

DOI: 10.3724/SP.J.1006.2011.00965

大豆基因组和转录组的核基因密码子使用偏好性分析

张 乐¹ 金龙国¹ 罗 玲¹ 王跃平¹ 董志敏¹ 孙守红² 邱丽娟^{1,*}

¹ 国家农作物基因资源与遗传改良重大科学工程 / 农业部作物种质资源利用重点开放实验室 / 中国农业科学院作物科学研究所, 北京 100081; ² 中国科学院遗传与发育研究所, 北京 100101

摘 要: 研究大豆核基因密码子的使用模式, 探讨影响其密码子组成和编码特点的因素, 为运用基因工程技术提高改良大豆提供理论依据。以大豆基因组的 46 430 个高置信编码基因和 2 071 条大豆全长转录本序列为数据来源, 应用 CodonW 软件对大豆全基因组密码子组成、同义密码子使用频率和全长转录组编码区密码子使用各项参数的计算和统计分析发现, 基因的表达水平与编码区 G+C 和 GC3s 含量均呈极显著正相关, 且 G+C 和 GC3s 含量越高的基因密码子使用偏好性越高, 并确定了 UCC 和 GCC 为大豆最优密码子。编码区长度分组分析表明, 密码子使用偏好性随编码区长度的增加而降低, 编码区较长的基因则趋向于随机使用密码子, 且在转录组数据范围内, 编码区长度介于 400~600 bp 的基因表达水平最高。大豆叶片和种子中特异表达基因的密码子使用偏好性和基因表达水平较为接近, 但种子特异表达基因的 G+C 和 GC3s 含量均显著高于叶片特异表达基因, 而其芳香族氨基酸含量则极显著低于叶片特异表达基因。

关键词: 大豆; 基因组; 转录组; 密码子

Analysis of Nuclear Gene Codon Bias on Soybean Genome and Transcriptome

ZHANG Le¹, JIN Long-Guo¹, LUO Ling¹, WANG Yue-Ping¹, DONG Zhi-Min¹, SUN Shou-Hong², and QIU Li-Juan^{1,*}

¹ National Key Facility for Crop Gene Resources and Genetic Improvement / Key Laboratory of Germplasm Utilization, Ministry of Agriculture / Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China; ² Chinese Academy of Sciences, Beijing 100101, China

Abstract: Research of soybean nuclear gene codon composition, usage pattern and influencing factors can provide theoretical basis for applying genetic engineering technology to improve soybean varieties. A total of 46 430 high confidence coding sequences predicted from soybean genome and 2 071 full-length transcripts were used to analyze the composition and characteristics of soybean nuclear gene codons. CodonW software was applied to calculate the nucleotide composition, relative synonymous codon usage and other parameters of soybean genome and transcriptome. The result indicted that gene expression level was significantly and positively correlated with G+C and GC3s contents, and genes with high G+C and GC3s contents had high codon preference. UCC and GCC were identified as optimal codons in soybean. Analysis of coding sequences with different length showed that codon preference reduced as the coding sequence (CDS) length increased, and longer CDS tend to select codons randomly. CDS length between 400 to 600 bp had the highest expression level among the transcriptome data. The preference and expression level were almost the same between leaf-specific and seed-specific genes. But seed-specific genes had significantly higher G+C and GC3s contents than leaf-specific genes, and the contents of aromatic amino acids encoded by seed-specific genes were highly significantly lower than these by leaf-specific genes.

Keywords: Soybean; Genome; Transcriptome; Codon

密码子是生命信息传递的基本单位, 编码同一种氨基酸同义密码子的不均衡使用称为密码子的使用偏好性, 其中被优先使用的某些密码子称为最优

密码子。生物体对同义密码子的选择不仅在基因表达水平方面发挥重要调节作用^[1], 而且有利于提高翻译的准确性和效率^[2]。已有研究发现影响密码子

本研究由国家高技术研究发展计划(863 计划)项目(2006AA10A110), 国家自然科学基金项目(30871621)和国家转基因生物新品种培育科技重大专项重点课题(2009ZX08009-088B)资助。

* 通讯作者(Corresponding author): 邱丽娟, E-mail: qiu_lijuan@263.net

第一作者联系方式: E-mail: zhangle_caas@yahoo.cn

Received(收稿日期): 2010-12-23; Accepted(接受日期): 2011-03-28.

使用模式的因素主要有细胞内同功 tRNA 丰度^[3-4]、基因组和 CDS (coding sequence)的 GC 含量^[5-7]、突变偏好性^[8-9]、基因在染色体上的位置^[10]、基因长度^[11]、氨基酸疏水性^[12]、蛋白质二级结构^[13]等。密码子的使用受到这些因素的选择压力,通常被认为是物种间发生分离,进而产生新物种的重要原因之一^[14-15]。

具有高油、高蛋白特点的大豆(*Glycine max* L. Merr.), 是世界上主要农作物之一,广泛用作人类食物、动物饲料和生物能源等。大豆基因组大小约 1 115 Mb^[16], 目前已完成并发表基因组序列,从中预测出具有高置信编码区域的基因共 46 430 个^[17], 为在全基因组水平对大豆核基因密码子的使用特性进行研究创造了条件。

本研究从基因组和转录组 2 个水平,解析大豆基因组的密码子组成,探讨大豆核基因的密码子使用偏好性和影响因素及叶片、种子特异表达基因的编码特点,为根据大豆基因特点优化和改造外源基因密码子,从而有效抑制转基因沉默现象^[18-19]和克服大豆遗传转化效率低的障碍,利用基因工程技术改良大豆提供理论依据。

1 材料与方法

1.1 数据的获取

基因组水平的基因数据是从大豆基因组数据库 Phytozome (<http://www.phytozome.net/soybean>)批量下载(batch download)的大豆全基因组的高置信蛋白编码基因序列,以及叶片和种子的特异表达基因序列。根据大豆基因组数据库对已知大豆基因功能的注释和未知基因通过同源比对等方法的功能预测,

选取大豆叶片中特异表达的 58 个基因和大豆种子中特异表达的 21 个基因为研究对象(表 1 和表 2),分别以这两类基因构建子数据集。

转录组水平的基因数据来自本研究。以我国大面积推广的栽培大豆品种绥农 14 为材料,构建了正常叶片、大豆花叶病毒侵染叶片、鼓粒期种子和混合组织(根、茎、叶、花、荚和种子)等 4 个全长 cDNA 文库,从中共鉴定出 2 071 条包含完整编码信息的大豆全长转录本^[20-21]。这些全长转录本被定位于大豆基因组的 20 条染色体上。

1.2 参数统计方法

利用 CodonW 1.4.2 (<http://codonw.sourceforge.net/>)对大豆基因组预测出的 46 430 个高置信蛋白编码序列和转录组 2 071 个全长基因序列,分别进行密码子组成和使用偏好性分析。密码子组成的度量指标包括 G+C 含量(鸟嘌呤和胞嘧啶含量); A3s、T3s、G3s、C3s (同义密码子在第 3 位上腺嘌呤、胸腺嘧啶、鸟嘌呤和胞嘧啶的出现频率); GC3s (同义密码子第 3 位的 G+C 含量); Aro (芳香族氨基酸频率)等。密码子使用偏好性的度量指标包括 RSCU (同义密码子相对使用度)、ENC (有效密码子数)、CAI (密码子适应指数)等。

同义密码子相对使用度(relative synonymous codon usage, RSCU)是对 59 个同义密码子(不包括 3 个终止密码子 TAG、TGG、TGA 和仅由一个密码子编码的甲硫氨酸 ATG 及色氨酸 TGG 密码子)的使用偏好性评估。该值等于同义密码子的实际观测值与同义密码子平均使用期望值的比值^[22-23]。如果密码子使用无偏好性,则 RSCU 值为 1; 如果该密码子比

表 1 叶片特异表达基因信息
Table 1 Information of leaf-specific genes

基因名称 Gene name	染色体 Chr.	功能注释 Function annotation
Glyma02g29170	2	Protein phosphatase 2C
Glyma09g17060	9	Protein phosphatase 2C
Glyma10g38910	10	Chlorophyll A-B binding protein
Glyma20g28890	20	Chlorophyll A-B binding protein
Glyma10g38910	10	Chlorophyll A-B binding protein
Glyma20g28890	20	Chlorophyll A-B binding protein
Glyma04g04450	4	Protein of unknown function (DUF3223)
Glyma14g08560	14	CYTOCHROME P450
Glyma17g36530	17	Glyma17g36530.1
Glyma05g01770	5	Aldehyde dehydrogenase family
Glyma11g09920	11	Domain of unknown function (DUF3411)
Glyma12g02260	12	Domain of unknown function (DUF3411)

(续表 1)

基因名称 Gene name	染色体 Chr.	功能注释 Function annotation
Glyma04g04160	4	Protein of unknown function (DUF3223)
Glyma06g04340	6	Protein of unknown function (DUF3223)
Glyma10g34150	10	Lactate/malate dehydrogenase, alpha/beta C-terminal domain
Glyma20g33380	20	Lactate/malate dehydrogenase, alpha/beta C-terminal domain
Glyma01g22610	1	ARSENICAL PUMP-DRIVING ATPASE-RELATED
Glyma02g11250	2	ARSENICAL PUMP-DRIVING ATPASE-RELATED
Glyma11g11870	11	Fructose-bisphosphate aldolase class-I
Glyma11g11900	11	Fructose-bisphosphate aldolase class-I
Glyma12g04150	12	Fructose-bisphosphate aldolase class-I
Glyma13g41370	13	Surface antigen
Glyma15g04030	15	Surface antigen
Glyma04g16980	4	Ankyrin repeat/zinc finger, C3HC4 type (RING finger)/protein tyrosine kinase
Glyma11g25680	11	Ankyrin repeat
Glyma09g41380	9	3-oxoacyl-[acyl-carrier-protein (ACP)] synthase III
Glyma15g00550	15	3-oxoacyl-[acyl-carrier-protein (ACP)] synthase III
Glyma18g44350	18	3-oxoacyl-[acyl-carrier-protein (ACP)] synthase III
Glyma01g22620	1	FORMIN-RELATED
Glyma02g11260	2	FORMIN-RELATED
Glyma09g06490	9	Ribosomal protein S12
Glyma01g38750	1	Manganese-stabilising protein/photosystem II polypeptide
Glyma02g06830	2	Manganese-stabilising protein/photosystem II polypeptide
Glyma11g06510	11	Manganese-stabilising protein/photosystem II polypeptide
Glyma16g25860	16	Manganese-stabilising protein/photosystem II polypeptide
Glyma05g30600	5	Photosystem II protein Y (PsbY)
Glyma08g13790	8	Photosystem II protein Y (PsbY)
Glyma11g37580	11	Photosystem II protein Y (PsbY)
Glyma18g01530	18	Photosystem II protein Y (PsbY)
Glyma03g00600	3	GTPase of unknown function
Glyma19g30210	19	GTPase of unknown function
Glyma13g03430	13	Rieske [2Fe-2S] domain
Glyma14g23860	14	Rieske [2Fe-2S] domain
Glyma02g16210	2	Methyltransferase domain
Glyma10g03590	10	Methyltransferase domain
Glyma05g28900	5	Rieske [2Fe-2S] domain
Glyma08g12070	8	Rieske [2Fe-2S] domain
Glyma05g29990	5	Domain of unknown function (DUF3406)/AIG1 family/GTPase of unknown function
Glyma08g13100	8	Domain of unknown function (DUF3406)/AIG1 family/GTPase of unknown function
Glyma07g03000	7	Ribosomal protein S15
Glyma08g23110	8	Ribosomal protein S15
Glyma05g32900	5	Pyridine nucleotide-disulphide oxidoreductase/pyridine nucleotide-disulphide oxidoreductase
Glyma08g00520	8	NADH DEHYDROGENASE-RELATED
Glyma11g16260	11	Domain of unknown function (DUF3411)
Glyma13g23110	13	Domain of unknown function (DUF3406)/AIG1 family/GTPase of unknown function
Glyma17g11770	17	Domain of unknown function (DUF3406)/AIG1 family/GTPase of unknown function
Glyma13g28180	13	Glutamine synthetase, catalytic domain/glutamine synthetase, beta-grasp domain

表 2 种子特异表达基因信息
Table 2 Information of seed-specific genes

基因名称 Gene name	染色体 Chr.	功能注释 Function annotation
Glyma11g16090	11	Seed maturation protein
Glyma10g04280	10	Cupin
Glyma13g18450	13	Cupin
Glyma09g28700	9	Bowman-Birk serine protease inhibitor family
Glyma09g28720	9	Bowman-Birk serine protease inhibitor family
Glyma09g28730	9	Bowman-Birk serine protease inhibitor family
Glyma09g39630	9	Bowman-Birk serine protease inhibitor family
Glyma09g39640	9	Bowman-Birk serine protease inhibitor family
Glyma14g26400	14	Bowman-Birk serine protease inhibitor family
Glyma14g26410	14	Bowman-Birk serine protease inhibitor family
Glyma16g33400	16	Bowman-Birk serine protease inhibitor family
Glyma18g46550	18	Bowman-Birk serine protease inhibitor family
Glyma18g46560	18	Bowman-Birk serine protease inhibitor family
Glyma18g46580	18	Bowman-Birk serine protease inhibitor family
Glyma01g29480	1	Small hydrophilic plant seed protein
Glyma03g07470	3	Small hydrophilic plant seed protein
Glyma10g03310	10	Seed maturation protein
Glyma10g39270	10	Seed maturation protein
Glyma20g28550	20	Seed maturation protein
Glyma05g22030	5	Late embryogenesis abundant protein 18

其他同义密码子使用更频繁，则其 RSCU 值大于 1；反之亦然^[24]。

有效密码子数(effective number of codons, ENC or Nc)是检测单个基因密码子非均衡使用的偏好程度，取值范围为 20 (每个氨基酸只使用一个密码子)到 61 (各个密码子被均衡使用)。其值越低，表明该基因的密码子使用偏好性越强^[25-26]。

密码子适应指数(codon adaption index, CAI)是对大豆基因编码区同义密码子与密码子最佳使用相符合程度的检测，其取值范围在 0~1 之间。表达量较高的基因具有较高的 CAI 值，表达量较低的基因具有较低的 CAI 值。CAI 值与基因表达水平的实际观测值非常接近，并已广泛应用于基因表达水平的预测^[27-28]。

最优密码子(optimal codon)是指被优先使用的密码子，其确定方法为计算所有基因的 ENC 值，并对这些值进行排序，取该有序数据集的上下限区域各 5%的序列数据，形成 2 个新的数据子集。比较 2 个数据子集中密码子的 RSCU 值，如果差异大于 0.3，且该密码子的 RSCU 值在高表达基因样本中大于 1，在低表达基因样本中小于 1，则将该密码子定义为

“最优”密码子^[29]。

最优密码子使用频率(frequency of optimal codons, FOP)指所使用的最优密码子占总密码子数的百分比。

运用 SPSS16.0 统计软件，绘制 ENC 与 GC3s 的关系图，并进行密码子组成和使用偏好性各参数(A3s、T3s、G3s、C3s、CAI、Fop、ENC、GC3s、G+C 含量和 Aro)间相关性的分析，以及叶片和种子特异表达基因密码子组成和使用偏好性各参数的差异显著性分析。

2 结果与分析

2.1 基因组和转录组的密码子组成和使用性参数比较

大豆基因组 46 430 个高置信蛋白编码基因 CDS 区序列的平均长度为 1 262.5 bp；G+C 含量范围为 24.0%~69.9%，平均为 44.5%，CAI 值以大豆核糖体核蛋白(RNP)基因为参考基因进行计算(表 3)。大豆转录组 2 071 个全长基因序列的平均长度为 887.5 bp；所包含基因的 G+C 含量的变化范围为 23.8%~67.0%，平均为 47.5%。由表 3 可知，转录组基因的同义密码

表 3 基于大豆基因组和转录组全长基因密码子的组成及使用参数
Table 3 Composition and parameters of codon usage in predicted genes from genome and identified gene from full-length cDNA transcriptome

密码子组成和使用性参数 Composition and usage parameters of codons	预测的高置信蛋白编码基因 Predicted genes		本实验鉴定的全长基因 Identified genes	
	变化范围	平均数±标准差	变化范围	平均数±标准差
	Variation range	$\bar{x} \pm SD$	Variation range	$\bar{x} \pm SD$
T3s	0.025–0.70	0.40±0.08	0.07–0.70	0.38±0.09
C3s	0.02–0.86	0.26±0.09	0.02–0.81	0.31±0.11
A3s	0.02–0.68	0.32±0.07	0.01–0.65	0.27±0.91
G3s	0–0.77	0.27±0.07	0–0.90	0.29±0.09
CAI	0.03–0.35	0.12±0.02	0.03–0.28	0.12±0.35
Fop	0.23–0.69	0.44±0.04	0.25–0.65	0.46±0.06
ENC	22.86–61.0	51.38±4.41	23.23–61.0	54.94±0.55
GC3s	0.09–0.91	0.41±0.10	0.14–0.89	0.47±0.12
G+C	0.24–0.69	0.45±0.05	0.24–0.67	0.48±0.51
Aro(%)	0–0.27	0.09±0.03	0–0.31	0.08±0.04

子第 3 位鸟嘌呤频率 G3s 和芳香族氨基酸含量这两个参数的变化范围大于基因组基因, 其余各参数变化范围都位于大豆基因组 46 430 个高置信蛋白编码区密码子使用参数的取值范围内。这从基因组和转录组 2 个水平验证了本研究所分析参数的正确性。

值得指出的是, 基于大豆基因组基因计算的 CAI 值大于 0.3 的基因共有 6 个, 分别为 *Glyma04g40720*、*Glyma06g05640*、*Glyma16g28590*、*Glyma20g29560*、*Glyma13g35050* 和 *Glyma12g35430*。其中, *Glyma04g40720*、*Glyma06g05640*、*Glyma13g35050* 和 *Glyma12g35430* 这 4 个基因目前功能未知; *Glyma16g28590* 基因是位于细胞壁中作为细胞壁组成成分的一种富含脯氨酸的伸展蛋白; *Glyma20g29560* 是种子贮藏蛋白的一种, 属于 LTP (lipid transfer protein, LTP) 基因家族, 具有蛋白酶抑制剂活性。后两个参与机体基本生长和代谢活动的功能基因具有较高的 CAI 值, 证明了应用 CAI 值估计未知基因密码子使用偏好性程度的可行性^[30]。

2.2 密码子使用参数的相关性分析

ENC 与 GC3s 描绘散点图(ENC-plot)的连续曲线, 反映了无选择压力条件下二者之间的关系(图 1), 而大多数基因位点的分布偏离期望曲线, 表明除核苷酸组成偏好外, 自然选择等其他因素对密码子的使用也具有一定影响。位于曲线下方的基因, 具有较高的 GC3s 含量, 趋向于使用较少的密码子(ENC 偏低), 具有较强的密码子使用偏好性; 而曲线上方的基因则倾向于随机使用密码子。

相关分析表明(表 4), 基因表达水平(CAI 值)与

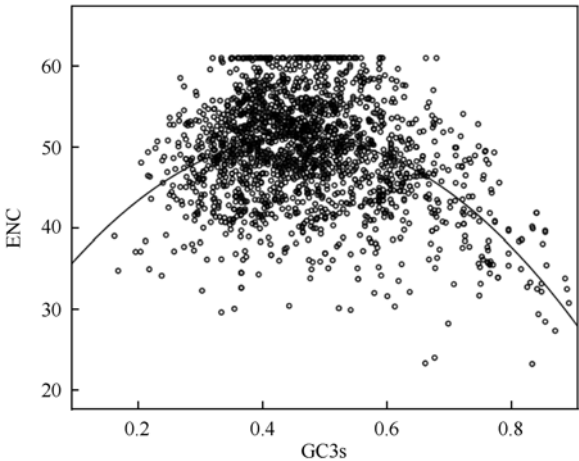


图 1 大豆密码子的 ENC-plot 曲线
Fig. 1 ENC-plot of soybean codons

同义密码子第 3 位碱基含量(T3s 和 C3s)、G+C 含量、GC3s、Fop 和芳香族氨基酸含量均呈极显著正相关($P<0.01$), 与同义密码子第 3 位碱基含量(A3s 和 G3s)和 ENC 均呈极显著负相关($P<0.01$)。有效密码子数(ENC)与同义密码子第 3 位碱基含量(C3s)、G+C 含量、GC3、Fop 均呈极显著负相关($P<0.01$), 与同义密码子第 3 位碱基含量(A3s 和 T3s)呈极显著正相关($P<0.01$), 与芳香族氨基酸含量呈显著正相关($P<0.5$)。由各参数间的相关性关系可知, 同义密码子第 3 位的碱基含量直接影响着基因的表达水平和密码子使用偏好性程度的大小。密码子使用偏好性越强(ENC 值较小)的基因, 越偏好于使用最优密码子(FOP 值较大)和 G+C 含量高尤其是以 G/C 结尾(GC3s 较大)的密码子, 同时其表达量(CAI 值)也越高。

表 4 转录组全长基因密码子使用性各参数之间的相关系数
Table 4 Correlation coefficients among the parameters of codon usage in full-length transcriptome

参数 Parameter	T3s	C3s	A3s	G3s	CAI	Fop	ENC	GC3s	G+C
C3s	-0.68**								
A3s	0.29**	-0.65**							
G3s	-0.31**	0.13**	-0.40**						
CAI	0.28**	0.12**	-0.07**	-0.18**					
Fop	-0.06**	0.35**	-0.31**	-0.02	0.81**				
ENC	0.07**	-0.07**	0.20**	-0.03	-0.08**	-0.11**			
GC3s	-0.76**	0.84**	-0.77**	0.56**	0.06**	0.25**	-0.12**		
G+C	-0.61**	0.66**	-0.68**	0.25**	0.18**	0.14**	-0.14**	0.74**	
Aro	0	0.11**	0.02	-0.11**	0.08**	0.02	0.04*	0	-0.17**

* 和 ** 分别表示各参数间相关性达 0.05 和 0.01 概率的显著。

* and ** indicate the significance of correlation at the 0.05 and 0.01 probability levels, respectively.

按照 ENC 值大小排序, 选取上限区域 5% 的 104 个基因编码区序列和下限区域 5% 的 104 个基因编码区序列为子数据集, 分别对其 RSCU 值进行计算和比较, 选择两者差异大于 0.3, 且 RSCU 值在高表达基因样本中大于 1 和低表达基因样本中小于 1 的密码子(图 2 中用“×”表示), 最终确定了 UCC 和 GCC 两个密码子为大豆基因组表达的“最优”密码子, 均为以 G/C 结尾的密码子。

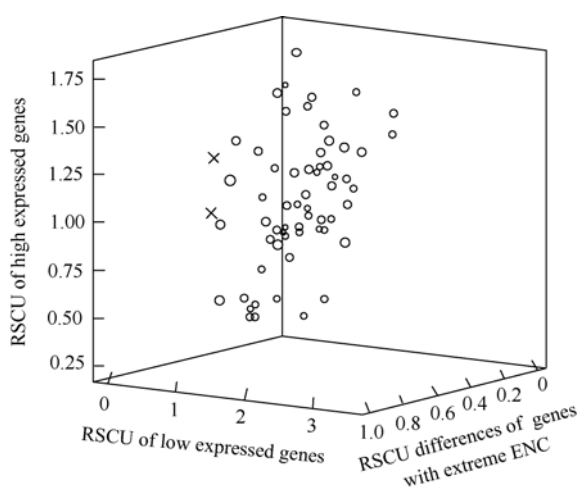


图 2 大豆最优密码子

Fig. 2 Soybean optimal codons

× 表示大豆最优密码子; ○ 表示其余的 57 个密码子。

× indicates soybean optimal codons; ○ indicates the remaining 57 codons.

编码区长度分组分析表明(表 5), 编码区长度与 ENC 间的关系基本是编码区长度越长, ENC 值越大, 即密码子使用的偏好性随编码区长度的增加而降低, 编码区较长的基因趋向于随机使用密码子。同时, 由于 CAI 和 GC3s 呈极显著正相关, 当编码区长度介于 400~600 bp 范围时, 二者均表现为 5 组数据的

最大值, 说明长度位于此范围的基因的表达式居 5 组中最高, 同时其密码子偏好于以 G/C 结尾。

2.3 组织特异性密码子分析

以大豆基因组 46 430 个高置信编码基因中, 与光合作用相关的 58 个叶片特异表达基因和与贮藏蛋白功能相关的 21 个种子特异表达基因为研究对象, 计算这两类组织特异表达基因中各密码子的组成和同义密码子相对使用度(RSCU 值), 并将其与整个基因组比较。结果表明, 叶片特异表达基因各氨基酸的同义密码子使用性与整个基因组的密码子使用特性相似, 而种子特异表达基因中 Phe、Tyr、His、Asn 和 Cys 偏好的密码子与全基因组和叶片特异表达基因的使用偏性不同, 且这 5 种氨基酸都具有 2 个同义密码子。在终止密码子的使用上, 全基因组和叶片特异表达基因偏好于使用 UGA, 而种子特异表达基因偏好于使用 UAA。对这两类组织特异表达基因的密码子组成和使用参数进行差异显著性分析发现, 种子特异表达基因的 G+C 含量和 GC3s 都显著高于叶片特异表达基因($P<0.05$), 叶片中的芳香族氨基酸含量极显著高于种子特异表达基因($P<0.01$), 而二者的 ENC 值和 CAI 值均无显著差异。这一结果说明, 叶片和种子的密码子使用偏好性和二者的基因表达水平较为接近, 但个别氨基酸偏爱使用的密码子有所不同。

3 讨论

3.1 大豆基因的实际测定与基因组预测相结合探索密码子的使用性

本研究中的大豆 46 430 个高置信基因编码序列来自 Phytozome (<http://www.phytozome.net/soybean>)

表 5 基于转录组全长基因的参数 ENC、GC3s 和 CAI 的比较
Table 5 Comparison of ENC, GC3s, and CAI among full-length genes with different lengths

分组 Group	基因长度 Length of genes (bp)	基因数目 No. of genes	平均数±标准差 $\bar{x} \pm SD$		
			ENC	GC3s	CAI
1	<200	95	45.50±9.286	0.482±0.123	0.105±0.034
2	200 and <400	884	48.85±6.882	0.387±0.116	0.120±0.036
3	400 and <600	812	51.16±6.178	0.484±0.124	0.141±0.034
4	600 and <800	244	50.98±4.439	0.453±0.103	0.129±0.028
5	800	36	52.12±3.823	0.474±0.856	0.132±0.278

表 6 大豆 59 种同义密码子在全基因组和叶片、种子特异表达基因中的使用频率
Table 4 Frequency of 59 synonymous codons in soybean genome genes, leaf-specific genes and seed-specific genes

氨基酸 Amino acid	密码子 Codon	全基因组 Genome		组织特异表达基因 Tissue specific expressed gene RSCU		氨基酸 Amino acid	密码子 Codon	全基因组 Genome		组织特异表达基因 Tissue specific expressed gene RSCU			
		数目 No.	RSCU	叶片 Leaf	种子 Seed			数目 No.	RSCU	叶片 Leaf	种子 Seed		
Phe	<u>UUU</u>	483265	1.15	1.06	0.58	Ser	<u>UCU</u>	444640	1.55	1.60	0.90		
	UUC	359232	0.85	0.94	1.42		UCC*	264218	0.92	1.10	1.20		
Leu	UUA	243950	0.76	0.65	0.27		<u>UCA</u>	392303	1.37	1.40	1.40		
	<u>UUG</u>	492148	1.54	1.55	1.66		UCG	105724	0.37	0.40	0.30		
	<u>CUU</u>	473815	1.48	1.46	1.40	Pro	<u>CCU</u>	360311	1.51	1.50	1.40		
	CUC	283483	0.88	0.95	1.50		CCC	170513	0.71	0.70	1.20		
	CUA	193081	0.60	0.63	0.39		<u>CCA</u>	334799	1.40	1.40	1.20		
	CUG	237084	0.74	0.76	0.77		CCG	91607	0.38	0.40	0.20		
	Ile	<u>AUU</u>	501771	1.44	1.56		1.18	Thr	<u>ACU</u>	337737	1.40	1.40	1.20
		AUC	261867	0.75	0.73		1.14		ACC	222512	0.92	1.20	1.30
AUA		282957	0.81	0.71	0.68	<u>ACA</u>	318915		1.32	1.20	1.10		
Met	<u>AUG</u>	468373	1.00	1.00	1.00		ACG	86986	0.36	0.30	0.40		
Val	<u>GUU</u>	511157	1.60	1.66	1.16	Ala	<u>GCU</u>	501373	1.56	1.60	1.10		
	GUC	204212	0.64	0.58	0.82		GCC*	256035	0.79	0.80	1.10		
	GUA	184447	0.58	0.52	0.35		<u>GCA</u>	420390	1.30	1.20	1.10		
	<u>GUG</u>	381973	1.19	1.23	1.67		GCG	111787	0.35	0.40	0.60		
Tyr	<u>UAU</u>	332610	1.19	1.15	0.98	Cys	<u>UGU</u>	204431	1.09	1.04	0.94		
	UAC	225579	0.81	0.85	1.02		UGC	169233	0.91	0.96	1.06		
TER	<u>UAA</u>	13811	1.00	0.76	1.44	TER	<u>UGA</u>	17747	1.28	1.20	0.60		
	UAG	9988	0.72	1.04	0.96		Trp	<u>UGG</u>	257147	1.00	1.00	1.00	
His	<u>CAU</u>	298424	1.21	1.18	0.69	Arg	CGU	121905	0.73	0.76	0.98		
	CAC	195560	0.79	0.82	1.31		CGC	102975	0.62	0.74	1.66		
Gln	<u>CAA</u>	411208	1.14	1.06	1.10		CGA	98611	0.59	0.50	0.57		
	C AG	310398	0.86	0.94	0.90		CGG	80714	0.48	0.47	0.35		
Asn	<u>AAU</u>	537146	1.19	1.18	0.77	Ser	<u>AGU</u>	293405	1.03	0.91	1.22		
	AAC	369067	0.81	0.82	1.23		AGC	215776	0.75	0.67	0.90		
Lys	AAA	568448	0.95	0.90	0.91	Arg	<u>AGA</u>	322929	1.93	1.87	1.00		
	<u>AGG</u>	29842	1.05	1.10	1.09		<u>AGG</u>	276899	1.65	1.66	1.44		
Asp	<u>GAU</u>	680853	1.35	1.35	1.12	Gly	<u>GGU</u>	378292	1.20	1.27	1.03		
	GAC	328756	0.65	0.65	0.88		GGC	224151	0.71	0.78	0.92		
Glu	<u>GAA</u>	655540	1.07	1.00	1.07		<u>GGA</u>	401778	1.28	1.17	1.30		
	GAG	571921	0.93	1.00	0.93		GGG	251915	0.80	0.78	0.75		

下画线表示该氨基酸同义密码子中偏好的密码子; *表示最优密码子。
Codons with underlines are the preferred ones among synonymous codons of an amino acid. * indicates the optimal codon.

大豆基因组数据库,是根据大豆基因组序列经基因预测获得的,而用于探讨影响大豆密码子使用性因素和各因素间关系的大豆转录组序列是经本实验室分离鉴定测序得到的真实存在的基因。目前,对于物种基因使用性的研究大多基于已知基因组序列预测出的CDS,未经实验验证其CDS及其对应蛋白的真实性,因此本研究将基因组预测出的全部CDS序列与实验中获得的全长基因序列相结合,提高了研究的准确性和实用性。目前进行基因预测的方法主要有两大类,一类是依赖于统计模型识别基因结构的重头计算法,另一类是利用已知表达序列(如EST、cDNA或同源序列)同源比对的预测方法^[31]。前者可以预测任何类型的全新基因,但准确性较低(在外显子水平,预测准确性约80%);后者准确性较高(在外显子水平,预测准确性可达90%以上),但不能预测未知表达序列的全新基因^[31-32],本文所用的从基因组序列中预测出的编码区序列的方法主要为后者。

3.2 大豆基因编码区长度对密码子使用偏性和基因表达水平的影响

本研究通过编码区长度分组分析表明,大豆基因密码子使用偏好性随编码区长度的增加而降低,该结论与大肠杆菌^[33-34]、拟南芥^[35]、线虫^[36-37]、果蝇^[38]等模式生物种的密码子使用规律相同,但不同于人类等哺乳动物^[39]。同时,经编码区长度分组分析还发现,基因的表达水平随编码区的长度先升高,后下降,在长度为400~600 bp时表达量最高,这可能与生物处于自然选择压力下,编码区较短的基因可以减少转录和翻译中物质、能量和时间的消耗,因此其表达水平高有关,对于物种本身的进化也更有利^[40],相反,编码区较长的基因往往含有数量较多或较长的内含子,其切割和加工过程耗能较多,同时较长基因的表达也需要更多的调控因子的协助^[41],因而表达量通常较低。综上所述,大豆与其他生物的不同之处在于,其基因表达水平并非始终随编码区长度增加而下降,而是在编码区长度小于400 bp时,其表达量随编码区长度的增加略有增加。

3.3 大豆密码子使用性分析对基因工程技术的指导作用

生物界蕴藏着丰富的基因资源。将不同物种中的优良基因,应用基因工程手段转入被改良的生物,从而使其获得有利性状,提高其生存竞争及利用效率。然而,将外源基因转入受体生物时,由于转入基

因的密码子和宿主基因组密码子使用性不同,易引起甲基化^[42],从而引发转基因沉默或转基因表达量降低。因此,在外源基因转入前,应按照宿主物种的密码子使用偏好性对其进行优化和改造。如将Bt毒蛋白基因转入棉花,先将富含AT(63%)密码子的cryIA(b)和cryIA(c)型毒蛋白基因(全长1 845 bp)的21%的密码子改造成植物偏爱的同义密码子,使G+C含量从37%提高到49%,最终使外源基因在棉花中的表达量分别提高10倍和100倍^[19]。

综上所述,通过大豆全基因组和转录组以及大豆叶片和种子特异表达基因的密码子使用特点分析,对于优化和改造外源基因并导入大豆,从而使大豆获得更多有利性状(如产量提高、品质改良)并对大豆品种进行定向改良具有十分重要的理论和现实意义。

4 结论

大豆的最优密码子为UCC和GCC,编码区长度越短的基因,其密码子使用偏好性越强,并偏爱于使用最优密码子和G+C含量高尤其是以G/C结尾的密码子,且其表达水平也越高。编码区长度位于400~600 bp范围的基因的密码子使用偏好性和基因表达量在参试全长转录本中最高,推测其可能为大豆的最佳编码区长度。叶片和种子特异表达基因的密码子使用偏好性和基因表达水平均较为接近,但个别氨基酸偏爱使用的密码子有所不同。

References

- [1] Carlini D B, Stephan W. *In vivo* introduction of unpreferred synonymous codons into the *Drosophila Adh* gene results in reduced levels of ADH protein. *Genetics*, 2003, 163: 239-243
- [2] Sharp P M, Matassi G. Codon usage and genome evolution. *Curr Opin Genet Dev*, 1994, 4: 851-860
- [3] Stenico M, Lloyd A T, Sharp P M. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucl Acid Res*, 1994, 22: 2437-2446
- [4] Olejniczak M, Uhlenbeck O C. tRNA residues that have co-evolved with their anticodon to ensure uniform and accurate codon recognition. *Biochimie*, 2006, 88: 943-950
- [5] Holmouist G P, Flipske J. Organization of mutations along the genome: a prime determinant of genome evolution. *Trends Ecol Evol*, 1994, 9: 65-69
- [6] Bernard I G. The human genome: organization and evolutionary history. *Annu Rev Genet*, 1995, 29: 445-476
- [7] Shi X-F(石秀凡), Huang J-F(黄京飞), Liu S-Q(柳树群), Liu C-Q(刘次全). The feature of synonymous codon bias and GC-content relationship in human genes. *Prog Biochem Biophys*

- (生物化学与生物物理进展), 2002, 29(3): 411–414 (in Chinese with English abstract)
- [8] Xia X. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. *Gene*, 2005, 345: 13–20
- [9] Chen S L, Lee W, Hottes A K, Shapiro L, McAdams H H. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA*, 2004, 101: 3480–3485
- [10] Romero H, Zavala A, Musto H, Bernaerdi G. The influence of translational selection on codon usage in fishes from the family Cyprinidae. *Gene*, 2003, 317: 141–147
- [11] Moriyama E N, Powell J R. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucl Acids Res*, 1998, 26: 3188–3193
- [12] Knight R D, Freeland S J, Landweber L F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol*, 2001, 2: RESERCH0010
- [13] Gupta S K, Majumdar S K, Bhattacharya T, Ghosh T C. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem Biophys Res Commun*, 2000, 269: 692–696
- [14] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 1994, 11: 725–736
- [15] Schmidt W. Phylogeny reconstruction for protein sequences based on amino acid properties. *J Mol Evol*, 1995, 41: 522–530
- [16] Arumuganathan K, Earle E D. Nuclear DNA content of some important plant species. *Plant Mol Rep*, 1991, 9: 208–218
- [17] Jeremy S, Steven B C, Jessica S, Ma J X, Mitros T, Nelson W, Hyten D L, Song Q J, Thelen J J, Cheng J L, Xu D, Hellsten U, May G D, Yu Y, Sakurai T, Umezawa T, Bhattacharyya M K, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S Q, Abernathy B, Du J C, Tian Z X, Zhu L C, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X C, Shinozak K, Nguyen H T, Wing R A, Cregan P, Specht J, Crimwood J, Rokhsar D, Stavey G, Shoemaker R C, Jackson S A. Genome sequence of the palaeopolyploid soybean. *Nature*, 2010, 463: 178–183
- [18] Cao H-Y(曹慧颖), Zhang R(张锐), Guo S-D(郭三堆). High level expression of human thymosin $\alpha 1$ concatemer in transgenic tomato plants. *Sci Agric Sin* (中国农业科学), 2009, 42(7): 2291–2296 (in Chinese with English abstract)
- [19] Zou Y-M(邹永梅), Shi J-S(施季森), Zhu-Ge Q(诸葛强), Huang M-R(黄敏仁). Reacting the silencing genes in the transgenic plants. *Mol Plant Breed* (分子植物育种), 2006, 4(1): 95–102 (in Chinese with English abstract)
- [20] Dong Z-M(董志敏), Li Y-H(李英慧), Zhang B-S(张宝石), Guan R-X(关荣霞), Chang R-Z(常汝镇), Qiu L-J(邱丽娟). An improved SMART method to construction full-length cDNA library for large clones. *Soybean Sci* (大豆科学), 2006, (5): 1–4 (in Chinese with English abstract)
- [21] Wang Y-P(王跃平), Li Y-H(李英慧), Chen X-T(陈雄庭), Chang R-Z(常汝镇), Qiu L-J(邱丽娟). Construction and characterization of the filling stage's seed cDNA library from Suinong14 (*Glycine max*). *Chin J Oil Crop Sci* (中国油料作物学报), 2008, 30(1): 40–45 (in Chinese with English abstract)
- [22] Sharp P M, Haney T M F, Mosurski K R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucl Acids Res*, 1986, 14: 5125–5143
- [23] Liu Q, Feng Y, Xue Q. Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. *Mitochondrion*, 2004, 4: 313–320
- [24] Sau K, Gupta S K, Sau S, Mandal S C, Ghosh T C. Factors influencing synonymous codon and amino acid usage biases in mimivirus. *Biosystems*, 2006, 85: 107–113
- [25] Wright F. The effective number of codons used in a gene. *Gene*, 1990, 87: 23–29
- [26] Gupta S K, Bhattacharyya T K, Ghosh T C. Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *Biomol Struct Dyn*, 2004, 21: 1–9
- [27] Peixoto L, Zavala A, Romero H, Musto H. The strength of translational selection for codon usage varies in three relatives of *Sinorhizobium meliloti*. *Gene*, 2003, 320: 109–116
- [28] Romero H, Zavala A, Musto H. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucl Acids Res*, 2000, 28: 2084–2090
- [29] Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci USA*, 1999, 96: 4482–4487
- [30] Sharp P M, Li W H. The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. *Nucl Acids Res*, 1987, 15: 1281–1295
- [31] Wei C, Brent M R. Using ESTs to improve the accuracy of de novo gene prediction. *BMC Bioinformatics*, 2006, 7: 327–337
- [32] Kwan A L, Li L, Kulp D C, Dutcher S K, Stormo G D. Improving gene-finding in *Chlamydomonas reinhardtii* Green Genie2. *BMC Genomics*, 2009, 10: 210–220
- [33] Sharp P M, Cowe E, Higgins D G, Shield D C, Wolfe K H, Wright F. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity. *Nucl Acids Res*, 1988, 16: 8207–8211
- [34] Stoletski N, Eyre-Walker A. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol*, 2007, 24: 374–381
- [35] Morton B R, Wright S I. Selective constraints on codon usage of nuclear genes from *Arabidopsis thaliana*. *Mol Biol Evol*, 2007, 24: 122–129
- [36] Cutter A D, Wasmath J D, Blaxter M L. The evolution of biased codon and amino acid usage in nematode genomes. *Mol Biol Evol*, 2006, 23: 2303–2315
- [37] Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*,

- and *Arabidopsis*. *Proc Natl Acad Sci USA*, 1999, 96: 4482–4487
- [38] Vicario S, Mason C E, White K P, Powell J R. Developmental stage and level of codon usage bias in *Drosophila*. *Mol Biol Evol*, 2008, 25: 2269–2277
- [39] D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J Mol Evol*, 1991, 32: 504–510
- [40] Moriyama E N, Powell J R. Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol*, 1997, 45: 514–523
- [41] Holmquist G P, Filipinski J. Organization of mutations along the genome: a prime determinant of genome evolution. *Trends Ecol Evol*, 1994, 9: 65–69
- [42] Perlak F J, Deaton R W, Armstrong T A, Fuchs R L, Sims S R, Greenplate J T, Fischhoff D A. Insect resistant cotton plants. *Biol Technol*, 1990, 8: 939–943

关于召开“中国作物学会 2011 年学术年会”的第一轮通知

“中国作物学会学术年会”是我国作物科学领域最高级别学术交流大会，是每年我国作物科学工作者相聚的盛会。大会汇聚了作物科学领域的科技工作者，并邀请作物科学界的两院院士和著名专家作大会学术报告。自 2002 年创办学术年会以来，已成功举办了 8 次，会议质量逐年上升，会议规模和影响力日益扩大，已成为中国作物学会的品牌会议。

2011 年会议主题：生物育种产业与粮食安全；会议时间：2011 年 10 月 18 日报到，2011 年 10 月 19–20 日开会；会议地点：四川省成都市金牛宾馆。同时将举行中国作物学会作物种子专业委员会成立大会。

一、会议形式及主要内容

会议将以大会报告、分会场报告、研究生论坛及墙报的形式进行广泛的学术交流。现场评选青年优秀学术报告和优秀墙报奖并在大会闭幕式上颁奖，获奖论文将免费刊登在《作物学报》上并注明为年会优秀论文。参选者须是在读研究生或者 35 周岁以下的青年科技工作者，第一作者和通讯作者的工作单位须在国内，并在 2009 年 1 月至 2011 年 10 月之间在国内完成主要工作。

分会场学术交流的主要内容：(1)生物技术——重要基因的克隆与功能分析、作物转基因技术与转基因育种、重要基因定位与标记开发、作物分子标记育种、分子设计育种理论与应用等。(2)作物遗传育种——主要农作物种质资源遗传多样性及核心种质构建、种质资源的鉴定与评价、基于基因组学的种质资源研究、主要作物种质改良与创新、主要农作物重要性状遗传规律与育种理论和方法研究、高产优质广适性农作物新品种培育与应用等。(3)作物栽培与耕作——作物可持续高产与超高产理论与技术、作物抗逆与中低产田技术、作物高产高效生态生理、作物品质形成与优质高产技术、精准栽培与轻简化技术、节水与抗旱栽培、秸秆还田与保护性耕作等。(4)种子科学技术与产业发展——种子生物学研究进展、作物种子生产的理论与技术、作物种子加工和处理的理论与技术、作物种子与种质保存的理论与技术、种子检验的理论与技术等。

二、论文摘要、分会场学术报告和墙报征集(征集截止日期为 2011 年 8 月 31 日)

论文摘要的征集：本次会议征集未公开发表过的论文摘要，要求针对会议交流的主要内容，每篇摘要正文字数在 1000 字以内。投稿方式：通过电子邮件附件形式将征文摘要发送至中国作物学会办公室，须注明所投论文的学科分类(参见会议学术交流内容)和交流方式(如分会场学术报告或墙报交流)，不接受邮寄的打印稿。

分会场学术报告的征集：参会的代表如希望在分会场交流学术研究成果或进展，请报告题目以电子邮件传至中国作物学会办公室。

墙报的征集：由于时间限制，只能挑选部分报告在分会场交流，其余以墙报形式进行交流。大会要求每位参会代表提交一篇论文摘要(1000 字左右)并制作成墙报进行展示，尺寸标准为：90 cm×120 cm，纵向排版。要求文字务必简明扼要，图文并茂。请在报名回执表中填写论文和墙报题目。

三、报名和联系方式

报名参会人员请从中国作物学会网站下载并填写报名回执表，并以电子邮件形式传送到中国作物学会办公室。

E-mail: ccss304@sina.com, 网址: <http://www.chinacrops.org/>, 电话: 010-82108616, 传真: 010-82108785

通讯地址：北京中关村南大街 12 号中国作物学会办公室，邮编：100081，联系人：杜娟，刘丹丹