原核生物基因组 DNA 链组成的非对称性

包其郁 杨焕明

(中国科学院遗传与发育生物学研究所 北京华大基因研究中心 中国科学院基因组信息学中心 北京 100101)

Genome DNA Strand Compositional Asymmetry of the Prokaryotic Organism

Bao Qiyu Yang Huanming

(*Institute of Genetics and Developmental Biology, Beijing Genomics Institute/Genomics and Bioinformatics Center, Chinese Academy of Sciences, Beijing* 100101, *China*)

关键词:原核生物,碱基组成,非对称性

中图分类号:Q933 **文献标识码**:A **文章编号**:0001-6209(2002)06-0755-04

迄今,已有 60 多个原核生物的基因组全序列被发表,我们因此有可能对它们基因组的构成有更深入的认识。绝大多数生物基因组 DNA 的 G 与 C 和 A 与 T 的含量相等[1]。但是,在许多原核生物基因组的先导链和后随链内存在 G 与 C 或 A 与 T 分布的不对称(GC shew 或 AT shew),原核生物 DNA 链的非对称性表现在碱基、密码子和基因水平。

1 碱基组成的非对称性(base composition asymmetry)

1.1 GC 分布不对称(GC skew)

Lobry[2] 于 1996 年通过对大肠杆菌(*Escherichia coli*)、枯草芽孢杆菌(*Bacillus subtilis*)和流感嗜血杆菌(*Haemophilus influenzae*)等 3 种细菌基因组的分析,发现它们 DNA 链不同区域的碱基组成非对称,前导链含有较多的 G 而后随链含有较多的 C(GC skew)。GC skew 的计算公式为(nG-nC)/(nG + nC),其中 nG(nC)为一特定大小 DNA 片段(窗口)内 G 或 C 的含量,窗口的大小一般设为 10kb,20kb 或 50kb[3]。对于大多数原核生物来说,它们先导链的 G 都多于 C,(nG-Gn)/(nG + nC)为正值,而后随链的 G 少于 C,(nG-nC)/(nG + nC)为负值。所以,在复制的起点和终点,会发生(nG-nC)/(nG + nC)的正负值之间转变。当以基因组的长度为横坐标,GC skew 为纵坐标作图时,起点在负值向正值转变处,接近或相当于 0 的位置;而终点在正值向负值转变处,同样接近或相当于 0 的位置。GC skew 在大多数真细菌如大肠杆菌、枯草芽孢杆菌、生殖道枝原体(*Mycoplasma genitalium*)、沙眼衣原体(*Chlamydia trachomatis*)、结核分支杆菌(*Mycobacterium tuberculosis*)、梅毒螺旋体(*Treponema pallidum*)、普氏立克次体(*Rickettsia prowazekii*)、流感嗜血杆菌、肺炎枝原体(*Mycoplasma pneumoniae*)和幽门螺杆菌(*Helicobacter pylori*)等中存在,并可据此对这些真细菌的单一复制起点和终点进行定位[4-6]。而在已测序的 11 种(株)古细菌中,通过 GC skew 预测存在单一复制起点的只有嗜酸热原体(*Thermoplasma acidophilum*)[7]。另外,硫磺矿硫化叶菌(*Sulfolobus solfataricus*)也籍此预测了一个复制起点(其可能有多个复制起点)[8];但其它古细菌如加氏甲烷球菌(*Methanococcus jannaschii*)、热自养甲烷杆菌(*Methanococcus thermoautotrophicum*)、发光古球菌(*Archaeoglobus fulgidus*)和火球菌(*Pyrococcus horikoshii*)等没有明显的链内 GC skew,可能有多个复制起点,不能用此法进行复制

作者简介:包其郁(1961-),男,浙江乐清人,副教授,中国科学院遗传与发育生物学研究所 1999 级基因组学方向博士生。

收稿日期:2001-12-26,修回日期:2002-02-04

起点定位[3,5]。

DNA 链碱基组成的非对称性也可以用于基因组为线性染色体的莱姆病病原体—伯氏疏螺旋体(*Borrelia burgdorferi*)复制起点的分析。线性染色体的复制可以是从一端开始,也可能从中间开始向两端复制。通过 GC skew 分析,预测伯氏疏螺旋体复制起点在染色体中部的 450 kb 处,后经实验得到证实[9]。某些大病毒的基因组也存在碱基组成的非对称性。对 10 个人疱疹病毒基因组 GC 分布研究结果表明,HHV6、HHV7 和 HCMV 存在 GC skew[5]。GC skew 还存在于叶绿体基因组[10]和质粒 DNA[9]。

在 GC skew 的基础上,Grigoriev[4] 建立了一种累计 skew(cumulative skew)的方法。这种方法是从 DNA 序列的任一位置开始,计算(nG-nC)/(nG + nC),并依次把相邻的(nG-nC)/(nG + nC)累计相加,最大值在复制终点,最小值在复制起点。它的优点是适用于一些 GC skew 不太明显的微生物,如肺炎枝原体的基因组序列,用一般的 GC skew 作图很难观察(nG-nC)/(nG + nC)正负值的转变点,但用累计(nG – nC)/(nG + nC)就很容易看出;另外,累计(nG-nC)/(nG + nC)的图形是一条'V'形的曲线,并非一般 GC skew 的为一上下波动的曲线,故而更直观(见图 1)。

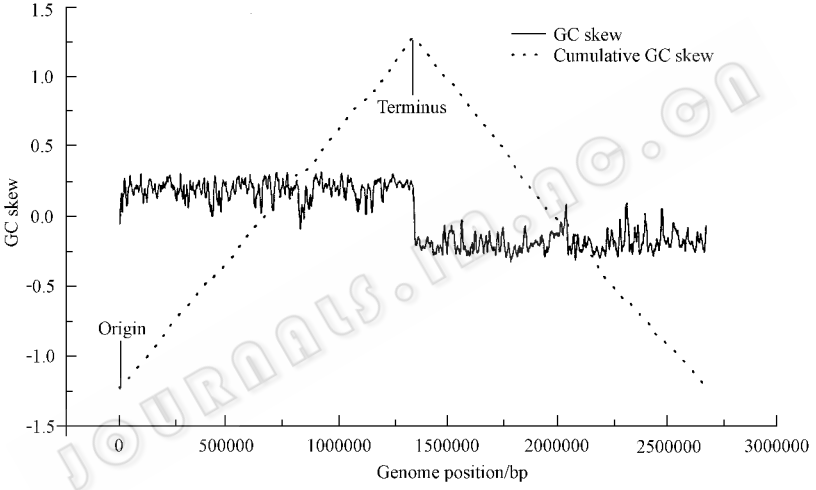

图 1 腾冲嗜热厌氧菌基因组的 GC skew 和累计 skew

1.2 AT 分布不对称(AT skew)

基因组 DNA 链内碱基分布的非对称性不仅局限于 G 与 C,而且在某些真细菌中还存在 AT skew[(A – T)/(A + T)]。对生殖道枝原体和枯草芽孢杆菌的分析发现,这些真细菌存在 AT skew,但比 GC skew 的要弱[5]。对近 40 种已全基因组测序的原核生物的分析表明,除上述两种真细菌外,还有 9 种真细菌的基因组也存在 AT skew。这 11 种真细菌在以 16S rRNA 为基础绘制的进化树上,分属于两支,其中腾冲嗜热厌氧菌(*Thermoanaerobacter tengcongensis*)、乳酸乳球菌(*Lactococcus lactis*)、肺炎链球菌(*Streptococcus pneumoniae*)、化脓性链球菌(*Streptococcus pyogenes*)、丙酮丁酸梭菌(*Clostridium acetobutylicum*)、耐盐芽孢杆菌(*Bacillus halodurans*)、金黄色葡萄球菌(*Staphylococcus aureus*)和枯草芽孢杆菌为一支,生殖道枝原体、肺炎枝原体和溶脲脲原体(*Ureaplasma urealyticum*)为另一支。值得一提的是,这些真细菌都有一个较大的基因方向性偏好(gene orientation bias),有 68% 以上的基因分布于前导链上,而其它真细菌和古细菌嗜酸热原体的基因在前导链上的分布都低于 68%(包其郁等待发表资料)。

由于分析基因组前导链和后随链的碱基分布的前提是能够通过 GC skew 等方法判定基因组的复制起点和终点。对于多复制起点的细菌如蓝细菌(*Synechocystis* sp.)和前述的古细菌加氏甲烷球菌等,目前还不能准确判断复制起点,用 GC skew 无法分析它们 DNA 链组成的非对称性[5]。同样,T4 噬菌体基因

组,某些真核生物染色体或染色体的一些区段,如整个酵母基因组,线虫基因组,果蝇染色体及人 T 细胞受体 β 位点(7 号染色体上的 670kb)等也未见碱基分布的非对称性[3]。

综上所述,不同原核生物基因组 DNA 链碱基组成的非对称程度相同或各有差异,GC skew 较 AT skew 强而普遍。造成这种前导链和后随链碱基分布不对称现象的原因可能有以下几个方面。

2　密码子使用偏好(codon usage bias)

对于绝大多数(尤其是基因组 GC 含量低于 55%)原核生物,无论是在先导链或后随链,编码序列都显著地偏好于密码子的第一位为嘌呤 A 或 G,第二位为 A 或 T,第三位则为 T 或 A(包其郁等未发表资料)。这说明 AAT、GAT、ATT、GTT、AAA、GAA、ATA 和 GTA 等密码子有更多的机会被使用。对 51 株已全基因组测序原核生物密码子使用统计结果显示,使用频率最高的前 4 位分别为 AAA(4.10%)、GAA(4.05%)、GAT(3.17%)、ATT(3.17%),AAT(2.63%)、GTT(2.19%)和 ATA(2.04%)分别列第 6、12、16 位,只有 GTA 使用较少(1.38%),列第 34 位(包其郁等未发表资料)。也有资料显示,基因组 GC 含量较高的梅毒螺旋体(52.8%),密码子第三位是 G 的机率要多于 A[11]。然而,在先导链,以 G 或 T 开头或结尾的密码子显著地多于后随链,常见的密码子有 GTG、GCG 和 GAG;而以 C 或 A 开头或结尾的密码子,如 CTC、GCC、CCC、ATC 和 ACC,常常在后随链多于先导链[6]。

3　基因方向性偏好(gene orientation bias)

基因方向性偏好在原核生物(除多个复制起点的古细菌无法判断外)是一种普遍现象。现已完成全基因组测序的 40 多种(株)真细菌和古细菌嗜酸热原体,前导链上编码的基因全部超过 50%。其中有 10 株超过 70%,超过 80%的 4 株为腾冲嗜热厌氧菌、乳酸乳球菌、生殖道枝原体和肺炎链球菌,最高的为腾冲嗜热菌,达 86.7%[12](包其郁等待发表资料)。如上所述,前导链上含有 68%以上基因的真细菌,都有 AT skew。不过,现在已知的基因分布最偏的还是硕大利什曼原虫(*Leishmania major* Friedlin)的 1 号染色体,它的 79 个基因中有 29 个分布于近左端粒的 79 kb 范围内,而另外 50 个基因分布在相邻的 180 kb 范围内的互补链上[13]。

4　基因密度和密码子使用的差别

在原核生物基因组,那些在密码子使用上与一般基因相差很大,与核蛋白体蛋白基因,翻译和转录相关基因,伴侣-降解蛋白基因等在密码子使用上高度相似的基因为高度表达基因。因此,在绝大多数原核生物基因组中高度表达基因,包括核蛋白体蛋白基因,与翻译和转录有关的因子基因,分子伴侣基因和与主要的能量代谢相关的基因[3,14]。对于快速生长的细菌,如大肠杆菌、霍乱弧菌(*Vibrio cholerae*)、枯草芽孢杆菌和流感嗜血杆菌,主要的糖酵解和三羧酸循环基因为高度表达基因。在产甲烷菌,与甲烷代谢有关的基因为高度表达基因[3,14]。高度表达基因大多编码于前导链,且通常都有密码子偏好,如核蛋白体蛋白基因密码子的第三位多为 G[3,14]。Morton[10] 对薄肌眼虫(*Euglena gracilis*)叶绿体基因组的分析表明,高度表达基因也主要编码于前导链,且在密码子的使用上具有明显的偏好,而后随链的密码子则没有偏好。另外,大于 80%的高度表达基因都伴有一主要由 G 组成的 SD 序列(核心为 GGAGG)[14]。

5　突变与修复偏差和信号序列的分布不同

5.1　转录伴随修复相关的突变偏差(bias)和脱氨基事件

转录伴随修复(transcription-coupled repair)是一种高度链特异而有效的核苷酸切除修复途径,它只对模板链起修复作用,对编码链不起作用[15,16]。所以,编码链的突变不能由此得到校正。有足够的证据表明[16,17],在 DNA 转录时,以单链形式存在的编码链更容易发生突变,胞嘧啶脱氨基变成胸腺嘧啶(C-T)的频率是双链 DNA 时的 100 倍以上,其结果是编码链胞嘧啶减少,胸嘧腺啶增加。

5.2 信号序列等寡核苷酸序列的分布不同

GC skew 或 AT skew 可能由于大量的 chi 或其它信号序列位于先导链上所致[5,12]。大肠杆菌的 chi 序列为富含 G 的 8 核苷酸序列(GCTGGTGG),它与 recBCD 复合体一起在 DNA 重组中起作用;该序列共有 1000 多拷贝,其中 75%位于前导链[18]。但是,如果除去大肠杆菌的 chi 序列,基因组的 GC 或 AT skew 并未发生显著变化,即使把大肠杆菌和幽门螺杆菌的各占基因组 10%和 6%的所有具有碱基偏好的 8 核苷酸全部去掉,结果虽然 GC skew 减低,但仍然存在。因此,Tillier 等认为这些 8 核苷酸可能与其它许多潜在的寡核苷酸序列共同引起链内碱基组成的非对称性[12]。

Uno 等[19]认为 DNA 链内碱基组成的非对称性并非寡核苷酸序列引起,相反,是由于 G 的偏态分布导致 chi 等寡核苷酸序列的偏态分布。枯草芽孢杆菌、流感嗜血杆菌和乳酸乳球菌等基因组中也存在大量富含 G 的 chi 等寡核苷酸序列,这些序列并不总是与复制的方向相关,偏好分布于先导链上。而 chi 等寡核苷酸序列的分布与 GC skew 密切相关,因为只有 GC 分布明显偏态的细菌基因组,如大肠杆菌和枯草芽孢杆菌,才有 chi 的偏态分布,而在 GC skew 不太明显的细菌基因组,如流感嗜血杆菌,chi 也不存在偏态分布[19]。除上述原因外,DNA 链内碱基组成的非对称性可能还与细胞周期中 dNTP 池(pool)的波动、操纵子结构不同、单碱基的不同[3]及邻位依赖突变偏好(context-dependent mutation bias)[20]等有关。由此看来,DNA 链组成的非对称性并非由单一因素所致。

致谢 中国科学院微生物研究所谭华荣和黄力两位研究员对本文提供许多宝贵意见,特此致谢。

参 考 文 献

[1] Fickett J W, Torney D C, Wolf D R. *Genomics*, 1992, **13**(4): 1056 ~ 1064.
[2] Lobry J R. *Mol Biol Evol*, 1996, **13**(5): 660 ~ 665.
[3] Karlin S. *Trends Microbiol*, 1999, **7**(8): 305 ~ 308.
[4] Grigoriev A. *Nucleic Acids Res*, 1998, **26**(10): 2286 ~ 2290.
[5] Mrazek J, Karlin S. *Proc Natl Acad Sci U S A*, 1998, **95**(7): 3720 ~ 3725.
[6] Rocha E P, Danchin A, Viari A. *Mol Microbiol*, 1999, **32**(1): 11 ~ 16.
[7] Ruepp A, Graml W, Santos-Martinez M L, *et al*. *Nature*, 2000, **407**(6803): 508 ~ 513.
[8] She Q, Singh R K, Confalonieri F, *et al*. *Proc Natl Acad Sci U S A*, 2001, **98**(14): 7835 ~ 7840.
[9] Picardeau M, Lobry J R, Hinnebusch B J. *Genome Res*, 2000, **10**(10): 1594 ~ 1604.
[10] Morton B R. *Proc Natl Acad Sci U S A*, 1999, **96**(9): 5123 ~ 5128.
[11] Lafay B, Lloyd A T, McLean M J, *et al*. *Nucleic Acids Res*, 1999, **27**(7): 1642 ~ 1649.
[12] Tillier E R, Collins R A. *J Mol Evol*, 2000, **50**(3): 249 ~ 257.
[13] Myler P J, Audleman L, deVos T, *et al*. *Proc Natl Acad Sci U S A*, 1999, **96**(6): 2902 ~ 2906.
[14] Karlin S, Mrazek J. *J Bacteriol*, 2000, **182**(18): 5238 ~ 5250.
[15] Proietti De Santis L, Garcia C L, Balajee A S, *et al*. *Mutat Res*, 2001, **485**(2): 121 ~ 132.
[16] Francino M P, Chao L, Riley M A, *et al*. *Science*, 1996, **272**(5258): 107 ~ 109.
[17] Beletskii A, Bhagwat A S. *Proc Natl Acad Sci U S A*, 1996, **93**(24): 13919 ~ 13924.
[18] Blattner F R, Plunkett G, Bloch C A, *et al*. *Science*, 1997, **277**(5331): 1453 ~ 1474.
[19] Uno R, Nakayama Y, Arakawa K, *et al*. *Gene*, 2000, **259**(12): 207 ~ 215.
[20] McVean G A, Hurst G D. *J Mol Evol*, 2000, **50**(3): 264 ~ 275.