，这本书已经出版，那时也将获得人类基因组的全部序列，所以，这就是比较基因组学和模式生物为什么如此重要的原因。

1.2 人类基因组序列和比较基因组学的历史

今天的社会是一个瞬息万变的社会，机会稍纵即逝。科学突破经常在媒体上被描绘成仿佛只是一夜之间发生的事情。但在实际生活当中，未必如此。同样，人类基因组测序也是一样。序列资料本身就是全世界的实验室，其中包括许多高度专业化的实验室通力协作的结果，其基础已经集上百年之大成，而且是在许多不同的科学分支上集合而成的。

很难确定比较基因组学的精确起点，但是，把施莱登（Matthius Scleiden，植物学家）和施旺（Theodor Schwann，生理学家）在 1830 年对于细胞理论最初的确立作为起始点的说法是恰当的。39 年后，化学家 Friedrich Miescher 分析了细胞提取物，并且发现其包含有蛋白质和一个不平常的含有磷酸的复合物，他把它称为核素（nuclein，这就是我们今天所知道的核酸）。第一个描述染色体的人据说是佛莱明（Flemming，1843～1905），当时他在研究蝾螈。和他同时代的孟德尔（Mendel），被人尊称为“遗传学之父”，作为一个神父，他业余时间是个园艺工作者。对孟德尔研究工作的再发现及证实，预示了 20 世纪早期遗传学领域中的一个大爆发现象的来临。当时，有许许多多的证实工作是在植物、蜻蜓和海鞘上完成的。这些生物数量稳定，保证供给，而且易于操作（模式生物的首要特征）。因此，尽管“模式生物”的概念和比较基因组学仍然有很长的路要走，但遗传学的这些特殊方面的应用从一开始就非常明确。1911 年，伴随着包含有 5 个基因的首张连锁图的文章的发表，所有模式生物的前辈——果蝇亮相了。T. H. Morgan 当时可能并没有意识到他的工作为果蝇基因组测序计划奠定了基础，而这个项目在 89 年之后才得以实施完成。

令人吃惊的是，“基因组”并不是个眼下才时髦的词汇，而是由 Winkler 在 1920 年就提出并发展起来的名词。当然它可能和现在的含义并非完全吻合，其原因在于在 1944 年之前一直没有发现 DNA 是细胞中的遗传物质。直到那时人们还坚信遗传物质必定是蛋白质，因为蛋白质是化学上复杂的化合物，核酸是简

单的，基因是复杂的，因此基因一定来自蛋白质。令人不可思议的是，遗传学的发展在20世纪后半期加快了速度。一个里程碑式的事件是1953年由Waston和Crick发现了DNA双螺旋结构，它的重要性远远超过了1956年由蒋友兴和Levan证实的“正常人具有46条染色体”这一结论（结束了多年对染色体数目为16～40条的争论）。1977年，进行了第一次人类基因的克隆，而在22年后的1999年，完成了第一条人类染色体的完整测序（Dunham et al. 1999）。

可以肯定地说，在过去的50年中，遗传学已经变成一门日益专业化的学科，在分子生物学中确实如此。然而，现在既然有了这么多可以利用的测序资料，那么工作的重点将转移到确定基因的功能上来，这将需要更多的多学科技术手段和对于“生物学”更广泛的评价。模式生物和比较基因组学研究硕果累累，成绩显著，这一切可以在本书后面的每一章节中通过对于它们所描述的广泛的应用而得到确证。

1.3 人类基因组序列

人类基因组测序在科学史上是一件奇妙卓越的丰功伟绩。它代表了全世界上百个实验室多年工作的最高水平。在最后的冲刺阶段，在私人公司塞莱拉和由美国NIH（美国国立卫生院）和英国的Wellcome托拉斯所领导的公众基金团体之间的完成测序工作的竞赛已经引起了强烈的社会反响，使遗传学成为令人瞩目的焦点。虽然媒体宣扬人类基因组序列对于普通大众来说具有非常重要的意义，而且大量涉及保险的令人生畏的事件广泛盛行，但是对于科学，人类基因组将是一个巨大的资源。它是脊椎动物中的第一个可利用的全基因组序列；它可以公开获得而且将为比较研究提供一个可参照的基因组。这还仅仅是个开始，属于不同生物和人类不同种族的其他的基因组将纷至沓来，这一切将告诉我们许许多多关于基因排列次序的重要性和不同种类之间、多态性之间的具体情况，以及在生物之间或其内部的深刻内容。而在过去一段时间变得似乎稍稍不合乎潮流的进化和数量遗传学以及生物学的众多领域，又将回到众人瞩目的中心。

1.4 序列与功能一致吗？

1.4.1 “未知”基因

我看到在一些大众出版物中，把人类基因组测序看成试图去阅读一部32册的大不列颠百科全书或是一本没有段落、标题或标点符号的圣经。这在某种程度上是件困难的事情！其实这不完全正确。基因预测计划已经有了很大发展，这种预测具有生物特异性，能够高效地识别假定的外显子和（或）基因（Claverie 1997，Burset and Guigo 1996）。在数据库中快速地一瞥将发现许多线虫的基因都标注为质粒ID，因此是“假定的基因”。这其中，一些表现出与数据库中其他具

有特征描述的基因的序列相似性，而且因此能够不被描述成一个“假定的”功能或被划归到一个基因家族中，这是一个开始，但是在此之后还会有多少情况我们不知道呢（表 1.1）?

表 1.1 已完成测序的基因组大小、预测的基因（ORF：可读框）数目和在数据库中的未知基因的百分率

生物	基因组大小/Mb	预测了的可读框	未知基因的百分率/%	参考文献
外生殖器支原体（*Mycoplasma genitalium*）	0.58	470	20	Fraser et al. 1995
嗜血流感病毒（*Haemophilus influenzae*）	1.83	1743	40	Fleischmann et al. 1995
大肠杆菌（*Escherichia coli*）	4.63	4288	38	Blattner et al. 1997
酿酒酵母（*Saccharomyes cerevisiae*）	12.1	6034	25	Botstein et al. 1997
线虫（*Caenorhabditis elegans*）	97	19 099	24	线虫测序协作组 1998
果蝇（*Drosophila melanogaster*）	120*	13 600	23	Adams et al. 2000

*参考该基因组测序的常染色体部分，而且不包括异染色质 DNA 中的另外的 60Mb。

这个表格简化了这种情况，即乍一看，好像在原核生物大肠杆菌（*E. coli*）和嗜血流感病毒（*H. influenzae*）的“未知基因”要比在线虫和果蝇的多。然而，值得注意的是这些数据是在该基因组刚刚发表时所得到的，而确定基因功能要追溯到当数据库还特别小而且有关功能的分析才刚刚起步的时候。

对于真核生物资料的更为接近的检查勾勒了更为详尽的图画。当酵母的基因组测序发表时，它的 60%基因并没有通过实验确定其功能。然而，其中的大部分表现出一些序列相似性或具有可能功能的基序（motif），剩下有差不多 25%没有任何线索（Botstein et al. 1997），因此这 25%记录在这个表格（表 1.1）中。线虫中所预测的基因中的 42%具有跨门类之间的匹配，其中绝大多数具推测性的功能信息。另有 34%只与其他的线虫序列相匹配（线虫测序协作组 1998）即大概是线虫或 *C. briggsae* 的 cDNA，其中少数已经具有功能性的特征，因此更恰当的“未知”基因的数量，估计应当是 58%。至于果蝇，大约有 23%的预测基因在已知数据库中没有相匹配部分，另外的 27%只与已表达序列标志（expressed sequence tag，EST）相当（Adams et al. 2000），其中的许多还没有得到很好的研究。所以，再一次讨论那些已知功能资料时，50%这样的修正数值大概更合适、更准确一些。人类基因组测序的状况与其他真核生物测序状况相类似，而且面临对假定的基因确定功能，仍然要有更多的工作去做。

1.4.2 可变剪接

基因鉴别只是测定功能的漫长道路中最初的一部分。能够预测基因和构象的计算机程序常常既可以通过对已表达序列标志数据库的搜寻，又可以通过对cDNA文库的筛选来加以证实。一个已表达序列标志序列的匹配证实了一个“假定的”基因是“真正的”基因。这些已表达序列标志序列常常只代表了一个cDNA克隆的不完整的单通道测序。显然，能够通过测序整个克隆来对结构进一步证实，但是这里另一个因子——可变剪接因子也参与其中。

目前的资料预测果蝇有 13 601 个基因，而这个数字还低于预测的线虫基因数（19 099 个）。然而，当前关于 cDNA 的资料表明尽管在果蝇基因组中只有13 601个基因，但是这些基因通过可变剪接至少编码 14 113 个转录产物，而且这个转录产物的数字实际上也被认为是低估了（Adams et al. 2000）。这种现象不仅仅在果蝇中发生。例如，*WT1* 基因，它参与了哺乳动物泌尿生殖系统的发育，在人类中至少编码 16 个不同的蛋白质同等型（Hastie 1994）。至此，据估计超过 30%的人类基因受到可变剪接的影响（Hanke et al. 1999，Mironov et al. 1999），这种情况还会由于可能出现的翻译后修饰而进一步复杂化，尚没有图表可以对此加以说明（Bork 2000）。*PTHrP* 基因（在河豚鱼章节中会更多谈到）在人类中有三个同等型（Yasuda et al. 1989a，Mangin et al. 1989），但是在其他哺乳动物和河豚鱼中却只有一个（Mangin et al. 1990，Yasuda et al. 1989b，Thiede and Routledge 1990，Power et al. 2000）。对于同等型世代在功能和进化上所起的作用才刚刚开始进行探索。这是由于在世界范围内不断增加的测序能力以及已表达序列标志测序项目的普及化的结果。翻译后修饰的问题将随着蛋白质功能研究的不断增加而大量涌现。

因此，对于“序列与功能一致吗?”这样一个问题，答案是明确的：“不是”。在我们有关基因功能的知识中，甚至是对于完整基因组的序列以及来自其他轶事一样的证据仍然有许多空白之处（Bork et al. 1998）。我们现在确定功能的能力还非常依赖于数据库注释以及计算机推测程序，特别是对于大量的测序资料的例行注释。Bork（2000）曾经估计对于序列的特征确定大约有 70%的精确度。这里主要的问题在于可利用的序列数据总量和蛋白质的试验性特征描述之间的差距在加大。序列数据只能够在一定程度上揭示基因功能，对于蛋白质的特征描述和试验有待于付出更多的努力。

1.5 基因组的非编码部分

人类基因组测序中的最大优点之一是它将包括所有的非编码序列。对于绝大多数生物来说，数据库中的序列数据的主要部分是以 cDNA 的形式存在，cDNA 是基因组的已表达部分。编码序列是重要的，同时在基因组的非编码区域发现有

调控因子。基因的预测程序是很先进的，但是设计解读调控因子和启动子区域以及5′和3′端未翻译区域（UTR）的程序才刚刚起步（Fickett and Hatzigeorgiou 1997）。

从基因组中对于基因控制所需要的DNA总量进行消减，仍然留下了没有功能作用的大部分，到现在对于这部分的了解仍然很少。尽管重复元件已经被无休止地分类，但被确定具有功能性意义的重复元件的数量还是为数很少。不稳定的三联体重复序列与亨廷顿氏病（Huntington's disease）和肌强直性肌肉营养不良这样一些遗传性疾病有关（Caskey et al. 1992）。现在了解到许多重复元件是属于反转录病毒起源的。这些反转录因子中的一些与基因组的进化以及基因组可塑性（Pickeral et al. 2000）密切相关。对于与主要组织相容性复合体（MHC）区域相关的方面已经进行了最深入细致的研究，而这一区域被认为是具有引起基因重排的能力，在进化中起了十分显著的作用（Abdulla et al. 1996，Kulski et al. 1997，Dawkins et al. 1999）。

许多人在一种“生物体越复杂，核DNA含量就越高”的误解中工作。虽然这种看法为当前的测序计划所支持，但仍然有相当多的、具有大量而且分散的基因组的生物与这些情况不同（表1.2）。为什么百合（lily）的一个特定种类具有比人类多15倍的DNA呢？所有的这些“额外的”基因的显著特点是什么？这种现象被称为“C值悖理”。直到现在我们才到达能够开始解答这个谜团的阶段。

表1.2 各种各样真核生物的多倍体染色体数量和DNA含量（仿自 Clark and Wall，1996）

种名	俗名	一个细胞核中DNA含量/pg	*n*
Frittillaria davisii	百合	98.4	12
Protopterus	肺鱼	50	19
Avena sativa	橡树	21.5	21
Triticum aestivum	（面包用）小麦	18.1	21
Allium cepa	洋葱	16.8	8
Homo sapiens	人	3.7	23
Mus musculus	小鼠	2.5	10
Drosophila	果蝇	0.1	4
Arabidopsis thaliana	拟南芥（鼠耳草）	0.07	5
Saccharomyces ceravisiae	酵母	0.026	15

几个完整的可参照基因组的可利用性将把研究扩大到“非编码（或‘无用’）DNA”这个以前被忽视的领域。它提供了回答“这些剩下的 DNA 真正做什么?”这个问题的可能性。

1.6 基因组测序计划

准确地说，人类基因组测序计划现在是备受媒体注意的焦点。这是一项令人惊奇的计划。然而，使这项计划成为可能的技术是在复杂程度较低的基因组中得到发展的。第一眼看上去好像基因组测序正变得更常规化：基因组在线数据库(GOLD1.0)(Kyrpides 1999)(http://igweb.integratedgenomics.com/ GOLD/)中列举了所有已经完成和正在进行的基因组项目。到 2000 年 9 月 3 日，在数据库中已经有 25 个种类完成了基因组测序，还有 106 个原核生物以及 31 个真核生物正在测序之中。在基因组测序尤其是在真核生物基因组测序上的这样一个高潮，反映出全基因组鸟枪测序的成功，这项技术首先在嗜血流感病毒 *Haemophilium influenzae* 中加以报道（Fleischmann et al. 1995），而且现在正应用在更复杂的生物体上试验（Adams et al. 2000）。

1.6.1 原核细胞的测序计划

第一个被测序的微生物基因组是大肠杆菌 ΦX174 的基因组，它有 5386 个碱基对（Sanger et al. 1978）。令人吃惊的是，这是在 Sanger 关于双脱氧测序方法学的论文发表仅仅一年之后完成的。在当时，只是对相对比较小的病毒基因组使用了当时可以利用的技术方法来进行测序（人工放射性测序）。而直到 1995 年，人们才使用全基因组鸟枪测序技术，完成了第一个细菌的基因组，1.83Mb 的嗜血流感病毒株的测序（Fleischmann et al. 1995）。这个方法对于那些排列紧密、基因丰富、不包含内含子和大量重复 DNA 的小基因组非常有效，因此无需通过较大的嵌入文库来提供详细的遗传图或复杂的构建支架，也能够对那些小基因组进行装配。

这样大量的测序资料的出现已经改变了微生物研究的焦点和手段。外生殖器支原体（*Mycoplasma genitalium*）的测序（至今在各种自由生活的生物中所记载的最小的基因组，580 kb）已经明确地说明了一个自我复制细胞所需要的一整套基因的最少数量（Fraser et al. 1995）。测序这些“小”基因组相对简单易行，这一点说明了比较研究走在了真核生物研究的前面（Perrière et al. 2000）。对这些生物的深层认识包括许多商业和医学上的应用。原核生物进化过程和系统发生关系过程的评估能够被用作确定药物作用靶点的范围的工具（Allsop 1998）。比较研究能够阐明病理发生的分子机制：识别单个基因的功能并且决定基因如何相互作用来形成诸如毒力等复杂性状（Field et al. 1999）。原核生物测序项目的目标之一就是对感染株的一整套基因与一个毒性减弱的实验室株进行比较来检查导

致毒力和寄主特异性的因子（Saunders and Moxon 1998）。在人类基因组可利用的情况下，是可能评价寄主的遗传背景中的病原体的（Field et al. 2000）。家畜具有巨大的经济利益，认识寄主-微生物相互作用对于家畜的疾病来说也是很重要的。从商业角度来说，在这些不断增加的知识以外的有用的副产品应当是更精确的药物靶向作用和新疫苗的发展（Allsop 1998）。

从生物化学和遗传学来看，大肠杆菌和其他许多微生物已经被研究了 50 年之久。所有生物所共有的基本生物化学通路基本一致而且对这方面的许多认识都在细菌基因组上一一兑现。尽管在原核生物和真核生物的基因和基因结构之间还存在着许多不同，但在人类和大肠杆菌之间的比较分析仍然能够提供基因功能上的一些信息。尽管在脊椎动物和大肠杆菌之间的种间同源基因的数量极为有限，但是单一的蛋白质结构域是保守的。这一点和后生动物辐射进化相关的基因爆炸理论相一致，并且与为多细胞生物进化所必需的多结构域细胞外和细胞表面蛋白质的构建的理论相一致，也为外显子混编所支持（Patthy，1999）。这个功能域的保守性已经让人们对那些鲜有特征性表现的脊椎动物基因有了认识。许多定位克隆的基因编码较大的多功能域蛋白质，而其中的一些包含具有未知功能的假定酶域。使用细菌基因（Mushegian et al. 1997）的基序（motif）检测和结构建模已经揭示了假定的功能位点，这些功能位点在过去用常规所使用的方法是检测不到的。在 Werner 综合征（一种具有早熟衰老特征的疾病）中，鉴别出与一个核酸酶、一个 3′-5′校正阅读外切核酸酶以及一个解旋酶具有同源关系的三个域，从而证实了这个蛋白质可能参与 DNA 的修复和加工过程（Mushegian et al. 1997）。这为剖析人类疾病的精确的分子本质提供了一个切入点。

这些相对简单的生物对于我们了解遗传学和进化做出了许许多多的贡献。基因组测序计划和接下来的分析在卫生保健和预防医学领域中具有重大的意义。

1.6.2 不断前进中的真核生物测序计划

在正在进行的真核生物测序计划中，只有两种生物（小鼠和人）是属于脊椎动物的，而其他的进行过测序的动物之一——果蝇的基因组测序工作在这本书的汇编过程中已经完成（Adams et al. 2000），而且在下一章中要给予详细的讨论。其他的还有原生生物（*Cryptosporidium parvum*，*Giardia lamblia*，*Leishmania major* 等）、真菌类（*Pneumocystis carinii*，*Neurospora crassa* 等）和植物（拟南芥、水稻、玉米等）基因组在线数据库（GOLD 1.0） （Kyrpides 1999）（http://igweb.integrategenomics.com/GOLD/）。在原生动物和真菌测序计划背后的原因与原核生物相似，是了解病理形成和疾病控制。而对植物基因组进行测序有着巨大的经济意义。

植物基因组学

在许多方面，植物基因组测序也类似于原核生物测序计划，因为它们倾向于

被划分为一个单独的领域，与更具声望的脊椎动物项目互不重叠。然而，植物基因组学有许多有意义的东西要提供。如果没有植物基因组学作为其一部分，比较基因组学也将变得不完整。植物基因组学的测序焦点是有着 120Mb 的最小基因组的拟南芥（*Arabidopsis thalina*）。农作物一般具有相对比较复杂一点的基因组，甚至比人类基因组还要大一些，例如大麦的单倍体含量为 5300 Mb。有几个机制已经用来解释一些植物的基因组的扩张，诸如基因组复制（小麦是六倍体）和重复元件的扩展，其中反转录因子起了相当大的作用（Bennetzen and Kellogg 1997）。

至今，拟南芥属（*Arabidopsis*）的 54.8 Mb 的测序工作已经完成。这些序列分布在它所有的 5 个染色体上，其中有 15.1Mb 的测序工作正处于完成的阶段。对 4 号染色体整个的序列所进行的详细分析说明，与其他测序的基因组相似的是，在拟南芥中只有 60%的基因具有已经确定的功能（Bevan et al. 1999）。所以序列产生与对于功能的理解之间的差距在植物中也同样具有。由拟南芥项目而产生的问题之一是一个生物学的问题。拟南芥是一种双子叶植物，而绝大部分的农作物是单子叶植物（小麦、大麦、水稻、玉米、高粱、橡树和甘蔗等）。因此，使用有赖于诸如基因共线性等因子的信息的这样一项技术的直接转移常常要通过水稻实现，水稻的基因组是较小的单子叶植物基因组（440Mb）之一。事实上水稻已经成为以日本为基地的一个首选测序工作的目标（Sasaki et al. 1996）而且计划把它变成植物中的第二个优秀模式基因组。

植物与动物相比较完全不同，但是从来自植物基因组测序计划的资料来看，植物就像一些原核生物那样，对于我们全面认识基因功能大有用处。由种间同源基因簇所构成的数据库中的植物基因内容能够帮助我们识别那些基于保守基序的基因功能，而且从多样的生物中汲取对基因功能的认识。这将把一个植物特异性生物功能的新领域添加到在其他生物的决定基因功能的过程中（Bevan and Murphy 1999）。

植物遗传学一直通过对大量物种在基础生物学、进化、适应及基因组研究等相当广泛的范围进行研究，并且把这些附加价值加到那些很少被选择做深入的基因组测序的那些种类上，从而使植物遗传学得到繁荣的发展。很明显，浏览现在的大量文献资料，可以注意到基本上所有的问题都是由动物比较遗传学家提出的（例如基因密度的同一性、基因组复制、同线性、保守的基因次序、种间同源基因的确定等），植物遗传学家也在详尽地研究这些情况。这两个领域并非泾渭分明，不容混淆，发现两者如何共同发展将会非常有趣。

事实上，Bennetz（1999）在他对于植物基因组学的观点中提出了一个植物比较基因组学的计划，对此，动物/脊椎动物遗传学家将会很好地予以考虑。他建议启动拟南芥和水稻这两种植物的基因组测序工作，它们将作为参照基因组，

并成为在所有植物中发现基因以及描述其特征的基础。一些物种的物理图谱将被构建起来而且他将这些种类命名为“节点”物种，因为它们具有相当小的基因组而且可以作为一个重要的具有系统发生多样化特征的植物家族的替代品。他使用高粱代替玉米，选择莲代替大豆来作为试验材料。较多种类将从属于中等程度（大约5000克隆，可能是50 000，译者注）的已表达序列标志项目，而这种方法对于基因发现以及等位基因多样性调查来说是最为经济的途径。此外，这些已表达序列标志将为基因表达的精确的DNA芯片分析提供种属特异性序列。

并不是所有的植物和动物种类都能够测序，但是可以肯定，通过对于少数种类的全基因组测序可能是最经济的途径，对那些具有经济价值或者在进化中占据枢纽位置的种类开展高密度图以及已表达序列标志的工作。

1.6.3 已经完成的真核细胞基因组测序项目：酵母（*Saccharomyces cerevisiae*）和线虫（*Caenorhabditis elegans*）

这两种真核生物的基因组都已完全测序（酵母基因组测序于1996年完成，而线虫基因组于1998年完成），这将有助于发展实现人类基因组测序计划的技术。已经发表的许多综述文章都谈到了基因组含量分析，其中有：Dojon (1996)；Oliver (1996)；线虫测序组 (1998) 是为数不多的其中几篇，因此在这里这个专题可能介绍得不够详细。那么这些相对简单的真核生物会告诉我们什么呢？

在这两个种类之间的种间同源基因的比较已经揭示了它们具有一整套为这两个真核生物所共有的，诸如DNA和RNA代谢、中间代谢、运输等核心生物学过程中高度保守的蛋白质（酵母可读框总量的40%，线虫可读框总量的19%）(Chervitz et al. 1998)。通过这样一个非常基本的相减法（subtractive method）就有可能鉴别那些具有生物特异性的基因。对于线虫而言，这种方法包括了参与多细胞化作用的基因。其中包括有诸如细胞程序化死亡机制和转录调节物（即核激素受体，它在酵母中不具有其种间同源物）这样一些过程。在调节和信号转导中所使用的蛋白质结构域的详细比较表明，尽管有相当一些功能域（片段）的共有现象，但这些蛋白质中的绝大部分都没有相应的种间同源物 (Chervitz et al. 1998)。这一点与细菌蛋白质的分析有关，而且进一步证实通过外显子或结构域的混编，实现多结构域蛋白质多样化这一理论 (Patthy 1999)。

不要忘记线虫（*C. elegans*）是线虫纲中的一种动物，而且当进行基因组测序时，所预测的基因的34%都可以和其他线虫纲动物序列相匹配（线虫测序组，1998）。因此线虫还能够使我们深刻地了解其他重要的线虫纲寄生虫，做到举一反三，触类旁通。这些动物还包括蛔虫（*Ascaris lumbricoides*），一个使全世界十亿人感染的大型肠道寄生虫，以及吸血圆线（虫）科寄生虫 (Blaxter 1998)。线虫基因组测序的完成，使人们对这种对于人类健康有着重大意义的线虫纲动物

有了较好的认识和了解。已经列入正在测序名单上的几种寄生虫就反映了这个问题的重要性。

酵母和线虫的确具有共同的基因，但是这样的情况能一直延伸到脊椎动物吗？Bassett 等人（1997）指导了一个在生物芯片上进行的实验来估计人类疾病基因的同源部分能够在模式生物中被发现的频率，将 84 个定位克隆的人类基因对照小鼠（*M. musculus*）、黑腹果蝇（*D. melanogaster*）、线虫（*C. elegans*）、酵母（*S. cerevisiae*）和大肠杆菌（*E. coli*）的蛋白质数据库进行搜索，除了小鼠，这个预期有高成功率的生物以外，与其他几种模式生物相比至少有一半人类基因具有很高的统计学意义，而另外的 30%则表现出中等程度的显著性意义。

被功能性特征描述的酵母基因与人类疾病基因的比较给研究者们已经提供了一个研究人类疾病机制的切入点。例如，遗传性共济失调（AT）是一种人类常染色体隐性疾病，这种疾病以对电离辐射过敏为特征，具有两个功能性重叠的酵母同源基因：*MEC1* 和 *TEL1*。在酵母中对这些基因的研究证实了它们也参与了一个 DNA 破坏检验点途径（Morrow et al. 1995）。有缺陷的酵母基因还能够被弥补而且通过人类种间同源基因的加入而恢复其功能。在 dbEST 中所进行的 cDNA 搜索工作，鉴别了一个人类蛋白质的 42% 与酵母 Spt4p 蛋白质（SUPT4H）相同。这是所认为的正常染色质结构和转录所需要的一套基因中的一个。它们以一个复合体的形式执行功能，可能是对组蛋白有修饰作用，装配核小体或者调节核小体与 DNA 或是与其他蛋白质之间的相互作用。当这个人类基因在酵母中被表达时，它与每一个 spt4 无效突变基因部分互补。这个互补作用部分是由于低表达水平所造成的，但是即便如此这个实验还是证实了一个保守功能。在人类中的高分辨率的细胞学研究证实，Spt4p 具有一个对于特异性核结构没有亚定位的核位置。这个情况说明了基因在细胞调节的基础过程中的更为普遍的作用，而且对于由非正常转录所引起的疾病来说是一个有潜力的候选基因（Hartzog et al. 1996）。

多细胞生物功能的基因代表能够在线虫中进行调查，例如，遗传性的多外生骨疣是在人类中发生的一个常染色体显性骨疾病。目前认为有三个基因与此疾病有关（*EXT1*、*EXT2* 和 *EXT3*），它们可能是作为肿瘤抑制子起作用。如果这些基因只在脊椎动物中进行研究，那么可以假设在骨生长中它们有一个专门的作用。线虫没有骨骼，但是数据库搜索显示至少有两个同源基因在线虫中存在，同时表明人类的 *EXT2* 基因可能具有更广泛的作用。许多线虫突变体已经被定位到围绕在这两个同源基因附近的区域，而且其中有几个会引起早期发育中细胞迁移以及分化的缺陷（Clines et al. 1997）。在线虫中通过互补试验所做的 EXT2 的未来的功能分析和在线虫和小鼠中所做的基因敲除试验可能会很好地证实这个潜在扩大的功能。