第 25 卷 第 2 期 华侨大学学报(自然科学版) Vol. 25 No. 2
2004 年 4 月 Journal of Huaqiao University (Natural Science) Apr. 2004

文章编号 1000-5013(2004)02-0192-05

敏捷气热菌密码子及 AUG侧翼序列保守性分析

张光亚 方柏山

(华侨大学材料科学与工程学院, 福建 泉州 362011)

摘要 比较敏捷气热菌和大肠杆菌等 9 种编码区 GC 含量,以及各异生物的密码子使用情况. 结果表明,与敏捷气热菌编码区 GC 含量比较接近的果蝇,在密码子使用上与之相差最小. 它们在分类学上分别属于不同的域,与敏捷气热菌同属古菌,但 GC 含量相差较大的强烈炽热球菌,则相差较大. 说明编码区 GC 含量对密码子使用偏好,比生物分类学(系统发育)地位更重要. 碱基 A,C 在高、低表达基因中出现的概率差别较大,尤其在 - 1, - 3, - 4 和 - 7 位点;碱基 A,C 在调控翻译起始效率中的作用,可能大于碱基 G,U. 因为碱基 G,U 在高表达和低表达基因中,其出现的概率差别不大. 高表达和低表达基因起始密码子侧翼序列中,某些位点保守性存在差异,其中高表达基因 - 4 位和 - 3 位可能与其高表达特性有关.

关键词 敏捷气热菌, 密码子使用偏好, AUG侧翼序列, 碱基概率分布, 保守性

中图分类号 Q 933 ∶Q 811.4 **文献标识码** A

由于遗传密码子的简并性,编码不同氨基酸的密码子的数目,除 5 个外,从 1 个到 6 个都有,不过它们使用的往往并不相同. 对于一种生物的基因,倾向于使用部分特定的同义密码子,此现象即称为密码子偏好(Codon Bias). 密码子偏好现象在生物中普遍存在. 敏捷气热菌(*Aeropyrum pernix*),是目前报导的第一个严格好氧的超嗜热古菌. 其最适生长温度为 90～95 ℃,最适 pH 接近中性,倍增时间为 200 min. 而且,敏捷气热菌为异养微生物,它们能利用蛋白胨、酵母膏和酪蛋白作为碳源、氮源和能源[1],使得培养它相对简单. 大规模发酵培养 1 d 菌体,其得率为每升湿重 1.0 g. 这些特征使得敏捷气热菌成为获得嗜热酶最有可能的生物来源[2]. 了解敏捷气热菌中密码子的使用情况,有助于改造自身基因或宿主细胞,获得具有广泛用途的嗜热酶. 同时,了解古菌密码子使用情况,对研究生物系统进化和密码子进化具有一定理论意义. 本文分析了敏捷气热菌密码子使用情况,并和大肠杆菌等 9 种编码区 GC 含量各异生物密码子使用情况作比较. 同时,计算了敏捷气热菌中高表达基因和低表达基因 AUG侧翼序列碱基出现概率的分布及保守性,获得了一些有价值的信息.

1 材料与方法

实验材料中,敏捷气热菌(*Aeropyrum pernix*)、*Sulfolobus tokodaii*、强烈炽热球菌(*Pyrococcus furiosus* DSM 3638)、大肠杆菌(*Escherichia col*)、铜绿假单胞菌(*Pseudomonas aeruginosa*)、蚜虫初生内共生菌(*Buchnera sp.* APS)、棒状节杆菌(*Caulobacter crescentus*)、酵母(*Saccharomyces cerevisiae*)、拟南芥(*Arabidopsis thaliana*)和果蝇(*Drosophila melanogaster*)的编码区 GC 含量(Coding GC,表示为 GC_c,下同)分别(%)为 57.46, 33.59, 41.09, 50.36, 67.14, 27.43, 67.68, 39.46, 44.58 和 53.93. 其中,前 3 种为古菌,中间 4 种为细菌,后 3 种为真核生物. 它们的密码子使用频率来自于密码子使用数据库(Codon Usage Database)[3]. 敏捷气热菌中 2 692 条 mRNA 起始密码子侧翼序列(从 - 20 位到 + 13 位)、基因长度、有效密码子数目(Effective Number of Codons) NC 值和 GC_{3s}值均来源于 Transterm 数据库[4]. 然后,再根据下列的条件进行手工筛选. (1) 基因长度大于 300 bp. (2) 包含完整的从 - 20 位到

收稿日期 2003-09-05

作者简介 张光亚(1975-),男,助教,硕士,主要从事生物信息及分子生态学的研究. E-mail: zhgyghh@hqu.edu.cn

+13 位 AUG侧翼序列,经筛选后共得到 2 419 条 mRNA 起始密码子侧翼序列. AUG侧翼序列的保守性采用 $M_1(l)$ 表示[5], $M_1(l)$ 值的取值范围在 0～3 之间,其数值越大表明保守性越强.

2　结果与分析

2.1　敏捷气热菌与 9 种生物密码子使用比较

敏捷气热菌与大肠杆菌等 9 种生物的密码子使用比较结果,如表 1 所示. 表中, n_1～n_9 分别表示敏捷气热菌与棒状节杆菌、铜绿假单胞菌、果蝇、大肠杆菌、拟南芥、强烈炽热球菌、酵母、*Sulfolobus tokodaii* 和蚜虫初生内共生菌 9 种生物密码子出现频率的比值. 一般认为, n 值大于 2.0 或者小于 0.5 时,差异显著[6],如表中带星号的数据. n_1～n_9 达显著水平的个数分别为 34,37,21,23,28,33,31,42 和 46. 这表明,敏捷气热菌密码子使用偏好与蚜虫初生内共生菌相差最大, *Sulfolobus tokodaii* 次之. 而与敏捷气热菌编码区 GC 含量比较接近的果蝇和大肠杆菌,在密码子使用偏好上相差较小. 尽管它们在分类学上分别属于不同的域(Domain). 这说明编码区 GC 含量对密码子使用偏好产生的影响,比生物分类学(系统发育)地位的影响更大,与其它报道[7]相吻合. 敏捷气热菌所有密码子中,AGG (Arg),AUU (Ile)和 AAA(Lys)与其它所有生物对应密码子的差异均达显著水平,AGG (Arg)比值均大于 2. 这可能是由于,一方面该超嗜热菌产生的嗜热酶中含有大量的 Arg,以保证它们在高温下的活性. 因为 Arg 比带同样电荷的氨基酸有更大的侧链,侧链所提供的疏水作用及离子间相互作用能提高蛋白的热稳定性[8]. 另一方面,敏捷气热菌编码区 GC 含量较高,在转录后的 mRNA 中 GC 含量也很高,Arg 的密码子偏向于使用 AGG,而不是其它同义密码子. AAA(Lys)比值大于 2 的,均为编码区 GC 含量高于敏捷气热菌的棒状节杆菌和铜绿假单胞菌. 这说明,在这 2 种微生物中,AAA 的使用频率更低. AAA(Lys)比值小与 2 的生物编码区, GC 含量均低于敏捷气热菌,更进一步证实了编码区 GC 含量对密码子使用偏好的影响. 终止密码子 UAG的使用频率,也明显高于其它几种生物.

表 2　敏捷气热菌与其它 9 种生物密码子使用比较

AA	Codon	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
Phe	UUU	1.44	3.56*	0.46*	0.27*	0.28*	0.23*	0.23*	0.19*	0.13*
Phe	UUC	0.69	0.64	0.98	1.34	1.04	1.18	1.18	1.57	4.99
Leu	UUA	11.85*	15.79*	1.08	0.33*	0.38*	0.24*	0.18*	0.10*	0.07*
Leu	UUG	1.04	0.82	0.44*	0.56*	0.34*	0.51	0.27*	0.71	0.70
Leu	CUU	2.71*	5.43*	1.87	1.40	0.70	0.70	1.39	0.90	1.40
Leu	CUC	2.25*	1.38	2.75*	3.78*	2.38*	2.25*	7.14*	6.83*	22.51*
Leu	CUA	12.87*	12.87*	2.20*	4.18*	1.82	1.01	1.35	1.10	1.96
Leu	CUG	0.41*	0.34*	0.73	0.59	2.83*	3.37*	2.69*	7.36*	13.98*
Ile	AUU	2.76*	2.76*	0.49*	0.27*	0.37*	0.25*	0.26*	0.20*	0.13*
Ile	AUC	0.31*	0.33*	0.54	0.53	0.67	0.91	0.73	1.36	1.30
Ile	AUA	52.70*	35.13*	3.36*	4.31*	2.53*	0.78	1.77	0.64	0.74
Met	AUG	0.91	0.97	0.84	0.75	0.81	0.90	0.94	0.96	0.92
Val	GUU	3.89*	7.78*	1.93	1.05	0.77	0.59	0.95	0.72	0.96
Val	GUC	0.58	0.78	1.62	1.59	1.76	2.13*	1.95	4.10*	6.64*
Val	GUA	12.22*	3.67*	2.33*	1.26	1.47	0.77	1.24	0.50	0.75
Val	GUG	0.96	0.87	1.03	1.20	1.66	2.09*	2.72*	3.48*	7.41*
Ser	UCU	5.33*	11.33*	1.30	0.85	0.36*	1.00	0.39*	0.47*	0.29*
Ser	UCC	1.79	1.27	0.79	1.67	1.39	2.44*	1.08	3.14*	5.13*
Ser	UCA	4.10*	11.60*	0.89	0.75	0.38*	0.68	0.37*	0.41*	0.39*
Ser	UCG	0.51	0.83	0.65	1.27	1.17	4.29*	1.26	3.25*	4.29*
Ser	AGU	4.77*	2.57*	0.58	0.66	0.48*	0.66	0.47*	0.40*	0.45*
Ser	AGC	1.57	0.99	1.26	1.71	2.30*	2.38*	2.67*	4.52*	8.04*
Pro	CCU	5.09*	8.48*	2.54*	2.36*	0.95	1.68	1.31	1.25	1.25
Pro	CCC	1.28	1.85	1.34	4.49*	4.54*	2.62*	3.56*	5.74*	11.47*
Pro	CCA	3.94*	4.12*	0.67	1.05	0.56	0.46*	0.50	0.53	0.80

续表

AA	Codon	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
Pro	CCG	0.43*	0.39*	0.82	0.64	1.54	4.36*	2.49*	3.85*	5.69*
Thr	ACU	6.41*	5.28*	0.93	0.85	0.51	0.61	0.44*	0.41*	0.43*
Thr	ACC	0.49*	0.49*	0.76	0.74	1.55	2.12*	1.28	3.67*	5.56*
Thr	ACA	7.58*	13.27*	0.97	1.09	0.68	0.67	0.60	0.62	0.54
Thr	ACG	0.67	1.76	0.76	0.81	1.44	1.94	1.40	2.70*	4.61*
Ala	GCU	3.02*	5.16*	1.72	1.44	0.87	1.17	1.17	0.98	1.26
Ala	GCC	0.46*	0.54	1.10	1.53	3.53*	2.68*	2.94*	7.06*	13.11*
Ala	GCA	4.23*	2.76*	1.06	0.64	0.77	0.55	0.84	0.65	0.71
Ala	GCG	0.42*	0.50	1.40	0.67	2.20*	3.32*	3.19*	5.02*	6.53*
Tyr	UAU	1.30	2.38*	1.18	0.71	0.85	0.59	0.67	0.35*	0.40*
Tyr	UAC	1.84	1.04	1.13	1.71	1.50	1.13	1.42	1.64	4.33*
Ter	UAA	1.43	2.85*	1.07	0.43*	0.95	0.71	0.86	0.43*	0.37*
Ter	UAG	2.42*	8.06*	4.03*	8.76*	4.84*	3.45*	5.13*	4.03*	6.04*
Ter	UGA	0.61	0.38*	1.82	0.89	0.83	0.57	1.36	0.83	3.03*
His	CAU	1.16	0.98	0.59	0.49*	0.45*	0.88	0.45*	0.64	0.33*
His	CAC	1.03	0.85	0.81	1.41	1.50	1.63	1.68	3.72*	4.83*
Gln	CAA	0.68	0.49*	0.20*	0.21*	0.16*	0.32*	0.11*	0.20*	0.11*
Gln	CAG	0.58	0.44*	0.44*	0.57	1.06	2.03*	1.32	3.08*	3.82*
Asn	AAU	0.10*	1.21	0.21*	0.21*	0.20*	0.25*	0.12*	0.13*	0.07*
Asn	AAC	0.84	0.71	0.61	0.75	0.76	0.94	0.64	1.16	1.61
Lys	AAA	2.88*	2.16*	0.47*	0.22*	0.25*	0.20*	0.18*	0.15*	0.09*
Lys	AAG	0.86	1.11	0.70	2.18*	0.85	0.65	0.90	0.99	3.47*
Asp	GAU	0.95	1.21	0.46*	0.39*	0.35*	0.46*	0.34*	0.35*	0.33*
Asp	GAC	0.59	0.61	1.05	1.36	1.51	1.63	1.29	2.74*	5.11*
Glu	GAA	0.75	0.49*	0.55	0.30*	0.34*	0.24*	0.25*	0.25*	0.23*
Glu	GAG	1.42	1.46	1.26	2.89*	1.68	1.35	2.85*	2.35*	10.87*
Cys	UGU	2.93*	3.22*	0.60	0.61	0.31*	0.87	0.40*	0.69	0.33*
Cys	UGC	0.95	0.68	0.46*	1.01	0.86	2.77*	1.31	3.39*	2.77*
Trp	UGG	0.93	0.88	1.32	0.95	1.05	1.06	1.27	1.30	1.45
Arg	CGU	0.47*	0.48*	0.43*	0.19*	0.42*	3.43*	0.58	2.69*	0.30*
Arg	CGC	0.16*	0.14*	0.37*	0.35*	1.77	6.11*	2.61*	13.44*	3.73*
Arg	CGA	0.69	1.09	0.31*	0.68	0.42*	2.18*	0.87	2.91*	0.44*
Arg	CGG	0.45*	0.49*	0.84	1.16	1.45	9.96*	4.03*	17.44*	13.95*
Arg	AGA	12.46*	22.42*	2.20*	2.88*	0.60	0.40*	0.53	0.43*	0.71
Arg	AGG	20.66*	22.72*	7.21*	20.22*	4.17*	2.15*	4.94*	3.82*	41.32*
Gly	GGU	1.64	1.80	1.13	0.59	0.67	1.20	0.62	0.63	0.66
Gly	GGC	0.50	0.53	1.22	1.22	3.54*	4.23*	3.37*	6.04*	7.95*
Gly	GGA	2.73*	2.86*	0.68	1.22	0.49*	0.32*	1.10	0.46*	0.50
Gly	GGG	2.24*	2.58*	5.44*	2.25*	2.51*	1.81	4.28*	3.50*	7.52*

2.2 起始密码子侧翼序列的碱基分布

将上述2 419条mRNA按NC值从低到高进行排序.取最两端的高表达基因(低NC值)和低表达基因(高NC值)样本各40个[6],分别统计从-20位到+13位各位点(λ)碱基的数量.然后,计算碱基出现的频率 $p(i)$值,i为A,C,G,U.高表达和低表达基因各位点碱基的分布图,如图1所示.起始密码子AUG第1个碱基A,在低表达基因中出现的概率(0.625)高于高表达基因(0.325),而高表达基因中碱基G出现的频率高于A.这表明,敏捷气热菌中高表达基因偏向于使用GUG为起始密码子,而低表达基因偏向于使用AUG.在起始密码子前-1,-3,-4和-7位点碱基A,其在低表达基因中出现的频率高于高表达基因,平均概率为0.235;高表达基因为0.179.在起始密码子后+4,+6和+9位低表达基因中,碱基A出现的概率均高于高表达基因,只在+11位高表达基因中,A出现的概率高于低表达基

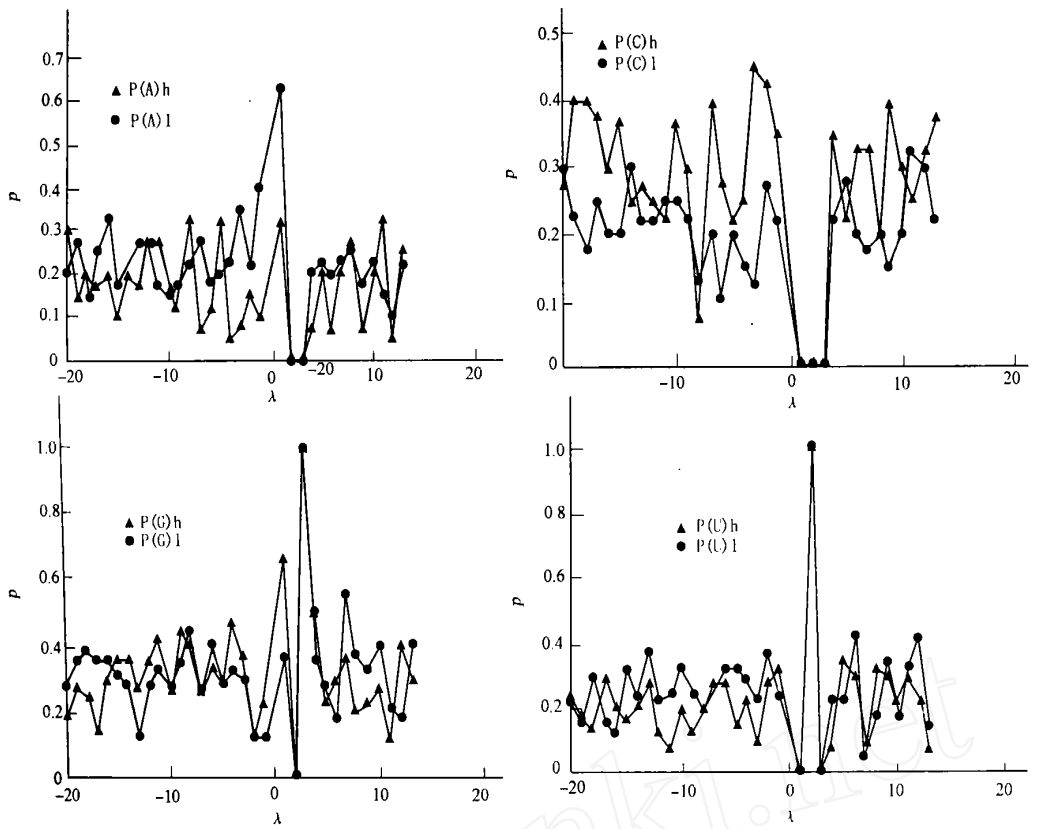

图 1　敏捷气热菌起始密码子侧翼序列各位点上各种碱基的概率分布

因. 起始密码子后 A ,在高表达基因中出现的平均概率为 0. 173 ,在低表达基因中为 0. 198. 起始密码子前 - 3 , - 7 和 - 18 位碱基 C ,在高表达基因中出现的频率明显高于低表达基因 ,平均概率为 0. 313 ,而低表达基因为 0. 211. 起始密码子后 + 7 , + 9 和 + 13 位碱基 C ,在高表达基因中出现的频率明显高于低表达基因 ,出现的平均概率分别为 0. 308 和 0. 228. 起始密码子前 - 4 和 - 13 位碱基 G ,在高表达基因出现的频率较高 ; - 17 和 - 18 位碱基 G 在低表达基因中出现的频率较高. 起始密码子前碱基 G ,在高表达和低表达基因中出现的概率分别为 0. 305 和 0. 294. 起始密码子后碱基 G ,在高表达和低表达基因出现的概率在一些位点 ,存在差异 ,总体的概率都较高 ,分别为 0. 293 和 0. 323. 碱基 U 在起始密码子前后某些位点的概率有些差异 , 但总体差别不大. 在高表达和低表达基因中 ,其平均概率分别为 0. 204 和 0. 260 ,起始密码子后碱基 U 出现的概率分别为 0. 228 和 0. 253. 总体而言 ,碱基 A ,C 在高与低表达基因中出现概率的差别较大 ,尤其在 - 1 , - 3 , - 4 和 - 7 位点. 这说明 ,这些位点在蛋白质翻译起始过程对翻译起始的效率可能有影响. 而碱基 A ,C 在调控翻译起始效率中的作用 ,可能大于碱基 G ,U. 因为碱基 G 和 U 在高表达和低表达基因中 ,概率差别不大.

3　结束语

本文分析了好氧超嗜热古菌敏捷气热菌密码子使用偏好 ,以及起始密码子侧翼序列的保守性 ,得到 3 点结果. (1) 敏捷气热菌编码区 GC 含量比较接近的果蝇 ,在密码子使用偏好上与之相差最小. 它们在分类学上分别属于不同的域 ,而与敏捷气热菌同属古菌. 但 GC 含量相差较大的强烈炽热球菌 ,则相差较大. 这说明编码区 GC 含量对密码子使用偏好 ,比生物分类学(系统发育) 地位更重要. (2) 统计了 4 种碱基 A ,C , G 和 U 在高表达和低表达基因中起始密码子侧翼序列(从 - 20 位到 + 13 位) 中概率 ,碱基 A ,C 在高、低表达基因中出现的概率差别较大 ,尤其在 - 1 , - 3 , - 4 和 - 7 位点 ;而碱基 A ,C 在调控翻译起始效率中的作用可能大于碱基 G ,U. 因为碱基 G 和 U 在高表达和低表达基因中概率的差别不大. (3) 计算表示起始密码子侧翼序列保守性的 $M_1(l)$ 值 ,并比较了高表达和低表达基因从 - 20 位到 + 13 位的 $M_1(l)$ 值. 发现两二者在某些位点的保守性存在差异 ,其中高表达基因中 - 4 位和 - 3 位可能与其

高表达的特性有关.

参 考 文 献

1 Pamela Chavez, Croocker Yoshihiko, Sako Aritsune Uchida. Purification and characterization of an intracellular heat-stable proteinase (pernilase) from the marine hyperthermophilic archaeon *Aeropyrum pernix* K1[J]. Extremophiles, 1999,(3):3～9

2 Yoshihiko Sakoa, Pamela Chavez, Croockera Yuzaburo Ishida. An extremely heat-stable extracellular proteinase (aeropyrolysin) from the hyperthermophilic archaeon *Aeropyrum pernix* K1[J]. FEBS Letters,1997,(415): 329～334

3 Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases: Status for the year 2000[J]. Nucl. Acids Res., 2000,28(1): 292～293

4 Grant H J, Oliver R, Peter AS, et al. Transterm: A database of mRNAs and translational control elements[J]. Nucleic Acids Res., 2002, 30(1): 310～311

5 陈颖丽,李前忠. *E. coli* 和 Yeast 基因起始与终止密码子临近序列碱基保守性、关联性的对比研究[J].内蒙古大学学报(自然科学版),2000,31(2):164～167

6 范三红,郭蔼光,单丽伟等. 拟南芥基因密码子偏爱性分析[J].生物化学与生物物理进展,2003,30(2):221～225

7 Chen Lingling, Zhang Chunting. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages[J]. Biochemical and Biophysical Research Communications, 2003,(306):310～317

8 马 挺,刘如林. 嗜热菌耐热机理的研究进展[J].微生物学通报,2002,29(2):86～88

9 Sharp P M, Cowe E. Synonymous codon usage in *Saccharomyces cerevisiae*[J]. Yeast, 1991,7:657～678

Analysing the Coden and the Conservative Character of AUG Context in Aeropyrum pernix

Zhang Guangya Fang Baishan

(College of Mater. Sci. & Eng., Huaqiao Univ., 362011, Quanzhou, China)

Abstract Among 10 organisms with different coding GC contents, *Aeropyrum pernix* was compared with *E. coli* and others in their coden usage bias. least difference in coden usage bias was shown between *A. pernix* and *Drosophila melanogaster*, these two organism also were close in coding GC contents but they belong to two different domains in taxonomy. This illustrates that coding GC is more important than plilogenetic lineage for their coden usage bias. The probability of the occurrence of bases A and C in highly expressed genes and lowly expressed genes differred fairly greatly, esp in -1, -3, -4 and -7 sites. Bases A and C might be more important than bases G and U in the control of initial efficiency of translation. The conservative character of AUG context in highly expressed genes was shown to be different from that of lowly expressed genes. Of all the context sites, the highly expressed genes -4 and -3 were probably associated with their highly expressed character.

Keywords *Aeropyrux pernix*, codon usage bias, AUG context, probability of base distribution, conservative character