

两种超嗜热菌基因组中同义密码子使用的分析

张光亚 葛慧华 方柏山

(华侨大学材料科学与工程学院, 福建 泉州 362021)

摘要 对两种超嗜热菌敏捷气热菌(*Aeropyrum pernix*)和风产液菌(*Aquifex aeolicus*)的密码子使用进行分析. 相对密码子使用值(RSCU)的计算表明, 它们在密码子使用上具有一定的相似性, 但也有不同的使用模式, 导致这一现象的机制可能存在差异. 密码子 RSCU 值对应分析表明, 两种菌中高表达基因均具有相对较高的 GC 含量, 但基因表达水平对密码子使用偏好的影响程度却不一样. 在 *A. pernix* 中, 它是造成整个基因组密码子使用偏好的最主要因素, 而在 *A. aeolicus* 中却不是.

关键词 敏捷气热菌, 风产液菌, 密码子使用, 表达水平, 对应分析

中图分类号 Q 939.1; Q 755

文献标识码 A

超嗜热菌是指能生活在 90 ℃ 以上高温, 最适生长温度高于 80 ℃ 的一类极端微生物, 是国际高温生物研究的热点. 从中分离出来的酶显示了独特的特性, 因此, 可以将其改造成具有新特性而加以应用^[1]. 目前主要的方法是通过基因工程技术, 将编码极端酶的基因在异源普通宿主菌(如大肠杆菌)中进行表达. 然而, 对于生物体而言, 基因或基因组中同义密码子的非随机使用(即密码子使用偏好)是一个普遍存在的遗传变异现象. 密码子偏好是多种因素相互作用的结果, 基因表达水平也是其中之一^[2]. 本文选择两种超嗜热菌敏捷气热菌(古菌)和风产液菌(细菌)作为研究对象, 两种菌的相关特性及用途请参见文[3, 4]. 利用它们的基因组数据, 通过图形和多元统计方法分析这两种基因组密码子使用的情况, 其目的在于分析和对比基因组中密码子使用的基本特征, 鉴定其使用偏好的模式. 同时, 着重于基因表达水平, 分析导致密码子使用偏好的影响因素.

1 材料与方法

1.1 基因组序列

从 Transterm 数据库^[5]下载敏捷气热菌(NC_000854)和风产液菌(NC_000918)基因组中所有非冗余的编码区(CDS)序列, 分别有 2 692 条和 1 550 条 CDS. 为了减少长度较短的基因变异所带来的误差, 根据国际通行方法, 用基因组中所有长度大于或等于 300 bp 的基因来分析, 最后分别得到 2 617 条和 1 513 条 CDS.

1.2 统计分析

为分析和对比两个基因组中密码子的使用偏好情况, 计算了每个 CDS 和每个基因组的密码子使用次数(*N* 值)、相对同义密码子使用值(RSCU)、有效密码子数目(ENC), 以及密码子自适应指数(CAI), 相关计算见文[6]. 对应分析(Correspondence Analysis, CA)是常用作对密码子使用和氨基酸使用的变异进行分析多元统计分析方法. 本文中 CA 分析采用 RSCU 值, 其计算方法见文[7]. 以上有关密码子使用的分析在 CodonW 软件(<ftp://www.molbiol.ox.ac.uk/cu>)和 GCUA 软件^[8]中完成, 统计分析由 SPSS 10.0 完成.

收稿日期 2005-10-02

作者简介 张光亚(1975-), 男, 讲师, 博士研究生, 主要从事酶与生物信息学的研究; 通信作者: 方柏山(1957-), 男, 教授, E-mail: fanbs@hqu.edu.cn

基金项目 国务院侨务办公室科研基金资助项目(05Q0018)

2 结果与分析

2.1 超嗜热菌基因组密码子使用偏好

表 1 列出了 *A. pernix* 和 *A. aeolicus* 两个基因组 RSCU 的情况. 由表 1 可以看见, *A. pernix* 和 *A. aeolicus* 基因组中, 存在明显的密码子使用偏好. 同时, 在每一种编码氨基酸的密码子使用上, 也存在一定的差异. 因此, 两者基因组在密码子使用偏好方面的分化程度较高, 可能使用不同的基因编码策略进行蛋白质的翻译. 总体而言, 两个菌基因组均偏向于使用第 3 位是 G/C 的密码子, 但也有 15 个密码子在使用上存在显著差异(表中 * 标出). 其中, 在编码 Arg, Gly 和 Glu 的 12 个密码子中, 共有 9 个

表 1 两个菌基因组密码子 RSCU 值

氨基酸	密码子	RSCU 值		氨基酸	密码子	RSCU 值		氨基酸	密码子	RSCU 值	
		<i>A. pernix</i>	<i>A. aeolcus</i>			<i>A. pernix</i>	<i>A. aeolcus</i>			<i>A. pernix</i>	<i>A. aeolcus</i>
Phe	UUU*	0.44	1.13	Ser	UCU	0.73	1.11	Val	GUU	0.96	1.52
	UUC	1.56	0.87		UCC	1.24	1.60		GUC	1.04	0.42
Leu	UUA*	0.25	1.00	UCA	0.56	0.89		GUA	0.67	1.28	
	UUG	0.38	0.46	UCG*	0.87	0.43		GUG	1.33	0.77	
	CUU	0.89	1.53	AGU	0.54	0.98	Asp	GAU	0.66	0.75	
	CUC	2.03	1.75	AGC	2.07	0.99		GAC	1.34	1.25	
	CUA	0.95	0.44	Pro	CCU	1.12	1.07	Gly	GGU	0.70	0.93
CUG	1.49	0.82	CCC		1.50	1.69	GGC*		1.53	0.48	
Cys	UGU	0.69	0.98		CCA	0.57	0.55		GGA*	0.56	2.00
	UGC	1.31	1.02	CCG	0.82	0.69	GGG*	1.21	0.59		
	Tyr	UAU*	0.76	0.37	Arg	CGU	0.30	0.19	Glu	GAA*	0.35
UAC		1.24	1.63	CGC*		0.53	0.15	GAG*		1.65	0.70
Ile	AUU	0.46	0.71	CGA*		0.21	0.07	Asn	AAU	0.44	0.61
	AUC	0.72	0.40	CGG*	0.55	0.09	AAC		1.56	1.39	
	AUA	1.82	1.88	AGA*	0.87	2.34	Lys	AAA*	0.44	—	
Ala	GCU	1.04	1.08		AGG	3.55	3.17	AAG	1.56	1.04	
	GCC	1.56	0.84	Thr	ACU	0.77	0.93	Gln	CAA*	0.32	0.72
	GCA	0.57	1.14		ACC	1.38	1.08		CAG	1.68	1.28
	GCG	0.83	0.93		ACA	0.91	0.83	His	CAU	0.64	0.34
					ACG	0.94	1.17		CAC	1.36	1.66

存在显著差异. *A. pernix* 中使用频率显著大的密码子有 UAU, UCG, CGC, CGA, CGG, GGC, GGG 和 GAG, 其中大部分位高 GC 含量的密码子. 这可能与该菌基因组 GC 含量高有关. 但 *A. aeolicus* 编码 Cys(UGU 和 UGC)同义密码子的 RSCU 值非常接近, 在使用上并不存在一定的偏好. 在这两种菌中, 编码 Arg 的 AGG 和 AGA 使用最为频繁, 而在其他许多生物中, 这两种密码子被认为是稀有密码子, 使用效率极低. 因此, 若包含有这两个密码子的基因在外源表达过程中, 需考虑替换这两种密码子.

2.2 密码子使用的非均一性

ENC 和 GC3s 常用来研究同一生物体内的不同基因间密码子使用的差异, 两个菌体中不同基因 ENC 值的分布, 如图 1 所示.

A. aeolicus 中 ENC 值范围从

31.56 到 61.00, 平均值为 47.25, 标准偏差为 4.11. *A. pernix* 中 ENC 值的范围从 28.01 到 61.00, 平均值为 44.67, 标准偏差为 6.3. 说明后者有更宽的密码子使用偏好的范围. 这种密码子使用的非均一性, 可从各基因密码子第 3 为 GC 含量(GC3s)的分布得到证实, 如图 2 所示. *A. aeolicus* 中 GC3s 范围从 16.4% 到 62.3%, 平均值 46.2%, 标准偏差 0.06. *A. pernix* 中 GC3s 的范围从 27.1% 到 90.9%, 平

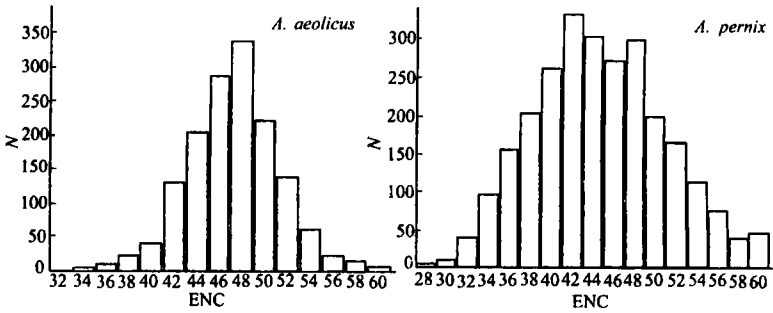


图 1 两个菌所有基因 ENC 值的分布

均值为 64.3%, 标准偏差为 0.1. 显然, 后者的 GC3s 变化范围更宽. 这暗示着除了碱基组成外, 还可能其他的因素影响了整个基因组中密码子的偏好. 由图 1, 2 可知, *A. aeolicus* 中 ENC 值在 47~49 之间的基因数量最多, 45~47 之间基因数量次之; GC3s 值在 0.45~0.50 之间的数量最

多, 0.50~0.55 之间基因数量次之, 但数量却相差 239. *A. pernix* 中 ENC 值在 41~43 之间的基因数量最多, 43~45 之间基因数量次之, 47~49 之间基因数量与之相差不多; 而 GC3s 值在 0.65~0.70 之间的基因数量最多, 0.60~0.65 之间基因数量次之. 两者在数量上仅相差 30.

2.3 密码子使用的对应分析

根据 RSCU 值, 利用 CA 分析两个菌中的所有基因. 对 *A. aeolicus* 的基因组, 其主要因子轴对变异方差的贡献仅为 8.77%. 第 2 和第 3 因子轴对方差的贡献分别为 4.79% 和 4.35%, 3 条轴因子合计占变异方差的 17.91%, 说明可能没有影响其密码子使用频率的主要趋势. *A. pernix* 情况则不同, CA 鉴定出一个影响密码子偏好的主要因子轴(轴 1), 它对变异方差的贡献达到 13.71%. 第 2, 3 因子轴对方差的贡献分别为 9.64% 和 4.81%, 3 条轴因子合计占变异方差的 32.99%. 说明存在影响其密码子使用频率的主要趋势. 由图 3 可知, *A. aeolicus* 中绝大部分的点都落在原点附近, 这说明这些基因都具有或多或少相似的

密码子使用偏好; 而少部分的点零散的分布轴 1 正值端的两端, 说明所对应的基因在密码子使用偏好上不相似. *A. pernix* 中点的分布情况则不同, 落在原点附近的点较少, 说明具有相似密码子使用偏好的基因数量较少. 另外, 根据轴 1 值可将所有的点分为两组, 而且这两组之间几乎没有重叠. 一组分布在轴 1 正值端, 另一组在负值端. 这可能与 2 组基因各偏向于使用某些特定的密码子有关. 第 1 条向量轴上每个基因的位置与该基因表达水平(CAI)的散点图, 如图 4 所示. *A. aeolicus* 中 CAI 与轴 1 的相关系数为 -0.821 , $p < 0.01$, 两者之间呈极显著负相关, GC3s 与之的相关系数为 -0.827 . 说明表达水平越高的基因密码子, 第 3 位的 GC 含量越高. GC3s 比基因表达水平对密码子使用偏好的影响略大, 其主要因子轴对变异方差的贡献为 8.77%. 两者均不是影响密码子使用的主要趋势.

A. pernix 中的情况则不同, 轴 1 与 CAI 的相关系数 $r = 0.77$, $p < 0.01$, 与 NC 的相关系数 $r = -0.196$, $p < 0.01$, 与 GC3s 的相关系数 $r = 0.637$, $p < 0.01$, 均达到极显著水平. 但 CAI 的相关性最强. 由此可见, 在该菌中基因表达水平是导致不同基因间密码子使用差异的主要原因, 而与轴 2 相关系数最大的是 GC3s ($r = 0.642$, $p < 0.01$), 其他均小于该值. 说明表达水平高的基因密码子第 3 位 GC 含量高, 而且基因表达水平是影响其密码子使用的主要趋势. 同时, GC3s 也是一个比较重要的因素.

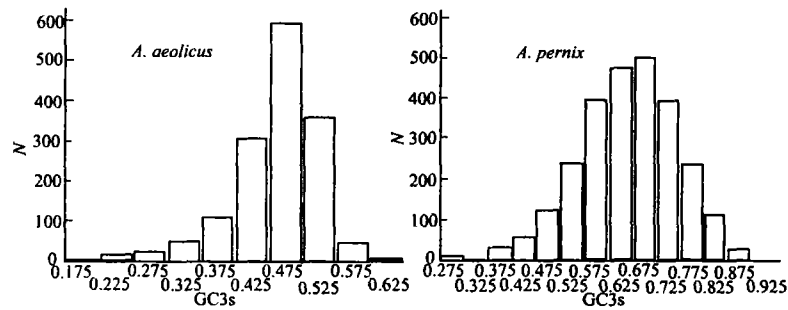


图 2 两个菌所有基因 GC3s 值的分布

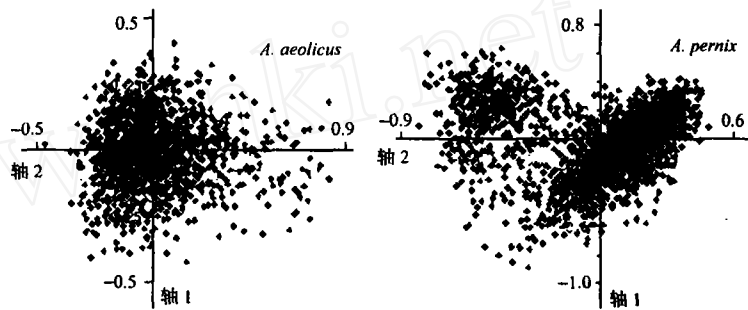


图 3 2 个菌 RSCU 值在两个主轴上的变化图

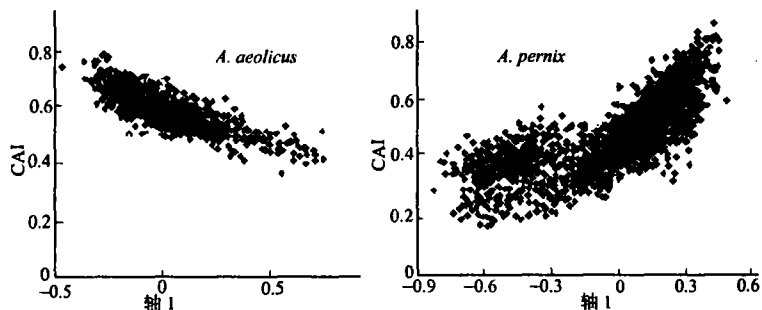


图 4 2 个菌第 1 条向量轴上的基因位置和对应 CAI 散点图

3 结束语

本文所分析的两个菌具有超嗜热的共性,但两者分属不同的域,分化时间很早,进化时间相差很长,导致其在密码子使用策略上存在比较大的差异.虽然在密码子使用上,两者有一定的相似性,但导致这一现象的原因似乎并不相同.由于 *A. aeolicus* 中 GC3s 平均值为 46.2%,属于 GC 含量中等的微生物;而 *A. pernix* 中 GC3s 平均值为 64.3%,属于高 GC 含量的微生物.后者偏向使用第 3 位为 G/C 的密码子是一个常见现象^[9,10],但前者就比较难以解释.究其原因,可能与其所处高温环境有关.除了 GC 含量之外,基因的表达水平与密码子使用偏好之间存在着一定的联系.在两种菌中,都存在高 GC 含量的基因倾向于使用有限的密码子,而且具有较高的表达水平.这一模式与很多生物的分析结果一致^[11,12].基因表达水平在 *A. aeolicus* 中不是导致密码子使用偏好的最主要因素,而在 *A. pernix* 中却是.

参 考 文 献

- 1 唐雪明,王正祥,诸葛健.具有工业应用价值的高热稳定性极端酶[J].食品与发酵工业,2001,5(27):65~70
- 2 钱 韦.两种植物病原黄单胞菌基因组中同义密码子使用的分析[J].植物病理学报,2004,34(2):97~106
- 3 张光亚,方柏山.敏捷气热菌密码子及 AUG 侧翼序列保守性分析[J].华侨大学学报(自然科学版),2004,25(2):192~196
- 4 吕 健,Rakhely G, Kovaes K L,等.嗜热菌 *Aquifex pyrophilus* 中 mbh2 基因簇的鉴定和部分基因的克隆及测序[J].微生物学报,2001,41(6):674~678
- 5 Grant H J, Oliver R, Peter AS, et al. Transterm: A database of mRNAs and translational control elements [J]. Nucleic Acids Res, 2002, 30: 310~311
- 6 侯卓成,杨 宁.影响链球菌肺炎球菌基因组密码子使用的因素分析[J].遗传学报,2002,29(8):747~752
- 7 王世峰,许洪林,陆柔剑,等.痘苗病毒基因组密码子使用频率分析[J].病毒学报,2002,3(18):227~234
- 8 McInerney J O. GCUA: General codon usage analysis [J]. Bioinformatics, 1998, 14: 372~373
- 9 Grocock R J, Sharp P M. Synonymous codon usage in *Pseudomonas aeruginosa* PA01 [J]. Gene, 2002, 189:131~139
- 10 Chen L L, Zhang C T. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages[J]. Biochemical and Biophysical Research Communications, 2003, 30(6):310~317
- 11 Gupta S K, Ghosh T C. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa* [J]. Gene, 2001, 273:63~70
- 12 Dutta P A, Das C J. Codon usage in highly expressed genes of *Haemophilus influenzae* and *Mycobacterium tuberculosis*: Translational selection versus mutational bias [J]. Gene, 1998, 215: 405~413

Analysis of Synonymous Codon Usage in Two Hyperthermophiles

Zhang Guangya Ge Huihua Fang Baishan

(College of Material Science and Engineering, Huaqiao University, 362021, Quanzhou, China)

Abstract Statistics of codon usage in the genomes of two heperthermophiles *Aeropyrum pernix* and *Aquifex aeolicus* were calculated based on the published genomic DNA sequences. It was found that they were similar in the usage of codon, while they also had different patterns, however the mechanism that caused this might be different. There exits great variation of codon usage among these genes. Correspondence analysis based on the RSCU showed that the higher of the GC content, the higher of the gene expression level, but the extent that gene express level influenced the codon usage bias was different. In *A. pernix*, it was the most important factor that shaped the codon usage patterns, but not in *A. aeolicus*.

Keywords *Aeropyrum pernix*, *Aquifex aeolicus*, codon usage, gene expression level, correspondence analysis