玉米同义密码子使用偏性分析

于金海
(黑龙江省农业科学院 作物育种研究所,黑龙江 哈尔滨 150086)

摘要:利用玉米 B73 全基因组 53 764 个基因的表达序列数据,使用多重变量分析软件 CodonW 对影响密码子用法的因素进行分析。结果表明:在 59 种同义密码子中有 27 种为玉米最优密码子。同时还指出,最优密码子的使用频率(FOP)与基因的 G+C 含量(GC),尤其是第 3 位密码子的 G+C 百分含量($GC3_S$)、密码子适应指数(CAI)和密码子偏爱指数(CBI)之间均呈现极显著的正相关(相关系数分别为 r=0.780 7、0.822 3、0.731 4和 0.986 3),而与有效等位基因数 ENc 存在显著负相关(相关系数 r=－0.681 5)。表明基因 GC 含量直接影响玉米密码子使用的偏性,同时基因的表达水平越高,对密码子的使用偏向性越强。27 个最优密码子的首次确定将对玉米转基因具有重要指导意义。

关键词:玉米;同义密码子使用偏性;最优密码子;GC 含量
中图分类号:S513.035.3;Q755　　**文献标识码**:A　　**文章编号**:1002-2767(2010)07-0001-04

遗传密码共有 64 种,代表 20 种不同的氨基酸和翻译终止信号,每个密码子被细胞质中的转运 RNA(tRNA)所识别,从而完成蛋白质的翻译过程。除了 3 个终止密码子外,构成生物体蛋白质的 20 种氨基酸由 61 种不同的密码子所编码,这就意味着存在几种不同的密码子编码同一种氨基酸的现象,将编码同一种氨基酸的密码子之间互称为同义密码子(synonymous codon)。这些同义密码子在翻译成蛋白质时并非均一使用的,而通常是某些密码子被优先使用,某一物种或某一基因通常会倾向于使用一种或几种特定的同义密码子,这些密码子称为最优密码子(optimal codon),此现象被称为密码子偏性(codon bias)。尽管密码子的改变对蛋白质的基本结构不会产生重大改变,但密码子的使用并不是随机的。有研究表明,从原核生物到真核生物,其基因组中同义密码子使用偏性的现象广泛存在[1-2],这一现象的产生与诸多因素有关,如基因的表达水平[3]、翻译起始效应[4]、基因的碱基组分[5]、基因的长度[3]、tRNA 的丰富度[6]等。密码子的偏向性使用主要是保证翻译的准确性和效率,是物种在长期的进化过程中受各种选择压力的影响所致[7-8]。遗传密码在进化过程中是极其保守的,尽管其仍处于进化过程中。密码子使用上的多样性与进化过程密切相关,通常认为在漫长的进化过程中密码子或氨基酸使用上的分歧是物种间发生分离进而产生新物种的重要原因之一[9-11]。在同一基因组内部,高表达的基因相对于其它低表达或中等表达的基因具有更为强烈的密码子使用上的偏好[12-13]。

近年来,随着很多物种全基因组测序的完成,使得研究者对密码子的使用问题在全基因组水平上有一个全面的了解,也成为近年来基因组学研究的热点。2009 年 11 月,美国科学家宣布玉米自交系 B73 的全基因组草图绘制已经完成[14],其数据可从公共数据库免费获得,这意味着玉米功能基因组的研究已经起步。同时,也为利用现代生物信息学的手段在全基因组水平上研究其功能基因的同义密码子使用偏性成为可能。现利用玉米全基因组数据,首次全面分析了玉米全部基因密码子使用情况,确定了玉米最优密码子的种类、使用频率以及影响密码子使用的因素,以期对该作物通过选择合适密码子而构建最佳的转基因表达系统,进而提高特定基因的表达效率,为今后玉米转基因育种提供有价值的参考和依据。

1　材料与方法

1.1　基因表达数据的获得

以玉米自交系 B73 全基因组表达序列为研究对象,全基因组数据库来自玉米测序网站 http://www.maizesequence.org/index.html,共获得了 53 764 个基因的表达序列。

1.2　分析软件的使用

密码子使用的相关分析采用 CodonW 分析程序(J. Peden, http://www.molbiol.ox.ac.uk/cu/),该软件采用多重变量分析的方法,对基因组内的密码子使用模式均会给出详细的分析结果,是分析密码子使用偏性的通用软件。该软件分析密码子使用偏性指标:(1)有效密码子数(Effective number of codon,ENc),该值是一个基因的密码子使用频率与同义密码子平均使用频率偏差

收稿日期:2010-04-06
作者简介:于金海(1971-),男,黑龙江省密山市人,学士,助理研究员,从事玉米遗传育种研究。E-mail:hnygy@163.com。

的量化值。ENc 值的范围介于为 20(每个氨基酸只使用一个密码子的极端情况)～61(各个密码子均被平均使用),越靠近 20 偏性越强。(2)密码子适应指数(Codon Adaptation Index,CAI),密码子适应指数常用于基因表达水平的测量。此值为 0～1,越接近 1 表示基因的表达水平越高。(3)密码子偏爱指数(Codon Bias Index,CBI),反应一个具体基因中高表达优越密码子的组分情况。(4)最优密码子使用频率(frequency of optimum codons,FOP),是指所使用的最优密码子占总数的百分比。(5)相对密码子使用度(Relative synonymous codon usage,RSCU),是指对于某一特定的密码子在编码对应氨基酸的同义密码子间的相对概率,如果密码子的使用没有偏好,则该密码子的 RSCU 值等于 1。当某一密码子的 RSCU 值大于 1,则表明该密码子的使用频率相对较高,反之亦然。

2 结果与分析

2.1 最优密码子的确定

表 1 中给出了玉米全部高表达和低表达基因的 59 种同义密码子(不包括 3 种终止密码子 TAG、TGG、TGA 和 1 个起始密码子 ATG 以及仅有 1 个密码子的色氨酸 TGG)以及各同义密码子平均使用频率的情况。由表 1 中可以看出玉米编码 18 种氨基酸(不包括起始密码子 ATG 编码的甲硫氨酸和只有一个密码子的色氨酸)的密码子的使用均存在偏好性,同义密码子数为 2 或 3 的均偏向使用 1 种密码子,而同义密码子数为 4 或 6 的则有 2 或 3 种偏向使用的密码子。从而可以确定在玉米所使用的 59 种同义密码子中 TTC、CTC、CTG 等 27 个密码子为最优密码子。而且最优密码子均为以 C 或 G 结尾的密码子,说明最优密码子的使用与密码子中第 3 位碱基组成密切相关。

表 1 59 种同义密码子在玉米高/低表达基因样本组的使用频率

AA	Codon	RSCU		AA	Codon	RSCU	
		High	Low			High	Low
Phe	TTT	0.03(477)	1.32(22346)	Ser	TCT	0.09(816)	1.68(25369)
	TTC *	1.97(30259)	0.68(11533)		TCC *	2.23(20724)	0.59(8835)
Leu	TTA	0.01(70)	0.97(14038)		TCA	0.07(668)	1.61(24249)
	TTG	0.10(1261)	1.40(20275)		TCG *	1.75(16215)	0.28(4286)
	CTT	0.09(1175)	1.59(23128)		AGT	0.04(374)	1.14(17158)
	CTC *	2.81(35875)	0.52(7489)		AGC *	1.82(16868)	0.70(10543)
	CTA	0.06(787)	0.73(10598)	Pro	CCT	0.14(1782)	1.62(18587)
	CTG *	2.93(37459)	0.80(11620)		CCC *	1.46(18112)	0.41(4733)
Ile	ATT	0.06(486)	1.44(23637)		CCA	0.13(1617)	1.70(19546)
	ATC *	2.87(22836)	0.60(9870)		CCG *	2.26(28023)	0.27(3083)
	ATA	0.07(578)	0.95(15599)	Thr	ACT	0.06(571)	1.53(17836)
Val	GTT	0.06(899)	1.73(25966)		ACC *	1.87(17455)	0.58(6720)
	GTC *	1.63(25944)	0.57(8552)		ACA	0.08(732)	1.65(19219)
	GTA	0.04(566)	0.81(12211)		ACG *	1.99(18560)	0.24(2856)
	GTG *	2.27(36077)	0.89(13330)	Ala	GCT	0.11(2890)	1.72(27325)
Tyr	TAT	0.03(318)	1.34(16892)		GCC *	1.92(50432)	0.52(8250)
	TAC *	1.97(21379)	0.66(8401)		GCA	0.09(2481)	1.53(24254)
His	CAT	0.07(693)	1.47(18698)		GCG *	1.88(49288)	0.24(3762)
	CAC *	1.93(18040)	0.53(6693)	Cys	TGT	0.03(254)	1.21(10755)
Gln	CAA	0.06(692)	1.11(21812)		TGC *	1.97(14563)	0.79(6976)
	CAG *	1.94(21605)	0.89(17404)	Arg	CGT	0.11(1072)	0.86(7257)
Asn	AAT	0.06(626)	1.35(29124)		CGC *	3.03(29706)	0.40(3342)
	AAC *	1.94(19669)	0.65(13907)		CGA	0.08(741)	0.69(5782)
Lys	AAA	0.06(767)	1.04(29752)		CGG *	1.82(17838)	0.45(3810)
	AAG *	1.94(25448)	0.96(27498)		AGA	0.06(574)	2.09(17541)
Asp	GAT	0.07(1413)	1.46(40137)		AGG	0.90(8810)	1.51(12639)
	GAC *	1.93(40921)	0.54(14981)	Gly	GGT	0.11(1865)	1.41(20563)
Glu	GAA	0.07(1314)	1.17(36100)		GGC *	2.76(48757)	0.64(9330)
	GAG *	1.93(36546)	0.83(25730)		GGA	0.14(2528)	1.31(19076)
					GGG *	0.99(17445)	0.64(9394)

注:高表达优越密码子被标记 *,卡方检验值为 $P<0.01$。

2.2　密码子差异性分析

图 1 给出 59 种同义密码子中分别以 A、C、G 和 T 碱基结尾的密码子在多重变量分析中前 2 个坐标向量的分布情况，可以看出，以 A 和 T 结尾的密码子聚集在坐标平面的左侧，而以 G 和 C 结尾的密码子聚集在右侧，仅发现 2 个以 G 结尾的密码子位于纵轴的左侧。因此目标向量的第一坐标轴（即纵轴）可用来区分 A、T 结尾的密码子和 G、C 结尾的密码子。还可以看出，以 A 结尾的密码子大部分位于横轴的上方，而以 T 结尾的密码子除了 2 个位于横轴上方外，其余均位于坐标轴的下方。以 G 结尾的密码子基本上位于横轴的上方，而以 C 结尾的密码子相对分布较分散，说明以 C 结尾的密码子在使用上比较均匀，没有明显的偏性存在，而其它 3 种密码子的使用上均存在一定的偏性。

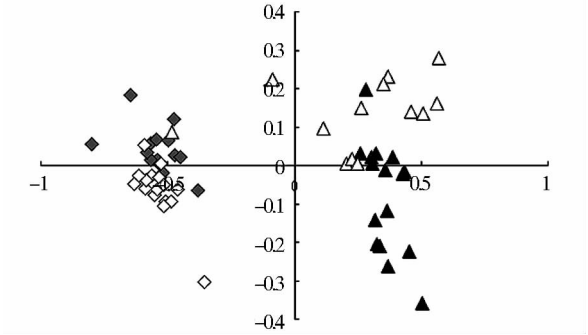

图 1　4 种碱基结尾的 59 种同义密码子分布图

◆A-ending, ▲C-ending, ΔG-ending, ◇T-ending

2.3　密码子使用偏性相关因素分析

对最佳密码子使用频率（FOP）和有效等位基因数（ENc）与基因的 GC 含量、第 3 位密码子 GC3s 含量、第 3 位密码子 A、C、G 和 T 碱基平均含量、密码子适应指数 CAI 和密码子偏爱指数 CBI 等之间的相关系数进行了估算。结果表明，玉米最优密码子使用频率 Fop 与 GC 含量、GC3s 含量、C3s 和 G3s 之间的相关系数分别为 r＝0.780 7、0.822 3、0.826 2 和 0.600 1，均达到极显著水平，说明最优密码子使用频率与基因编码序列组成 GC 含量尤其是第 3 位密码子的 GC 含量存在明显正相关。而与 T3s 和 A3s（相关系数 r＝－0.7805 和－0.8034）之间存在明显负相关。同时，最优密码子使用频率与有效等位基因数之间存在明显负相关（相关系数 r＝－0.6815），使用的有效等位基因数越多，其最优密码子使用频率越低。此外，最优密码子使用频率与密码子适应指数 CAI 和密码子偏爱指数 CBI（相关系数 r＝0.7314 和 0.9863）之间、CAI 与 CBI（相关系数 r＝0.6782）之间均存在显著的正相关，这说明基因的表达水平越高，其密码子使用偏向性越强。

表 2　主要参数相关系数显著性检验

CBI	Fop	ENc	GC	GC3s	T3s	C3s	A3s	G3s	CAI	CBI
Fop	1.0000									
Nc	－0.6815	1.0000								
GC	0.7807	－0.7391	1.0000							
GC3s	0.8223	－0.7955	0.9274	1.0000						
T3s	－0.7805	0.7692	－0.9165	－0.9700	1.0000					
C3s	0.8262	－0.7522	0.8369	0.9364	－0.9069	1.0000				
A3s	－0.8034	0.7471	－0.8959	－0.9500	0.8586	－0.8865	1.0000			
G3s	0.6001	－0.6301	0.7234	0.8333	－0.7881	0.6268	－0.7849	1.0000		
CAI	0.7314	－0.4415	0.3544	0.4823	－0.3842	0.5655	－0.5086	0.3635	1.0000	
CBI	0.9863	－0.6880	0.7970	0.8339	－0.7978	0.8300	－0.8214	0.5992	0.6782	1.0000

注：表中数值均达到极显著水平（$P<0.01$）。

3　讨论

密码子是核酸携带信息和蛋白质携带信息间对应的基本原则，是生物体内信息传递的基本环节。一个物种的基因在长期进化过程中会逐渐适应宿主的基因组环境，采用符合宿主基因组的密码子用法。研究表明，不同的物种之间对密码子的使用具有很强的特异性，进化过程中相近的物种具有相似的密码子使用模式，说明物种的密码子使用特点在进化过程中具有保守性，这种变化直接反映在进化使核苷酸的组成发生了变化[15]。密码子的使用偏性产生的生物学基础目前还不清楚，但随着人类以及各种动植物全基因组测序的相继完成，对密码子使用情况将有更进一步的了解。近几年来，研究者通过模式生物的基因组碱基组分和密码子用法特征研究表明，除了物种本身的原因，影响密码子使用偏性的因素很多，而基因组环境的碱基组分、尤其是基因编码区 GC 含量在其同义密码子使用偏性的产生方面占有较大的优势[3-4,16-17]。

该研究表明，基因编码区的碱基组分（GC 含量）强烈影响其同义密码子使用偏性，二者呈高度正相关。作为单子叶植物的玉米其基因编码序列 GC 含量为 56.1%，明显高于双子叶植物的拟南芥（基因 GC 含量为 43.2%），前者偏向于使用以 G 或 C 结尾的密码子，而拟南芥则偏好于使用 A 或 T 结尾的密码子[18]。尤其是基因第 3 位密码子的 GC 含量（GC3s）与最适密码子的平均使用

频率之间正相关关系更为显著。

研究还发现,密码子的使用偏性还与基因的表达强弱有关,常常在一些高表达的基因中密码子发生偏离的程度越大,而在低表达基因中则相反。这与 Stenico 等和 McInerney 等的研究结果一致,高效表达的基因所使用的密码子相比较那些低表达的基因具有明显不同的密码子使用频率,高效表达的基因较其它非高效表达基因在密码子使用上具有更为严重的偏向性趋势,它们通常使用一套偏好性同义密码子[19-20]。此外,这种偏好性还与细胞质中可利用的 tRNA 密切相关,发生偏好性使用的密码子直接对应于含量最丰富的 tRNA,这样可以使基因翻译成蛋白质时达到最佳的翻译效率[6]。因此,对物种密码子使用偏性进行研究,可以为转基因育种时构建合适的表达系统提供依据,从而提高转基因的表达效率[21]。

该研究中确定了玉米的 27 个最优密码子,对于今后玉米转基因过程中构建合适的转基因表达系统具有重要的指导意义,从而提高特定蛋白的表达量,增强转基因育种的效果。

参考文献:

[1] 赵翔,霍克克,李育阳. 毕赤酵母的密码子用法分析[J]. 生物工程学报,2000,16(3):308-311.

[2] 王艳,马文丽,郑文岭. SARS 冠状病毒的密码子偏好性分析[J]. 生命科学研究,2003,7(3):219-223.

[3] Duret L, Mouchiroud D. Expression pattern and surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila and Arabidopsis [J]. Pro Nat Acad Sci USA, 1999, 96(8):4482-4487.

[4] Sajau H, Washio T, Saito R, et al. Correlation between sequence conservation of the 5 untranslated region and codon usage bias in Musmusculus genes [J]. Gene, 2001, 276(2):101-105.

[5] Esley M, Heizer Jr, Douglas W, et al. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: A whole-genome analysis [J]. Mol Biol Evol, 2006, 23:1670 -1680.

[6] Kanaya S. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Badillus subtilis tRNA: Gene expression level and species-specific diversity of codon usage based on multivariate analysis [J]. Gene, 1999, 238:143-155.

[7] Gouy M, Gautier C. Codon usage in bacteria correlation with gene expressivity [J]. Nucleic Acid Research, 1982(10): 7055-7074.

[8] Akashi H. Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA[J]. Genetics, 1995, 139:1067-1076.

[9] Goldman N, Yang Z H. Codon based model of nucleotide substitution for protein coding DNA sequences [J]. Molecular Biology and Evoution, 1994(11):725-736.

[10] Nesti C, Poli G, Chicca M, et al. Phylogeny inferred from codon usage pattern in 31 organisma [J]. Computer Applications for the Biosciences, 1995(2):167-171.

[11] Schmidt W. Phylogeny reconstruction for protein sequences based on amino acid properties [J]. Journal of Molecular Evolution, 1995, 41:522-530.

[12] Chiapello H, Lisacek F, Caboche M, et al. Codon usage and gene function are related in sequences of Arabidopsis thaliana [J]. Gene, 1998, 209:1-38.

[13] Epstein R J, Lin K, Tan T W. A functional significance for codon third bases [J]. Gene, 2000, 245:291-298.

[14] Patrick S, Schnable D W, Robert S F, et al. The B73 Maize Genome: Complexity, Diversity and Dynamics [J]. Science, 2009, 326:1112-1115.

[15] David J L, Gregory A C S, Donal A H. Synonymous codon usage is subject to selection in thermophilic bacteria [J]. Nucleic Acids Research, 2002, 30:4272-4277.

[16] Paul M S, Laura R E, ZENG Kai. Forces that influence the evolution of codon bias [J]. Phil Trans R Soc B, 2010, 365:1203-1212.

[17] LV Hui, ZHAO Wei-ming, ZHENG Yan, et al. Analysis of Synonymous Codon Usage Bias in Chlamydia [J]. Acta Biochim Biophys Sin, 2005, 37:1-10.

[18] 张晓峰,薛庆中. 水稻和拟南芥 NBS. LRR 基因家族同义密码子使用偏好的比较[J]. 作物学报,2005,31(5):596-602.

[19] Stenico M, Lloyd A T, Sharp P M. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational bases [J]. Nucleic Acids Res, 1994, 22: 2437-2446.

[20] McInerney J Q. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi [J]. Proc Nat Acad Sci USA, 1998, 95:10698-10703.

[21] 张雅丽,杨国庆,郭英,等. 密码子优化的牛精蛋白基因在大肠杆菌中的表达[J]. 高技术通讯,2002,12(4):42-46.

Analysis of Synonymous Codon Usage Bias in Maize

YU Jin-hai

(Crop Breeding Institute of Heilongjiang Academy of Agricultural Sciences, Harbin, Heilongjiang 150086)

Abstract: By using the whole genome sequences of maize inbred B73, twenty-seven optimal codons of maize were identified in 59 synonymous codons and a detailed relative analysis results were obtained by means of a multivariate analysis program-CodonW. The results showed that the frequency of optimal codon (FOP) was present positive correlation with G+C content of genes(GC, r=0.780 7), G+C content at the third position of synonymous codons($GC3_S$, r=0.822 3), codon adaptation index(CAI, r=0.731 4)of species and Codon Bias Index (CBI, r=0.9863). But had negative correlation with effective number of codons(ENc, r= −0.6815). The results indicated that the codon usage of maize genes influenced mostly by GC content of genes with higher expression showed more significant codon usage bias. Twenty-seven codons identified firstly as optimal codons in B73 may provide some more useful information for maize gene-transformation.

Key words: maize; synonymous codon usage bias; optimal codons; GC content