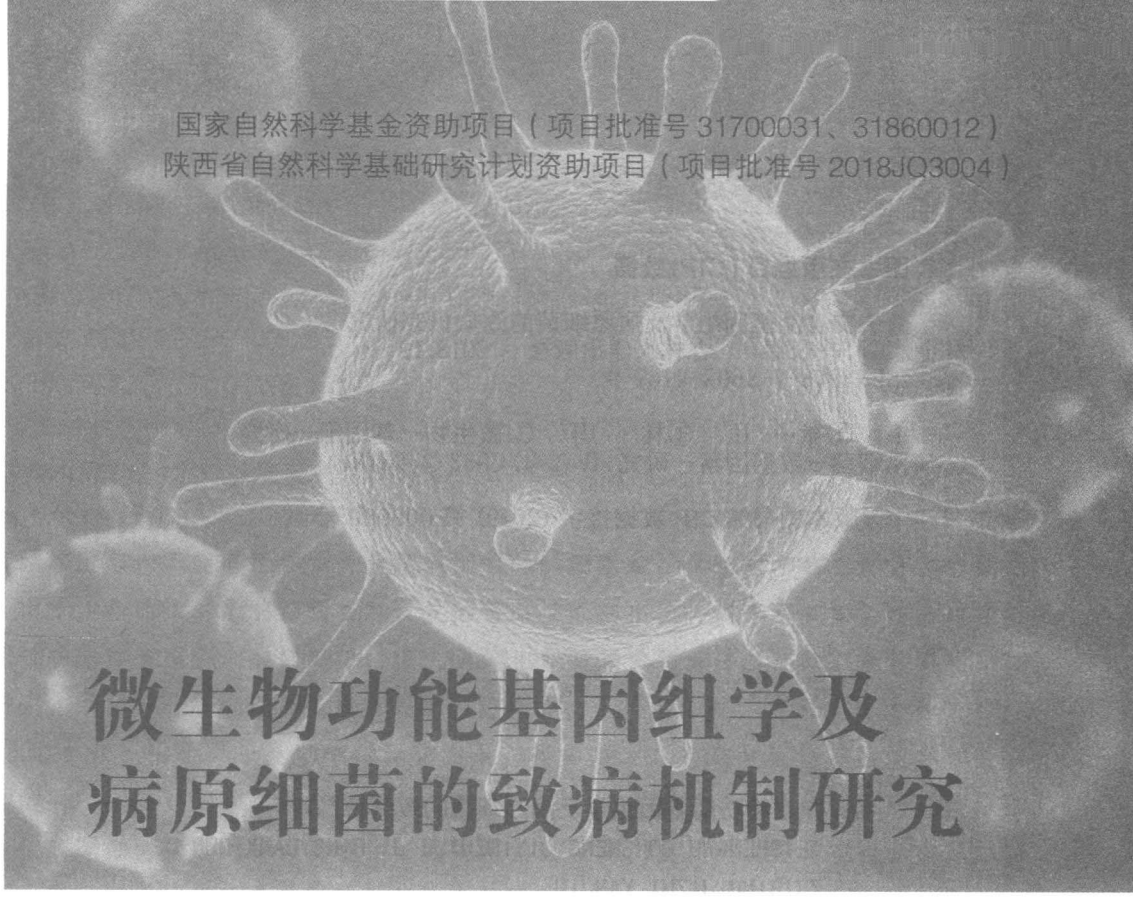
国家自然科学基金资助项目（项目批准号 31700031、31860012）
陕西省自然科学基础研究计划资助项目（项目批准号 2018JQ3004）

林金水 著

微生物功能基因组学及病原细菌的致病机制研究

吉林大学

国家自然科学基金资助项目（项目批准号 31700031、31860012）
陕西省自然科学基础研究计划资助项目（项目批准号 2018JQ3004）



微生物功能基因组学及病原细菌的致病机制研究

林金水　著

吉林大学出版社

图书在版编目(CIP)数据

微生物功能基因组学及病原细菌的致病机制研究 / 林金水著. —长春 : 吉林大学出版社, 2018.12
ISBN 978-7-5692-4167-9

Ⅰ. ①微… Ⅱ. ①林… Ⅲ. ①微生物—基因组—研究 ②病原细菌—致病因素—研究 Ⅳ. ① Q933 ② R378

中国版本图书馆 CIP 数据核字 (2019) 第 009080 号

书　　名：微生物功能基因组学及病原细菌的致病机制研究
WEISHENGWU GONGNENG JIYINZUXUE JI BINGYUANXIJUN DE ZHIBING JIZHI YANJIU

作　　者：林金水　著
策划编辑：邵宇彤
责任编辑：邵宇彤
责任校对：郭一鹤
装帧设计：优盛文化
出版发行：吉林大学出版社
社　　址：长春市人民大街 4059 号
邮政编码：130021
发行电话：0431-89580028/29/21
网　　址：http://www.jlup.com.cn
电子邮箱：jdcbs@jlu.edu.cn
印　　刷：三河市华晨印务有限公司
开　　本：185mm × 260mm　1/16
印　　张：14.25
字　　数：329 千字
版　　次：2019 年 3 月第 1 版
印　　次：2019 年 3 月第 1 次
书　　号：ISBN 978-7-5692-4167-9
定　　价：59.00 元

前言

病原细菌与人类健康密切相关，病原细菌的研究则是医学细菌学的主要内容和精髓所在。由于诸多因素的影响，新发和复发的传染性疾病出现得越来越频繁，越来越多的动物源性病原体有感染人类的趋势，使人类的生命健康受到威胁。因此，重视对病原细菌的研究，了解其致病机理，采取有效的预防措施，是摆在细菌学家面前的重要任务。

近30年来，由于分子生物学的迅猛发展，特别是近几年来微生物基因组学和蛋白质组学的兴起，使对病原细菌的研究如虎添翼，为人类重新认识病原细菌、研究病原细菌和防控病原细菌引起的传染病提供了新的视角和手段，使人类无论是在揭示病原细菌的致病机理研究方面，还是在预防细菌性疫苗研究方面，成绩斐然。

本书很好地整合了微生物功能基因组学的知识要点，如基因组学概述、如何利用基因组数据、基因功能预测的技术方法、DNA微阵列技术及其在基因表达数据分析等。同时，随着现代生物技术的完善和渗透，极大地推动了病原生物研究及其相应临床工作的开展，并由此产生了许多新的诊断治疗病原疾病的理论和技术。本书介绍了一些有关病原细菌研究的新进展，以提供人类抗击病原细菌危害的新技术、新方法和新手段。我们殷切希望本书有助于病原细菌学的研究，可以作为从事细菌学和其相关领域研究的研究人员、技术人员以及生物学和医学专业本科生或研究生的参考用书。

尽管本书在编写过程中尽量收集相关的信息以及最新的研究进展，但是由于相关文献浩如烟海，新的研究又层出不穷，再加上编者水平所限，不足之处在所难免，敬请各位读者批评指正，不胜感谢。

林金水

2018年5月

目录

第 1 章　基因组学概述

基因是许多生物性状控制的源头，因此在疾病、疫苗、生物化学等相关领域的研究都需要全基因组序列的研究支持和帮助。在基因组控制性状方面的研究正在世界各地的实验室进行着，并使用遗传、生物化学、代谢物组、基因组、蛋白质组和计算方法等多种科学手段和角度对它进行着研究。随着科学方法的使用和研究程度的加深，更多具有价值基因组测序的数据被发现，相关的基因组学逐渐成为炙手可热的生物学学科之一，而相关基因组技术也随之发展起来。

1.1　基因组学的定义和分类

基因组学（genomics）这个术语已经在相关研究领域广泛使用，得到科学界认可的同时常常出现在社会各处而引起人们的关注。1986 年，Thomas H.Roderick 第一次提出基因组学（genomics），用于称呼对基因组进行相关研究活动的新学科。除了这个词汇，我们还需要知道另外一个词汇——基因组（genome），它指的是具体某个细胞或者生物体内，整套基因和染色体等遗传物质的集合。Thomas 提出的基因组学（genomics）就是根据基因组（genome）这一词派生而来的。1920 年，H.Winkler 首次从“基因（GENes）”一词和“染色体（chromosOMEs）”一词中各取一部分，组成并使用了基因组（genome）这一词。之后，这一词逐渐开始被接受和使用。在 1987 年，一个新期刊还以此词命名并创立。随着科学的不断进步，对基因组的分析不再是测序扩展和图谱绘制。到了 1995 年，基因功能分析也出现了。此时，基因组这一称谓便不足以描述相关领域研究了，所以出现了更综合和全面的称谓——基因组学，也就是以基因组为研究对象，探究其物质结构以及生物功能的学科。

我们已经知道了基因组学所研究的内容和工作，此词已广泛被接受和使用，但是基因组学这一术语的定义仍没有被明确。现在，基因组学多被用来代表基因组的图谱绘制、测序和分析的内容。在其他相关细分领域的研究则配以具体能代表研究内容的学术词进行代表，如在基因组序列和蛋白质的功能研究领域，在学术报告、论文等文献中往往被称作结构基因组学、功能基因组学与蛋白质组学等。此外，随着基因组学研究的深入和其在药理学、医学和生理学等其他不同领域的应用结合，更多生物学的新分支已经延伸出来。例如，已经出现了毒物基因组学（toxico genomics）、药物基因组学（pharmaco genomics）、医学基因组学（medical genomics）、生理基因组学（physiological genomics）、生态基因组学（ecological

genomics）等与基因组学相关的各种术语。由此可见，基因组是一个涵盖以基因组以及其相关为研究对象的相关领域的术语，所以基因组学在这里是指在基因组水平研究基因和生物性状表现、生物物质分子结构与功能等领域的分子基础生物学学科，所运用的技术是全基因组测序信息和高通量基因组技术等。此处定义的关键词是分子基础，也就是基因组学研究是基于分子水平来认识生物系统的结构和功能。与传统分子生物学方法相比，基因组学的研究是在基因组学的基础上进行的，所采用的技术是基因组序列信息和高通量基因组技术。这里所说的生物系统指的不是生态研究中的群落、种群和生态系统，而是细胞和亚细胞的系统，如某个细胞、组织、器官或整个生物体。

基因组是生物体性状控制的源头，其研究可以涉及众多生物学领域，所以基因组学科之下延伸的分支学科很多。下面，我们将基因组研究的分类以及各种相关基因组学术语的概念进行简要介绍。

1.1.1 根据系统特性分类

生物系统是物质结构和生物功能高度结合、平衡的内环境系统。根据一般系统理论，任何系统中最基本的两个因素就是结构和功能。因此，从一般系统理论的观点来划分，基因组学可以分为结构基因组学和功能基因组学。

对基因、蛋白质以及其他生物大分子的泛基因组结构所进行的研究，我们称之为结构基因组学，所包括的内容有基因组图谱测序、绘制和组织以及蛋白质结构的描述。科学界将对基因组的图谱绘制、预测和组织，蛋白质结构的描述以及相关工作称作基因组分析研究，而现在结构基因组学这一术语的出现和应用后，已经逐渐涵盖和超越基因组分析研究这一术语指代的研究工作了。在分子层面，多个层次和角度研究生物体的系统功能的工作叫作功能基因组学，包括基因功能和调节网络。基因功能之下能够从不同的功能层次和角度划分出具体功能研究定义，如生物化学功能（如消化道内蛋白酶的合成与其功能关系）、细胞功能（如在细胞生长过程中 DNA 制造蛋白等）、发育功能（如细胞分化过程中遗传物质的复制和性状表达）或适应功能（基因突变与有性繁殖过程中基因对新性状产生对环境适应的关系）。功能基因组学的本质研究内容是探究特定基因组序列和生物体某项性状表现之间的关系，所以功能基因组学研究在进行基因组序列测序等工作的同时，需要进行大规模的实验，并对大量实验结果进行分析统计，最终得到结果。功能基因组研究对人类而言，有希望了解许多人类遗传疾病以及许多疾病的病理特点，对生物学来说有利于了解生物体基因与功能的规律和科学机制。虽然在当下科学界的各类学术期刊、会议论文等文献中已经广泛使用基因组学、功能基因组学这类术语，但是有时使用得并不恰当。目前，在 mRNA 测量及基于微阵列的基因表达相关领域的论文或者期刊中才会使用这些术语。

目前，我们已经能够通过全基因组测序技术得到某生物体的基因组信息，但这是远远不够的。基因序列尽管作为生物系统性状控制的核心，但是基因如何通过一系列过程生成蛋白等生命物质和其如何发挥生物作用仍然披着朦胧的面纱。目前，我们知晓的是，在生物系

统内基因（染色体等遗传物质上的 DNA 序列）先代谢出信使 RNA，这一过程叫作转录，接着通过信使 RNA 生成各种生物系统内的蛋白质。这些具有各种各样功能的蛋白质又组成了新的制造代谢需要物质的机器，由此遗传物质、蛋白质以及代谢物质等组成了生物系统的基本结构。有了物质结构还不够，还需要一个适合它们存在的环境。生物系统（细胞、器官、组织或生物体）的内环境始终处于一个适合内部生命活动的平衡状态中，各种遗传物质、蛋白质和代谢物等生命物质在这里保持最佳状态并发挥其功能。内环境与其中各个生物的物质种类和数量还会随着生物系统的发育情况、生理变化和外部环境的改变而变化。正是因为生物系统内部生命活动的复杂性、严密性和神秘性，所以对此的研究意义分毫不亚于基因组本身的研究。所以，仅是检验 DNA 和 mRNA 是不够的，功能基因学的研究必须对遗传物质、蛋白质、代谢产物以及内环境系统进行全面的研究。从这一观点出发，生物系统的蛋白质组（proteome）（蛋白质组学）和代谢物组（metabolome）（代谢物组学）研究也应当是功能基因组学研究的组成部分。从生物系统内部遗传信息传递的过程来看，我们将功能基因组学可以具体分为转录物组学（transcriptomics）、蛋白质组学（proteomics）和代谢物组学（metabolomics）。有了基因所控制的遗传性状过程的研究，功能基因组研究才完整。

我们将生物系统比作一个复杂的建筑，基因就是它的蓝图，通过一定的建造方式和建筑程序，才将整个建筑完成，而基因在生物系统中的建造方式和建造程序也是我们了解生物功能的重要内容。根据目前的实验成果可知，生物系统中某种蛋白质在蛋白质组含量的高低是会变化的，决定其变化的因素有很多，其中重要的因素之一就是它对应的 mRNA 在 mRNA 组中的含量。我们知道，搭载某种基因的 mRNA 的含量水平与生物系统的生命状态息息相关。所以，完成建筑的过程中，mRNA 对建筑程序和建筑方式的影响是必须进行深入研究的。转录物组（transcriptome）是指在一个细胞中全部 mRNA 的集合。转录物组学是指高通量研究转录物组的表达动力学。基因并不是生物系统的功能实体，同样 mRNA 也不是，所以转录物组研究成果只能用来作为数据，间接支持基因功能的研究。生物系统中某种蛋白质在蛋白质组的含量，其重要决定因素就是对应的 mRNA 在 mRNA 组中的含量。除此以外，翻译过程和蛋白质装饰因素的影响也是十分重要的。补充一点，基因（DNA）在生物系统中含量是在一定范围内的，所以其含量和蛋白质含量并没有线性关系。因此，蛋白质表达动力学和蛋白质相互作用的研究对阐明基因功能是至关重要的，并且是转录物组学的补充。蛋白质组是指一个基因组所编码的全部蛋白质。蛋白质组学是指利用直接测定和鉴别蛋白质的高通量方法，大规模研究蛋白质组的表达动力学和蛋白质相互作用。

生物系统所进行的代谢过程的最终产物，我们称之为代谢物。代谢物在生物系统中也是功能实体，与蛋白质一样。这些代谢物存在于生物系统的细胞内，也存在于细胞之间，并且随着生物系统内外状态改变而时刻发生变化。所以，代谢物的含量和状况是生物系统的生长阶段、发育状态和外部环境等因素影响的最终结果的表现。由此，我们可以通过测定和分析代谢物相关数据得出有利于基因功能性研究的重要信息。我们知道，蛋白质组指的是生物系统内所有蛋白质的集合，转录物组就是所有转录物的集合，那么类似的代谢物组就是指生

物系统内所有代谢物的集合。代谢物组学是指大规模研究代谢物组的动力学和相互作用的学科。

1.1.2 根据同其他科学学科的关系分类

随着基因组信息的逐渐被确定以及相关研究的不断进步，以基因为基础的生物学研究活动的开展条件越加充足。此时，按照应用属性方面来划分，可以将基因组研究分为基础基因组学（basic genomics）和应用基因组学（applied genomics）。基础基因组学的研究目标是站在泛基因组层次去了解认识生物系统的生物过程，应用基因组学的研究目标更加倾向于应用和解决实际问题。两者在研究技术上也不同，前者主要运用基因组学技术以及全基因组序列数据进行研究，后者主要是基于前者的基因组序列信息，运用相关的高通量技术进行研究。

基础基因组学可以再划分为生化基因组学（biochemical genomics）、遗传基因组学（genetical genomics）、进化基因组学（evolutionary genomics）、生理基因组学（physiological genomics）、生态基因组学（ecological genomics）和计算基因组学（computational genomics）。生化基因组学的研究内容是找出所有基因中拥有生物化学活性的部分，如能够生成蛋白酶的基因序列。学者 Martzen 等就是从事此方面的研究，从而找到了在酵母菌中发挥发酵作用的酵母基因。遗传基因组学的研究内容贴近遗传学研究，但其是基于基因组进行的，在某个没有外来基因注入的群体，运用高通量基因组工具研究基因的遗传和变异现象。进化基因组学，顾名思义，是研究物种的进化过程和不同种类的亲缘关系，当然也是利用基因相关技术。生理基因组学研究内容是基因在生物体内的表达过程以及各种蛋白质的功能和影响。计算基因组学更加类似一个工具，基于已有的基因组数据建立模型，然后通过模拟生物过程的特定算法进行相关生物活动的预测以及研究。计算基因组学是在进行实验基因组学研究中必不可少的工具。

生态基因组学这个术语有几种不同的含意。当前，对于生态基因组学这一称谓，不同的人存在不同的理解。学者 Nevo 认为，生态基因组学的主要研究内容是基因遗传和生态学现象的关系，是生态遗传学在基因学领域的分支。持有不同观点的 Cary 和 Chisholm 认为，生态基因组学的内容应当是以生态系统为研究目标，利用应用基因组科学等方法，解析生态系统的机制和影响。本书将生态基因组学理解为在生态学范畴内针对测定生物系统的组成、结构、功能和动力学的分子基础和机理的泛基因组认识研究。从生态系统研究水平来分，生态基因组学可以进一步划分为种群基因组学（population genomics）、群落基因组学（community genomics）和生态系统基因组学（ecosystem genomics）。群落基因组学的研究内容是以某个种群为研究对象，对其中的泛基因组遗传变异现象进行研究。与生态学中对群落的结构、功能、多样性和进化等特征研究不同的是，群落基因组学的研究是在基因层次上开展的。了解了群落基因组学，我们也就能轻易理解生态系统基因组学的内容了。所以，生态系统基因组学的研究对象是某个生态系统，了解解析基因遗传学因素在生态系统中对能量和物质在生命间的传递过程、各类生物和整体的稳定性、适应性变化机制的影响和作用。上面提到的两种

学科的研究能够解决我们很多疑问，下面是其中某些问题示例。第一，我们知道在成熟的生态系统中，不同的物种间相互影响、相互促进，从而有了物种的多样性，那么在基因水平如何解释这一现象呢？第二，生物系统具有代谢功能，这一个体功能对整体生态系统具有什么影响，有着怎样的联系？第三，遗传基因对生物体的环境适应性以及自身性状的稳定性有哪些影响？第四，假如能够掌握生态系统中某种群的状态和系统的代谢情况，是否能够由此进行整个生态系统的预测？第五，基于对物种的认识和控制，在特定的环境中，能否通过一定方式将某种生态系统控制在平衡和谐的状态？

基因组序列信息和基因组技术也已经广泛地用于解决一些与医学、工业、农业和环境相关的问题。基因组学的研究在其他相关领域也能起到非常重要的作用，如在工业基因组学、医学基因组学、农业基因组学和环境基因组学等。在相关的不同学科里，和基因组学相关产生了新的术语，如在医学领域，有癌基因组学（cancer genomics）、毒物基因组学（toxicogenomics）和药物基因组学（pharmacogenomics）。从这些称谓来看也许互不相干，但是它们的研究领域和范围存在重合和交叉。在这里，癌基因组就是说在癌症研究领域引入基因组学相关技术进行研究。毒物基因组学就是说采用基因组学相关技术和方法，对相关的化学物质进行药性药理分析，识别它的毒性。

目前，环境基因组学在科学界的概念并不清晰。有些学者将它使用在环境和健康的研究中。它也表示是影响环境的生物体的基因组的研究。更为广泛接受的是将环境基因组学认为是关于人类健康和环境健康有重大影响的基因组的相关研究。

1.1.3　根据所研究生物体的种类分类

根据所研究的生物体的种类来分类，基因组学可以划分成微生物基因组学（microbial genomics）、植物基因组学（plant genomics）、动物基因组学（animal genomics）和人类基因组学（human genomics）。根据所研究的微生物，微生物基因组学还可以进一步划分为病毒基因组学（viral genomics）、古菌基因组学（archaeal genomics）、细菌基因组学（bacterial genomics）和真菌基因组学（fungal genomics）。另一个经常在文献中使用的术语是比较基因组学（comparative genomics）。它一般是指来自各种微生物的基因组信息（即基因组序列、mRNA 和蛋白质表达形式）的比较，其目的是应用高通量的计算和实验方法来获得对生物过程和现象的泛基因组水平的了解。

1.2　基因组学的历史回顾和挑战

1.2.1　基因组学的历史回顾

基因组学研究最早出现在 1985 年，和物理学、天文学等学科相比是一个新兴领域。到

了1986年，人类基因组测序被科学界所提出。在随后的4年里，相关科学家进行了激烈的讨论，最后确定了最终的人类基因组计划（Human Genome Project，HGP）（见表1-1）。1986年2月，美国的一个机构第一次宣布会对人类基因组研究计划进行自主研究。1988年，美国国家科学院和国家研究委员会（NRC）委托人类基因组委员会发表报告并签署了这个计划，并为研究院和能源部的第一个联合计划提供了基础条件。该委员会决定，在15年内完成人类基因组计划，每年的费用是2亿美元，并建议先绘制人类基因组图谱，然后测序。该委员会建议在人类基因组计划执行的同时，进行模式生物（如酵母和鼠）基因组测序计划。1990年，人类基因组计划在美国正式启动，到2005年，人类基因组计划的测序工作已经完成。

表1-1　基因组学中重要的里程碑和事件

年	里程碑和事件	参考文献
1985—1988	美国国家科学院和国家研究委员会进行讨论、辩论和制订人类基因组计划（HGP）	Alberts et al.（1988）
1990	人类基因组计划在美国正式开始	Burris et al.（1998）
1995	第一个自由存活生物，Haemophilus influenzae细菌的基因组用鸟枪测序法测序完成	Fleischmann et al.（1995）
1996	一个国际小组完成了芽殖酵母Saccharomyces cerevisiae基因组的全部序列测定，标志着第一个真核生物基因组完全测序	Goffeau et al.（1996）
1998	C.elegans测序协会完成了第一个多细胞生物（Caenorhabditis elegans）基因组的完全测序	C.elegans测序协会（1998）
1998	微阵列和基因组学一起列入了十项最高科学突破	新闻和编辑委员会（1998）
2000	利用全基因组鸟枪测序法完成了果蝇Drosophila melanogaster的基因组测序。这是第二个，并且是最大的动物基因组测序	Adams et al.（2000）
2000	美国前总统克林顿宣布即将完成的人类基因组序列（3 000Mb）是“人类描绘的最奇妙的图谱”	Marshall et al.（2000）
2000	一个国家小组发表了开花植物Arabidopsis thaliana的基因组，标志着第一个植物基因组测序完成	The Arabidopsis Genome Initiative（2000）
2001	人类基因组序列草图的两个版本在《Science》和《Nature》杂志发表，这是以基因组为基础的生物学的基石，在生物学研究历史中提供最丰富的知识资源	Venter et al.（2001）；Lander et al.（2001）

续　表

年	里程碑和事件	参考文献
2002	发表了两个主要的水稻（Oryza sativa）亚种的基因组序草图。这在农业研究中是一个里程碑，并且是第一个经济上很重要的谷物的基因组序列信息	Yu et al.（2002）；Goff et al.（2002）

在人类基因组计划进行的同一阶段，美国相关机构对其他生物的基因组测序工作已经开始了，它的意义在于这将对某些行业起到巨大的推动作用。1995 年 7 月，基因组研究所发表了流感嗜血杆菌（Haemophilus influenzae）Rd（一种非寄生微生物）的全基因组序列（约 1.8Mb）。这代表了利用鸟枪测序法成功地测出了第一个全基因组的序列，该方法是获得基因组序列的快速而有效的方法。这个基因组测序的完成用了一年的时间，标志着开启了基因组学新纪元。此后，有 100 多种微生物基因组测序完成，并有 200 多种微生物基因组测序计划正在进行中。

酿酒酵母（Saccharomyces cerevisiae）是一种重要的能替代其他生物进行基因组功能分析的物种。与人类和其他真核生物不同，它是一个单细胞生物，并且基因组尺寸较小（约 12Mb）。它能够在合成培养基上生长，因此它生长的化学条件和物理条件可以完全控制。S.cerevisiae 有一个对典型的遗传分析非常适合的生命循环。已经开发的酵母的有效遗传工具可以将任一基因替换为突变等位基因或将它从基因组中完全去除。1996 年，一个国际研究小组完成了 S.cerevisiae 基因组的全序列测定。这是第一个完成的真核生物基因组的全序列测定。

秀丽线虫（Caenorhabditis elegans）（一种杆形线虫）是一种在遗传上易于掌控的生物模型，广泛地用在遗传、发育和其他生物过程的研究。1998 年，发表了该生物的基因组序列（97Mb）。这是第一个被完全测序的多细胞生物。

在生物学中，黑腹果绳（Drosophila melanogaster）是研究最广泛的生物之一，它常作为研究高等生物（包括人类）发育和细胞过程的模型系统。2000 年，利用全基因组鸟枪测序法完成了该生物的基因组（180Mb）测序。这成为一个里程碑，标志着 20 世纪的结束并预示着生物学探测和分析的新纪元的开始。至此，这是第二个，并且是最大的动物基因组测序。这也是在对这种生物研究的 90 年中最近的一块里程碑。

开花植物的组织和生理特征与其他多细胞生物（如 C.elegans 和 Drosophila）大不相同。植物的全基因组序列信息对我们认识植物和动物遗传基础的不同以及植物基因调节和功能表征是非常有用的。拟南芥（Arabidopsis thaliana）是植物基因组分析的一个重要模型系统，因为它的生长期短、繁殖量大并且基因组较小（约 125Mb）。它为研究所有植物（包括主要农作物）的遗传方式提供了一个通道。2000 年，一个国际合作团体完成了 Arabidopsis 基因组的测序，标志着第一个植物基因组测序工作的完成。

2000 年 6 月 26 日，美国前总统比尔·克林顿将即将完成的人类基因组序列（3 000Mb）

形容为“人类描绘的最奇妙的图谱”，而英国首相托尼·布莱尔预言，基于基因组的研究将领导“一场在医学中的革命，它的影响将远远超过抗生素的发现”。2001 年 2 月，人类基因组序列草图的一个版本发表在《Science》杂志上，其作者是由 Celera Genomics 的 Craig Venter 领导的一组研究人员。而另一个版本发表在《Nature》杂志上，它的作者是由 Francis Collins 领导的公开资助实验室联盟的国际人类基因组测序协会。这个基本完成的人类基因组序列是以基因组为基础的生物学的基石，在生物学研究历史中提供最丰富的知识资源。与人类第一次登上月球和第一颗原子弹爆炸相似，人类基因组序列的测定有许多象征性的重要意义，因为它从根本上改变了我们对自己的看法。这是我们首次看到人类内在的遗传骨架。人类基因组测序的完成对认识人类生物学及其进化有重要历史意义，并且打开了生物学的纪元。

水稻是世界上最重要的谷类作物，占世界农业总产量的 60 %。大部分大米被人类直接消费，并且有三分之一的人口依靠大米提供 50 % 以上的热量。水稻也是一个研究植物基因表达与调控的重要模式物种。与其他重要的谷物（如高粱、玉米、大麦和小麦）相比，水稻的基因组最小（约 430Mb）。2002 年，科学家发表了两个主要的水稻亚种的基因组序列草图。这在农业研究中是一个里程碑，同时新表明重要经济谷物基因组序列信息首次可以被利用。

全基因组序列的可用性极大地推动了基因组序列功能分析的基因组技术的发展和应用，如微阵列（microarrays）。1995 年，科学家首次提出在玻璃片上用高速机器人印刷互补 DNA 来定量测定相关基因。在这个成功应用之后，DNA 和寡核苷酸微阵列都被用在了监测基因表达和测定特异表达基因，包括在酵母中的，不同人类细胞、组织中的以及人类病变的细胞、组织中的基因。1998 年，《Science》杂志的新闻和编委将微阵列列入了与基因组学一起出现的十项突破性技术。类似微处理器加速了计算过程，基于微阵列的基因组技术使生物系统的基因分析发生了巨大的变化。微阵列技术是一种功能强大的新工具，研究人员可以用它从全面的、动态的分子视角来观察各种生理状态下的活细胞。

人类和其他生物基因组的成功测序也极大地促进了蛋白质组技术的发展和应用，如质谱（MS）。20 世纪 90 年代，质谱仪器和方法所取得的进展为蛋白质化学带来了革命性的进步，并且从根本上改变了蛋白质分析方法。尽管在 20 世纪 80 年代末期，在蛋白质组学研究方面已经取得了两项重要的技术突破：电喷雾离子化（ESI）和基质辅助激光解吸附离子化（MALDI）。但是，直到得到全基因组序列之后，生物质谱才成为蛋白质研究的强大的分析工具。由于得到全基因组序列成为可能，鉴定从特定细胞分离的蛋白质不再需要重新测序，而只需要将短肽的分子量或短氨基酸序列与序列数据库中推测的蛋白相关联。生物质谱的快速发展和许多不同生物全基因组序列测定的完成，标志着生物学新纪元的开始。

1.2.2 研究功能基因组学的挑战

功能基因组学的最终目标是利用基因组序列信息和相关基因组学技术，将序列与功能跟表型联系起来，并了解自然界中生物系统不同水平的功能。这个目标面临的挑战如下。

1. 阐释基因功能

尽管根据序列相似性比较的计算分析在阐释基因功能上是有用的，但是它只能提供基因功能的一些线索，因为序列相似并不等于功能相似，并且所预测的可读框（ORFs）有相当一部分在功能上也是未知的。另外，由于在不同基因间复杂的进化和结构 - 功能关系，通过相似性比较进行基因的功能分配有时候会引起误导或错误。换句话说，基因与酶的关系既有一对一的关系，也有多对一的关系，还有一对多的关系。如此复杂的基因与酶的关系表明根据序列解释功能困难重重。因此，认识基因产物的生物作用必须进行实验分析。然而，在一个复杂的细胞机构和调节网络中，阐释每个基因的作用是一项艰难的实验任务。

2. 鉴别和表征生命的分子机构

细胞是所有生命系统的基本工作单元，它们是生物“工厂”，利用由不同蛋白质和其他分子构成的分子“机器”来执行和整合成千上万的分散而高度专业化的过程。蛋白质很少单独工作，它们经常组合成庞大的多蛋白复合体。一些这样的组合体很像复杂的机器，执行着基本的细胞功能和代谢过程（如 DNA 复制、转录、翻译、蛋白质降解），在细胞内、细胞间及其所处的环境间传递着信息流（如信号传导、能力转化、细胞运动）或建造细胞结构。许多蛋白质机器在组成和功能上表现出高度保守性，因此对一种生物性质的认识，也可以应用到其他生物。尽管蛋白质机器的不同类型的具体数量还不清楚，但是全基因组序列分析推测其数量是有限的，可能每个细胞有几千个左右。

全基因组测序和蛋白质结构测定的新进展，为我们提供了模式生物中许多蛋白质的组成信息，但是这里存在的最大挑战是全面认识和表征存在于生物系统中分子机构的全部功能，并且认识蛋白质是如何使细胞具有它们的性能、结构和高度有序性的。

3. 描绘基因调节网络

所有的生命系统都有能力在空间和时间上通过快速变化来响应细胞内外环境的刺激。这些能力是通过许多个体蛋白质和蛋白质组合物在不同调节控制下进行复杂的相互协作得到的。个体蛋白质和其他大分子用“电线”连接起来像电子线路，形成复杂的遗传调节网络，它是细胞的“大脑”。该网络通过传导途径从细胞内外接收信息、处理信息做出“决定”，决定哪些基因表达以及合成多少产物，启动合适的细胞和生理响应。遗传调节网络的协调行动决定了活细胞的生理学特性，并且对细胞存活、生长和繁殖至关重要。

最近的基因组序列比较表明，一个人的基因数量只有一只简单虫子的 2 ~ 3 倍，只有一个单细胞微生物的 5 ~ 10 倍。然而，人类却具有很多种不同类型、不同功能的细胞、组织和器官。因此，仅从基因数量上简单的差别，不能解释存在于人与虫子或微生物之间巨大的表型差别。这种表型差别主要是归因于遗传调节网络的结构和复杂性。简单生物进化为多细胞生物，可能是通过发展更复杂独特的能够巧妙控制复杂的组合基因表达模式的调节网络，同时基因的功能和数量有适当的扩展。如果这个假设正确，在这样的调节网络中连接和接点的变化可能戏剧性地改变生物系统的生理性质和行为。然而，认识作为一个整体的基因组功能如何使生命体在复杂的自然历史过程中进行生活，是一项更大的挑战。

4. 系统水平上而非单个细胞水平上对生物系统的认识

生物彼此间以及与它们的环境之间的相互作用产生了比较复杂的生物系统，如种群、群落、生态系统和生物圈。将基因组引入生命研究所面临的挑战有利于基因组序列信息去认识从微米到大洲际空间范围内遗传能力和相互作用的结果，在生物系统中将机理研究的遗传水平与功能执行的系统水平联系起来（如个体、群落、生态系统和生物圈）。

已知生物中微生物的种类最多，在地球上任何可以想象的环境中都有微生物存在。地球上细菌细胞的总量估计有（4 ~ 6）$\times 10^{30}$ 个，这些细胞的碳含量有（3.5 ~ 5.5）$\times 10^{14}$kg。细菌也含有最大量的生物氮（1.3×10^{14}kg）和生物磷（1.4×10^{14}kg）。然而，对于植物和动物来说，微生物多样性的范围大部分还是未知的。由于在自然界中发现的大部分微生物处于不可培养状态，关于它们的遗传性质、代谢特征和功能知之甚少。因此，认识和表征不可培养的微生物是一个巨大的挑战。

生物系统有两个重要特性，一个是功能的稳定性，一个是环境适应性。在群落生态系统的研究中，人们对群落中物种多样性的重要性研究一直都在进行，也在探讨其和稳定性、适应性的关系。生物多样性与生态系统功能性之间关系的讨论和研究，始终是生态研究领域最受关注、最具有争议的内容之一。他们所探讨的核心问题是生物多样性程度为多少时，才能够保持某个生态系统的稳定和功能，也就是生态系统中有多少物种就能够实现功能的稳定。产生争论的原因在于从实际实验结果中得到了两组相左的数据。一些学者认为，导致物种多样性的原因之一在于物种之前相互影响和依赖；另一些学者认为，即使是物种并不丰富的群落也能够和多样性丰富的群落具有一样的代谢效率。很多优秀的研究人员已经在此方面进行过交流，但是由于群落相关数据无法得到，所以始终没有得到解答。

群落研究内容的一个重要问题是，在群落中多种物种之间不同的生态功能是如何相互影响和作用成为一个整体，进而保证了整个群落的稳定性、适应性和功能性。目前的普遍观点认为，某种种群的遗传基因和代谢多样性是个体生物系统的环境适应性和功能稳定性的决定性因素。在基因组研究起步时期，遗传信息的取得和解析无法实现，也就无法通过技术手段验证此观点正确与否。随着基因组研究步入快速发展时期，内容更加细化和延伸，通过基因组研究技术和成果，使群落研究有希望更近一步。

5. 计算上的挑战

生物系统内的各种生命活动，包括基因表达过程、新陈代谢以及内部运行机制等，都能够从基因组序列的解释中找到解答。研究中一个关键的难题就是如何将生物系统中各项生物活动的物质单位进行确定和划分，然后模仿生物系统中各个物质单位的机构和功能创造一个生物系统模型。

在群落生态系统中，影响和驱动生态系统功能性的是多个物种（个体）的共同作用。所以在生态系统中，个体生物（物种）是研究群落功能性和稳定性等的基本单元。另一个关键的难题是，选择何种合适的算法，才能尽可能真实地模拟细胞内复杂的生命活动，进而才能够在种群、群落以及生态系统层次研究其机制。

6. 多学科协作

如果想要将生物系统中的功能和机制全面认识清楚，那么必须在各个层次都有着明确的了解。但是，想要将其了解全面是一项涉及多个专业领域、工作量巨大的工作，可能需要一代人甚至两三代人共同努力才能实现。另外，由于涉及专业领域众多，研究内容十分庞大，目前尚没有机构有足够的人力物力投入到此项工作中。在人力方面，需要聚集相关领域的专业人才；在物力方面，需要具有实验条件的群落（生态系统）并且检测工具、实验仪器。所以，这是一项巨大的挑战。

1.3　基因组学的研究范围和方法

1.3.1　结构基因组学

基因组测序的工作量非常巨大。例如，人类基因组测序正式启动于 1990 年，到 2003 年才完成所有的工作。但是，基因测序对于生态系统研究领域来说，仅是一个关键起步阶段的内容。很明显，如果想要全面研究生态系统里的基因多样性，势必要对其中所有物种的遗传基因进行了解，也就需要进行基因测序。所以在结构基因组学里，全基因组序列的测定和基因功能注释是核心内容。物种多样性的原因是遗传基因多样性，所以了解物种的基因组序列并比对差异，是提高物种多样性对生态系统功能性、稳定性和适应性影响的认识的重要环节。在具有一定亲缘关系的不同物种之间，基因组序列差异内容的研究能够提供更多有关遗传基因与生物性状之间的信息。亲缘关系远的物种基因序列对研究基因功能的制约性很有价值；亲缘关系近的物种基因序列的比较有利于研究基序（motif）和机理。此外，基因序列信息中还能够提取出基因组组织和结构的相关数据。

随着科学的进步，基因组测序的费用越来越低，很多生物系统的基因组序列能够得到测序。另外，高通量毛细管测试机器的出现和使用使这一过程更加便捷。但是，仅有这些还是不够的，我们还需要更加先进的工具来完成基因组测序工作，因为这是一项十分巨大的工作，其中测序比对、分析以及功能注释都是非常关键的环节。

在基因组学研究中，经常采用的方法是破坏某个基因，然后与未破坏的情况进行比对，获得此基因序列的信息；控制某个基因片段，然后对比试验结果从而得出结论。科研工作者采取基因敲除方法使某基因片段失效，从而达到控制其性状的目的，或者利用某个基因片段替代试验基因片段，从而影响其表达。这两种方式都能够达到控制基因的目标。一种方法是基于转座子的随机插入突变，同时将一个抗生素耐性药盒基因随机地插入一个基因组；另一种方法是通过同源重组在结构中删除目标基因，该基因在结构中被完全删除或部分删除。同源重组一般是用不复制或自杀来实现的，质粒或带有条件活性复制子的质粒可作为传送系统。