

结球甘蓝核基因组密码子使用偏爱性分析

仪登霞^{1,3}, 杨丽梅^{1,*}, 方智远¹, 刘玉梅¹, 庄木¹, 张扬勇¹, 刘基生¹,
刘博¹, 武剑¹, 耿丽丽², 张杰², 张振贤^{3,*}

(¹中国农业科学院蔬菜花卉研究所, 北京 100081; ²中国农业科学院植物保护研究所, 北京 100193; ³中国农业大学农学与生物技术学院, 北京 100193)

摘要: 根据甘蓝测序结果, 利用 CodonW 软件对结球甘蓝的 48 000 个蛋白质编码基因序列进行了分析, 计算出同义密码子相对使用频率, 确定了 TCT、CCT、AGA、GTT、GCT、GAT 等 6 个高频密码子。将甘蓝与大肠杆菌、拟南芥、棉花、水稻的密码子使用频率进行了比较, 发现甘蓝密码子使用偏爱性与同为双子叶植物的拟南芥、棉花基本一致, 而与大肠杆菌及单子叶植物水稻均具有较大差异。以甘蓝密码子用法分析结果为依据, 对 Bt *cryIC* 抗虫基因进行了密码子的改造, 得到了具有甘蓝密码子使用特点的 *cryIC* 基因序列, 为甘蓝的转基因研究奠定了分子基础。

关键词: 甘蓝; 密码子用法; 高频密码子; 密码子偏爱性

中图分类号: S 635.1

文献标志码: A

文章编号: 0513-353X (2013) 10-1927-08

Analysis of Nuclear Genome Codon Preference in Cabbage

YI Deng-xia^{1,3}, YANG Li-mei^{1,*}, FANG Zhi-yuan¹, LIU Yu-mei¹, ZHUANG Mu¹, ZHANG Yang-yong¹,
LIU Ji-sheng¹, LIU Bo¹, WU Jian¹, GENG Li-li², ZHANG Jie², and ZHANG Zhen-xian^{3,*}

(¹Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China; ²Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, China; ³College of Agriculture and Biotechnology, China Agricultural University, Beijing 100193, China)

Abstract: Forty-eight thousand coding DNA sequences of cabbage were used to analyze the relative frequency of synonymous codon (RFSC). Six high-frequency codons including TCT, CCT, AGA, GTT, GCT, and GAT were revealed by CodonW. When the frequency of codon usage of cabbage was compared to *Escherichia coli*, *Arabidopsis thaliana*, and *Oryza sativa*, we found that the codon preference was considerably different between cabbage and *Escherichia coli*, *Oryza sativa*, while the codon preference was identical between cabbage and *Arabidopsis thaliana*, *Gossypium hirsutum*. The codons of Bt *cryIC* gene sequence have been reconstructed according to the results. The foreign gene containing the optimal codons of cabbage is helpful to researches involving cabbage transformation.

Key words: cabbage; codon usage; high-frequency condon; condon preference

生物的遗传密码具有简并性, 每一种氨基酸可由 1~6 个密码子编码, 对于不同物种间, 由于

收稿日期: 2013-05-03; 修回日期: 2013-06-14

基金项目: 国家自然科学基金项目 (31071697); 国家‘863’计划项目 (2008AA10Z105); 国家大宗蔬菜产业技术体系项目 (nycyt-x-35-gw01); ‘十二五’国家科技支撑计划项目 (2012BAD02B01); 农业部园艺作物生物学与种质创制重点实验室项目

* 通信作者 Author for correspondence (E-mail: yanglimei@caas.cn; zhangzx@cau.edu.cn)

tRNA 丰度、(G + C) % 和 wobble 摆动效应等因素的差异, 致使在密码子的使用上存在着偏爱性, 这种现象会对克隆的外源基因在宿主生物细胞中的表达造成影响 (Bennetzen & Hall, 1982; Ikemura, 1985)。因此, 有必要对一些常用生物的密码子偏爱性加以分析, 找出其高频使用密码子, 为改造外源基因提供依据, 从而实现外源基因的高效表达。

在结球甘蓝 (*Brassica oleracea* L. var. *capitata* L.) 中抗虫种质资源较为缺乏, 通过常规育种途径很难获得抗虫性强的品种。许多研究者致力于采用转基因技术将外源抗虫基因转入甘蓝, 获得了在 DNA 水平、RNA 水平、蛋白质水平均表达的转基因植株, 且抗虫效果明显 (Metz et al., 1995; Jin et al., 2000; Bhattacharya et al., 2002; 李汉霞 等, 2006; 崔磊 等, 2009; Rafat et al., 2010; Yi et al., 2011)。但有些外源基因在转基因植株中的表达情况不尽如人意, 造成这种结果的原因有很多, 外源基因使用的密码子不是甘蓝偏爱密码子可能是重要原因之一 (Matzke et al., 2000)。

日本 Kazusa DNA 研究所根据 NCBI-GenBank 中的基因资源建立了密码子用法数据库 (<http://www.kazusa.or.jp/codon/>), 该数据库首次以密码子使用频率为参数列出了甘蓝的密码子使用表。但是由于用于分析的样本数较少 (该数据库仅对结球甘蓝的 18 个蛋白质编码基因序列的 7 441 个密码子进行了分析), 所以得到的密码子使用频率不能很好地反映真实情况。目前, 结球甘蓝全基因组测序已经完成 (<http://brassicadb.org/brad/>), 为系统研究甘蓝基因组密码子的偏爱性提供了便利。本研究中利用 CodonW 软件对结球甘蓝的全基因组序列进行了分析, 首次全面地明确了结球甘蓝的高频使用密码子, 并与模式生物的偏好密码子进行了比较, 选用甘蓝偏爱密码子对 *Bt cryIC* 抗虫基因进行了针对性改造, 为实现其在甘蓝中的高效稳定表达奠定基础。

1 材料与方法

1.1 甘蓝基因的获取

根据甘蓝测序结果, 获取甘蓝基因组序列 (<http://brassicadb.org/brad/>)。目的基因序列的筛选原则为: (1) 编码蛋白质的基因, 并且编码序列完整; (2) 编码基因的序列长度大于 300 bp; (3) 在细胞质中翻译, 排除细胞器中翻译的基因; (4) 排除位于质粒、转座子、细菌、真菌和病毒上的基因; (5) 对多拷贝基因只统计一次 (Paul & Elizabeth, 1991)。

1.2 甘蓝全基因组密码子的使用分析

首先利用 CodonW 软件对筛选得到的甘蓝基因序列进行密码子使用频率进行分析: (1) 将软件 CodonW 解压到某一指定文件夹, 把待分析的甘蓝序列文件 (需为 FASTA 格式) 放在此文件夹内; (2) 运行 CodonW 软件程序, 依次选择 4 (codon usage indices), 12 (select all), 再选择 x 返回主菜单。此时指定文件夹中就含有甘蓝基因组序列的密码子使用频率的文本。其次计算甘蓝蛋白质编码基因序列的同义密码子相对使用频率 (relative frequency of synonymous codon, RFSC)。RFSC 反映的是样本中各个同义密码子的使用频率。计算式为某一密码子在样本中的实际观察数与该密码子对应的氨基酸在样本中的实际观察数的比值。它避免了由样本选择中各基因的长短与氨基酸丰度所造成的误差。最后根据 RFSC 的计算结果, 利用高频密码子分析法 (林涛 等, 2002) 来筛选甘蓝基因的高频使用密码子。若某一密码子相对同义密码子使用频率单值超过 60% 或者超过该组同义密码子平均占有频率的 1.5 倍的密码子即为高频密码子 (High-frequency codon, HF)。

1.3 甘蓝和其他生物密码子偏爱性比较

用 1/1 000 (即在 1 000 个密码子中出现的次数) 代表密码子在编码基因总密码子中出现的频率。

统计甘蓝各密码子的 1/1 000 并与从日本 Kazusa DNA 研究所的密码子用法数据库 (<http://www.kazusa.or.jp/codon/>), 截止到 2013 年 2 月 4 日该数据库已对大肠杆菌的 8 087 个基因共 2 330 943 个密码子, 拟南芥的 80 395 个基因共 31 098 475 个密码子, 棉花的 557 个基因共 190 383 个密码子、水稻的 92 188 个基因共 34 132 283 个密码子进行了分析) 中的大肠杆菌、拟南芥、棉花、水稻等几种代表性生物的数据进行了比较。计算甘蓝各密码子分别与其他 4 种生物的比值, 如比值在 0.5 ~ 2.0 之间, 表示两物种对该密码子的偏爱性较为接近, 反之则差别较大。

1.4 Bt *cryIC* 基因的密码子改造

以筛选出的甘蓝高频密码子为依据, 对 Bt *cryIC* 抗虫基因进行密码子的改造。

2 结果与分析

2.1 甘蓝高频密码子 (HF) 的确定

结合甘蓝测序结果, 获得甘蓝基因组序列, 经筛选得到 48 000 个基因共 12 242 981 个密码子。利用 CodonW 软件对其进行密码子 RFSC 分析 (表 1), 共筛选出高频密码子 6 个, 它们是 TCT (编码 Ser)、CCT (编码 Pro)、AGA (编码 Arg)、GTT (编码 Val)、GCT (编码 Ala)、GAT (编码 Asp)。

表 1 甘蓝蛋白质编码序列的同义密码子的相对使用频率和高频密码子

Table 1 The relative frequency of synonymous codon (RFSC) and high-frequency codon (HF) of coding sequence of cabbage

氨基酸 Amino acid	密码子 Codon	密码子数 No. of codon	相对使用频率/% RFSC	高频密码子 HF
Phe	TTT	237 581	47.00	
	TTC	268 987	53.00	
Ser	TCT	293 910	26.50	HF
	TCC	161 134	14.50	
	TCA	212 604	19.17	
	TCG	119 546	10.67	
	AGT	161 757	14.50	
	AGC	163 271	14.67	
Tyr	TAT	153 630	46.00	
	TAC	182 040	54.00	
Cys	TGT	117 146	54.00	
	TGC	99 914	46.00	
Leu	TTA	132 165	11.50	
	TTG	244 026	21.30	
	CTT	272 898	23.80	
	CTC	227 660	19.80	
	CTA	120 892	10.50	
	CTG	146 777	12.80	
Trp	TGG	152 195	1.00	
Pro	CCT	232 931	38.50	HF
	CCC	77 873	12.75	
	CCA	185 439	30.50	
	CCG	110 599	18.25	
His	CAT	157 515	56.50	
	CAC	121 661	43.50	
Gln	CAA	218 928	51.50	
	CAG	208 139	48.50	
Arg	CGT	104 386	15.50	
	CGC	57 463	8.50	
	CGA	67 126	10.00	
	CGG	58 667	8.67	
	AGA	229 695	34.17	HF
	AGG	155 324	23.17	

续表 1

氨基酸 Amino acid	密码子 Codon	密码子数 No. of codon	相对使用频率/% RFSC	高频密码子 HF
Ile	ATT	218 039	36.00	
	ATC	237 124	39.30	
	ATA	149 036	24.67	
Met	ATG	305 573	100.0	
Thr	ACT	199 720	31.50	
	ACC	138 903	22.00	
	ACA	181 411	28.75	
Asn	ACG	112 361	17.75	
	AAT	221 968	43.50	
	AAC	290 075	56.50	
Lys	AAA	340 515	44.50	
	AAG	424 803	55.50	
Val	GTT	313 341	37.50	HF
	GTC	176 420	21.00	
	GTA	111 238	13.25	
	GTG	234 172	28.00	
Ala	GCT	340 053	42.00	HF
	GCC	141 795	17.50	
	GCA	201 974	25.00	
	GCG	127 573	15.75	
Asp	GAT	410 694	62.50	HF
	GAC	244 067	37.50	
Glu	GAA	385 795	46.50	
	GAG	440 590	53.50	
Gly	GGT	253 031	31.50	
	GGC	129 757	16.00	
	GGA	280 543	34.75	
	GGG	143 131	17.75	
Ter	TAA	12 228	34.67	
	TAG	7 720	21.67	
	TGA	15 452	43.67	

2.2 甘蓝与其他生物的密码子偏爱性比较

根据表 1 数据统计甘蓝各密码子的 1/1 000 数据, 并与日本 Kazusa DNA 研究所密码子用法数据库中的大肠杆菌、拟南芥、棉花和水稻等的数据比较(表 2)。结果显示: 甘蓝与同为双子叶植物的拟南芥、棉花各密码子的使用频率比值均在 0.50 ~ 2.00, 说明其密码子偏爱性基本相同; 而甘蓝与原核生物大肠杆菌的密码子使用频率比值中有 7 种密码子的比值小于等于 0.50 或者大于等于 2.00, 与单子叶植物水稻的也有 6 种, 表明甘蓝的密码子偏爱性与大肠杆菌和水稻具有不同程度的差异。

表 2 甘蓝与其他生物密码子偏爱性比较
Table 2 Comparison of codon preference between cabbage and other organisms

氨基酸 Amino acid	密码子 Codon	密码子使用频率/1/1 000 Frequency of codon usage					密码子偏爱性比较 Comparison of codon preference			
		甘蓝 Cabbage (C)	大肠杆菌 <i>E. coli</i> (E)	拟南芥 <i>A. thaliana</i> (A)	棉花 <i>G. hirsutum</i> (G)	水稻 <i>O. sativa</i> (O)	C/E	C/A	C/G	C/O
Phe	TTT	19.4	24.4	21.8	20.4	13.1	0.80	0.89	0.95	1.48
	TTC	22.0	13.9	20.7	22.2	22.4	1.58	1.06	0.99	0.98
Ser	TCT	24.0	13.1	25.2	17.5	12.7	1.83	0.95	1.37	1.89
	TCC	13.2	9.7	11.2	12.9	16.3	1.36	1.18	1.02	0.81
	TCA	17.4	13.1	18.3	15.7	12.4	1.33	0.95	1.11	1.40
	TCG	9.8	8.2	9.3	6.7	12.3	1.20	1.05	1.46	0.80
	AGT	13.2	13.2	14.0	11.8	8.8	1.00	0.94	1.12	1.50
	AGC	13.3	14.3	11.3	12.3	16.0	0.93	1.18	1.08	0.83

续表 2

氨基酸 Amino acid	密码子 Codon	密码子使用频率/1/1 000 Frequency of codon usage					密码子偏爱性比较 Comparison of codon preference			
		甘蓝 Cabbage (C)	大肠杆菌 <i>E. coli</i> (E)	拟南芥 <i>A. thaliana</i> (A)	棉花 <i>G. hirsutum</i> (G)	水稻 <i>O. sativa</i> (O)	C/E	C/A	C/G	C/O
Tyr	TAT	12.6	21.6	14.6	17.4	10.0	0.58	0.86	0.72	1.26
	TAC	14.9	11.7	13.7	15.5	15.1	1.27	1.09	0.96	0.99
Cys	TGT	9.6	5.9	10.5	9.2	6.2	1.63	0.91	1.04	1.55
	TGC	8.2	5.5	7.2	9.1	12.4	1.49	1.14	0.90	0.66
Leu	TTA	10.8	17.4	12.7	10.0	6.1	0.62	0.85	1.08	1.77
	TTG	20.0	12.9	20.9	20.6	14.7	1.55	0.96	0.97	1.36
	CTT	22.3	14.5	24.1	23.4	15.2	1.54	0.93	0.95	1.47
	CTC	18.6	9.5	16.1	14.0	25.8	1.96	1.16	1.33	0.72
	CTA	9.9	5.6	9.9	7.2	7.7	1.77	1.00	1.38	1.29
Trp	CTG	12.0	37.4	9.8	10.8	21.0	0.32 [#]	1.22	1.11	0.57
	TGG	12.4	8.2	12.5	14.7	13.8	1.51	0.99	0.84	0.90
Pro	CCT	19.0	9.5	18.7	20.4	13.6	2.00 [#]	1.02	0.93	1.40
	CCC	6.4	6.2	5.3	9.3	12.1	1.03	1.21	0.69	0.53
	CCA	15.2	9.1	16.1	18.2	14.2	1.67	0.95	0.84	1.07
	CCG	9.0	14.5	8.6	6.1	18.0	0.62	1.05	1.48	0.50 [#]
His	CAT	12.9	12.4	13.8	14.6	11.3	1.04	0.93	0.88	1.14
	CAC	9.9	7.3	8.7	9.2	13.8	1.36	1.14	1.08	0.72
Gln	CAA	17.9	14.4	19.4	20.5	13.5	1.24	0.92	0.87	1.33
	CAG	17.0	26.7	15.2	14.9	20.8	0.64	1.12	1.14	0.82
Arg	CGT	8.5	15.9	9.0	8.0	7.2	0.53	0.94	1.06	1.18
	CGC	4.7	14.0	3.8	4.2	16.1	0.34 [#]	1.24	1.12	0.29 [#]
	CGA	5.5	4.8	6.3	6.1	6.4	1.15	0.87	0.90	0.86
	CGG	4.8	7.9	4.9	4.1	13.4	0.61	0.98	1.17	0.36 [#]
	AGA	18.8	7.1	19.0	14.1	10.5	2.65	0.99	1.33	1.79
Ile	AGG	12.7	4.0	11.0	13.0	16.0	3.18 [#]	1.15	0.98	0.79
	ATT	17.8	29.6	21.5	24.0	14.2	0.60	0.83	0.74	1.25
	ATC	19.4	19.4	18.5	18.2	19.4	1.00	1.05	1.07	1.00
	ATA	12.2	13.3	12.6	10.7	8.8	0.92	0.97	1.14	1.39
Met	ATG	25.0	23.7	24.5	25.4	23.8	1.05	1.02	0.98	1.05
Thr	ACT	16.3	13.1	17.5	18.4	10.6	1.24	0.93	0.89	1.54
	ACC	11.4	18.9	10.3	14.3	14.9	0.60	1.11	0.80	0.77
	ACA	14.8	15.1	15.7	14.7	11.6	0.98	0.94	1.01	1.28
	ACG	9.2	13.6	7.7	5.0	11.4	0.68	1.19	1.84	0.81
Asn	AAT	18.1	29.3	22.3	22.3	15.1	0.62	0.81	0.81	1.20
	AAC	23.7	20.3	20.9	22.1	18.5	1.17	1.13	1.07	1.28
Lys	AAA	27.8	37.2	30.8	27.1	16.0	0.75	0.90	1.03	1.74
	AAG	34.7	15.3	32.7	34.7	32.3	2.27 [#]	1.06	1.00	1.07
Val	GTT	25.6	21.6	27.2	27.1	15.5	1.19	0.94	0.94	1.65
	GTC	14.4	13.1	12.8	12.7	20.1	1.10	1.13	1.13	0.72
	GTA	9.1	13.1	9.9	8.2	6.8	0.69	0.92	1.11	1.34
	GTG	19.1	19.9	17.4	18.0	24.3	0.96	1.10	1.06	0.79
Ala	GCT	27.8	18.9	28.3	30.8	19.6	1.47	0.98	0.90	1.42
	GCC	11.6	21.6	10.3	16.0	30.8	0.54	1.13	0.73	0.38 [#]
	GCA	16.5	23.0	17.5	19.0	17.3	0.72	0.94	0.87	0.95
	GCG	10.4	21.1	9.0	5.8	26.6	0.49	1.16	1.79	0.39 [#]
Asp	GAT	33.6	33.7	36.6	34.5	25.3	0.99	0.92	0.97	1.33
	GAC	19.9	17.9	17.2	17.7	28.1	1.11	1.16	1.12	0.71
Glu	GAA	31.5	35.1	34.3	32.8	21.6	0.90	0.92	0.96	1.46
	GAG	36.0	19.4	32.2	30.0	38.6	1.86	1.12	1.20	0.93
Gly	GGT	20.7	23.7	22.2	25.0	14.8	0.87	0.93	0.83	1.40
	GGC	10.6	20.6	9.2	12.8	29.5	0.51	1.15	0.83	0.36 [#]
	GGA	22.9	13.6	24.2	22.0	15.9	1.68	0.95	1.04	1.44
Ter	GGG	11.7	12.3	10.2	13.3	17.1	0.95	1.15	0.88	0.68
	TAA	1.0	2.0	0.9	1.2	0.7	0.50 [#]	1.11	0.83	1.43
	TAG	0.6	0.3	0.5	0.5	0.8	2.00 [#]	1.20	1.20	0.75
	TGA	1.3	1.1	1.2	1.2	1.2	1.18	1.08	1.08	1.08

注: 标注有“#”的数据表示比值小于等于 0.50 或者大于等于 2.00。

Note: Data with “#” means that the ratio is ≤ 0.50 or ≥ 2.00 .

2.3 Bt *cry1C* 基因的密码子改造

以确定的甘蓝高频密码子为依据, 对 Bt *cry1C* 抗虫基因序列进行密码子的改造, 获得了具有甘蓝密码子使用特点的 *cry1C* 基因序列, 结果见图 1。

原始序列 Original	AAGCTTCTA	GACCCGGGCC	TATTTTACA	ACAATTACCA	ACAACAACAA	ACAACAAACA	ACATTACAAT	TACTATTTC	AATTACAACC
改造序列 Modified	AAGCTTCTA	GACCCGGGCC	TATTTTACA	ACAATTACCA	ACAACAACAA	ACAACAAACA	ACATTACAAT	TACTATTTC	AATTACAACC
原始序列 Original	ATGGAGGAGA	ACAACCAGAA	CCAGTGCATC	CCTTACAAC	GTTTGAGCAA	CCCAGAGGAG	GTGCTTTTGG	ATGGGGAGAG	GATCTCCACT
改造序列 Modified	ATGGAGGAGA	ACAACCAGAA	CCAGTGCATC	CCTTACAAC	GTTTGAGCAA	CCCTGAGGAG	GTGCTTTTGG	ATGGGGAGAG	GATCTCCACT
原始序列 Original	GGAATCTCCT	CCATCGACAT	CTCCCTTTCC	CTTGTTCAAT	TCCTCGTGTC	CAACTTCGTG	CCAGGTGGAG	GATTCTTTGT	TGGATTGATC
改造序列 Modified	GGAATCTCCT	CCATCGACAT	CTCCCTTTCC	CTTGTTCAAT	TCCTCGTGTC	CAACTTCGTG	CCAGGTGGAG	GATTCTTTGT	TGGATTGATC
原始序列 Original	GACTTCGTGT	GGGGAATCGT	TGGACCATCC	CAATGGGAGC	CATTCTTTGT	GCAAATCGAG	CAGTTGATCA	ACGAGCGGAT	CGCAGAGTTC
改造序列 Modified	GACTTCGTGT	GGGGAATCGT	TGGACCTTCC	CAATGGGATG	CATTCTTTGT	GCAAATCGAG	CAGTTGATCA	ACGAGAGAAT	CGCTGAGTTC
原始序列 Original	GCAAGGAACG	CTGCGATCGC	TAACCTGGAG	GGATTGGGAA	ACAACCTCAA	CATCTACGTG	GAGGCATTCA	AGGAATGGGA	GGAGGACCCA
改造序列 Modified	GCTAGGAACG	CTGCTATCGC	TAACCTGGAG	GGATTGGGAA	ACAACCTCAA	CATCTACGTG	GAGGCATTCA	AGGAATGGGA	GGAGGATCCT
原始序列 Original	AACAACCCAG	CAACCCAGGAC	CAGGTGATC	GACAGATTCA	GAATCCTTGA	TGGACTCCTT	GAGAGGGACA	TCCCTTCGTT	CAGGATCTCC
改造序列 Modified	AACAACCCAG	CAACCCAGGAC	CAGAGTGATC	GATAGATTCA	GAATCCTTGA	TGGACTCCTT	GAGAGGGACA	TCCCTTCGTT	CAGGATCTCC
原始序列 Original	GGATTGAGG	TGCCACTTTT	GTCGTTTAC	GCACAAGCAG	CTAACCTTCA	TCTTGCAATC	TTGAGAGAT	CCGTGATCTT	CGGAGAGCGC
改造序列 Modified	GGATTGAGG	TGCCCTTTT	GTCGTTTAC	GCTCAAGCAG	CTAACCTTCA	TCTTGCTATC	TTGAGAGATT	CCGTGATCTT	CGGAGAGAGA
原始序列 Original	TGGGGATTGA	CCACTATCAA	CGTTAACGAG	AACTACAACA	GACTTATCAG	GCATATCGAC	GAGTACGCGG	ATCATTGTGC	AAACACTTAC
改造序列 Modified	TGGGGATTGA	CCACTATCAA	CGTTAACGAG	AACTACAACA	GACTTATCAG	GCATATCGAT	GAGTACGCTG	ATCATTGTGC	AAACACTTAC
原始序列 Original	AACAGGGGAT	TGAACAACCT	GCCAAAGTCC	ACTTACCAAG	ATTGGATCAC	CTACAACAGG	TTGAGGCGAG	ATTGACCTT	GACTGTGTTG
改造序列 Modified	AACAGGGGAT	TGAACAACCT	GCCAAAGTCC	ACTTACCAAG	ATTGGATCAC	CTACAACAGG	TTGAGGAGAG	ATTGACCTT	GACTGTGTTG
原始序列 Original	GACATCGCGG	CTTCTTTCCC	AAACTACGAC	AACAGGAGAT	ACCCAATCCA	ACCAGTAGGA	CAACTTACCA	GGGAGGTTTA	CACTGACCCA
改造序列 Modified	GACATCGCTG	CTTCTTTCCC	AAACTACGAT	AACAGGAGAT	ACCCAATCCA	ACCAGTTGGA	CAACTTACCA	GGGAGGTTTA	CACTGATCCA
原始序列 Original	TTGATCAACT	TCAACCCACA	ATTGCAATCC	TGGGCACAAT	TGCCAACCTT	CAACGTAATG	GAGAGCAGCG	CAATCAGGAA	CCCGCATTTC
改造序列 Modified	TTGATCAACT	TCAACCCACA	ATTGCAATCC	TGGGCTCAAT	TGCCTACCTT	CAACGTTATG	GAGAGCAGCG	CAATCAGAAA	CCCTCATTTC
原始序列 Original	TTCGACATCT	TGAACAACCT	TACCATCTTC	ACTGACTGGT	TCAGCGTGGG	ACGCAACTTC	TACTGGGGAG	GACATAGGGT	GATCTCCAGC
改造序列 Modified	TTCGACATCT	TGAACAACCT	TACCATCTTC	ACTGATTGGT	TCAGCGTTGG	AAGAACTTC	TACTGGGGAG	GACATAGGGT	GATCTCCAGC
原始序列 Original	CTTATCGGAG	GAGGAAACAT	CACCTCCCCC	ATCTACGGAA	GAGAGGCAAA	CCAGGAGCCT	CCACGAAGTT	TCACCTTCAA	CGGACCAGTG
改造序列 Modified	CTTATCGGAG	GAGGAAACAT	CACCTCCCCC	ATCTACGGAA	GAGAGGCAAA	CCAGGAGCCT	CCAAGATCTT	TCACCTTCAA	CGGACCAGTG
原始序列 Original	TTCAGGACTT	TGTCCAACCC	TACTTTGAGG	TTGTTGCAAC	AGCCTTGGCC	AGCACCACCA	TTCAACTTGA	GAGGAGTAGA	GGGAGTGGAG
改造序列 Modified	TTCAGGACTT	TGTCCAACCC	TACTTTGAGG	TTGTTGCAAC	AGCCTTGGCC	AGCACCACCA	TTCAACTTGA	GAGGAGTTGA	GGGAGTGGAG
原始序列 Original	TTCTCCACCC	CTACCAACAG	CTTCACTTAC	AGGGGACGTG	GAAGTGTAGA	TTCTTGACT	GAGTTGCCAC	CTGAGGACAA	CAGCGTGCCA
改造序列 Modified	TTCTCCACCC	CTACCAACAG	CTTCACTTAC	AGGGGAAGAG	GAAGTGTAGA	TTCTTGACT	GAGTTGCCAC	CTGAGGATAA	CAGCGTGCCA
原始序列 Original	CCTCGCGAGG	GATACAGCCA	TAGATTGTGT	CATGCAACTT	TCGTTCAAAG	ATCCGGAACC	CCTTTCCTTA	CCACTGGAGT	GGTGTCTTCC
改造序列 Modified	CCTAGAGAGG	GATACAGCCA	TAGATTGTGT	CATGCAACTT	TCGTTCAAAG	ATCCGGAACC	CCTTTCCTTA	CCACTGGAGT	GGTGTCTTCC
原始序列 Original	TGGACTCATA	GAAGCGCAAC	TCTTACCAAC	ACCATCGACC	CAGAGAGGAT	CAATCAGATC	CCATTGGTGA	AGGGATTTCAG	AGTCTGGGGA
改造序列 Modified	TGGACTCATA	GAAGCGCAAC	TCTTACCAAC	ACCATCGATC	CAGAGAGGAT	CAATCAGATC	CCTTTGGTGA	AGGGATTTCAG	AGTTTGGGGA
原始序列 Original	GGCACCTCCG	TTATCACCAG	ACCAGGATTC	ACCGGAGGAG	ACATCCTTAG	GAGGAACACC	TTCCGAGACT	TCGTGTCCCT	TCAAGTCAAC
改造序列 Modified	GGCACCTCCG	TTATCACCAG	ACCAGGATTC	ACCGGAGGAG	ACATCCTTAG	GAGGAACACC	TTCCGAGATT	TCGTGTCCCT	TCAAGTTAAC
原始序列 Original	ATCAACTCCC	CAATCACCCA	ACGATACCGT	TTGCGGTTCA	GATACGCTTC	CAGCAGGGAT	GCAAGGGTTA	TCGTGTTGAC	CGGAGCAGCA
改造序列 Modified	ATCAACTCCC	CAATCACCCA	AAGATACAGA	TTGAGATTCA	GATACGCTTC	CAGCAGGGAT	GCAAGGGTTA	TCGTGTTGAC	CGGAGCAGCA
原始序列 Original	TCCACCGGAG	TGGGAGGACA	AGTCAGCGTG	AACATGCCAC	TTCAAAAGAC	TATGGAGATC	GGAGAGAACT	TGACCTCCAG	GACCTTCCGG
改造序列 Modified	TCCACCGGAG	TGGGAGGACA	AGTTAGCGTG	AACATGCCTC	TTCAAAAGAC	TATGGAGATC	GGAGAGAACT	TGACCTCCAG	GACCTTCAGA
原始序列 Original	TACACCGACT	TCAGCAACCC	TTTCTCCTTC	AGAGCTAACC	CAGACATCAT	CGGAATCAGC	GAGCAACACC	TTTTCCGAGC	AGGATCCATC
改造序列 Modified	TACACCGATT	TCAGCAACCC	TTTCTCCTTC	AGAGCTAACC	CAGACATCAT	CGGAATCAGC	GAGCAACCTC	TTTTCCGAGC	AGGATCCATC
原始序列 Original	AGCAGCGGAG	AGCTTTACAT	CGACAAGATC	GAGATCATCC	TTGAGATGC	AACCTTCGAG	GCAGAGTCCG	ATTGAGAGAG	AGCACAAAAG
改造序列 Modified	AGCAGCGGAG	AGCTTTACAT	CGATAAGATC	GAGATCATCC	TTGAGATGC	AACCTTCGAG	GCAGAGTCCG	ATTGAGAGAG	AGCACAAAAG
原始序列 Original	GCAGTGAACG	CTCTTTTTCAC	TTCTTCCAAC	CAGATCGGAT	TGAAGACCGA	TGTGACTGAC	TCCGAGAAGG	ATGAACCTTG	ATAAGGTACC
改造序列 Modified	GCAGTGAACG	CTCTTTTTCAC	TTCTTCCAAC	CAGATCGGAT	TGAAGACCGA	TGTGACTGAT	TCCGAGAAGG	ATGAACCTTG	ATAAGGTACC
原始序列 Original	CTCGAGGAGC	TCTCCGAATT	C						
改造序列 Modified	CTCGAGGAGC	TCTCCGAATT	C						

图 1 Bt *cry1C* 基因密码子改造结果

Fig. 1 The results of codons modification of Bt *cry1C* gene

3 讨论

分析密码子用法的方法主要有两种: 高频密码子分析法(林涛 等, 2002)和高表达优越密码子分析法(Frank, 1990)。高表达优越密码子分析法需要先计算每个基因的密码子有效数(Effective Number of Codons, Nc)和相对同义密码子用法(Relative Synonymous Codon Usage, RSCU), 根据 Nc 值确定高表达和低表达样本组, 然后计算出这 2 个样本组中各个密码子的 RSCU 值, 最后通过 *t* 检验确定出高表达基因的优越密码子。高表达优越密码子分析法步骤繁琐, 并且对于目的生物的遗传背景要求较高, 另外, 若蛋白质编码基因序列数量很大时, 采用该方法进行分析, 工作量是巨大的。高频密码子分析法克服了高表达优越密码子分析法的缺点, 该方法准确、简便、快速(林涛 等, 2002; 胡桂兵 等, 2006)。林涛等(2002)将高频密码子分析法筛选出的酵母(*Y. lipolytica*)高频密码子和传统的高表达优越密码子分析法筛选出的优越密码子进行比较, 证明了高频密码子分析法的可靠性, 且这两种方法的结果是吻合的。本研究利用 CodonW 软件, 采用高频密码子分析法对结球甘蓝的全基因组序列进行了分析, 计算出同义密码子相对使用频率, 确定了 TCT、CCT、AGA、GTT、GCT、GAT 等 6 个甘蓝高频密码子。

甘蓝密码子的偏爱性与大肠杆菌具有较大的差异, 与同为双子叶植物的拟南芥、棉花的密码子偏爱性基本一致, 而与单子叶植物的水稻具有一定的差异, 表明物种间的亲缘关系越远, 密码子偏爱性差异越大; 亲缘关系越近, 密码子偏爱性差异越小。因此, 在进行甘蓝转基因研究时, 应该注意密码子偏爱性的影响。由于甘蓝和拟南芥的密码子偏爱性基本相同, 在模式植物拟南芥中克隆的基因可以不经密码子优化而直接在甘蓝中表达; 同时, 设计的含有拟南芥偏爱密码子的外源基因除了可在拟南芥中高效表达外, 在甘蓝等近缘种中也有可能实现高效表达(Murray et al., 1989; 胡桂兵 等, 2006)。这种“通用性”可以节约资源、减少时间, 是十分有意义的。

密码子使用偏爱性是物种进化现象的反映(张文娟, 2006)。密码子使用偏好性的获得可以用来探究许多基本生物学问题, 如利用甘蓝与其他近缘物种密码子使用模式的差异, 结合其他系统发育分析方法, 可以判断基因重组或基因水平转移的发生。

通过密码子改造提高外源基因在动物、植物或微生物中的表达已有不少成功的例子(Paul & Elizabeth, 1991; Jeffrey et al., 2003; 柏锡 等, 2004), 但在甘蓝中尚未见报道。本文的分析结果可为来源于原核生物的外源基因(如 Bt 抗虫基因)在甘蓝中的表达、甘蓝基因在原核生物中的表达以及利用 PCR 技术克隆甘蓝功能基因时提供参考。小菜蛾、菜青虫、甜菜夜蛾和甘蓝夜蛾等是十字花科作物最重要的害虫, 且难以防治。本课题组一直致力于结球甘蓝抗虫转基因研究, 并已获得对小菜蛾和菜青虫具有较强抗性的转基因植株, 现已利用甘蓝密码子用法分析结果优化并重新合成了对鳞翅目夜蛾类害虫具有较强毒杀作用的新的 Bt *cry1C* 基因, 准备进行遗传转化, 希望在甘蓝抗虫基因工程育种方面取得突破。

References

- Bai Xi, Xu Jian-xia, Li Lin, Guo Zheng, Li Jie, Zhu Yan-ming. 2004. Analysis of codon usage in potato and its application in the modification of t-PA gene. *Hereditas*, 26 (1): 75 - 83. (in Chinese)
- 柏 锡, 徐建霞, 李 琳, 郭 政, 李 杰, 朱延明. 2004. 马铃薯密码子用法分析及其在 t-PA 基因密码子改造上的应用. *遗传*, 26 (1): 75 - 83.
- Bennetzen J L, Hall B D. 1982. Codon selection in yeast. *The Journal of Biological Chemistry*, 257 (6): 3026 - 3031.
- Bhattacharya R C, Viswakarma N, Bhat S R, Kirti P B, Chopra V L. 2002. Development of insect-resistant transgenic cabbage plants expressing a synthetic *cryIA(b)* gene from *Bacillus thuringiensis*. *Current Science*, 83: 146 - 150.

- Cui Lei, Yang Li-mei, Liu Nan, Lang Zhi-hong, Liu Yu-mei, Zhuang Mu, Zhang Yang-yong, Zhang You-jun, Huang Da-fang, Fang Zhi-yuan. 2009. Transformation and expression of Bt gene *cryIIa8* in cabbage. *Acta Horticulturae Sinica*, 36 (8): 1161 - 1168. (in Chinese)
- 崔磊, 杨丽梅, 刘楠, 郎志宏, 刘玉梅, 庄木, 张扬勇, 张友军, 黄大昉, 方智远. 2009. *Bt cryIIa8* 抗虫基因对结球甘蓝的转化及其表达. *园艺学报*, 36 (8): 1161 - 1168.
- Frank W. 1990. The effective number of codons usage used in a gene. *Gene*, 87: 23 - 29.
- Hu Gui-bing, Zhang Shang-long, Xu Chang-jie, Lin Shun-quan. 2006. Analysis of codon usages in citrus. *Journal of Fruit Science*, 23 (3): 479 - 485. (in Chinese)
- 胡桂兵, 张上隆, 徐昌杰, 林顺权. 2006. 柑橘密码子用法分析. *果树学报*, 23 (3): 479 - 485.
- Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organism. *Molecular Biology and Evolution*, 2 (1): 13 - 34.
- Jeffrey R P, Erminia S, Etsuko N M, Jennifer M G, Adalgisa C. 2003. Analysis of a shift in codon usage in *Drosophila*. *Journal of Molecular Evolution*, 57: S214 - S225.
- Jin R G, Liu Y B, Tabashnik B E, Borthakur D. 2000. Development of transgenic cabbage (*Brassica oleracea* var. *capitata*) for insect resistance by *Agrobacterium tumefaciens*-mediated transformation. *In Vitro Cellular & Developmental Biology-plant*, 36: 231 - 237.
- Li Han-xia, Yin Ruo-he, Lu Ya-chun, Zhang Jun-hong. 2006. Inheritance and resistance to insect in CryIA(c) transgenic cabbage. *Journal of Agricultural Biotechnology*, 14 (4): 546 - 550. (in Chinese)
- 李汉霞, 尹若贺, 陆芽春, 张俊红. 2006. CryIA(c)转基因结球甘蓝的抗虫性研究. *农业生物技术学报*, 14 (4): 546 - 550.
- Lin Tao, Ni Zhi-hua, Shen Ming-shan, Chen Liang. 2002. High-frequency codon analysis and its application in codon analysis of tobacco. *Journal of Xiamen University: Natural Science*, 41 (5): 551 - 554. (in Chinese)
- 林涛, 倪志华, 沈明山, 陈亮. 2002. 高频密码子分析法及其在烟草密码子分析中的应用. *厦门大学学报: 自然科学版*, 41 (5): 551 - 554.
- Matzke M A, Mette M F, Matzke A J M. 2000. Transgene silencing by the host genome defense: Implications for the evolution of epigenetic control mechanisms in plants and vertebrates. *Plant Molecular Biology Reporter*, 43: 401 - 415.
- Metz T D, Dixit R, Earle E D. 1995. *Agrobacterium tumefaciens*-mediated transformation of broccoli (*Brassica oleracea* var. *italica*) and cabbage (*Brassica oleracea* var. *capitata*). *Plant Cell Reports*, 15: 287 - 292.
- Murray E E, Lotzer J, Eberle M. 1989. Codon usage in plant gene. *Nucleic Acids Research*, 17: 477 - 498.
- Paul M S, Elizabeth C. 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast*, 7 (7): 657 - 678.
- Rafat A, Aziz M A, Rashid A A, Abdullah S N A, Kamaladini H, Sirchi M H T, Javadi M B. 2010. Optimization of *Agrobacterium tumefaciens*-mediated transformation and shoot regeneration after co-cultivation of cabbage (*Brassica oleracea* ssp. *capitata*) cv. KY Cross with *AtHSP101* gene. *Scientia Horticulturae*, 124: 1 - 8.
- Yi D X, Cui L, Liu Y M, Zhuang M, Zhang Y Y, Yang L M. 2011. Transformation of cabbage (*Brassica oleracea* L. var. *capitata*) with Bt *cryIa3* gene for control of diamondback moth. *Agricultural Sciences in China*, 10 (11): 1693 - 1700.
- Zhang Wen-juan. 2006. Codon analysis and its application in bioinformatics and evolutionary studies[Ph. D. Dissertation]. Shanghai: Fudan University. (in Chinese)
- 张文娟. 2006. 基于密码子水平的生物信息学分析及进化研究[博士论文]. 上海: 复旦大学.