文章编号：1008－3464 (2001) 02－0129－04

单细胞微生物基因组学研究

陆光涛，何勇强，唐纪良
(广西大学 农业部农业分子遗传重点开放实验室，广西 南宁 530005)

摘要：本文介绍了单细胞微生物基因组测序，全基因组序列的注释，以及基因组进一步研究的主要内容。
关键词：单细胞微生物；基因组学；测序
中图分类号：Q751　　**文献标识码**：A

Genomics research of the single cell microbes

LU Guang－tao，HE Yong－qiang，TANG Ji－liang
(The Key Laboratory of Agric· Molecular Genetics of the Chinese Ministry of Agric·，Guangxi Univ·，Nanning 530005，China)

Abstract：It is a review paper which introdues the single cell microbes' complete genome sequencing，the sequence annotation and the main contents of the further research on genomics·
Key words：single cell microbes；genomics；sequencing

1995 年 7 月，《Science》首次刊登了 TIGR (The Institude of Genomic Research) 采用全新测序方法完成的流感嗜血杆菌 (*H aemop op hilus inf luenz ae* Rd) 的全基因组测序与组合的论文[1]，这是人类完成的第一个单细胞微生物的全基因组序列的测定，标志着基因组时代的真正开始。

单细胞微生物基因组学研究主要源于人类基因组计划的实施。1990 年 10 月开始实施的人类基因组计划，是希望用 15 年时间，完成人类全部 23 对染色体的遗传图谱、序列图谱和转录图谱，同时还对大肠杆菌 (*Escherichia coli*)、酵母 (*Saccharomyces cerevisiae*)、美丽线虫 (*Caenorhabditis elegans*)、果蝇 (*Drosop hila melanogaster*) 和老鼠 (*Mus musculus*) 等 5 种模式生物的全基因组序列进行测定。

人类基因组约有 30 亿个 bp，编码 5－10 万个基因，以当时的技术水平要在 15 年内完成这一计划，是相当艰巨的。如何采用新技术和新方法进行 DNA 序列测定以及基因功能的表达模式的确定，成为人们积极探索的问题。以 J· C· Venter 领导的基因组研究所 (TIGR) 的科学家们于 1991 年首先提出了 EST (expressed sequence tags) 的概念，并利用这一方法成功地寻找和鉴定了在细胞组织中表达的基因，即通过构建细胞组织的 cDNA 文库，随机选择 cDNA 克隆，利用载体引物一次测定插入片段的 3′端和 5′端约 300～500 bp，这些 cDNA 的部分序列即为 EST，通过与 GeneBank 等数据库所收集到的基因的 EST 比较，可鉴定出细胞组织中的功能基因[2,3]。由于专门计算机软件的发展，已能够处理大量的 DNA 资料并进行高质量的序列组合，科学家们考虑能否借鉴 EST 的策略来进行微生物的全基因组序列测定。1994 年 4 月，Johns Hopkins 大学的 H· O· Smith 和 J· C· Venter 等人合作，开始利用基因组鸟枪测序法对 *H · inf luenz ae* 进行全基因组序列测定，不到 1 年时间，即完成了这一项工作[1]。接着，

[20] **收稿日期**：2000－08－17
基金项目：国家高技术 "863" 计划资助项目 (863－101－02－07－01)
作者简介：陆光涛 (1968－)，男，广西钦州人，广西大学助理研究员，硕士，在职博士研究生。

TIGR 的 Fraser 等人利用这一方法用不到 6 个月时间也完成了尿道支原体(*Mycoplasma genitalium*) 的全基因组测序工作[4]。这表明鸟枪测序法是成功的、快速准确的测序策略。

1 全基因组的测序

鸟枪测序法基本原理是在基因组随机文库的基础上,直接对随机文库的各个克隆进行测序,即对质 粒载体中所插入的DNA片段,同时从3′和5′两端进行测序,获得大量的两端序列已知而中间部分未知的克隆群,然后通过专门的计算机软件将测得的序列拼接成连续的序列图。各克隆片段中间未测定的总碱基数,即缺口(gap) 与测定的总碱基数有关,其规律遵从泊松(poisson) 公式的一个推论:$P_0 = e^{-m}$, P_0 为基因组中某个碱基未被测定到的概率, m 为所测定的碱基总数与基因组碱基总数相比的倍数。m 越大, P_0 值越小。那么,当所测定的碱基数越大,基因组中未被测定到的碱基数就越少。当 $m = 1$ 时, $P_0 = e^{-1} = 0.37$ 即基因组中有 37%的碱基未测定到。当 $m = 5$ 时, $P_0 = e^{-5} = 0.0067$, 即当所测定的碱基数是基因组总碱基数的 5 倍时, 基因组中有 0.67%的碱基未被测定到。同时, 当基因组DNA总长度为L,测定的随机克隆的插入片段数为n时,总缺口长度为Le^{-m},每个缺口平均大小为L/n。

基于以上原理, 单细胞微生物基因组鸟枪测序法的主要过程为: 第一, 基因文库的构建。基因文库的高度随机性是测序的基础, 构建的文库克隆数要达到一定数量, 以保证经末端测序的克隆片段的碱基总数大于基因组碱基总数的 5 倍以上。第二, 测序。利用正向及反向引物, 在质粒模板上进行测序反应, 对所测定的克隆的插入片段通过一次反应对 3′和 5′两端测序。第三, 序列拼接。将所测得的序列通过专门的计算机软件进行序列组合, 序列片段按严格标准连接成数个连锁群 (contig), 然后对连锁群进行排序和缺口填补。物理缺口(没有模板DNA 与之对应的缺口) 的填补有 4 种策略: 印迹法、肽链连接法、λ克隆排序法和 PCR 确定连锁群等。序列缺口用引物步移法来填补。在填平缺口, 获得全基因组序列之后, 再用专门软件进行建立序列的图形交互界面、基因组数据组合编辑等工作[1,5]。与传统测序法相比, 鸟枪法测序省去了许多中间步骤, 能快速准确地对基因组进行测序。传统测序法是采用克隆到克隆 (Clone - by - clone) 的策略, 即是对基因组 BAC 文库中各克隆进行测序, 在每个克隆的测序过程中, 先对每一个克隆进行亚克隆, 然后对每个克隆的各个亚克隆进行直接测序, 将各个亚克隆的序列拼接成一个克隆的序列图, 然后再将各克隆的序列拼接成一个连续的序列图。这一方法中间步骤多, 测序速度慢, 在测序前要首先对拟测序的区域进行物理图谱的构建。

2 全基因组序列的注释

在基因组序列测序完成, 得到全基因组序列图谱后, 工作的重点即是分析这一条由数百万个 4 种碱基对线性排列而成的长链所包含的遗传信息。目前的工作主要是以下两个方面:

2.1 G+C 含量的分析

获得全基因组序列图谱后, 首先是分析基因组中 GC 含量百分比, 然后考察各个区域 DNA 的 GC 含量, 不同 DNA 区域的 GC 含量并非一致, GC 富含区或 AT 富含区可能意味着该区域具有特殊功能。在 *H. influenzae* 中, GC 富含区内对应着 6 个 rRNA 操纵子(operon) 和一个隐藏的类似 Mu 噬菌体[1]。在 *M. genitalium* 中, rRNA 操纵子的 GC 含量为 44%, tRNA 操纵子 GC 含量为 52%, 均较其基因组平均 GC 含量的 32%要高得多[4]。这表明富含 GC 对 rRNA 和 tRNA 形成正确的二级结构是必需的。在以后的单细胞微生物基因组分析中, 也发现类似情况。嗜热高温菌 *Aquifex aeolicus* 和嗜热菌 *Thermotoga maritima* 的 16S - 23S - 5S rRNA 操纵子的 GC 含量比其基因组 GC 含量高得多, 16S - 23S - 5S rRNA 高 GC 含量是嗜热细菌的特征之一[6,7]。在 *B. subtilis* 的几个 GC 含量低的区域, 即 AT 富含区, 则含有前噬菌体或其它插入序列[8]。

考察 DNA 区域的 G - C/(G + C) 比率, 当这一比率发生显著变化时, 表明该区域可能含有 DNA 复制起点。这在 *B. subtilis*[8] 和古细菌 *M. jannaschii*[9] 中均得到证实。这表明在复制的先导链和后续

链间核苷的组成不均匀。但在 *T. maritima* 中却没有在具有这一特征的区域发现复制始点[7]。

通过对基因组分析，发现基因组内存在重复序列、插入因子、前噬菌体和前噬菌体部分残余 DNA。*B. subtilis* 基因组中至少包含 10 个前噬菌体或前噬菌体的部分残余 DNA[8]。*E. coli* 的基因组也发现多个前噬菌体[10]。基因组中的前噬菌体已经失去溶源生长所必需的基因，但仍然携带有具有一些其它功能的基因。这表明在物种进化过程中前噬菌体对基因水平转移起重要作用[8,10]。基因组中的重复单位的大小可以是由数十个 bp 至数百个 bp 不等，重复次数也是大小不同[7,8]。

2.2 ORF 的分析

在进行某种生物基因组的转录翻译水平分析前，首先对该物种已知的基因所编码的蛋白质进行分析，考察其密码表，设定起始密码子和终止密码子，然后利用专门软件，从整个基因组中寻找开放阅读框(ORF) 即蛋白质的可能编码区域。在对 *H. influenzae* 的 ORF 的考察中获得令人惊讶的结果，在 1 743 个 ORF 中，通过与 GenBank 数据库中其它物种的已知基因比较，仅有 1 007 个 ORF 的功能是已知的，另有 736 个 ORF 所编码的蛋白质功能是未知的，这些未知的 ORF 一部分是在 GenBank 等数据库找到相应的蛋白质序列与之匹配，但蛋白质功能未知，另一部分则是在数据库中找不到相应的蛋白质与之匹配[1]。在以后的基因组 ORF 分析中，均有相当一部分 ORF 功能未知。*E. coli* 的 4 288 个 ORF 有 1 630 个是功能未知的，占其 ORF 总数的 38%[10]；而在 *B. subtilis* 的 4 100 个 ORF 中，有 1 722 个 ORF 功能未知，占其 ORF 总数的 42%[8]；即使是基因组最小的 *M. genitalium*，在其 470 个 ORF 中，也有 96 个在 GenBank 中没有找到任何其它生物体的已知的蛋白质序列与之匹配[4]。对于已知功能的 ORF，根据其生物学功能，按 Riley 分类法[1,4]进行功能类群分类，共分为 14 个类群，或者按照 COD 法(Clusters of Orthologous Groups) 将所有 ORF 分为 18 个功能类群。

在 ORF 的起始密码子中，使用频率最高的是 ATG，78%的 *B. subtilis* 和 85%的 *E. coli* 的 ORF 均以 ATG 为起始密码子，TTG、GTG 的使用频率较低，在 *B. subtilis* 中这两种起始密码子的使用频率分别为 13%和 9%，在 *E. coli* 中则分别为 3%和 14%。另外，*E. coli* 中还发现有 15 个 ORF 使用稀有起始密码子 ATT 和 CTG[8,10]。

3 基因组的进一步研究

人们通过对基因组的研究，可以了解生物体各种代谢过程，遗传机制和生命活动所需的基本条件以及生物特殊功能如致病性的遗传基础。在微生物全基因组序列的测定与注释完成后，基因组学的研究工作主要集中在以下几方面内容：

3.1 基因功能的研究

在全基因组测序和分析完成后，人们最关心的自然是基因的功能问题。在所有的已完成测序的单细胞微生物基因组的 ORF 中，都有相当部分 ORF 功能是未知的，其中的一部分 ORF 没有在 GenBank 中找到任何与之匹配的蛋白质。如此之多的 ORF 所编码的蛋白质结构和功能不为人所知，究竟是这些 ORF 所编码的蛋白质在生物体内存在的时间极短，很快就被降解掉？还是这些基因所编码的蛋白质及其功能一直未被人们发现呢？如果是后者，则表明尽管人们已经对细胞的结构与功能进行近一个世纪的研究，但还有将近一半的细胞生物学和生物化学的功能仍未被人们所认识。全基因组序列的测定与分析以及功能基因组的研究，为人们研究基因功能提供极好机会和手段。目前通过基因缺失和平行分析的方法，正对 *S. cerevisiae* ORF 的功能进行深入研究[11]。

3.2 系统进化研究

基因组学研究的一个重要内容就是基因组间的比较研究。人们在 DNA 水平上对不同生物体进行比较研究，可以了解生物物种间的进化关系以及不同生物体生命活动的异同。70 年代以前，生物主要分为原核生物和真核生物两大类，1977 年，C. R. Wose 等通过对 200 多种原核生物 16S rRNA 和真核生物 18S rRNA 的寡核苷酸序列分析比较，从中发现了生命的第三种形式——古细菌。1978 年，R. H. Whittaker 等提出了三原界学说，将生物分为三个原界：古细菌原界、真细菌原界和真核生物原界。

按照这一学说，在生物进化过程的早期，存在一类各种生物的共同祖先，由它分三条路线进化分别形成三个原界。微生物全基因组的分析和基因组的比较研究，为三个原界学说提供有力支持。通过对古细菌 *M · jannaschii*[9]、真细菌 *H · influenzae*[1]和真核生物 *S · cerevisiae*[12]基因组的比较研究，发现古细菌 *M · jannaschii* 在产能、固氮以及细胞分裂方面的有关基因与真细菌 *H · influenzae* 有很高的同源性，而在转录翻译系统以及分泌系统方面的有关基因与真核生物 *S · cerevisiae* 的关系更近。这说明古细菌与真核生物的亲缘关系较与原核生物的亲缘关系更近。可以预计随着越来越多的全基因组被分析和基因组间的比较研究，人们将可以重新绘制生物的系统进化树。

3.3 病原物致病性的研究

基因组学的研究，使人们更全面深入地了解微生物致病性和致病机理。*H · influenzae* 全基因组的分析研究发现，编码细胞膜上脂多糖的 DNA 序列中至少有 9 个有关的位点含有多个衔接重复的 4 核苷酸，这些重复序列的缺失或增加，可导致转录信号或阅读框架的改变，使其可以逃避人体免疫监视系统[1]。病原菌 *N · meningitidis* 和 *H · influenzae* 基因组的比较研究发现，尽管两种病原菌均生长于人的鼻咽并能引起髓膜炎，但它们的代谢方式有很大不同。*H · influenzae* 缺乏 TCA 循环和 ED 途径，缺少吸收离子的运输系统，与 *N · meningitidis* 相比，仅有少部分的基因与电子传递有关，但却有相当多的基因与氨基酸和碳水化合物的吸收有关。这表明 *H · influenzae* 更多地依赖底物磷酸化途径而不是氧化磷酸化途径获取生命活动所必需的能量。代谢途径的显著差别，可能与这两种病原菌在人体寄主不同生理条件下而表现不同致病能力有关[1,13]。

3.4 基因水平转移的研究

基因水平转移即基因在两个同时存在的物种之间转移，但一直没有令人信服的证据。单细胞微生物全基因组测序的完成为人们研究基因水平转移提供了极好的研究机会。基因水平转移在多种微生物的基因组学研究中已得到证实。基因组中具有一些基因水平转移的辅助证据：基因的 GC 含量突然与相邻区域的 DNA 序列的 GC 含量非常不同，基因组中残存着插入序列的部分序列以及含有前噬菌体和前噬菌体残存的部分序列。通过对 *N · Meningitidis* 基因组的分析，已鉴定了三个主要的基因水平转移区域，其中两个包含有与其致病性有关的基因[13]。对 E · coli 基因组的进一步研究发现，其 4 288 个 ORF 中的 755 个即与基因水平转移有关[14]。

T · maritima 是从高温环境下的海底淤泥中分离到的一种进化非常缓慢的嗜热真细菌，通过对其基因组的比较研究，发现其 52%的基因与真细菌非常相似，24%的基因与古细菌非常相似，并且其中的 81 个基因聚集在大小为 4～20 kb 的 15 个区域内，其中的 7 个区域中保守的基因排列顺序，以前仅见于古细菌中，这表明在细菌和古细菌的祖先在进化过程中，曾发生过基因水平转移[7]。

基因组学的研究，揭示越来越多的生命本质，它使得常规的以基因为研究对象的分子生物学研究产生了一次飞跃。截止 2000 年 7 月底，已有约 50 种微生物完成了基因组的测序，其中一部分尚未发表相关研究论文，另有 100 余种微生物目前正进行基因组的测序，我国于 1999 年底也启动了数项微生物基因组测序计划。随着越来越多生物体完成基因组测序，以及基因组的进一步分析研究即蛋白质组研究的开始，人类对生命本质认识将更为全面与深入。

参考文献：

[1] FLEISCHMANN R D, ADAMS M D, SMITH H O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd [J]. Science, 1995, 269: 496-512.

[2] ADMAS M D, KELLEY J M, VENTER J C, et al. Complementary DNA sequencing: expressed sequence tags and human genome project [J]. Science, 1991, 252: 1 651-1 656.

[3] ADAMS M D, DUBNICK M, VENTER J C, et al. Sequence identification of 2 375 human brain genes [J]. Nature, 1992, 355: 632-634.

(下转第 153 页)

Manceaux, 1909

刚地弓形虫 *T. gondii*(Nicolle et Manceaux, 1908)

宿主与寄生部位:猪。血液、淋巴液、体腔液及有核细胞。

5.2 无类锥体纲 Aconoidasida Methora, Peters et Haberkorn, 1980

5.2.1 血孢子虫目 Haemospororida Danilewsky, 1885

(1) 疟原虫科 Plasmodiidae Mesnil, 1903

① 住白虫属 *Leucocytozoon* Sambon, 1908

卡氏住白虫 *L. caulleryii* Mathis et Leger, 1909

宿主与寄生部位:鸡。白细胞。

沙氏住白虫 *L. sabrazesi* Mathis et Leger, 1910

宿主与寄生部位:鸡。白细胞。

5.2.2 梨形虫目 Piroplasmorida Wenyon, 1926

(1) 巴贝斯科 Babesiidae Poche, 1913

① 巴贝斯属 *Babesia* Starcovici, 1893

双芽巴贝斯虫 *B. bigemina* (Smith et Kiborne, 1893)

宿主与寄生部位:黄牛、水牛、奶牛。红细胞。

6 纤毛门 Ciliophora Doflein, 1901

6.1 动基裂纲 Kinetofragminophorasida de Puytorac, 1971

6.1.1 毛口目 Trichostomatorida Butschli, 1889

(1) 小袋科 Balantidiidae Doflein et Reichenow, 1858

① 小袋属 *Balantidium* Claparede et Lachmann, 1858

结肠小袋虫 *B. coli*(Malmsten, 1857) Stein, 1862

宿主与寄生部位:猪。结肠。

(未完待续)

(责任编辑　胡春柳)

(上接第 132 页)

[4] FRASER C M, GOCAYNE J D, VENTER J C, et al. The minimal gene complement of *Mycoplasma genitalium* [J]. Science, 1995, 270: 397 - 403.

[5] VENTER JC, SMITH H O, Hood L. A new strategy for genome sequencing [J]. Nature, 1996, 381: 364 - 366.

[6] DECKERT G, WARREN P V, SWANSON R V, et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus* [J]. Nature, 1998, 392: 353 - 358.

[7] NELSON K E, CLAYTON R A, VENTER J C, et al. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima* [J]. Nature, 1999, 399: 323 - 329.

[8] KUNST F, OGASAWARA N, MOSZER I, et al. The complete genome sequence of the gram - positive bacterium *Bacillus subtilis* [J]. Nature, 1997, 390: 249 - 256.

[9] BULT C J, WHITE O, VENTER J C, et al. Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*[J]. Science, 1996, 273: 1 058 - 1 073.

[10] BLATTNER F R, BLOCH C A, SHAO Y, et al. The complete genome sequence of *Escherichia coli* K - 12 [J]. Science, 1997, 277: 1 453 - 1 474.

[11] NINZELER E A, SHOEMAKER D D, ASTROMOFF A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis [J]. Science, 1999, 285: 901 - 906.

[12] GOFFEAU A, BARRELL B G, BUSSEY H, et al. Life with 6000 genes [J]. Science, 1996, 274: 546 - 567.

[13] TETTELIN H, SAUNDERS N J, VENTER J C, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC 58[J]. Science, 2000, 287: 1 809 - 1 815.

[14] LAWRENCE J G, OCHMAN H. Molecular archaeology of the *Escherichia coli* genome [J]. Proc Natl Acad Sci, 1998, 95: 9 413 - 9 417.

(责任编辑　梁健)