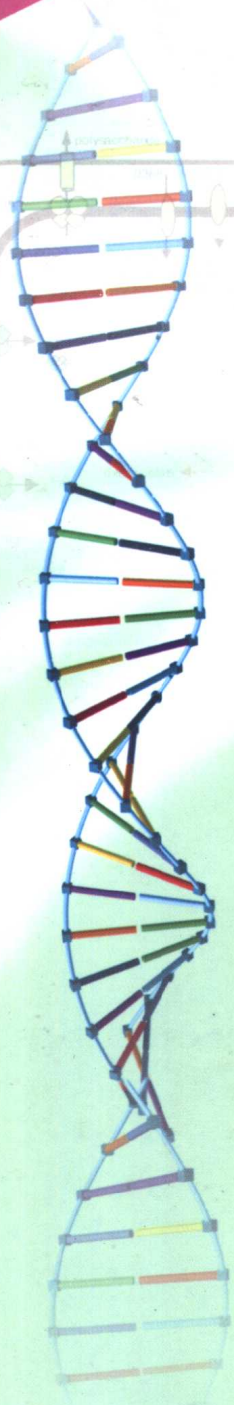
WEISHENGWU JIYINZUXUE

微生物基因组学

主编 胡福泉



人民军医出版社
PEOPLE'S MILITARY MEDICAL PUBLISHER

微生物基因组学

WEISHENGWU JIYINZUXUE

主　编　胡福泉

副主编　饶贤才

编　者　（以姓氏笔画为序）

王嘉丽　毛旭虎　邓国宏　叶姜瑜
丛延广　朱永红　朱军民　刘俊康
安　静　张克斌　余　全　胡晓梅
胡福泉　胡廷漱　饶贤才　郭　刚
谭银玲　黎　庶

人民军医出版社

People's Military Medical Publisher

北　京

图书在版编目(CIP)数据

微生物基因组学/胡福泉主编．—北京：人民军医出版社，2002.10
ISBN 7-80157-554-7

Ⅰ.微… Ⅱ.胡… Ⅲ.微生物—基因组—研究 Ⅳ.Q933

中国版本图书馆 CIP 数据核字(2002)第 027332 号

人民军医出版社出版
(北京市复兴路 22 号甲 3 号)
(邮政编码：100842 电话：68222916)
人民军医出版社激光照排中心排版
潮河印刷厂印刷
春园装订厂装订
新华书店总店北京发行所发行
*
开本：787×1092mm1/16 · 印张：35.125 · 彩页 18 面 · 字数：1153 千字
2002 年 10 月第 1 版(北京)第 1 次印刷

印数：0001～3500 定价：98.00 元

内容提要

本书是国内第一部关于微生物基因组的专业著作。全书共分20章，简述了微生物基因组学的现状、微生物基因组的测序战略及注释；详细介绍了嗜盐古细菌、嗜热甲烷杆菌、大肠埃希菌、脑膜炎奈瑟菌、梅毒螺旋杆菌、衣原体、支原体等19种微生物基因组的概述、一般生物学特性，并分别列出了与其DNA复制、修复、重组、翻译及其他功能有关的基因。全书内容新颖，资料详实，层次清晰；是微生物学、遗传学、生物化学、分子生物学、生物信息学等与微生物基因组研究相关的医学工作者的高级参考书。

责任编辑　姚　磊

前言

耗资30亿美元，测序工作量达30亿碱基对的人类基因组计划（Human Genome Project，HGP）是一项举世闻名的跨国科学研究计划。然而，对目前正如火如荼进行的微生物基因组计划（Microbial Genome Projects，MGP），除了专业人士之外，恐怕就没有多少人知道了。由于病毒基因组一般都很小，大都在发现之时即很快被测序。故微生物基因组计划通常是指对基因组达百万 bp 的广义菌类的基因组展开的测序与注释工作。微生物基因组计划与人类基因组计划从来就是密切联系、相得益彰的。由于微生物基因组相对较小，易于操作，它的研究往往先行一步，起到“先行官”的作用。早在人类基因组计划启动之前，1977 年人类就完成了全长 5.3 kb 的 φX174 噬菌体基因组的测序，1990 年完成了 230 kb 的人巨细胞病毒的全基因组测序，1995 年完成了第一个细菌 1.8 Mb 的流感嗜血杆菌基因组的测序……微生物基因组学研究方面所取得的理论和技术进展，为人类基因组计划提供了极有益的借鉴。一些模式微生物，如大肠杆菌和酿酒酵母菌，本身就是人类基因组计划的研究内容。反之，人类基因组计划的强大资金投入和在人类基因组计划中发展和完善起来的生物信息学技术又极大地促进了微生物基因组计划的飞速发展。由于微生物种类的多样性，可以估计，今后微生物基因组计划的总测序工作量将会超越人类基因组计划。

目前，大部分医学微生物都已有代表株完成测序或正在被测序之中。对于一种微生物基因组测序和注释的完成，人们对它的基本生物学特性、致病性、免疫性等重要问题的理解，不管是宏观的还是微观的，都将不再是传统意义上的认识。本书介绍的内容将向你展示，随着各微生物基因组计划的完成，微生物学的知识正在被完全更新。

在安排本书的写作计划时（2001 年初），公共数据库中发布的已经完成测序与注释的微生物（细菌及古细菌）基因组为 37 个（31 个种），其中有的细菌（如大肠杆菌、结核杆菌、衣原体等）测定了多个不同的株。本书选择性介绍了 19 个不同的代表种。到本书完成校样时，完成测序的微生物基因组已达到 80 个，尚有 123 个在进行之中，可见进展之快。这些资料分布在国际著名杂志以及国际数据库中。为了将这些信息及时介绍给读者，我们编写了这本《微生物基因组学》，鉴于它所涉及的知识的全新性和编者们在该领域内知识的有限性，尽管我们做了最大努力，相信书中瑕疵仍然在所难免，恳请专家和读者批评与斧正。

本书中所用的图多为彩图，但为排版、印刷方便起见，将彩图集中放到书后，在图原位置保留了黑白图。由此给读者带来了不便，谨致歉意。

主　编　胡福泉

2002 年 2 月 14 日

目 录

第一章　微生物基因组学研究概况

一、微生物基因组计划进展

人类基因组计划(human genome project, HGP)是一项耗资约 30 亿美元，测序工作量约 30 亿碱基对的举世闻名的跨国科研计划。然而，对目前正如火如荼进行的微生物基因组计划(microbial genome projects, MGP)，除了专业人士之外，恐怕就没有多少人知道了。这不等于说微生物基因组计划是无足轻重的计划。实际上，目前微生物基因组计划已经涉及到 160 多个细菌，测序工作量已达到约 5 亿多碱基对。由于微生物种类的多样性，微生物基因组计划正在以惊人的速度扩展。可以估计，在若干年之后，微生物基因组计划的总投入和测序工作量都将超越人类基因组计划。它对人类产生的深远影响也将是难以估计的。

微生物基因组研究与人类基因组计划是相得益彰的。由于微生物基因组相对较小，易于操作，它的研究往往先行一步，起到"先行官"的作用。例如，在人类基因组计划启动之前，早在 1977 年完成了第一个生物基因组测序就是全长 5.3 kb 的 φX174 噬菌体基因组；1990 年完成了全长 230 kb 的人巨细胞病毒的全序列测序工作。在这期间完成了许多病毒基因组的测序。1995 年完成了第一个细菌流感嗜血杆菌(haemophilus influenzae Rd)的全基因组测序。微生物基因组学研究方面所取得的理论和技术进展，为人类基因组计划提供了极有益的借鉴。一些作为模式生物的微生物(如大肠杆菌和酵母菌)本身就是人类基因组计划的研究内容。反之，人类基因组计划的巨大资金投入和在人类基因组计划中发展和完善起来的生物信息技术又极大地促进了微生物基因组计划的飞速发展。从近 1 年来的进展(表 1-1)足可看出其发展速度。

表 1-1　最近 1 年微生物基因组计划进展情况

	Revised Oct. 20 2000	Revised July 25 2001
Completed	37	55
Annotation in progress	14	17
Sequencing in progress	76	93
In total	127	165

表 1-2 是最近一次公布的已完成的微生物基因组计划一览表。

表 1-2　目前已经完成的微生物基因组计划

Name of bacteria	No. of bp	No. of Proteins
[A]Aeropyrum pernix	1 669 695 bp	2 694 proteins
[B]Aquifex aeolicus	1 551 335 bp	1 522 proteins
[A]Archaeoglobus fulgidus	2 178 400 bp	2 420 proteins
[B]Bacillus halodurans C-125	4 202 352 bp	4 066 proteins
[B]Bacillus subtilis	4 214 814 bp	4 100 proteins
[B]Borrelia burgdorferi	910 724 bp	850 proteins

（续 表）

Name of bacteria	No. of bp	No. of Proteins
[B]Buchnera sp. APS	640 681 bp	564 proteins
[B]Campylobacter jejuni	1 641 481 bp	1 654 proteins
[B]Caulobacter crescentus	4 016 947 bp	3 737 proteins
[B]Chlamydia pneumoniae CWL029	1 230 230 bp	1 052 proteins
[B]Chlamydia pneumoniae AR39	1 229 853 bp	997 proteins
[B]Chlamydia pneumoniae J138	1 228 267 bp	1 070 proteins
[B]Chlamydia muridarum	1 069 412 bp	818 proteins
[B]Chlamydia trachomatis D/UW-3/CX	1 042 519 bp	894 proteins
[B]Clostridium acetobutylicum	3 940 880 bp	3 672 proteins
[B]Deinococcus radiodurans	2 648 638 bp	2 580 proteins
[B]Escherichia coli K12	4 639 221 bp	4 289 proteins
[B]Escherichia coli O157:H7 EDL933	5 528 970 bp	5 349 proteins
[B]Escherichia coli O157:H7	5 498 450 bp	5 361 proteins
[B]Haemophilus influenzae	1 830 138 bp	1 709 proteins
[A]Halobacterium sp. NRC1	2 014 239 bp	2 058 proteins
[B]Helicobacter pylori 26695	1 667 867 bp	1 566 proteins
[B]Helicobacter pylori J99	1 643 831 bp	1 491 proteins
[B]Lactococcus lactis	2 365 589 bp	2 266 proteins
[A]Methanobacterium thermoautotrophicum	1 751 377 bp	1 869 proteins
[A]Methanococcus jannaschii	1 664 970 bp	1 715 proteins
[B]Mesorhizobium loti	7 036 074 bp	6 752 proteins
[B]Mycobacterium tuberculosis	4 411 529 bp	3 918 proteins
[B]Mycobacterium tuberculosis CDC1551	4 403 836 bp	4187 proteins
[B]Mycobacterium leprae	3 268 203 bp	1 605 proteins
[B]Mycoplasma genitalium	580 073 bp	484 proteins
[B]Mycoplasma pneumoniae	816 394 bp	677 proteins
[B]Mycoplasma pulmonis	963 879 bp	782 proteins
[B]Neisseria meningitidis MC58	2 272 325 bp	2 025 proteins
[B]Neisseria meningitidis Z2491	2 184 406 bp	2 121 proteins
[B]Pasteurella multocida	2 257 487 bp	2 014 proteins
[B]Pseudomonas aeruginosa	6 264 403 bp	5 565 proteins
[A]Pyrococcus abyssi	1 765 118 bp	1 765 proteins
[A]Pyrococcus horikoshii	1 738 505 bp	1 979 proteins
[B]Rickettsia prowazekii	1 111 529 bp	834 proteins
[B]Sinorhizobium meliloti	3 654 135 bp	3 341 proteins
[B]Staphylococcus aureus N315	2 813 641 bp	2 595 proteins
[B]Staphylococcus aureus N315	2 878 134 bp	2 697 proteins
[B]Streptococcus pneumoniae	2 160 8371 bp	2 094 proteins
[B]Streptococcus pyogenes	1 852 451 bp	1 696 proteins
[A]Sulfolobus solfataricus	2 992 245 bp	2 977 proteins
[B]Synechocystis PCC6803	3 573 470 bp	3 169 proteins
[A]Thermoplasma acidophilum	1 564 906 bp	1 509 proteins

（续　表）

Name of bacteria	No. of bp	No. of Proteins
[A]Thermoplasma volcanium GSS1	1 585 104 bp	1 499 proteins
[B]Thermotoga maritima	1 860 725 bp	1 846 proteins
[B]Treponema pallidum	1 138 011 bp	1 031 proteins
[B]Ureaplasma urealyticum	751 719 bp	613 proteins
[B]Vibrio cholerae	4 033 464 bp	3 827 proteins
[B]Xylella fastidiosa	2 679 306 bp	2 766 proteins

注：[A] = Archara，[B]= Bacteria

我国正在进行之中的微生物基因组计划有：

1. Thermoanaerobacter tengcongensis，2. 7Mb，Beijing Center，HGP；

2. Staphylococcus epidermidis strain ATCC12228，2. 40Mb，Chinese National Human Genome Center at Shanghai；

3. Leptospira interrogans serovar icterohaemorrhagiae Lai，4. 8Mb，Chinese National Human Genome Center at Shanghai。

二、微生物基因组测序的战略

基因组测序是一个庞大的线性工程，就像修建一条很长的铁路。人类基因组的任务约 30 亿个碱基，细菌基因组一般为数百万。如果用一台自动测序仪从头至尾测序，完成一个测序反应（通常可读出 500～600 bp），再根据读出序列设计新测序引物，往前一步一步走，按此笨办法，测出人类基因组大约需要 900 多年，测出大肠杆菌基因组约需 30 多年，这是不切实可行的。因此，基因组测序必须有一套巧妙的战略。

对于人类基因组，需要先绘制遗传图、物理图谱，尤其是以 STS（sequence-tagged site）为标记（landmark）的物理图谱，然后才可以将基因组“打碎”，在各大实验室分段包干，采用很多仪器同时进行测序，再将测序结果整合起来，成为基因组的序列图谱。序列图谱是基因组计划要完成的分辨率最高的基因组图谱。对于微生物基因组，通常采用 Fleischman 等人在测定流感嗜血杆菌基因组中建立的“鸟枪随机测序”法（shotgun random sequencing）测序，再将测序结果进行整合的战略（图 1-1）。

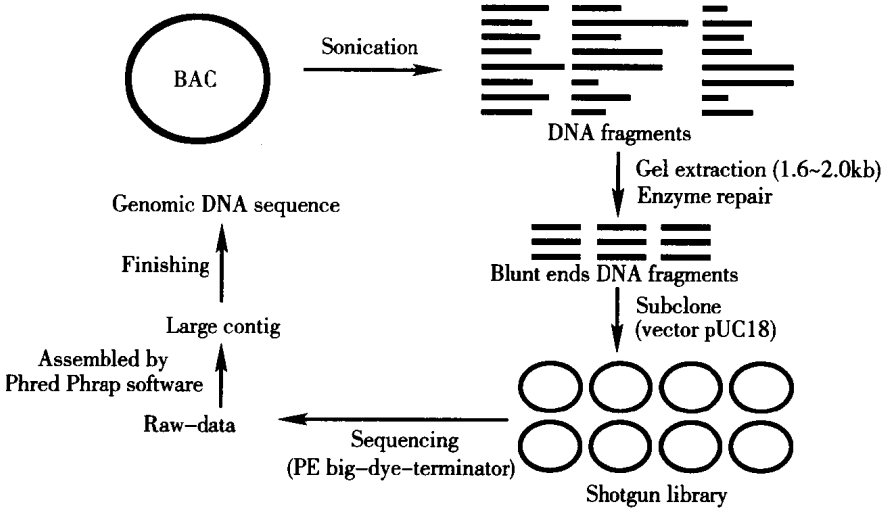


图 1-1　鸟枪随机测序战略示意图

这套战略，包括下列步骤：

1. 基因组 DNA 的制备　制备高质量 DNA。这里的高质量是从 DNA 分子的完整性和纯度来讲的。然后用超声波将 DNA 分子随机剪切为长约 2kb 的片段，并用琼脂糖凝胶电泳分离纯化 2kb 左右的片段，用于建立小片段文库。同时，采用不同的超声条件，再构建一个含大片段（约 15～20 kb）DNA 的文库。

2. 大片段和小片段 DNA 文库的建立　将大片段和小片段 DNA 分别插入测序载体，转染细菌，建立两个独立的 DNA 文库；并对文库的随机质量和容量进行鉴定。制备随机文库 DNA 片段的另一种方法就是不完全酶切法，用此法可以不用构建大、小两个文库，而只需构建一个文库即可。

3. 高流通量测序　最大限度随机化地从构建的文库中挑取克隆制备测序模板。使用多台自动化测序仪进行高流通量测序。每个克隆只需在两端各进行一个测序即可。

4. 组装　不断随机挑取克隆进行测序，并将测序结果存入计算机，phred－phrap 软件根据重叠序列可以确定相邻序列的位置，从而将相互比邻的序列融合成重叠群（contig）. 随着测序量的增多，重叠群的数量会越来越少，但每个重叠群的碱基数越来越大。通常，测序量达到目的基因组的 6～8 倍覆盖量时，就可能覆盖百分之九十几的基因组范围。重叠群形成和组装的原理见图 1-2。

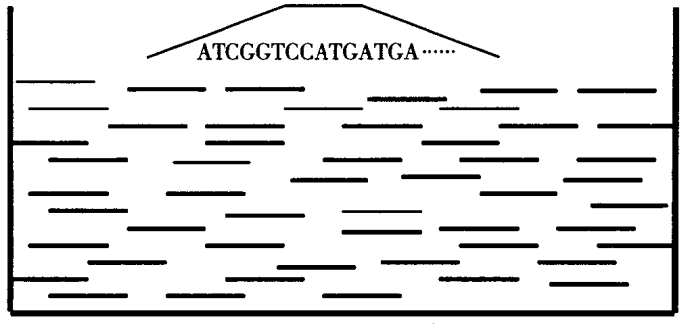


图 1-2　重叠群组装示意图

5. 缺口补平　对于物理缺口，可根据缺口两边的序列，设计引物，以完整 DNA 为模板，用 PCR 扩增，制备补缺模板，再用 primer walking 方法补缺。对于序列缺口和低质量薄弱区域，直接根据两侧已测出序列，设计测序引物，重复测序，直到获得满意结果。

6. 编辑　审查整个序列，对一些不能确定的序列和移框（frame-shifts）进行处理。

7. 注释　使用计算机软件，鉴定出基因组中的各种可以解读其意义的区域，如 ORFs，编码序列（coding sequences，CDS），调控序列，重复序列，复制起始点，已知基因的识别等。

三、微生物基因组的注释

基因组学研究中，测序问题还只是一个技术性问题，只要有相应设备，建立了技术方法，就只是一个工作量的问题了，只要有一批经过训练的熟练工，按照技术方案做下去，就可以完成任务。相对讲，注释问题要复杂得多，没有一批造诣高深的专家组成的智囊团，只靠一两位专家是无法胜任的。测序工作的完成还仅仅是解决了从字母到字母（from letter to letter）排列问题，得到的还只是一部没有空键、没有标点符号、没有词（words）及词组（phrase）、没有句子（sentence）、没有段落（paragraph）的“天书”。要读懂这部天书的意义，就是要解读构成这部天书的“词”、“词组”、“句子”、“段落”乃至“全文”的意义。这一过程就是对基因组的注释过程。可以想见，这一工作不是轻而易举的。通常，对一个微生物基因组进行注释的专家智囊团要包括微生物学、遗传学、生物化学、分子生物学、生物信息学等各个学科领域内的专家。例如，对于铜绿假单胞菌基因组的注释就是由 64 位著名专家共同完成的。细菌基因组研究的论文，署名作者通常都是几十位作者，枯草杆菌基因组论文的作者达到 151 位。他们通常都是参与注释工作的专家而不是测序工作的操作者。

目前，已经发表的那些有关微生物基因组全序列的文章，对于基因组的注释都还是只“部分注释”或“相对注释”，还有许多“词”、“词组”、“句子”、“段落”的意义没能读懂，还得继续读下去。因此，对于一个微生物

基因组的“完全注释”绝不是一朝一夕的。有些问题，在目前阶段还是无法读懂的，它有赖于新技术、新软件、新理论的出现与积累才能解决。“完全注释”也许需要几代人的努力才能完成的任务。显然，对于这样的天书无法待到全部读懂之日才发表。那么，在现阶段，一个微生物基因组完成测序之后，通常应该进行哪些注释工作呢？

（一）碱基组成分析

碱基组成分析是最基本的工作，也是最容易完成的工作。目前，已经有现成的软件来完成这种统计工作。通过这种分析，可以弄清一个基因组的G+C含量。G+C含量是一个物种的特征，在微生物分类学中常常把G+C含量作为分类参数之一。应该注意到的是，在微生物基因组中，碱基的分布常常可能是不均一的。衡量碱基分布均一性有一个参数叫“GC倾斜度”(GC skew = G−C/G+C)。对于基因组中G+C含量的不均一性，有两种解释：一是认为具有明显不同G+C含量区段的DNA，可能在近代进化中有着不同的来源。遗传物质的侧向转移(lateral transfer)或水平转移(horizontal transfer)与重组现象在微生物基因组中是非常常见的。这可能是微生物物种多样性的基础和进化的源泉，但有时又使得分类学上物种的界限变得模糊。第二种解释是，某些特殊的碱基不均一性区域往往是特殊功能所在区，如TATA框常常预示着启动子(promoter)之所在；又如复制起始点和tRNA基因都有它们特征性的碱基不均一性。利用这些特征，可以鉴定基因组中某些功能区段的位点。

（二）密码子使用偏嗜性(codon usage bias)分析

20种氨基酸各有其特定的密码子，许多氨基酸具有多个密码子，这种现象称为简并密码。对于存在于不同物种体内的同一个蛋白质，尽管它们氨基酸序列相同，但在基因水平上，核苷酸序列可能不一样。这是不同物种在编码同一氨基酸时对密码子使用的偏嗜性不一样使然。密码子使用的偏嗜性是物种的特征。对基因组中某些基因的密码子偏嗜性进行统计分析，有可能揭示微生物基因组中通过侧向转移而获得的基因。

（三）开放阅读框的鉴定

开放阅读框(open reading frame，ORF)的鉴定是基因组注释中一项非常重要且必不可少的工作。完成这一工作可以有多种现成软件可供借助，例如NCBI网站中的ORF finder就是其中之一(http://www.ncbi.nlm.nih.gov/gorf/gorf.html)。ORF的鉴定是根据起始密码和终止密码来进行的。在输入分析序列之后，有几个分析参数需要选择：

一是根据序列的物种来源选择遗传密码，数据库中共有22套不同物种的遗传密码可供选择。比如标准码选“1”，细菌码选“11”。

二是限定最短ORF的长度。默认值是100，这时给出的结果中都是大于100个核苷酸的ORFs。通常作为基因的ORF很少是小于100个核苷酸的，但在实际中可能会有例外，特别是编码一些小肽的基因。如欲检出更小的ORFs，可以根据需要选定此项参数。在分析结果的报告中，将会给出6框(six frame)分析结果(表1-3)。

表1-3　ORFs分析结果的报告举例

Frame	from　to	lenth	Frame	from　to	lenth
−3	25 457…28 627	3 171	−1	38 575…40 692	2 118
+1	13 465…15 174	1 710	+2	5 270…　6 394	1 125
−2	32 736…34 277	1 542	+3	19 032…19 592	561
……					

这里的“Frame +1”是指该ORF位于正链上，表明的“读框位置”是从第1号碱基开始依次进行记数。“Frame +2”是指该ORF位于正链上，表明的“读框位置”是从第2号碱基开始依次进行记数。“Frame +3”是指该ORF位于正链上，表明的“读框位置”是从第3号碱基开始依次进行记数。“Frame −1”是指该ORF位于负链上，表明的“读框位置”是从负链上的第1号碱基开始依次进行记数。余依此类推。由此可见，从报

告结果中，可以知道各 ORF 的起、止点和它们位于哪一条链上以及是第几框分析的结果。值得注意的是，不同软件中如果使用不同检索参数，如使用的起始密码子（ATG，GTG，TTG）或终止密码子（TGA，TAA，TAG）不同，输出的结果将有一定差异。由于不同物种对密码子使用有一定偏嗜性，在 ORFs 检测中应该使用何种起始密码及终止密码子应该加以考虑。

（四）移框（frame-shift）检测与校正

“移框”是指 OFR 上的某一点上增加（或减少）了 1 个或 2 个碱基，从而导致改变发生点下游整个氨基酸序列的改变。这种情况即可发生在自然情况下 DNA 复制之时（移码突变），也可发生在基因组测序及组装的过程中。后者是人为造成的，并不反映基因组的本来情况，需要予以矫正。移框检测可以在网站（http://genio. informatik. uni-stuttgart. de/GENIO/frame ）上通过计算机软件来完成。

（五）编码序列（coding sequences，CDSs）分析

编码序列的分析也有专门的分析软件。这类分析软件在确定一段序列为编码序列的过程中，除了考虑起始密码与终止密码之外，还要考虑：① 编码序列上、下游的“语法结构”（grammatical structure）特征。例如，作为前核生物的编码序列的 ORF，在其上游 －4～－19 区域内通常应有核糖体结合位点（Ribosomal Binding Site ，RBS），及通常所说的启动子（promoter）。② 编码序列以及非编码序列中使用的核苷酸语汇（nucleottide words，）如 6 核苷酸（hexamer）语汇的差异进行判断。编码序列的分析有助于确定哪些 ORFs 可能是真正的基因。CDSs 的分析鉴定也有现成的软件，例如 Genemark（网址：http://genemark. Bio logy . gatech. edu/Gene mark /heruristic. cgi/）就是其中之一。

（六）tRNA 基因检索

tRNA 有特殊的“三叶草”结构特征，通常具有“四环”（二氢尿嘧啶核苷环即 DHU 环，反密码子环，额外环及 TΨC 环），三柄（DHU 柄，反密码子柄，TΨC 柄），两臂（可变臂及氨基酸臂），且氨基酸臂的 3′端均为“－C－C－A－OH”结构。根据这些结构特征，tRNA 基因是不难鉴定的，现已有专门用于鉴定 tRNA 基因的软件。在法国 Pasteur 实验室网站（http://bioweb. pasteur. fr/seqanal/），选择 Fast tRNA analysis method 可以完成这一工作。利用 DNAStar 软件中的 GeneQuest 还可进行 tRNA 的二级结构折叠模拟。

（七）rRNA 基因的鉴定

前核生物的 rRNA 具有不同于真核生物 rRNA 的结构特征，利用细菌 rRNA 的序列结构特征，已经发展起来了专门技术用于细菌的鉴定与分类，网站 http://www. midilabs. com 就是专门从事这一服务的。该网站的 MicroSeq 16S rRNA Gene Kits 提供了 16S rRNA 基因的序列资料。一旦基因组全序列测序完成后，可以用现在已知的 16S rRNA 序列鉴定基因组中含有的 rRNA 基因。但是由于 rRNA 基因的多样性，目前鉴定 rRNA 基因的方法远不如鉴定 tRNA 基因那么成熟。

（八）重复序列、插入序列等特殊元件的检索

重复序列（repeated sequence，RS）、插入序列（inserted sequence，IS）、转座子（transposon）、致病岛（pathogenicity island）等特殊序列是不同生物基因组中的常见现象。如重复序列、插入序列几乎存在于各种基因组中，它们有着复杂的生物学意义。微生物基因组不像真核基因组有那么多的冗余序列，编码序列通常占用了全基因组序列的 85％以上（大肠杆菌为 87.5％，绿脓杆菌为 89.4％），相对讲重复序列不像真核生物那么多。而转座子及致病岛则是细菌基因组中常有的特殊结构。这些序列在结构上极富多样性，这给这些序列的鉴定增加了难度。但目前已有许多针对不同特殊序列的软件（尤其是检索重复序列的软件），用户可根据自己的特殊目的进入不同网站选用不同软件。如检索串联重复序列可选用 tandem repeat finder（ http://c3. biomath. mssm. edu/trf. html ）。检索插入序列可进入（http://www. embl-heidelberg. de/～seqanal/single. html）检索转座子可进入 http://nucleus. cshl. org/protarab/TnAnnotation. html 网站。

（九）复制原点鉴定

原核生物的复制原点有其特殊结构特征，如大肠杆菌的复制原点称为 oriC，为一段 245 bp 的序列，其中包括 4 个核苷酸 9 聚体“TTATCCACA”，4 个 9 聚体有两个按同一方向排列，另外两个按相反方向排列。9 聚体序列是 dnaA 蛋白的结合位点，9 聚体结构称为 dnaA 框，根据这些特征，可以鉴定复制原点。基因组序列中的“1”号碱基通常从复制原点开始。

(十)同源性基因检索

对于一个编码序列的ORF,一开始可能不知道它是否是真正的基因或是什么基因。通过同源性(homology)检索,如果发现它与某一已知功能的基因有高度类似性(similarity),就可以推定它可能是什么基因。因此,对于基因组中的每一个ORF都有必要进行特同源性检索。这一工作通常采用美国国立卫生研究院的BLAST软件(http://www.ncbi.nlm.nih.gov/)来完成。BLAST是一个"局域相似性比对分析工具"(Basic Local Alignment Search Tool,BLAST)。可以采用两种查询方式:一种是核酸序列比对核酸序列(nucleotide-nucleotide BLAST,blastn)检索;第二种是用核酸序列比对蛋白序列(Nucleotide query-Protein ,blastx)。两种方法均以FASTA格式输入核酸序列,分析结果输出的报告形式如彩图1-1。

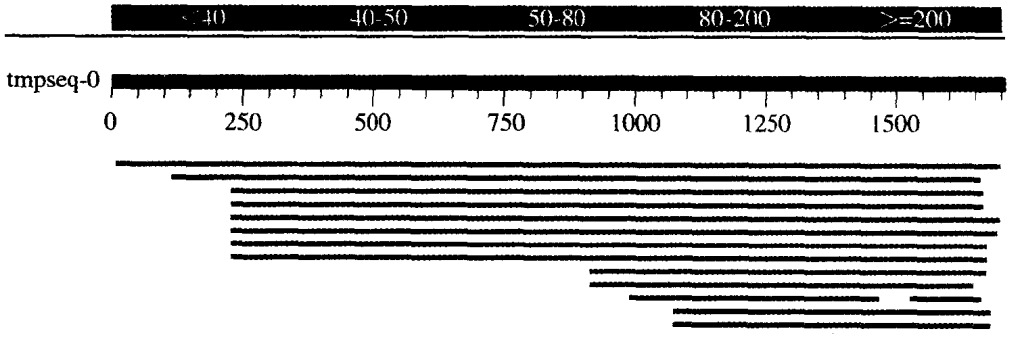


图 1-1 Color Key for Alignment Scores(彩图见彩页)

Sequences producing significant alignments:	Score	E Value
gb|AAD41902.1|AF159357_1 (AF159357) DNA primase [Enterobact...	156	6e−37
ref|NP_064752.1| RPprimase/helicase [Roseophage SIO1] >gi|9...	148	2e−34
ref|NP_052088.1| helicase [Bacteriophage phi−YeO3−12] >gi|6...	121	3e−26
ref|NP_052087.1|DNA primase/helicase [Bacteriophage phi−Ye...	121	3e−26
ref|NP_039642.1| 4B [Bacteriophage T3] >gi|15698|emb|CAA351...	118	2e−25
ref|NP_039641.1| 4A [Bacteriophage T3] >gi|130906|sp|P20315...	118	2e−25
ref|NP_041977.1| gene 4B/helicase [14,15] [Bacteriophage T7...	112	1e−23
ref|NP_041975.1| gene 4A, primase/helicase [14,15] [Bacteri...	112	1e−23
pdb|1CR1|A Chain A, Crystal Structure Of The Helicase Domai...	77	6e−13
pdb|1E0J|A Chain A, Gp4d Helicase From Phage T7 Adpnp Compl...	76	1e−12
gb|AAF83171.1|AE003888_4 (AE003888) replicative DNA helicas...	41	0.038
pir||T36274 probable epimerase − Streptomyces coelicolor >g...	34	4.9
pir||T38994 tetratricopeptide repeat protein − fission yeas...	33	8.4
gb|AAC03120.1| (AF047464) Tpr1 [Schizosaccharomyces pombe]	33	8.4

图 1-3 Color key for Alignment Scores

在这一例子中,图中的彩色横条表明检测到14个对象。从图例颜色可知命中对象的相似性程度:黑色表示相似性分值<40,蓝色相似性分值约40～50,绿色相似性分值约50～80,玫瑰红色相似性分值约80～200。图下面为说明性文字,从上到下按相似性大小,描述了各命中对象。分值(score)越高,相似性越大。以第1个命中对象为例,其E值(expect value,检索时选择参数"1")为"6e−37",这就是说当检索了$6e^{37}$个记录这样大小的数据库时,可望发现1个具有分值为156的对象,显然其概率非常小,也就是说相似性程度非常非常高。如果降低其相似性标准,就可能在更小的检测范围内即可发现更多的相似性对象。检索结束后,需要作出一个判断,什么样的对象才可以认为是同源性序列呢?目前没有一个明确的判定界线,专家需

要结合专业知识作出判断。这里提三点供参考：

1. 一般认为 E 值大于 $1e^{-5}$ 的对象是同源性基因的可能性不大。

2. 如果两个基因是同源性基因的话，在 BLAST 所比对的局域氨基酸的“一致性”(identity)一般应大于30%，且所比对的两个局域片段的小的一个一般不应小于大的一个的 60%。

3. 分值<60，是同源性类似物的可能性不大；分值>100 的命中对象一般可判断为同源性对象；分值在60～99 之间，为可疑同源性对象，需结合其他证据进行判断。

(十一)垂直同源蛋白质的聚类(Cluster of Orthologous Groups of Proteins)分析

这是通过对基因或蛋白质的序列分析来确定它们的种系发生关系(phylogenetic analysis)的聚类方法，对确定基因功能是非常有用的。种系发生分析也是通过同源比较来实现的。在种系发生分析聚类分析中将同源性基因分为垂直同源性基因(orthologus gene)和平行同源性基因(paralogous gene)。这是两个在一般词典中还找不到明确解释，也是非常容易混淆概念的两个名词。在此有必要对它们做一个概念上的交代。

先看看 Waler Fitch(1970，Systematic Zoology 19：99－113)最初对这两个名词的定义：Where the homology is the result of gene dupication so that both copies have descended side by side during the history of an organism，(for example，alpha and beta hemoglobin)，the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects history of the species (for example alpha hemoglobin in man and mouse)，the genes should be called orthologous. 作者认为，如果同源性基因是进化史上种内平行传递的两个复本基因(如 α 和 β 血红蛋白)，就称为“平行同源基因”；如果是存在于不同种间的同源性基因，基因的进化史反映的是物种的进化史，这样的基因就称为“垂直同源基因”。

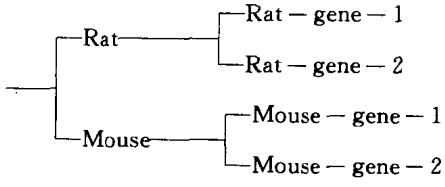
下面是 orthologous genes 和 paralogous genes 简单、明了的定义：

Two genes are to be orthologous if they diverged after a speciation event.

Two genes are to be paralogous if they diverged after a duplication event

意思是说，如果两个同源性基因是源于进化上种的分歧，那么它们是垂直同源基因；如果两个同源性基因是种内复本(duplication)，那么它们是“平行同源基因”。

请看下面的示意：



这里，如果一个祖先基因，经过进化，形成上述 4 个同源性基因，则：

- Mouse－gene－1 and Mouse－gene－2 are paralogous genes
- Rat－gene－1 and Rat－gene－2 are paralogous genes
- Rat－gene－1 is orhologous to Mouse－gene－1 and to Mouse－gene－2
- Rat－gene－2 is orthologous to Mouse－gene－1 and to Mouse－gene－2
- Mouse－gene－1 is orhologous to Rat－gene－1 and to Rat－gene－2
- Mouse－gene－2 is orhologous to Rat－gene－1 and to Rat－gene－2

人们常常容易混淆“orthology” 和"functional equivalence"的含义，这是需要特别注意的。如 Koonin 等人关于这两个名词概念就有这样的描述(Trends in Genetics 1996，12：334－336)：" By definiion，orthologs are genes that are related by vertical descent from a common ancestor and encode proteins with the same function in different species. By contrast，paralogs are homologous genes that have evolved by duplication and code for proteins with similar，but not identical functions". 这意思是说：orthologs 是具有共同祖先、垂直传递下来、且在不同物种中编码具有“相同功能”的蛋白质；而 paralogs 是已进化出来的同源性复本基因，编码具有“类似功能”，但不是功能完全相同的基因”。如果按照这种理解，问题就出来了，比如：上述的 Rat－gene－1

是与 Mouse－gene－1 相同呢还是与 Mouse－gene－2 相同？

实际上，种系发生分析的本质是要确定是否是“垂直同源性”关系，而不是确定功能是否相同。事实上不管是“垂直同源基因”还是“平行同源性基因”，它们都是“类似功能”的基因，而不是功能相同的基因。

需要指出的是，这里将英文"duplication"一词译为“复本”，在中文里不管是使用“复本”还是“副本”“完全相同”，但对于“复本”基因而言，在序列上肯定是有差异的，有时甚至差别还较大。

各 COGs 的聚类是通过“全基因组对全基因组比较(all-against-all sequence comparison)”来实现的. 将来自某一基因组的蛋白，分别一个一个的对另一个基因组来源的每一个蛋白进行比较，找出“最佳命中(best hit)”的蛋白。如此循环在各个已完成全基因组测序的微生物基因组中进行，每一个都需找到“彼此最佳命中(reciprocal best hit)”蛋白。那些彼此最佳命中蛋白就构成一个 COG 类。但是，构成一个 COG 中的任何一个蛋白都必须至少出现在 3 个种系发生的微生物中，才能构成一个 COG 类。同一个 COG 类中的蛋白质是垂直进化的关系(verical evolutionary descent)。值得注意的是，这种垂直进化关系不仅仅指“一对一(one-to-one)”的传递关系，当进化产生出“谱系特异性基因复本(lineage-specific gene duplications)”时，这种关系也可表现为“一对多(one-to-many)”及“多对多(many-to-many)”的关系。

在 COGs 分类中使用的是“COGNITOR”软件(网址：http://www. ncbi. nlm. nih. gov/COG). COGs 数据库及其配置的“COGNITOR”分析软件是作为新测序基因组注释的工作平台和研究基因组进化的平台。在微生物基因组注释中，将各 ORFs 翻译为蛋白质后进行 COG 归类。有些蛋白质具有不同结构域，各结构域还可以有不同的功能。因此结构域数目远远大于蛋白质数目。为管理和分类的方便，又将涉及某一广义生物活动过程的 COGs 或结构域进行更大的归类，共分为 18 大类，其中第 17 类是一类“仅具有一般推定功能的基因(General function prediction only)”。此外，将功能尚不清楚的基因归到第 18 类。此类中的某些基因，尽管根据同源性分析，可能认为是某一基因，但并不清楚该基因在确定的细菌体内是否表达出功能。第 17、18 两大类中的基因/蛋白质或结构域的鉴定依据是不充分的。对于每一大类都赋予一个代码，如将涉及“翻译、核糖体结构和生物合成”的 COGs 及结构域归为一个大类，赋予代码“J”。表 1-4 反映的是 COGs 分类系统的归类情况及各大类中目前已有的 COGs 谱系数目和结构域数目。

表 1-4 COGs 大类归类表

Code	COGs	domains	description
		Information storage and processing	
1J	213	6 739	*Translation, ribosomal structure and biogenesis*
2K	124	5 469	Transcription
3L	181	5 501	DNA replication, recombination and repair
		Cellular processes	
4D	31	842	Cell division and chromosome partitioning
5O	106	3 302	Posttranslational modification, protein turnover, chaperones
6M	148	4 077	Cell envelope biogenesis, outer membrane
7N	121	2 825	Cell motility and secretion
8P	152	5 286	Inorganic ion transport and metabolism
9T	82	3 702	Signal transduction mechanisms
		Metabolism	
10C	220	5 765	Energy production and conversion
11G	157	5 628	Carbohydrate transport and metabolism
12E	227	8 721	Amino acid transport and metabolism
13F	85	2 461	Nucleotide transport and metabolism
14H	152	4 185	Coenzyme metabolism
15I	75	2 721	Lipid metabolism
16Q	62	2 895	Secondary metabolites biosynthesis, transport and catabolism
		Poorly characterized	
17R	432	12 471	General function prediction only
18S	705	6 245	Function unknown

从"J"到"Q"各大类中，已经明确的 COGs 共有 2 136 类，此外还有 432 个 COGs 是仅具有推测功能的，还有 705 个 COGs 功能不清楚。"R"及"S"两大类是功能鉴定尚不充分的 COGs. 功能明确的加上功能鉴定尚不充分的，COGs 总数已有 3 273 个，随着加入 COG 数据库中的基因组数目的增加，COGs 的总数还可能会有变化。

自 2000 年 1 月 COG 数据库发布信息以来，目前又加进去了十多个完成测序的基因组的数据，使 COG 数据库中已经收集的基因组达 44 多个，包括细菌、古细菌和酿酒酵母。最近又加入了一个古细菌 Aeropyrum pernix 及两个多细胞真核基因组，即线虫(caenorhabditis elegants)和果蝇(drosophila melanogaster). 这些数据的加入是很有意义的，这向数据库中提供了一些原来缺乏的进化谱系的数据。由于这些信息的加入，使得基因组注释中，"有把握的功能预测(confident functional predication)"提高到约 50%。另一方面，有相当一部分原来注释为"基因"的 ORFs，在目前的数据库中没有任何相似性的蛋白质，这强烈提示那些 ORFs 不是真正的基因。

在目前已经完成测序的基因组中，有的一个种内测定了多株细菌，如大肠杆菌就测定了 3 株(K-12 型的 MG1655 株，在 O157:H7 中又测定了两个株，EDL933 及 Sakai)。目前 COG 数据库中收集的 44 个基因组分别属于 26 个种。每个种赋予一个代码，详见表 1-5，该表还可反应出每个细菌基因组编码的总蛋白数及其可在 COG 中归类的蛋白数。

表 1-5　目前在 COG 数据库中分析了的生物

Code	Name	Proteins	in COG
A	Archaeoglobus fulgidus	2 420	1 872
O	Halobacterium sp. NRC−1	2 605	1 701
M	Methanococcus jannaschii	1 786	1 330
	Methanobacterium thermoautotrophicum	1 873	1 388
P	Thermoplasma acidophilum	1 482	1 230
	Thermoplasma volcanium	1 499	1 243
K	Pyrococcus horikoshii	1 800	1 378
	Pyrococcus abyssi	1 768	1 456
Y	Saccharomyces cerevisiae	5 955	2 290
	Candida albicans	9 168	2 720
Q	Aquifex aeolicus	1 560	1 329
V	Thermotoga maritima	1 858	1 527
D	Deinococcus radiodurans	3 187	2 226
R	Mycobacterium tuberculosis	3 927	2 585
	Mycobacterium leprae	1 605	1 134
L	Lactococcus lactis	2 267	1 618
	Streptococcus pyogenes	1 697	1 211
B	Bacillus subtilis	4 118	2 870
	Bacillus halodurans	4 066	2 878
C	Synechocystis	3 167	2 159
E	Escherichia coli K12	4 275	3 414
	Escherichia coli O157	5 315	3 662
	Buchnera sp. APS	575	568
F	Pseudomonas aeruginosa	5 567	4 392
G	Vibrio cholerae	3 835	2 820
H	Haemophilus influenzae	1 714	1 542
	Pasteurella multocida	2 015	1 751

（续　表）

Code	Name	Proteins	in COG
S	Xylella fastidiosa	2 831	1 589
N	Neisseria meningitidis MC58	2 080	1 506
	Neisseria meningitidis Z2491	2 065	1 498
U	Helicobacter pylori	1 576	1 096
	Helicobacter pylori J99	1 491	1 075
	Campylobacter jejuni	1 634	1 302
J	Mesorhizobium loti	7 275	4 959
	Caulobacter crescentus	3 737	2 628
X	Rickettsia prowazekii	835	697
I	Chlamydia trachomatis	895	631
	Chlamydia pneumoniae	1 054	648
T	Treponema pallidum	1 036	716
	Borrelia burgdorferi	1 637	696
W	Ureaplasma urealyticum	614	406
	Mycoplasma pneumoniae	689	425
	Mycoplasma genitalium	484	381
	Total	112 878	75 725

将 26 种微生物的代码排列起来，得到如下一个固定排列，用做分析种系发生的模型：

aompkzyqvdrlbcefghsnujxitv

对于每一个 COG 谱系，都可以分析它所包括的蛋白质在哪些细菌中有表达，如果表达就写出该微生物的代码，如果不表达就在相应位置用一个省略号“－”，这就构成了该 COG 谱系的种系发生模式。表 1-6 为几个 COG 谱系的种系发生模式举例。

由表 1-6 可以看出，ATP-依赖的转录调节子，该蛋白质涉及的生化过程归入代码为“K”的大类，该蛋白在微生物结核杆菌(r)、大肠杆菌(e)、绿脓杆菌(f)及生殖支原体(g)中表达，而在目前已测定的其他微生物中都未见表达。表 1-6 第一栏中的数字表明该 COG 类中包含的来源于真核细胞的蛋白，ATP-依赖的转录调节子没有真核细胞来源的。目前在 COG 数据库中的真核基因组仅有 3 个：即 1 个酵母菌、1 个线虫和 1 个果蝇基因组。第二栏中的数字表明该 COG 类中来源于前核细胞的蛋白总数，ATP-依赖的转录调节子共有 8 个蛋白，分别来源于 4 种微生物。

表 1-6　COGs 蛋白谱系种系发生模式举例

种系发生模式	蛋白	代码	编号	蛋白质全名
- 8 ---r---efg---	MalT	[K]	COG2909	ATP-dependent transcriptional regulator
- 10 a-mpkz----	Ssh10b	[K]	COG1581	Archaeal DNA-binding protein
- 13 a--k--d-lb-efghs-	BirA	[KH]	COG1654	Biotin operon repressor
9 66 -o---qvdrlb-efghsn-jx--	CspC	[K]	COG1278	Cold shock proteins
- 65 ---drlbcefg---i	CsgD	[K]	COG2771	DNA-binding HTH domains
5 48 -o--yqvd-lbcefghsnuj-itw	VacB	[K]	COG0557	Exoribonucleases
- 21 --pkzy--lb--h--u---	TenA	[K]	COG0819	Putative transcription activator
- 4 ---efg---	Rsd	[K]	COG3160	Regulator of sigma
2 48 aompkzyqvdrlbcefghsnujxitw	NusG	[K]	COG0250	Transcription antiterminator
- 50 ---vdrlb-efghsnujxitw	GreA	[K]	COG0782	Transcription elongation factor

（胡福泉）