DOI：10.13523/j.cb.19970507
《生物工程进展》1997, Vol. 17, No. 5

专　　论

单细胞自由生物基因组全序列的测定和比较基因组研究

毕高峰　朱立煌
(中国科学院遗传研究所　北京　100101)

摘要　近几年来五种单细胞生物的基因组计划得以完成。本文介绍了从五种生物的全基因组序列获得的一些成果, 包括全基因组鸟枪法测序 基因组分析和新比较基因组等三个方面, 并对生物基因组计划的研究方法作一些探讨。

关键词　基因组计划　全基因组鸟枪法测序　基因组分析　比较基因组学

致力于揭示生物全部遗传信息的基因组计划近几年来取得了很大进展。1995 年 J. C. Venter 领导的基因组研究所(TIGR, The Institute of Genomic Reseach)完成了第一个单细胞自由生物全基因组分析, 即流感嗜血杆菌(Haemophilus influenzae Rd)基因组全序列测定。这项工作也是 TIGR 发展的新的测序策略—— 全基因组鸟枪法测序成功的范例[1]。1996 年 TIGR 及其合作者用此种基因组测序策略又完成了两种生物的基因组全序列测定, 它们是迄今所知的具有最小基因组的单细胞生物尿殖道支原体(Mycoplasma genitalium), 和一种不同于原核、真核生物的单细胞生物产甲烷古细菌(Methanococcus jannaschi)[2,3]。其后德国 Richard Herrmann 等发表了肺炎支原体(Mycoplasma pneumoniae)基因组全序列的测定分析工作, 与此同时, 历时七年(1989—1996 年)的第一个真核生物酿酒酵母(saccharomyces cevevisiae)基因组计划在欧共体及美、日、加、英等各国实验室共同努力下得以完成[4,5]。这些生物基因组全序列的完成比人们原先预期的要快得多。TIGR 的工作则被誉为这一领域的一匹黑马[6]。预计 1997 年大肠杆菌(Escherichia coli K)的基因组计划将完成, 美丽隐杆线虫(caenothabditis elegans)的基因组计划将于 1998 年完成[7]。最受瞩目的人类基因组计划(HGP, Human Genome Project)正如期进行, 以目前的进展速度, 肯定能够于 2005 年前完成[8]。

以基因组为研究对象的基因组计划是常规的以基因为对象的分子生物学研究的一次飞跃。生物全部基因组信息的获得使得人们开始对生物的认识更为全面更为深入, 从而使基因遗传学跨越到基因组遗传学。

本文试对上述的研究进展的主要成果作一概况的综述。

一、新的测序策略—— 全基因组鸟枪法测序

TIGR 创立的全基因组鸟枪法测序特别适合基因组为 2Mb 以下的微生物的基因组计划。运用该策略对基因组直接进行序列分析, 绕过传统的测序程式中物理图谱的构建, 利用高速计算机完成数量庞大的基因组数据处理工作, 从而能在较短时间内完成基因组全序列的测定[1]。

其基本原理和程序如下: 就某个生物的基因组而言, 在对基因组文库全部克隆片段进行末端序列测定中采测到的碱基数, 即缺口

(Gap)与测定的总的碱基数有关,随着测定的总碱基数的数值增大,缺口的总碱基数目会迅速减小。其公式是泊松公式的一个推论,即 $P=e^{-m}$。其中 P 为基因组中某个碱基未被测定的概率,m 为所测定的碱基数与基因组大小相比的倍数。m 越大 P 值越小。由此公式推断 m 值达到 5 时,即随机测定的碱基数达到基因组 5 倍时,基因组中未测定的碱基数为基因组总碱基数的 0.67%($e^{-5}=0.0067$)。对于象流感嗜血杆菌这样大的基因组(1.83Mb),可以推算会产生 128 个缺口,每个缺口平均长度为 100bp[1]。

根据以上原理,TIGR 采用以下步骤进行全基因组鸟枪法测序:第一,建立高度随机、插入片段大小为 2kb 左右的基因组文库。基因组文库高度随机性是进行鸟枪法测序的基础。因此,建立文库时须用机械切割的办法把基因切成小片段。如流感嗜血杆菌文库中插入片段的大小在 1.6～2.0kb 左右[1],产甲烷古细菌插入片段在 2.5kb 左右[3]。所有插入片段都经过琼脂糖凝胶电泳分离回收,再进行转化。文库中克隆数要达到一定数量,即经末端测序的克隆片段的碱基总数应达到基因组 5 倍以上。第二,高效、大规模的末端测序。对文库中每一个克隆,进行两端测序,TIGR 在完成流感嗜血杆菌的基因组时,使用了 14 台测序仪,用三个月时间完成了必需的 28,463 个测序反应,测序总长度达 6 倍基因组大小。第三,序列集合。TIGR 发展了新的软件 TIGR—ASSEMBLER 来完成这部分工作。同时 TIGR 修改了序列集合规则,使形成克隆连锁群(contig)的条件更为严格,以最大限度地排除错误的连锁匹配。TIGR-ASSEMBLER 将上万个序列数据进行集合。将各个片段连接成数个连锁群。他们是用内存为 512M 的计算机 SPARCenter2000 完成全部的运算。在集合运算中,对所有的连锁群都进行了相互的搜索,以达到最大限度的连锁匹配。第四,缺口的填补。在序列集合完成之后,用 TIGR 发展的软件 ASM-ASSIGN 将未定位的连锁群进行排序定位,转换成两种待填补的缺口,一是没有模板 DNA 与之对应的物理缺口,一是有模板 DNA 但未测序的序列缺口。为此他们又建立插入片段大小为 15-20kb 的 K文库以备缺口填补,并对 K文库中的克隆进行末端测序。主要应用 4 个步骤来填补物理缺口,包括印迹法,肽链连接,K克隆排序和 PCR 确定连锁群测序等。K克隆是缺口填补的主要片段来源,K 文库可以提供一些在小文库中致死的 DNA 模板,并且 K克隆末端的序列信息非常适合缺口的填补,序列缺口是用引物步移的方法填补的,即从每一个模板的两端进行 PCR,用以填补序列缺口。在缺口填平,获得整个基因组序列之后,再用 TIGR 发展的软件 TIGR-EDITOR 纠正读码框 ORF,建立序列的图形交互界面,基因组数据集合编辑等工作[1]。

TIGR 创立的随机测序策略使得生物学家可以在短时间内获得生物基因组的全序列信息,解决了微生物基因组计划中最关键的问题。

二、全基因组序列分析——基因组学的新内容

全基因组序列是由无标点的线性排列的双链 DNA 碱基对组成。如何分析这条 DNA 长链,对获得的全序列进行诠释,是基因组研究的关键,即基因组分析。上述五个生物完整基因组计划提供了对基因组全序列进行解读的经验。由于此类工作面对的是庞大的数据,全部分析必须在计算机上进行,又称计算机辅助基因组分析[4]。具体有以下内容。

1. 数据存放。基因组全序列是一庞大的信息。其数据管理是将其存放于计算机网络中的数据库。如 TIGR 完成的三种生物存放于 GDSB(基因组序列数据)[1,2,3]。肺炎支原体存放于 GenBank[4]。由于计算机网络提供的便利,基因组序列的宝贵资源可以供全世界使用,任何人通过网络可以在其中对之进行诠释、检查、纠正和补充,以充实所获得的信息[1]。

2. 碱基百分含量分析。碱基百分含量的考察是通过软件 WINDOW 进行,即以一定大小的窗口如 5000bp 来对基因组各段进行 GC 百分含量的考察[4]。与全基因组平均的 GC 百分

含量相比, 基因组中不同区域的 GC 百分含量并不一致。无论是 GC 富含区还是 AT 富含区, 都可能是一些特殊功能的区域。肺炎支原体的 GC 百分含量高和 GC 百分含量低的区域都是开放读码框(ORF) 密度低的区域[4]。流感嗜血杆菌的 GC 富含区有一个隐藏的 mu-噬菌体, 此区 GC 百分含量比基因组平均值高[1]。酵母基因组中GC 百分含量高的区域对应染色体的重组值较高的区域, 而 AT 含量高的区域对应重组值较低的区域, 包括着丝粒和端粒[5]。尿殖道支原体GC 百分含量最低的区域对应着复制起始点(ori) , 而 GC 百分含量高的区域对应着 rRNA 和 tRNA[2]。流感嗜血杆菌 GC 百分含量高的区域也对应着 6 个 rRNA 基因[1]。从这点来看, GC 含量高是构成 tRNA 和 rRNA 二级结构的前提条件[2]。

3. ORF 的考察。在生物的基因组全序列测定之后, 最关键的工作是确定基因组中各个基因的功能。首先要找到基因组中全部的 ORF。一般这项工作是借助于软件如 FRAMES 完成[4]。其原理是通过此种生物的已知基因考察其密码表, 设定起始密码子和终止密码子, 再经过计算机分析确定每一个ORF。其次, 用 GENEMARK[4]来估测每一个ORF 的编码可能。将每个预计编码序列在蛋白质数据库中进行搜索, 如蛋白质数据库 NRBP 和 PIR 等, 以严格的匹配原则考察预计编码序列与已存在的蛋白质的同源性。每一个 ORF 根据其与已知蛋白质序列的比较结果可以分为三类[1]。一类是通过比较确知其功能的, 另一类是在数据库中有相匹配的蛋白质序列, 但不知其功能。最后一类是在数据库中找不到任何匹配的蛋白质序列。

对第一类, 即已确定功能的 ORF 可以按生物功能归类, 以便进行功能分析和基因组之间的比较。TIGR 和 Richard Herrmann 参考 Riley 提出的微生物 14 个生物学功能类群进行 ORF 的功能分类[1, 2, 3, 4]。这 14 个功能类群[1]包括: (1) 氨基酸代谢、(2) 酶、辅基和转运蛋白的合成、(3) 细胞包装、(4) 细胞进程、(5) 关键中介物代谢、(6) 能量代谢、(7) 脂酸和磷脂代谢、(8) 嘌呤、嘧啶、核苷和核苷酸的代谢、(9) 调控系统、(10) 复制、(11) 转录、(12) 翻译、(13) 转运蛋白和结合蛋白、(14) 其他。这 14 个类群包括 102 种生物学功能。将基因组全部 ORF 归入这 14 个功能类群中, 再对每一类群的ORF 进行分析。可以对生物的各种代谢过程, 遗传机制, 基本生命所需条件等进行描述, 同时, 还能对此种生物特有的功能如致病性、感染性进行分析[1]。对这五种生物的ORF 分析揭示了许多新的知识, 以下仅举三例:

(1) 、通过流感嗜血杆菌能量代谢类群的 ORF 分析, 了解到在这种生物中缺乏三羧酸循环(TCA) 中必需的三个酶的基因, 它们是柠檬酸合成酶基因、异柠檬酸脱氢酶基因和顺乌头酸酶基因。因此推断流感嗜血杆菌 TCA 缺失, 不能合成谷氨酸, 因为谷氨酸的供体是 TCA 的中间产生物 A-戊二酸[1]。(2) 、在尿殖道支原体基因组中有一个特殊的 ORF, 属于一个重复序列, 称为 MgPa。MgPa 编码 29kD 的蛋白质。通过全基因组考察, 共发现有 9 个与 MgPa 同源的重复序列, 这些重复序列之间可以发生重组, 这可诱导尿殖道支原体群体中抗原性改变, 为逃避宿主中的免疫攻击提供条件[2]。(3) 、遗传冗余性是指生物中存在着两套或更多套基因编码同类的蛋白质, 遗传冗余是新基因进化的原材料。酵母染色体端部特有的一个性质就是遗传冗余性。如酵母第三染色体的两端和第五、第十一染色体的端部 ORF 的 DNA 序列相互之间同源。另外一些着丝粒附近的 ORF 也具有同源性。还有两个同源的 ORF 编码柠檬酸合成酶等。这些遗传冗余性现象有助于阐明酿酒酵母基因组的进化历程[5]。

4. 基因组全部遗传信息按一定结构排列在染色体上, 成为一个有机的整体。通过五种生物全基因组序列的研究, 已经获得了关于原核生物和真核生物的染色体结构的一些基本知识。如在原核生物基因组分析中对原核环状基因组的结构单元, 如复制起始位点、重复序列、基因家族和基因调控系统等结构的序列特征已有所了解。从真核酵母的基因组全序列的分析中揭示了真核生物染色体的结构特点, 对真核生物特有的着丝粒序列

和端粒序列进行了分析。总之, 通过全基因组分析可望阐明生物基因组是如何组织和协调各基因的功能, 以完成生命进程的。

三、比较基因组学的新时代

生物全基因组信息的获得还开辟了比较基因组学的新时代, 使得不同生物之间的比较基因组的研究可以直接在 DNA 序列的水平上进行, 从而使我们对生物及生命的了解更为深刻。下面从三方面的工作来阐明这类工作的重要意义。

1. 最小的有细胞生物——尿殖道支原体基因组

尿殖道支原体有已知最小的基因组。将他与其他相对较大基因组的比较可以对生命赖以维持的一套最小的基因进行探讨。由此可以确定能自我复制的细胞必需的一套最少的核心基因。两方面的比较工作已经开展, 即 TIGR 关于尿殖道支原体与流感嗜血杆菌基因组的比较[2] 和 Herrmann 关于尿殖道支原体与系统关系较近的肺炎支原体的比较基因组研究[9]。

流感嗜血杆菌的基因组为 1.83Mb, 而尿殖道支原体的基因组只有 0.58Mb, 二者相差 3 倍多。这就提出了一个问题, 基因组大小到底影响了什么, 是基因数目减少, 还是基因尺度缩小? 通过对两个生物 ORF 的考察, 可以看到, 流感嗜血杆菌基因大小平均 900bp, 尿殖道支原体的基因为 1040bp, 基因大小差不多; 而两者的基因密度也没有明显差异, 流感嗜血杆菌为 1 个基因/ 1042bp, 尿殖道支原体为 1 个基因/ 1235bp。可见基因组尺寸减小并不引起基因密度的增加和基因本身尺寸的减小。二者差别在于基因的数目上, 流感嗜血杆菌基因组有 1743 个 ORF, 尿殖道支原体只有 470 个 ORF。对已经确认其功能的 ORF, 按功能分类分别对两生物进行比较, 可以看到流感嗜血杆菌分布在各个功能类群的 ORF 都多于尿殖道支原体。

尿殖道支原体中有 90 个已确认功能的 ORF 是流感嗜血杆菌所没有的。其中 60% 基因与格兰氏阳性菌或其他支原体生物同源, 这表明这些基因编码的蛋白质与系统分化有关。另外还有 96 个 ORF 在基因库中与任何已知序列都不匹配, 可能是此种生物或同类生物的新基因[2]。

Herrmann 对两种支原体的全基因的比较也许能从另一个角度回答关于最小基因组的问题。已经知道, 两种支原体生活在同一种生活环境中, 具有相同的烧瓶样形态和血清交叉反应。但它们也有不同之处, 如对人类不同组织的不同致病性。

从基因组整体比较, 二种生物有一定程度不同, GC 百分含量相差 8%, 引起 GC% 不同的主要原因是三联体码中第三个位置的碱基不同。同时 GC% 的不同影响到了 AT 或 GC 在密码子第一或第二位置的氨基酸分布。基因组组织结构的比较说明二者的基因次序是不保守的, 虽然两种生物的基因组都可分为 6 个区, 但这 6 个区的次序在两种生物中是不同的, 这与肺炎支原体的基因组的几个重复序列 RepMP1、RepMP2/3、RepMP4 和 RepMP5 有关。在尿殖道支原体基因组中仅可以看到这些重复顺序的痕迹, 即前面提到的 MgPa。除 RepMP1 外, 其余几个重复序列与 MgPa 有高度同源性。由此可以推论尿殖道支原体在这些重复顺序中发生了基因组各区的重组, 这导致两个支原体基因组各区顺序的排列不同。

对两种基因组比较的结果表明肺炎支原体基因组包含着所有尿殖道支原体的 ORF。在基因的组成上二种生物存在值得注意的一致性, 表现在一些功能系统的基因数量和同源性上。这涉及一些基本的生命过程, 如 DNA 复制、转录转译、基因表达的调节、蛋白分泌系统和能量保持系统。这些功能系统在两种生物中涉及的基因数量的一致说明尿殖道支原体确实在这些功能上集合了一套最小量的基因, 若再加以减小可能会影响细胞的生存。两种生物类似还表现在几个基因和几种功能在两生物中的共同缺失上。

研究者认为存在于肺炎支原体基因组中的尿殖道支原体没有的遗传信息是解释两种生物学上不同表现的关键, 也是确定最小的自我复制的细胞中必需功能的关键。肺炎支原体有

209 个 ORF 在尿殖道支原体基因组中不存在(有的 ORF 在肺炎支原体的拷贝数比尿殖道支原体要多,这可能会引起二者基因组大小的不同但不会引起新的功能的出现)。这些 ORF 重要功能包括 hsd 型限制修饰系统、两个磷酸烯醇式丙酮酸: 碳氢化物磷酸转移酶系统、NADP 依赖的乙醇脱氢酶系统以及全套的精胺酸双氢酶生化途径上所有的酶。另外还有一些包含有重复 DNA 序列的 ORF。这些 ORF 编码的蛋白是导致肺炎支原体具有区别于尿殖道支原体的特性[9]。

尿殖道支原体有已知最小的基因组,在概念上相当于通过遗传工程失活或删除非必需基因后的最小细胞生物。不过限定一个最小基因组必须限定这个细胞的生存条件。因为如果有适当的营养,基因组虽去除一些蛋白质也可以存活,但不能在原来条件下存活。细胞最小化的研究[9,10]可以有两个途径,其一是通过尿殖道支原体基因组,逐步减少基因;另一种途径[10]是现在已尝试的通过对尿殖道支原体与流感嗜血杆菌这两个亲缘关系较远的生物基因组的比较选取其共同的基因(共 240 个),再加上一些其他基因,最后组成一套含 256 个基因的最小基因组。后一种途径会因实验条件的控制而使必需基因缺失,因为一些必需基因在不同生物中并不具有同源性。两种支原体生物基因组的比较则能较好地回答一个最小的、自我复制细胞所必需的功能问题,因此可以更加准确地了解一个自我复制生物的基因组成,在此基础上对生命的多样性也会加深了解。

2、既非原核、也非真核的古细菌——古细菌的基因组

产甲烷球菌是高温高压条件下生活的一种自养的单细胞生物。其全基因组序列的测定为全面考察此生物在进化网络上的位置提供了可能。对它的比较基因组研究是借助于与流感嗜血杆菌和尿殖道支原体两种原核生物的全基因组序列,以及同酵母等真核生物基因组的序列比较进行的。在产甲烷古细胞的基因组中较多的 ORF 无法确认其功能,与现代生物同源性相对较小。一些结论只能从已确定功能的 ORF 中得出。

比较结果表明,古细菌产甲烷球菌与原核生物有着共同的染色体组织与结构,如环状基因组、基因操纵子等,它在能量产生和固氮方面的基因与原核生物有很高的同源性。产甲烷球菌基因组中 20 多个编码无机离子运输蛋白的 ORF 与细菌的多糖合成酶基因同源,而且这些基因在基因组中组成基因家族统一调控,这一点也与原核生物类似。另外产甲烷球菌与细胞分裂有关的蛋白质与细菌类似。至于产甲烷球菌代谢途径的关键酶大部分还没有识别出来,可能采用与现代生物完全不同的形式。

然而产甲烷球菌在细胞遗传信息传递,尤其是转录和翻译系统,以及分泌系统方面与真核生物同源。这说明产甲烷古细菌与真核生物亲缘关系更近。这有以下几方面的证据:

1) 所有的真核生物和原核生物共同具备的核糖体亚单位在产甲烷球菌中都具备,而且,古细菌有真核生物特异的核糖体蛋白,而没有原核生物特异的核糖体蛋白。

2) 翻译延伸因子 EF-12, EF-2 及氨酰—RNA 合成酶与真核生物有很高的同源性。

3) 11 个 RNA 合成酶中 5 个与所有生物同源,6 个只与真核同源。

4) 翻译起始系统与真核生物类似,而与原核生物区别很大。

5) 在复制系统中,没有细菌 DNA 合成酶 I 和流感嗜血杆菌和尿殖道支原体相类似的酶。而具有真核生物 A和 E DNA 复制酶和细菌 DNA 合成酶Ⅱ,以及几个与古细菌同源的酶。

6) 在古细菌中发现了 5 个组蛋白基因,与真核的组蛋白 H2a, H2b, H3 和 H4 及真核转译相关的 CAAT 结合因子 CBF-A 同源,这表明古细菌在基因表达和 DNA 超螺旋动力学方面与真核生物类似。

7) 在产甲烷球菌基因组中一些基因编码内含子蛋白(intein)。

以上比较基因组学提供的结果表明,在进化系统树上,古细菌与真核生物亲缘关系比原核生物更近。由此可见,在自养生物的三个分

支,细菌、古细菌和真核生物中,细菌的分化发生较早[5]。

3、最简单的真核生物——酿酒酵母的基因组

单细胞真核生物酿酒酵母的基因组大小为12,068kb,比单细胞的原核生物和古细菌大一个数量级。但酿酒酵母的基因组是最小的真核生物(无论是单细胞还是多细胞生物)基因组。可见真核生物的生物学复杂性远大于原核生物和古细菌。这可从酿酒酵母与另外四种基因组全序列已测定的生物之间的基因组比较看出。酿酒酵母基因组共有5887个ORF,这比原核生物和古细菌要多很多。酿酒酵母的基因密度为1个基因/2kb,比原核生物如流感嗜血杆菌和尿殖道支原体等要小。

然而与其它酵母、真菌以及多细胞真核生物的基因组相比,酿酒酵母的基因组具有作为最小真核基因组的特征。这体现在两个方面,其一是基因密度。酿酒酵母基因组是最致密的真核生物基因组,据目前已知的真核生物数据,酿酒酵母基因组密度高于同类的裂殖酵母(Schizosaccharomyces pombe)的基因密度1个基因/2.3kb。已知简单的多细胞生物线虫的基因密度为1个基因/6kb,而人类的基因密度大约为1个基因/30kb。可见在真核生物中有着与原核生物不同的规律,即基因组的大小影响基因密度的大小。其二是酿酒酵母只有4%的编码基因有内含子,而裂殖酵母则有40%编码基因有内含子[5]。

结　语

基因组计划无论是在方法上还是在生物学基础理论上都取得了很大的成就。从根本上讲,基因的阐明最终要在基因组中完成。有了基因组的数据,生物学家可以摆脱以往的由点到面的思路,从新的角度进行研究。尤为重要的是,随着各种生物基因组计划的相继完成,都将在方法策略和基因诠释方面促进人类基因组计划的完成,与此同时,生物技术又将登上一个新的制高点——基因组工程。

参考文献

[1] Fleichmann R D, Adam M D, White O et al. Science, 1995, 269: 496- 512

[2] Fraser C M, Gocayne J D, White O et al. Science, 1995, 270: 397- 403

[3] Bult C J, Whit O, Olsen G J et al. Science, 1995, 273: 1058- 1073

[4] Himmelreich R, Hlbert H, Plagens H et al. Nucleic Acids Res., 1996, 24: 4420- 4449

[5] Goffeau A, Barrell B G, Bussey H et al. Science, 1996, 274: 546- 567

[6] Nowak R. Science, 1995, 269: 468- 470

[7] 茅矛, 况少青, 陈国强等. 生物工程进展, 1996, 16: 2-6

[8] Schuler G D, bofuski M S, Stewart E A et al. Science, 1996, 274: 540- 546

[9] Himmelreich R, Plagens H, Hlbert H et al. Nucleic Acids Res., 1996, 24: 4420- 4449

[10] Mushegian A R and Koonin E V. Proc. Natl. Acad. Sci., 21: 10268- 10273

Whole-genome Sequencing and the Comparative Genomics of Free Living Single Cell Organisms

Bi Gaofen　Zhu Lihuang

(Institute of Genetics, Chinese Academy of Sciences, Beijing　100101)

Abstract　Five genome projects on single cell microbes have been finished in recent years. Some achievements obtained from these studies are described, including whole-genome shotgun sequencing, genomic analysis, and comparative genomics. A further discussion is also given on the methodology of genome projects.

Key words　Genome project, Whole-genome sequencing, Genome analysis, Comparative genomics