人类基因同义密码子偏好的特征以及与基因 GC 含量的关系

石秀凡 黄京飞 柳树群 刘次全*

(中国科学院昆明动物研究所，细胞与分子进化重点实验室，昆明 650223)

摘要 对人类的 728 个基因，按其编码区中 GC 的含量分成四组 (从 GC< 0.43 到 GC> 0.58)，分别考察了这四组样本对同义密码子偏好的特征，发现在全部样本中都呈现 NTG (N 代表四种碱基中的任一种) 特受偏爱和 NCG 尽量避免的特征. 基因环境中 GC 含量与 C3/G3 含量 (密码子第三位 C 和 G 的含量) 的相关分析，以及四组样本对密码子的偏好都支持以 C 结尾的密码子在编码中有特殊的优势，这种优势有利于保证翻译的准确性. 还考察了各种氨基酸含量随编码区 GC 含量不同而变化的趋势.

关键词 密码子偏好，同义密码子，人类基因，C 结尾密码子，NTG 密码子

学科分类号 Q755

在先前对密码子的研究中，我们指出人类基因显示出对以 C 或 G，尤其 C 结尾的同义密码子的偏好. 同时鉴于以 C/G 结尾的密码子比以 A/U 结尾的同义密码子具有更强的密码子-反密码子结合能，我们推测偏好 C/G 结尾的密码子有利于提高翻译的准确度[1]. 对各种动物的基因组碱基组分的研究指出，从低等生物到冷血脊椎动物，4 种碱基在基因组中呈均一分布，虽然在不同物种间碱基的比例可能有较大的差别；但从温血脊椎动物开始，基因组中大部分是 GC 贫乏区，其余是 GC 丰富区和 GC 极端丰富区，并且基因大量集中在 GC 丰富区和 GC 极端丰富区. 人类基因的 38% 位于 GC 丰富区，28% 位于 GC 极端丰富区，而这两区只分别占人类基因组的 31% 和 3%. 占大部分基因组的 GC 贫乏区却只含基因总体的 34%[2]. 编码区的碱基组分和对同义密码子的偏好都与其所在的基因组环境的碱基组分相关[3,4]. 在先前工作的基础上，我们进一步考察了不同 GC 含量的基因对同义密码子的偏好，分析基因所在区域 GC 含量对密码子偏好的影响和不同 GC 含量的基因所编码的蛋白质中各种氨基酸含量的差别.

1 材料和方法

首先由网上 (http://www.expasy.ch/sprot/hpi/) 下载和 8，9，10，11，X 五条人染色体相关的 SWISS-PROT 蛋白质文件，然后由 EMBL 站点下载所有列在 SWISS-PROT 蛋白质文件上的基因序列，再由这些基因序列中挑选包含完整编码序列 (CDS) 并具有最长非编码序列 (NCDS) 的序列作为该蛋白质的样本. 由于 GC 贫乏的基因很少，我们又加入了先前收集到的 52 个 GC 较少的样本，共得 728 个样本以供分析. 该样本集虽较小，但与含 1 400 个 CDS 的样本[4]相比，前者 CDS 中和密码子第三位的 GC 含量范围分别为 34.52% ~ 75.60% 和 26.14% ~ 94.56%，后者为 33% ~ 77% 和 27% ~ 97%，即我们样本的覆盖面仅稍差于 GC 极端丰富区.

样本集的处理：首先用本组编的 PickupCDS 软件将全部样本的 CDS 与 NCDS 分离，然后依据 CDS 中 GC 的含量将样本分为四组. 1 组：GC< 0.4300，共 93 个样本；2 组：0.4300 ⩽ GC ⩽ 0.5000，共 212 个样本；3 组：0.5000 < GC ⩽ 0.5800，共 227 个样本；4 组：GC> 0.5800，共 196 个样本. 同义密码子相对使用度 (RSCU，其值大于、等于、或小于 1 意味偏好、正常、或避免使用该密码子) 和氨基酸的出现频率用本组编的 CodonB 软件实现. 统计分析用 STATISTICA 5.0 (StatSoft, Inc.).

2 结果和讨论

2.1 特受偏爱的 NTG 和尽量避免的 NCG

表 1 列出了四组不同 GC 含量的样本 RSCU 值. 从表 1 明显可见同义密码子偏好的普遍趋势，

* 通讯联系人.

Tel: 0871-5195183，E-mail: tbrg@public.km.yn.cn

收稿日期：2001-08-21，接受日期：2002-01-31

Table 1 The bias of synonymous codons in four groups with various GC contents

Amino acid	Codon	RSCU			
		Group1	Group2	Group3	Group4
Val	GTT	**1.30**	**1.03**	0.59	0.26
	GTG	**1.18**	**1.45**	**2.00**	**2.47**
	GTC	0.65	0.85	**1.10**	**1.09**
	GTA	0.87	0.67	0.31	0.18
Ile	ATT	**1.52**	**1.36**	0.94	0.56
	ATC	0.72	**1.09**	**1.74**	**2.29**
	ATA	0.76	0.55	0.33	0.16
Leu	CTT	**1.25**	**1.05**	0.67	0.38
	CTG	**1.25**	**1.72**	**2.62**	**3.43**
	CTC	0.65	0.98	**1.36**	**1.47**
	CTA	0.64	0.55	0.37	0.21
	TTG	**1.12**	**1.05**	0.73	0.44
	TTA	**1.08**	0.65	0.25	0.07
Phe	TTT	**1.33**	**1.11**	0.64	0.52
	TTC	0.67	0.89	**1.26**	**1.48**
Ala	GCT	**1.57**	**1.37**	**1.13**	0.71
	GCG	0.15	0.23	0.36	0.64
	GCC	0.87	**1.18**	**1.71**	**2.10**
	GCA	**1.40**	**1.22**	0.80	0.55
Thr	ACT	**1.46**	**1.24**	0.91	0.51
	ACG	0.22	0.29	0.46	0.71
	ACC	0.85	**1.11**	**1.71**	**2.05**
	ACA	**1.48**	**1.35**	0.93	0.73
Pro	CCT	**1.51**	**1.44**	**1.17**	0.76
	CCG	0.17	0.25	0.42	0.64
	CCC	0.67	0.85	**1.32**	**1.77**
	CCA	**1.65**	**1.46**	**1.10**	0.86
Ser	TCT	**1.59**	**1.37**	**1.03**	0.66
	TCG	0.13	0.20	0.34	0.57
	TCC	0.75	**1.09**	**1.51**	**1.65**
	TCA	**1.37**	**1.08**	0.71	0.52
	AGT	**1.30**	**1.12**	0.88	0.53
	AGC	0.86	**1.15**	**1.54**	**2.07**
Gly	GGT	**1.04**	0.80	0.67	0.47
	GGG	0.61	0.78	0.98	**1.12**
	GGC	0.72	1.00	**1.40**	**1.87**
	GGA	**1.63**	**1.42**	0.95	0.54
Arg	CGT	0.64	0.63	0.58	0.42
	CGG	0.48	0.77	**1.26**	**1.78**
	CGC	0.37	0.63	**1.17**	**1.95**
	CGA	0.85	0.80	0.69	0.49
	AGG	**1.25**	**1.34**	**1.29**	0.98
	AGA	**2.41**	**1.84**	1.00	0.38
Cys	TGT	**1.29**	**1.08**	0.84	0.55
	TGC	0.71	0.92	**1.16**	**1.45**
Asp	GAT	**1.39**	**1.13**	0.87	0.54
	GAC	0.61	0.87	**1.13**	**1.46**
Asn	AAT	**1.30**	**1.11**	0.81	0.47
	AAC	0.70	0.89	**1.19**	**1.53**
His	CAT	**1.33**	**1.05**	0.74	0.43
	CAC	0.57	0.95	**1.26**	**1.57**
Tyr	TAT	**1.28**	**1.10**	0.83	0.47
	TAC	0.72	0.90	**1.17**	**1.53**
Glu	GAG	0.66	0.88	**1.28**	**1.60**
	GAA	**1.34**	**1.12**	0.72	0.40
Gln	CAG	**1.10**	**1.27**	**1.57**	**1.77**
	CAA	0.90	0.73	0.43	0.23
Lys	AAG	0.80	0.99	**1.30**	**1.59**
	AAA	**1.20**	1.01	0.70	0.41
stop	TGA	**1.22**	**1.24**	**1.50**	**1.85**
	TAG	0.53	0.59	0.83	0.76
	TAA	**1.25**	**1.17**	0.67	0.39

The values of synonymous codon usage (RSCU) greater, or equal or less than 1 means the trend of bias, or normal or avoid in codon usage. Bold faces indicate the values greater than 1.

即随基因编码区中 GC 含量的增加，以 GC 结尾的同义密码子的使用度增大. 但最突出的二点例外是 NTG 和 NCG 密码子 (N 代表四种碱基中的任一种). 除编码亮氨酸 (Leu) 的 6 个密码子中的 TTG 在 3, 4 组中小于 1 外, NTG 在四组样本中的 RSCU 值都大于 1, 且在 2, 3, 4 组中都是最偏好密码子. 而 NCG 即便在 GC 最丰富的第 4 组中, RSCU 值最高也只有 0.7 左右. 我们在网上 (http://www.kazusa.or.jp/codon/) 查看了大肠杆菌 (*Escherichia coli*), 酵母 (*Saccharomyces cerevisiae*), 果蝇 (*Drosophila melanogaster*) 和非洲爪蟾 (*Xenopus laevis*) 的密码子使用频率, 虽然这些物种的编码序列中 GC 的含量相差较大, 分别为 51.11%, 39.72%, 53.99% 和 47.51%, 但对 NUG 都显示出较大的偏好. 对 NCG 的避免在酵母和非洲爪蟾中也十分明显. 这反映了基因组对 TG 二核苷酸的普遍偏好和某些物种基因组对 CG 二核苷酸的避免. 不同物种基因组对各种二核苷酸的偏好程度已见报道[5], 这种偏好直接影响到对同义密码子的挑选, 但其生物含义还不清楚.

2.2 以 C 结尾的密码子的特殊优势

因为偏好密码子与基因中的 GC 含量密切相关, 我们将基因编码区与其非编码区的 GC 含量作配对统计比较 (表 2), 发现即便在 GC 最低的 1 组, C3 平均值也略高于非编码区中 C 的平均值

Table 2 The results of paired *t* test for comparing C3/G3 with CN/GN (the C/G content of non-coding region) in group 1 and 2 where GC< 50%

	CN	C3	GN	G3
Group1(SD)	0.1818 (0.03)	0.1924 (0.04)	0.1911 (0.02)	0.1974 (0.03)
P		0.1736		0.3442
Group2(SD)	0.1878 (0.04)	0.2427 (0.04)	0.1987 (0.03)	0.2394 (0.04)
P		10^{-6}		10^{-6}

Standard deviation is in the parentheses.

(虽无显著性差异), 在 GC 低的 2 组, 其非编码区的 GC 含量与 1 组相近, 而 C3, G3 显著高于非编码区中的 C 和 G ($P< 10^{-6}$), 表明在 GC 含量最低的样本中 C3/G3 与 CN/GN (非编码区的 G 或 C 含量) 也无显著性差异, 并且在 GC 低的环境中还包含某些 GC 含量较高的编码区. 由图 1 的相关曲线可见, C3 与 CN 之间, G3 与 GN 之间都呈明显

线性正相关，但C3与CN间的相关系数更大，上升的速率也更快．再结合表1中的数据都显示以C结尾的密码子在人类基因中有特殊的优势．C结尾密码子的优势在其他文献中也见报道[6,7]．这种优势很可能与密码子-反密码子结合能的强弱有关．以C结尾的密码子具有最强的结合能，有利于保证翻译的准确性[1]．

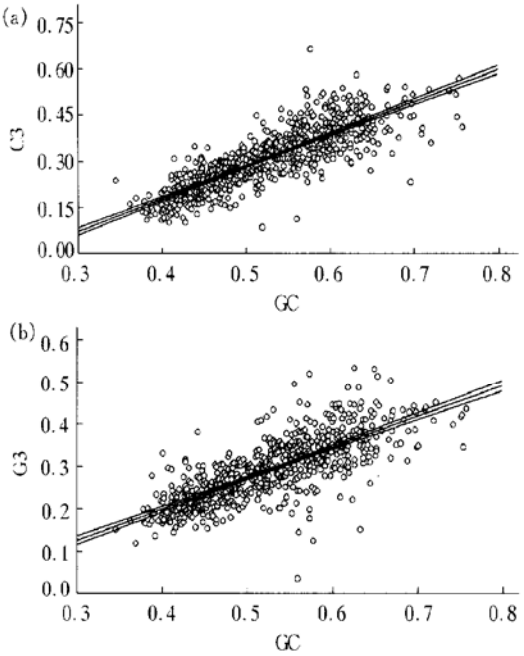

Fig. 1 Correlation curves between C3 / G3 (the content of C/G in the 3rd sites of codons) and the content of C+ G in the coding sequences

Both C3 and G3 show positive linear correlation with GC content, but C3 shows higher correlation coefficient and obviously higher slope. (a) C3= − 0.2426+ 1.0538 × GC. correlation: r = 0.83761. (b) G3= − 0.0922+ 0.73140 × GC. correlation: r = 0.77256.

2.3 氨基酸的偏好

基因编码区GC的含量也影响到蛋白质中氨基酸成分的差别．表3列出了20种氨基酸在四组样本中出现的频率．其中Ala、Gly和Pro三种疏水氨基酸分别由三种四重简并的GC型密码子（即密码子的第一、第二位碱基为G或C）编码，其频率明显地随基因中GC含量的增加而升高．Phe和Ile这二种疏水氨基酸都由AT型密码子（即密码子的第一、第二位碱基为A或T）编码，其频率随基因中GC含量的增加而下降．极性和带电氨基酸中，由AT型密码子编码的Asn和Lys明显地随基因中GC含量的增加而下降，其余也随基因中GC含量的增加而呈下降趋势，只有Arg和Thr呈上升趋势．总之，随基因中GC含量的增加，疏水氨基酸比例上升，亲水氨基酸比例下降．六重简并编码的三种氨基酸，Leu、Ser和Arg分别属于疏水、极性和带电氨基酸，Leu和Ser在四组样本中出现的频率相近并很高，分别为0.09和0.07以上，提示这二种氨基酸在各类蛋白质中都占较大比例，六重简并编码可适应不同GC含量的基因对这些氨基酸的需求．

Table 3 Frequencies of 20 amino acids appearing in the four groups with various GC contents

Amino acid	Group 1	Group 2	Group 3	Group 4
Gly	**0.0520**	0.0613	0.0740	**0.0789**
Ala	**0.0548**	0.0618	0.0708	**0.0832**
Pro	**0.0455**	0.0524	0.0594	**0.0791**
Ile	**0.0572**	0.0524	0.0469	**0.0351**
Phe	0.0427	0.0408	0.0388	0.0334
Leu	0.0955	0.0928	0.0961	0.0996
Val	0.0585	0.0603	0.0620	0.0641
Trp	0.0108	0.0200	0.0211	0.0226
Cys	0.0205	0.0200	0.0211	0.0226
Met	0.0236	0.0242	0.0243	0.0201
Ser	0.0781	0.0764	0.0764	0.0743
Thr	0.0546	0.0555	0.0526	0.0658
Asn	**0.0523**	**0.0430**	**0.0360**	**0.0288**
Gln	0.0441	0.0461	0.0436	0.0439
Tyr	0.0313	0.0310	0.0287	0.0260
Lys	**0.0753**	0.0676	0.0573	**0.0431**
His	0.0227	0.0242	0.0241	0.0237
Glu	0.0775	0.0719	0.0704	0.0607
Asp	0.0547	0.0518	0.0503	0.0430
Arg	0.0464	0.0513	0.0528	0.0597

Bold faces indicate the amino acids that show obviously different frequencies in the four groups.

参 考 文 献

1 石秀凡，黄京飞，梁宠荣，等．人类基因中同义密码子的偏好与密码子-反密码子间的结合强度密切相关吗？科学通报，2000，**45**（23）：2520~2525
Shi X F, Huang J F, Liang C R, *et al*. Chinese Science Bulletin, 2001, **46** (12): 1015~1019

2 D'Onfrio G, Mouchiroud D, Aïssani B, *et al*. Correlation between the compositional properties of human genes, codon usage, and amino acid composition of protein. J Mol Evol, 1991, **32** (6): 504~510

3 Bernardi G. The isochore organization of the human genome. Annu Rev Genet, 1989, **23**: 637~661

4 Bernardi G. The human genome: organization and evolution history. Annu Rev Genet, 1995, **29**: 445~476

5 Karlin S, Campbell M A, Mrózek J. Comparative DNA analysis across diverse genomes. Annu Rev Genet, 1998, **32**: 185~225

6 Musto H, Romero H, Zavala A, *et al*. Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. J Mol Evol,

1999, **49** (1): 27~ 35
7 Thomas L K, Dix D B, Thompson R C. Codon choice and gene expression: Synonymous codons differ in their ability to direct aminoacylated-transfer RNA binding to ribosomes *in vitro*. Proc Natl Acad Sci USA, 1988, **85** (12): 4242~ 4246

The Features of Synonymous Codon Bias and GC-content Relationship in Human Genes

SHI Xiu-Fan, HUANG Jing-Fei, LIU Shu-Qun, LIU Ci-Quan*
(*Cellular and Molecular Evolutionary Laboratory, Kunming Institute of Zoology, The Chinese Academy of Sciences, Kunming* 650223, China)

Abstract 728 human genes were divided to four groups according to the GC contents of their coding sequences (from GC< 0.43 to GC> 0.58). Examination of synonymous-codon bias in the 4 groups show that NTG (N represents any base of T, A, C, G) is most favored and NCG is most avoided in all four groups. Statistical correlation analysis of GC content in genetic environment with C3/G3 content (the C or G in the 3rd position of codons) and the codon bias in the four groups suggest that C-ending codons are special preferred. This is in favor of accurate translation. Frequency of each amino acid in the four groups was also examined.

Key words codon bias, synonymous codons, human genes, C-ending codons, NTG codons

* Corresponding author. Tel: 86-871-5195183, E-mail: tbrg@public.km.yn.cn
Received: August 21, 2001 Accepted: January 31, 2002