16个完整基因组中核糖体蛋白基因排列顺序保守性的研究*

王 宁 陈润生**
(中国科学院生物物理研究所, 北京 100101)
王永雄**
(Department of Statistics, University of California at Los Angeles, CA 90095-1554, USA)

摘要 在16个完整基因组中, 对70个核糖体蛋白基因的排列顺序进行了分析. 这些基因在每个基因组中平均构成9~14个操纵子. 结果显示: (1) L3和L14操纵子中包含的20多个核糖体蛋白的排列顺序在古细菌和真细菌这两个不同界的基因组中都非常保守; (2) 有些操纵子结构分别是真细菌或古细菌所特有的; (3) 在每一界中, 有些操纵子中的核糖体蛋白的基因排列顺序在不同的物种中存在一定的差异, 这种差异可以用来推测物种之间的亲缘关系. 这种方法为研究古老物种的起源和进化提供了一条新途径.

关键词 完整基因组 操纵子 核糖体蛋白 基因排列顺序

随着越来越多的完整基因组测序的完成, 人们开始在整个基因组的水平上研究物种中的基因是如何组织排列在一起的[1], 并试图通过比较基因的排列顺序来研究物种的系统发育关系. 本文针对一类最古老和保守的蛋白——核糖体蛋白, 在16个完整基因组中进行了分析. 这16个基因组包含12个真细菌和4个古细菌, 通过对这些古老物种中的核糖体蛋白基因的组织、排列进行分析, 为研究生命的起源和这些物种之间的进化关系提供可能.

1 数据及来源

数据取自公开发表的数据库, 截止到本工作开始, 在NCBI的GenBank中一共可以找到16个完整基因组. 这16个完整基因组包括12种真细菌和4种古细菌. 12种真细菌分别是: *Aquifex aeolicus* (*A. aeo*)[2], *Borrelia burgdorferi* (*B. bur*)[3], *Treponema pallidum* (*T. pal*)[4], *Chlamydia trachomatis* (*C. tra*)[1)], *Escherichia coli* (*E. coli*)[5], *Haemophilus influenzae* Rd (*H. inf*)[6], *Helicobacter pylori* (*H. pyl*)[7], *Mycoplasma genitalium* (*M. gen*)[8], *Mycoplasma*

1999-03-09收稿, 1999-05-19收修改稿
*国家自然科学基金(批准号: 39392900, 39830070和19890380)和DMS-9703918 from NSF资助项目
**联系人
1) Stephens R S, Kalman S, Lammel C J, et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*

pneumoniae (*M. pneu*)[9], *Bacillus subtilis* (*B. sub*)[10], *Mycobacterium tuberculosis* (*M. tub*)[11], *Synechocystis PCC*6803 (*S. pcc*)[12]; 4 种古细菌分别是: *Methanococcus jannaschii* (*M. jan*)[13], *Methanobacterium thermoautotrophicum* (*M. the*)[14], *Archaeoglobus fulgidus* (*A. ful*)[15] 和 *Pyrococcus horikoshii* (*P. hor*)[16]. 这些物种的名称、分类及基因组的大小列在表 1 中.

表 1 16 个完整基因组的名称、分类及大小

基因组名称(简称)	基因组的分类	基因组大小/ bp
Aquifex aeolicus (*A. aeo*)	Eubacteria; Aquificales; Aquificaceae; Aquifex	1 551 335
Borrelia burgdorferi (*B. bur*)	Eubacteria; Spirochaetales; Spirochaetaceae; Borrelia; Borrelia burgdorferi group	910 724
Treponema pallidum (*T. pal*)	Eubacteria; Spirochaetales; Spirochaetaceae; Treponema	1 138 011
Chlamydia trachomatis (*C. tra*)	Eubacteria; Chlamydiales; Chlamydiaceae; Chlamydia	1 042 519
Escherichia coli (*E. coli*)	Eubacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae; Escherichia	4 639 221
Haemophilus influenzae (*H. inf*)	Eubacteria; Proteobacteria; gamma subdivision; Pasteurellaceae; Haemophilus	1 830 138
Helicobacter pylori (*H. pyl*)	Eubacteria; Proteobacteria; epsilon subdivision; Helicobacter	1 667 867
Mycoplasma genitalium (*M. gen*)	Eubacteria; Firmicutes; Low G+C gram-positive bacteria; Mycoplasmas and walled relatives; Mycoplasmatales; Mycoplasmataceae; Mycoplasma	580 073
Mycoplasma pneumoniae (*M. pneu*)	Eubacteria; Firmicutes; Low G+C gram-positive bacteria; Mycoplasmas and walled relatives; Mycoplasmatales; Mycoplasmataceae; Mycoplasma	816 394
Bacillus subtilis (*B. sub*)	Eubacteria; Firmicutes; Low G+C gram-positive bacteria; Bacillaceae; Bacillus	4 214 814
Mycobacterium tuberculosis (*M. tub*)	Eubacteria; Firmicutes; Actinomycetes; Mycobacteria; Mycobacteriaceae; Mycobacterium	4 411 529
Synechocystis PCC6803 (*S. pcc*)	Eubacteria; Cyanobacteria; Chroococcales; Synechocystis	3 573 470
Methanococcus jannaschii (*M. jan*)	Archaea; Euryarchaeota; Methanococcales; Methanococcaceae; Methanococcus	1 664 970
Methanobacterium thermoautotrophicum (*M. the*)	Archaea; Euryarchaeota; Methanobacteriales; Methanobacteriaceae; Methanobacterium	1 751 377
Archaeoglobus fulgidus (*A. ful*)	Archaea; Euryarchaeota; Archaeoglobales; Archaeoglobaceae; Archaeoglobus	2 178 400
Pyrococcus horikoshii (*P. hor*)	Archaea; Euryarchaeota; Thermococcales; Thermococcaceae; Pyrococcus	1 738 505

2 方法

在这 16 个完整基因组中, 平均每个基因组有 50 多个核糖体蛋白, 这些核糖体蛋白大多数是以操纵子(operon)的形式出现. 其中, *E. coli* 是目前实验室研究得最清楚的一个, 它的许多启动子和转录起始位点可以在 GenBank 的注释中找到, 即可以很容易找到它精确的操纵子结构, 但对其他 15 个基因组来说, 这些信息却是未知的.

2.1 在 15 个基因组中确定操纵子的结构

在寻找包含核糖体蛋白的操纵子时, 我们发现操纵子中除了包含核糖体蛋白基因外, 通常还包括一些其他的非核糖体蛋白基因. 因此, 在预测包含核糖体蛋白的操纵子时, 制作了一些

限制作为形成操纵子的判据．我们认定如果一些核糖体蛋白基因(中间允许插入一些其他基因)在顺序上连续(基因间的间隔序列小于 70 个碱基)，而且开读框架方向相同，则这些基因形成一个操纵子．通过这种方法，我们在每个基因组中，预测出包含核糖体蛋白的每个操纵子的结构．

2.2　利用动态规划算法确定同源基因

在比较 16 个完整基因组中每个操纵子的基因排列顺序时，我们发现，由于 GenBank 基因命名的混乱，一些同源的非核糖体蛋白基因(甚至包含几个核糖体蛋白)在不同基因组中常常具有不同的基因名，比如：对于同一段编码区，它的基因名在 *B. sub*，*B. bur* 和 *T. pal* 3 个物种中分别为 *ylqC*，*BB0696* 和 *TP0906*．所以，为了保证结果的准确性，我们利用动态规划算法[17]，对操纵子中的所有基因(包括核糖体蛋白和非核糖体蛋白)进行基因序列的相似性比较，来确定这些具有相同或不同名称的基因在不同基因组中是否为同源基因．这样，虽然增加了大量的计算，但保证了同源基因匹配的精确性．我们在氨基酸序列的水平上做序列对齐，如果不同基因组中的两个序列的相似比大于 25%，则认为这两个基因为同源基因[18]．

2.3　在 **16** 个完整基因组中比较每个操纵子中的基因排列顺序

对于每个操纵子中的每一个基因，确定其在其他基因组中的同源基因后，我们在 16 个基因组中比较每个操纵子中的基因排列顺序，如果这些基因的排列顺序在大多数基因组中都是一致的(允许有基因的插入或缺失)，我们则认为该操纵子的基因排列顺序是保守的，并以该操纵子中的第 1 个核糖体蛋白的名称命名为操纵子名．

3　结果

通过对 16 个基因组的分析，我们在每个基因组中平均预测出 9～14 个包含核糖体蛋白的操纵子．在这些操纵子中，共包含 70 个核糖体蛋白和 25 个非核糖体蛋白，结果列于表 2．在表 2 中，我们可以清楚地看到在古细菌和真细菌两界中所共有及各自特有的操纵子结构，同时在真细菌的不同物种中，还存在一些基因排列顺序的差异，这些差异可以反映出物种间的亲缘关系．

3.1　在古细菌和真细菌两界中都保守的操纵子

在所有的包含核糖体蛋白的操纵子中，我们发现 L3 和 L14 操纵子中的 20 多个核糖体蛋白的排列顺序在真细菌和古细菌两界中的每一个基因组中都是保守的．这有力地支持了真细菌和古细菌是起源于共同的祖先(表 2)．

3.2　在不同界中存在其特有的操纵子

从表 2 我们也可以清楚地看出，在同一界中，物种的核糖体蛋白的排列顺序是相对保守的．有些操纵子结构是真细菌这一界所特有的，而有些操纵子结构是古细菌这一界中所特有的．比如在真细菌中的 L35，L21 和 S16 操纵子结构，在古细菌中找不到相应的结构，还有核糖体蛋白 L16 只出现在每一种真细菌的 L3 操纵子中，并不存在于古细菌中；相反，核糖体蛋白 L18E 和 L24E 也只出现在每一个古细菌中．

3.3　在同一界中，有些核糖体蛋白排列顺序的差异能反映出物种间的亲缘关系

虽然核糖体蛋白的操纵子结构在每一界中是保守的，我们仍然可以在不同的物种中找到一定的差异，这些差异可以很好地反映出物种的亲缘关系．比如：

表2 在16个完整基因组中预测出的包含核糖体蛋白的操纵子结构[a)]

	真细菌（Eubacteria）											古细菌（Archaebacteria）			
基因名	*A. aeo*	*B. bur*	*T. pal*	*C. tra*	*E. coli*	*H. inf*	*H. pyl*	*M. gen*	*M. pneu*	*B. sub*	*M. tub*	*S. pcc*	*M. jan*	*M. the*	*A. ful*	*P. hor*
rpL28					√	√					√					
rpL33	√	√	√		√	√	√			√	√	√				
secE	—	√	√	√	√	√	—			√	√	√				
nusG	√	√	√	√	√	√	√			√	√	√		√		
rpL11	√	√	√	√	√	√	√	√	√	√	√	√		√		
rpL1	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL10	√	√	√	√	√	√	√	*pth*	*pth*	√	√	√	√	√	√	√
rpL12	√	√	√	√	√	√	√	*yacA*	*yacA*	√	√	√	√	√	√	√
ybxB	—	—	—	—	—	—	—	*nadph*	*nadph*	√			*rpoA*	*rpoA*	*rpoA*	*rpoA*
rpoB	√	√	√	√	√	√	√	*lgt*	*lgt*	√	√		*L30E*	*L30E*	*L30E*	*L30E*
rpoC	√	√	√	√	√	√	—			√	√		*nusA*	*nusA*	*nusA*	*nusA*
rpS12	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpS7	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
fusG	—	—	—	√	√	√	√	√	√	√	√	—	√	—	√	√
rpS6	√	√	√	√	√	√	√	√	√	√	√					
ssb	√	√	√	—	*priB*	*priB*	√	√	√	√	√					
rpS18	√	√	√	√	√	√	√	√	√	√	√					
rpL9	—	√	√	√	√	√	—	√	√		√			√		
rpS10	√	√	√	√	√	√	√	√	√	√	√	√		√	√	√
rpL3	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL4	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL23	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	—
rpL2	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpS19	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL22	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpS3	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL16	√	√	√	√	√	√	√	√	√	√	√	√	—	—	—	—
rpL29	—	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpSUI1	—	—	—	—	—	—	—	—	—	—	—	—	√	√	—	—
hpy	—	—	—	—	—	—	—	—	—	—	—	—	√	√	√	√
rpS17	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

表 2(续 1)

	真细菌 (Eubacteria)												古细菌 (Archaebacteria)			
基因名	*A. aeo*	*B. bur*	*T. pal*	*C. tra*	*E. coli*	*H. inf*	*H. pyl*	*M. gen*	*M. pneu*	*B. sub*	*M. tub*	*S. pcc*	*M. jan*	*M. the*	*A. ful*	*P. hor*
rpL14	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL24	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpS4E	—	—	—	—	—	—	—	—	—	—	—	—	√	√	√	√
rpL5	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpS14	√	√	√	—	√	√	√	√	√	√	√	—	—	√	√	—
rpS8	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL6	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL32E	—	—	—	—	—	—	—	—	—	—	—	—	√	√	√	√
rpL19E	—	—	—	—	—	—	—	—	—	—	—	—	√	√	√	√
rpL18	√	√	√	√	√	√	√	√	√	√	√	√	√	—	√	√
rpS5	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL30	√	√	—	—	√	√	—	—	—	√	√	—	√	√	√	√
rpL15	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
secY	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
adk	√	—	—	—	—	—	—	√	√	√	√	√	√	√		√
map	√	—	—	—	—	—	√	√	√	√	√	—				
infA	√	—	—	—	—	—	√	√	√	√	√	√				
rpL36	√	√	√	—	√	√	√	√	√	√	√	√		√→3		
rpS13	√	√	√	√	√	√	√	√	√	√	√	√	√→*S4E*	√→*S4E*	√→*S4E*	*S4E*→*1*
rpS11	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpS4	√	—	—	—	√	√	√	—	—	—	√	—	—	—	—	—
rpoA	√	√	√	√	√	√	√	√	√	√	√	√	*rpoD*	*rpoD*	*rpoD*	*rpoD*
rpL17	√	√	√	√	√	√	√	√	√	√	√	√	*L18E*	√	*L18E*	√
truA	—	—	—	—	—	—	—	—	—	—	—	√	—	—	—	—
rpL13	√	√	√	√	√	√	√	√	√	√	√	√	√	—	√	√
rpS9	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
rpL31	—	—	—	—	—	—	—	—	—	—	—	√	—	—	—	—
infC		√	√	√	√	√	√	√	√	√	√	*radC*				
rpL35		√	√	√	√	√	√	√	√	√	√	√				
rpL20		√	√	√	√	√	√	√	√	√	√	√				
rpL21		√	√	√	√	√	√	√	√	√	√	√				
ysxB		√	√	—	—	—	—	√	√	√	—	—				
rpL27		√	√	√	√	√	√	√	√	√	√	√				

表2(续2)

	真细菌（Eubacteria）											古细菌（Archaebacteria）			
基因名	*A. aeo*	*B. bur*	*T. pal*	*C. tra*	*E. coli*	*H. inf*	*H. pyl*	*M. gen*	*M. pneu*	*B. sub*	*M. tub*	*S. pcc*	*M. jan*	*M. the*	*A. ful*	*P. hor*
rpS15							√	√								
ATP-							√	√								
rpL28							√	√								
rpS16		√	√	√	√	√	√	√	√	√	√	√				
ylgC		√	√	—	—	—	√	—	—	√	√	—				
ylgE		√	√	—	√	√	√	—	—	√→1	√	—				
trmD		√	√	√	√	√	√	√	√	√	√	√				
rpL19		√	√	√	√	√	√	√	√	√	√→1	√				
rpL18E													√	√	√	√→1
reL13P													√	—	√	√
rpS9P													√	√	√	√
rpL24A													√	√→1	√	
rpL11													√	√	√	
rpS27A													√	√	√	
rpS24													√	√	√	
rpL34E													√	√		√
cmk													√	√		√
rpL14E													√	√		—
rpL31E													—	√	√	
rpL39E													√	√	√	
hyp													√	√	—	
hyp													√	√	√	
rpS19E													√	√	√	
rpL24E													√	√	√	—
rpS28E													√	√	√	√
rpL7AE													√	√	√	√
rpL44													√		√	
rpS27													√		√	
rpS3AE													√		√	
hyp													√		—	
rpS2P													√		—	
rpL15E													√		√	

a) 最左边一列为基因名，其中 *rp* * 代表核糖体蛋白的名称. 16个物种根据亲缘关系从左至右排列. √标记在该基因组中存在该种基因，而且是按照最左侧基因的排列顺序分布的；—标记该基因组中不存在该种基因；→代表在此基因后有其他基因插入，其中后面的数字为插入基因的个数；*hyp* 代表通过理论方法预测出的，不知确切功能的基因；双线标记着操纵子的边界. 在 *rpS13* 这行中，古细菌中的 *S4E* 标记着有核糖体蛋白 S4E 插入在这里

（1）在表 2 中，*M. gen* 和 *M. pneu* 具有完全相同的操纵子结构，包括位于 L11 和 S12 两个操纵子之间的 *pth*，*yacA*，*nadph* 和 *lgt* 4 个基因，而且这 4 个基因是这两个物种所特有的，在其他真细菌 *B. bur*，*T. pal*，*H. pyl*，*B. sub* 和 *M. tub* 中，这 4 个基因被 L10 操纵子所代替．

（2）*E. coli* 和 *H. inf* 具有完全相同的操纵子结构，而且在 S6 操纵子中，*priB* 基因代替了其他物种中的 *ssb* 基因．不仅如此，这两个物种在 *ysxB*，*ylqE* 和 *ylqC* 这 3 个基因是否出现在 L21 和 S16 操纵子中的状态也是完全一致的．所以，从这个表中，可以明显地看出这两个物种具有最近的亲缘关系．

（3）*A. aeo* 明显不同于其他基因组．首先，核糖体蛋白 L35 和 L20 连续出现在其他 11 个真细菌的基因组中，而在 *A. aeo* 中是散布的．对于 L21 和 S16 操纵子也是同样．其次，两个大的操纵子 L3 和 L14 在其他真细菌基因组中，或者融合在一起，或者彼此相邻，而在 *A. aeo* 中，它们却被 500 kb 的序列所隔开．值得一提的是，这两个操纵子在所有 4 个古细菌中也是融合在一起的，这与 *A. aeo* 代表着现存细菌中的最古老的祖先的假设是一致的．

（4）*P. hor* 明显与其他古细菌属于一界，但与其他 3 种古细菌相比，它缺少 L24A，S27A，和 L39E 操纵子．这意味着它代表着古细菌中的另一分支．

3.4　不同的操纵子在基因组中经常保守地聚集或融合在一起

从表 2 中可以看出：

（1）L10 操纵子与 L11 操纵子在 *A. aeo*，*B. bur*，*T. pal* 和 *C. tra* 基因组中融合在一起；

（2）L14 操纵子与 L3 操纵子在 *B. bur*，*T. pal*，*C. tra*，*H. pyl*，*M. gen*，*M. pneu*，*B. sub*，*M. tub* 和 *S. pcc* 基因组中融合在一起；

（3）S13 操纵子在除了 *E. coli* 和 *H. inf* 之外的其他 10 个真细菌基因组中与 L3 和 L4 操纵子融合在一起；

（4）在 *B. sub* 中，8 个操纵子 L33，L11，L10，S12，L3，L14，S13 和 L13 聚集在基因组中 47 kb 的一段区域中，而且这些操纵子的排列顺序与 *B. bur* 基因组中的操纵子排列顺序相同；

（5）在古细菌的 4 个基因组中，两个比较大的操纵子 L3 和 L14 也是融合在一起的．

最后，我们注意到核糖体蛋白这种保守的排列方式在真核基因组中并不存在．通过检查酵母（*S. cerevisae*）和线虫（*C. elegan*）的基因组，我们发现核糖体蛋白是散布分布的，最多只有两三个聚集在一起．可见，真核生物与原核生物的基因组组成是完全不同的．

4　讨论

这 16 个完整基因组跨越了真细菌和古细菌两个界，包含了亲缘关系各异的物种．比如，*M. gen* 和 *M. pneu* 是最近的一组，*H. inf* 和 *E. coli* 也是非常相近的；而 *A. aeo*，*E. coli* 和 *S. pcc* 则是亲缘关系非常远的．这 16 个物种覆盖了微生物的许多分支，在物种的分歧时间上跨越了很大的时间尺度．所以，这些物种的进化关系可以从核糖体蛋白的基因排列顺序上反应出来是件很有意义的事情．

4.1　为不同界的分类提供了一个新的指标

Woese[19] 根据 16S rRNA 将原来生命的两界系统（原核生物和真核生物）重新划定为三界，即将原核生物分为古细菌和真细菌．在本文中，通过比较包含核糖体蛋白的操纵子的基因排列顺序，我们发现有些操纵子结构是古细菌这一界生物所特有的，而有些操纵子结构是真细菌

这一界生物中所特有的. 这些不同界中特有的操纵子结构为物种的分类提供了又一新的分类指标.

4.2 为比较物种间的亲缘关系提供了进化依据

从比较核糖体蛋白基因的排列顺序, 可以得出近到 *M. gen* 和 *M. pneu*, 远到 *M. gen* 和 *A. aeo* 不同物种间的亲缘关系. 由于这个信息并不是基于比较某个基因的序列对齐的结果, 所以它为研究分子进化提供了一条新的途径.

许多文献已经报道在 *E. coli* 中, 核糖体蛋白与其他一些相关基因, 比如 *EF-G* 和 *rpo* 一起构成操纵子, 这样的一些操纵子聚集在一起, 而且一些核糖体蛋白会参与到自身合成的翻译反馈调控中, 这方面的综述详见文献[20] , 所以包含这些核糖体蛋白的操纵子应该是非常保守的. 上面的分析中, 我们找出了所有的包含核糖体蛋白的操纵子, 其中 L3 和 L14 操纵子中的 20 多个核糖体蛋白的排列顺序在真细菌和古细菌两界中都是保守的. 这种保守的原因我们并不十分清楚, 也许会有两三个操纵子被共同调控的可能.

参 考 文 献

1 Koonin E V. Big time for small genomes. Genome Research, 1997, 7: 418 ~ 421

2 Deckert G, Warren P V, Gaasterland T, et al. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature, 1998, 392: 353 ~ 358

3 Fraser C M, Casjens S, Huang W M, et al. Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*. Nature, 1997, 390: 580 ~ 586

4 Fraser C M, Norris S J, Weinstock G M, et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science, 1998, 281: 375 ~ 388

5 Blattner F R, Plunkett Ⅲ G, Bloch C A, et al. The complete genome sequence of *Escherichia coli* K-12. Science, 1997, 277: 1 453 ~ 1 462

6 Fleischmann R D, Adams M D, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, 1995, 269: 496 ~ 512

7 Tomb J F, White O, Kerlavage A R, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature, 1997, 388: 539 ~ 547

8 Fraser C M, Gocayne J D, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. Science, 1995, 270: 397 ~ 403

9 Himmelreich R, Hilbert H, Plagens H, et al. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res, 1996, 24: 4 420 ~ 4 449

10 Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature, 1997, 390: 249 ~ 256

11 Cole S T, Brosch R, Parkhill J, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature, 1998, 393: 537 ~ 544

12 Kaneko T, Sato S, Kotani H, et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. Ⅱ. Sequence determination of the entire genome and assignment of potential-coding regions. J DNA Res, 1996, 3: 109 ~ 136

13 Bult C J, White O, Olsen G J, et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science, 1996, 273: 1 058 ~ 1 073

14 Smith D R, Doucette-Stamm L A, Deloughery C, et al. Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H:

functional analysis and comparative genomics. J Bacteriol, 1997, 179: 7 135 ~ 7 155

15 Klenk H P, Clayton R A, Tomb J, et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature, 1997, 390: 364 ~ 370

16 Kawarabayasi Y, Sawada M, Horikawa H, et al. Complete sequence and gene organization of the genome of a hyperthermophilic archaebacterium, *Pyrococcus horikoshii* OT3. J DNA Res, 1998, 5: 147 ~ 155

17 Pearon W R. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol, 1990, 183: 63 ~ 98

18 Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins, 1991, 9: 56 ~ 68

19 Woese C R. Archaebacteria. Sci Amer, 1981, 244: 98 ~ 125

20 Nomura M, Gourse R, Bauthman G. Regulation of the synthesis of ribosomes and ribosomal components. Ann Rev Biochem, 1984, 53: 75 ~ 117