

• 药材与资源 •

刺五加功能基因密码子偏好性的分析

邢朝斌, 吴 鹏, 修乐山, 周 秘

河北联合大学生命科学学院, 河北 唐山 063000

摘要:目的 分析刺五加功能基因密码子的使用方式及其影响因素。方法 以刺五加的17条功能基因为材料,利用CodonW和SPSS软件进行多元统计分析和对应性分析。结果 刺五加功能基因密码子3个位置的GC量依次为51.03%、41.23%和40.04%,三者均与整个编码区的GC量显著相关($P < 0.05$),GC12与GC3的相关系数为0.262,未达到显著水平。同义密码子的相对使用频率大于1的密码子共27个,其中22个以A或T碱基结尾。对应性分析的结果表明,第1轴显示了22.78%的差异,与GC3、密码子适应指数和密码子偏好指数均极显著相关($P < 0.01$),相关系数分别为0.786、0.686和0.617,与有效密码子数的相关性未达显著水平。第2轴显示了19.28%的差异,且仅与有效密码子数极显著相关($r = 0.635$)。确定了刺五加功能基因的17个最优密码子。结论 刺五加的功能基因偏好使用以A或T碱基结尾的密码子,其使用模式受选择和突变的共同影响。

关键词: 刺五加; 密码子用法; 功能基因; 最优密码子; GC量; 偏好指数

中图分类号: R282.12

文献标志码: A

文章编号: 0253-2670(2013)05-0598-06

DOI: 10.7501/j.issn.0253-2670.2013.05.020

Analysis on bias of functional gene codon of *Eleutherococcus senticosus*

XING Zhao-bin, WU Peng, XIU Le-shan, ZHOU Mi

College of Life Science, Hebei United University, Tangshan 063000, China

Abstract: **Objective** To analyze the codon usage of functional genes in *Eleutherococcus senticosus* and their influencing factors. **Methods** The multivariate statistical analysis and correspondence analysis were carried out using CodonW and SPSS software with codon of 17 genes selected from the functional genes of *E. senticosus*. **Results** GC contents at the three positions of functional gene codons in *E. senticosus* was 51.03%, 41.23%, and 40.04%, all of which had a significant correlation coefficient ($P < 0.05$). The correlation coefficients with GC12 and GC3 were 0.262, and both were insignificant. The relative synonymous codon usage of 27 codons was greater than 1, among which 22 codons ended with A or T base. In the corresponding analysis, the first axis showed the variation of 22.78%, and there were significant correlation coefficients in codon adaptation, codon bias index, and GC3 ($P < 0.01$). The correlation coefficients were 0.686, 0.617, and 0.786, respectively, but they were insignificantly related with the effective number of codons (ENC). The second axis showed the variation of 19.28% and it was only significantly related with ENC ($r = 0.635$). Seventeen optimized codons in functional genes of *E. senticosus* were defined. **Conclusion** All codons of functional genes in *E. senticosus* prefer to ending with A or T base. The codon usage bias is formed under the effects of mutation and selection.

Key words: *Eleutherococcus senticosus* (Rupr. et Maxim.) Maxim.; codon usage; functional genes; optimized codon; GC content; bias index

依据“中心法则”,密码子作为遗传信息从碱基序列到氨基酸序列的桥梁中发挥着不可替代的作用,它不仅具有几乎所有生物共用同一套密码子表的特点,而且具有同义密码子现象,即2个或2个以上密码子编码同一种氨基酸的现象^[1]。同义密码

子家族的大小在核基因组中一般为2、3、4和6^[2]。在没有选择压力和中性突变的情况下,编码同一个氨基酸的几个同义密码子的使用频率应该相同。但实际翻译过程中,同义密码子的使用概率并不相同,这种现象被称为密码子偏好性。不同物种因各自的

收稿日期: 2012-10-18

基金项目: 国家自然科学基金资助项目(30701086); 河北省自然科学基金项目(C2009001252); 河北省自然科学基金-石药集团医药联合研究基金项目(H2012401006)

作者简介: 邢朝斌(1975—),男,副教授,研究方向为分子生药学、药用植物细胞工程。

Tel: (0315)3726238 Fax: (0315)3726341 E-mail: xzbheuu@126.com

基因表达水平、GC 量、基因长度等因素的影响,在长期进化过程中,最终形成各自不同的密码子偏好特点^[1,3]。通过对密码子偏好性进行分析,找出特定物种密码子使用的一些规律,对转基因中目的基因的高效表达、未知基因的克隆和从分子水平阐明物种的起源及遗传分化等研究具有重要意义^[3]。

刺五加 *Eleutherococcus senticosus* (Rupr. et Maxim.) Maxim. 是我国医药珍品,具有多种生理活性和药理作用,需求量较大^[4]。本实验室在已经克隆出刺五加的环阿屯醇合酶(cycloartenol synthase, CAS)^[5]、甘油醛-3-磷酸脱氢酶(glyceraldehyde-3-phosphate dehydrogenase, GAPDH)^[6]和鲨烯合酶(squalene synthase, SS)^[7-8]基因等共计 14 条功能基因及周晨光^[9]从刺五加愈伤组织中克隆的 3 条基因的基础上,对可能影响刺五加功能基因密码子使用偏好性的各种因素及其密码子使用模式进行了分析,以期为提高刺五加功能基因体外表达效率、重要性状调控基因的克隆及其分子进化研究等提供重要的理论支持。

1 材料

以本实验室克隆的刺五加 CAS (GenBank 登录号: JQ400139)、GAPDH (Genbank 登录号: HQ184063)、SS1 (GenBank 登录号: HQ456918)、SS2 (GenBank 登录号: JN714465)、鲨烯环氧酶(squalene epoxidase, SE)基因的 SE1 (GenBank 登录号: JF818129)、SE2 (GenBank 登录号: JN228206)、 β -香树酯醇合成酶(β -amyrin synthase, bAS)基因的 bAS1 (GenBank 登录号: JF818130)、bAS2 (GenBank 登录号: JF818131)、法尼基焦磷酸合酶(farnesyl diphosphate synthase, FPS)基因 (GenBank 登录号: JQ178346)、3-羟基-3-甲基戊二酰 CoA 还原酶(3-hydroxy-3-methylglutaryl coenzyme A reductase, HMGR)基因 (GenBank 登录号: JQ905593)、甲羟戊酸焦磷酸脱羧酶(mevalonate diphosphate decarboxylase, MDD)基因 (GenBank 登录号: JQ905594)、铜锌型超氧化物歧化酶(copper/zinc superoxide dismutase gene, Cu/Zn SOD)基因 (GenBank 登录号: JQ365625)、亚精胺合成酶(spermidine synthase, SPDS)基因 (GenBank 登录号: JQ365624)、钙调蛋白(calmodulin, CaM)基因 (GenBank 登录号: JQ365623)和周晨光^[9]从刺五加中克隆的转录因子 AP21 基因、木葡聚糖内糖基转移酶(xyloglucan endotransglucosylase/hydrolase)基因 XET1 和磷酸

盐转运蛋白(phosphate transporter)基因 PT1 的核苷酸序列作为分析样本。各序列均以 ATG 为起始密码子,以 TAA、TAG 或 TGA 为终止密码子,且编码区长度大于 300 bp (100 个氨基酸)。

2 方法

2.1 同义密码子使用偏好性分析

利用 CodonW 1.4.4 程序进行同义密码子偏好性分析,并进行对应性分析。计算指标包括:样本的总体 GC 量以及密码子第 1、2 和 3 位核苷酸上的 GC 量,分别记为 GC、GC1、GC2 和 GC3;同义密码子的相对使用频率(relative synonymous codon usage, RSCU);有效密码子数(effective number of codons, ENC);密码子偏好指数(codon bias index, CBI);密码子适应指数(codon adaptation index, CAI)。通过 SPSS 17.0 软件进行相关性检验。将各基因的 GC1 和 GC2 的平均值记为 GC12,以 GC3 为横坐标,GC12 为纵坐标,进行中性绘图分析,确定密码子 3 个位点的碱基组成之间的相关性及其影响密码子偏好性的因素。若 GC3 与 GC12 显著相关,则说明密码子 3 个位点上的碱基组成无差异,密码子的偏好受突变的影响。若两者的相关性不显著,说明密码子不同位点的碱基组成不同,基因组 GC 量高度保守,密码子的偏好特点受选择的影响比重较大。以 ENC 为纵坐标,GC3 为横坐标,进行 ENC 绘图分析。并利用公式 $ENC = 2 + GC3 + 29 / [GC3^2 + (1 - GC3)^2]$ 计算获得的密码子偏好性仅由碱基组成决定时的基因预期值,以此值构建标准曲线。当某基因的密码子偏好性仅受突变影响时则分布于曲线附近,若其受选择影响较大则分布于曲线下方较远的位置^[10]。

2.2 最优密码子分析

以 ENC 为标准,将刺五加的功能基因核苷酸序列中 ENC 值居于最高和最低的 10% 样本分别作为高、低表达组。分别计算其对应的 RSCU 值,若两者间的 RSCU 差值($\Delta RSCU$) >0.08 ,则确定其为高表达最优密码子^[3,10]。

3 结果与分析

3.1 密码子偏好性分析

利用在线工具 CUSP 分析刺五加 17 条基因组功能基因编码区,结果见表 1。全部基因编码区的 GC 量为 44.10%,高于其叶绿体基因组的 GC 量(37.95%)^[11]。密码子 3 个位置的碱基组成表现为 GC1 (51.03%) $>$ GC2 (41.23%) $>$ GC3 (40.04%),其中 GC1 和 GC3 的 GC 量与编码区和整个基因组

表1 刺五加功能基因密码子不同位置的 GC 量

Table 1 GC contents in different positions of functional gene codon of *E. senticosus*

基因	GC 量 / %				ENC	基因	GC 量 / %				ENC
	GC1	GC2	GC3	GC			GC1	GC2	GC3	GC	
AP21	49.45	43.27	41.09	44.61	54.29	FPS	52.19	28.28	41.98	40.82	55.99
XET1	43.99	43.30	51.20	46.16	55.10	CAS	51.52	41.11	37.94	43.52	53.13
PT1	50.47	41.02	45.56	45.68	57.51	HMGR	53.59	44.83	49.56	49.33	57.57
SS1	48.08	34.62	41.59	41.43	55.48	MDD	52.26	40.86	39.90	44.34	54.76
SS2	48.19	34.70	44.58	42.49	57.17	GAPDH	50.24	42.18	46.45	46.29	40.64
SE1	51.35	43.06	36.04	43.48	53.87	SOD	64.71	51.63	36.60	50.98	47.05
SE2	51.01	42.54	35.73	43.09	53.11	SPDS	55.39	38.32	37.13	43.61	54.05
bAS1	48.90	37.90	36.43	41.08	53.85	CaM	59.33	28.67	57.33	48.44	55.62
bAS2	48.54	38.29	35.85	40.89	52.88	合计	51.03	40.04	41.23	44.10	55.08

的 GC 量差别较大, GC2 与其相近。即在刺五加的基因组中, 以 A 或 T 碱基结尾的密码子使用频率高于 G 或 C 碱基结尾的密码子。刺五加基因组功能基因编码区的 ENC 取值区间在 40.64~57.57, 其中 94.12% 基因的 ENC 值大于 45。ENC 的取值为 20~61, ENC 值越靠近 20 表示同义密码子偏性越大, 越靠近 61 表示同义密码子偏性越小^[10], 说明刺五加功能基因编码区的偏性较弱。

刺五加功能基因编码区密码子各位置的 GC 量、ENC 和密码子数目的相关关系如表 2 所示。整个编码区的 GC 量与 GC1 之间的相关性达到了极显著水平 ($P<0.01$), 与 GC2、GC3 之间也均达到了显著水平 ($P<0.05$)。密码子 3 个位置 GC 量之间的相关性均未达到显著水平, 说明刺五加功能基因密码子 3 个位置的碱基组成间存在显著差别。ENC 值与密码子 3 个位置 GC 量及总 GC 量的相关性均未达显著水平, 说明密码子的碱基组成对密码子偏好性影响较小。ENC 和密码子数的相关系数为 0.414, 未达到显著水平, 表明密码子数目对 ENC

的影响很弱, 可以排除基因自身序列长度大小对密码子偏好性分析造成的干扰。

RSCU 值是某个特定密码子的实际观察值与理论观察值之间的比值。若 $RSCU>1$, 则密码子出现的频率比其他同义密码子高, 反之, 出现则低, 若 $RSCU=1$, 则密码子使用无偏好性^[3,10]。刺五加功能基因编码区的 RSCU 值如表 3 所示。起始密码子 ATG 和色氨酸的密码子 TGG 的 RSCU 值为 1。RSCU>1 的密码子共 27 个, 其中 22 个以 A 或 T 碱基结尾, 这些是刺五加功能基因的偏好密码子。RSCU<1 的密码子共 35 个, 其中 25 个以 G 或 C 碱基结尾, 这些是刺五加功能基因的非偏好密码子 (表 3)。

3.2 中性绘图分析

刺五加各功能基因的中性绘图分析如图 1 所示。多数功能基因均分布于对角线上方或沿对角线分布, 仅 4 条基因分布于对角线下方。GC3 的分布在 0.357 3~0.573 3, GC12 的分布在 0.402 4~0.581 7, 两者分布范围均较小。GC12 与 GC3 的相关系数为 0.262, 双尾检验未达到显著水平, 回归曲线斜率为 0.167 9, 说明 GC12 与 GC3 的相关性很弱, 密码子 3 个位置的碱基构成不同, 各位置的 GC 量高度保守, 突变在密码子偏好性形成过程中作用较弱, 选择在此过程中的作用较强。

3.3 ENC 和 GC3 的相关性分析

以刺五加各功能基因的 GC3 值为横坐标, ENC 值为纵坐标将各基因分布于平面图 (图 2) 中。由图可知, 刺五加各功能基因均分布于期望值曲线的下方, 但多数距离期望值曲线较近, 存在一定的偏性。说明刺五加功能基因的密码子偏好性更多地受

表2 各基因相关参数的相关性分析

Table 2 Correlation analysis of related parameters of each gene

相关参数	GC1	GC2	GC3	GC	ENC
GC2	0.164				
GC3	-0.007	-0.309			
GC	0.624**	0.511*	0.496*		
ENC	-0.241	-0.402	0.183	-0.256	
密码子数	-0.394	0.171	-0.357	-0.346	0.414

*和**分别表示各参数间相关性在 0.05 和 0.01 水平差异显著

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

表 3 刺五加功能基因蛋白编码序列同义密码子的相对使用频率

Table 3 RSCU of protein coding sequence of functional genes in *E. senticosus*

氨基酸	密码子	RSCU			氨基酸	密码子	RSCU			
		基因组	高表达	低表达			基因组	高表达	低表达	
Phe	TTT*	1.11	1.15	1.05	Tyr	TAT	1.15	1.51	1.55	
	TTC	0.89	0.85	0.95		TAC	0.85	0.74	0.89	
Leu	TTA*	0.69	0.80	0.47	TER	TAA	0.88	1.38	1.25	
	TTG	1.43	1.49	1.39		TAG	0.18	0.53	0.59	
	CTT	1.70	1.60	1.72	His	CAT	1.32	1.42	1.37	
	CTC	0.74	0.68	0.82		CAC	0.68	0.58	0.62	
	CTA	0.68	0.75	0.66	Gln	CAA*	1.14	1.53	1.34	
	CTG	0.76	0.68	0.95		CAG	0.86	0.47	0.67	
	Ile	ATT	1.68	1.67	1.64	Asn	AAT*	1.26	1.65	1.56
		ATC	0.62	0.57	0.62		AAC	0.74	0.61	0.79
		ATA	0.71	0.77	0.74	Lys	AAA*	0.95	1.38	1.29
Met	ATG	1.00	1.00	1.00	AAG		1.05	0.36	0.36	
Val	GTT*	1.61	1.77	1.42	Asp	GAT*	1.39	1.71	1.52	
	GTC	0.75	0.67	0.82		GAC	0.61	0.62	0.84	
	GTA	0.70	0.75	0.67	Glu	GAA*	0.98	1.39	1.15	
	GTG	0.95	0.81	1.09		GAG	1.02	0.28	0.48	
	Ser	TCT*	1.51	1.23	1.06	Cys	TGT*	1.06	1.12	0.90
		TCC	0.82	0.77	0.94		TGC	0.94	0.88	1.10
		TCA	1.36	0.75	1.33	TER	TGA	1.94	2.25	1.33
	TCG	0.61	0.00	0.33	Trp		TGG	1.00	1.00	1.00
	Pro	CCT	1.40	1.36	1.25	Arg	CGT*	0.81	0.91	0.68
CCC		0.61	0.64	0.75	CGC		1.01	1.05	1.08	
	CCA*	1.45	1.19	0.86	CGA	0.93	1.00	0.99		
	CCG	0.54	0.81	1.14	CGG	0.50	0.40	0.59		
	Thr	ACT	1.70	1.31	1.13	Ser	AGT*	0.90	1.14	0.68
		ACC	0.70	0.69	0.87		AGC	0.81	0.69	1.03
		ACA*	1.27	1.07	0.82	Arg	AGA*	1.45	1.48	1.17
	ACG	0.34	0.93	1.18	AGG		1.29	1.16	1.49	
	Ala	GCT	1.61	1.40	1.37	Gly	GGT	1.24	1.21	1.28
GCC		0.74	0.60	0.63	GGC		0.70	0.67	0.67	
GCA*		1.27	1.05	0.78	GGA*		1.26	1.36	1.19	
GCG		0.38	0.95	1.22	GGG		0.80	0.76	0.86	

*最优密码子

* refers to optimized codon

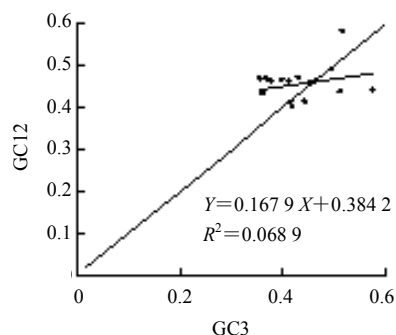


图 1 中性绘图分析

Fig. 1 Neutrality plot analysis

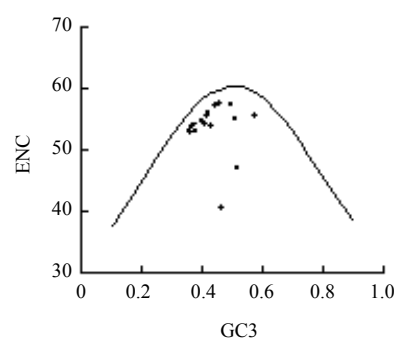


图 2 ENC 和 GC3 的相关性分析

Fig. 2 Correlation analysis of ENC and GC3

到选择效应的影响,而碱基突变作用对密码子偏好性的作用相对较弱。

3.4 对应性分析

基于 RSCU 的对应性分析结果显示,第 1、2、3 和 4 轴分别显示了 22.78%、19.28%、12.52% 和 9.07% 的差异,前 4 轴差异合计 63.65%,以第 1、2 轴为主。第 1 轴与 GC3、CAI 和 CBI 呈极显著相关关系 ($P < 0.01$),相关系数分别为 0.786、0.686 和 0.617,均达极显著水平 ($P < 0.01$),而第 1 轴与 ENC 的相关系数为 -0.131,未达显著水平。第 2 轴与 GC3、CAI 和 CBI 的相关系数分别为 -0.204、-0.045 和 -0.273,均未达显著水平,但第 2 轴与 ENC 的相关系数为 0.635,达极显著水平。说明 GC3 和 ENC 对刺五加功能基因密码子的偏好性作用较大,同时其他因素对刺五加功能基因密码子的偏好性也有一定影响,其使用模式形成过程较为复杂。

将刺五加各基因根据功能分为 3 类,分别为愈伤组织差异表达基因(包括 AP21、XET1、和 PT1)、萜类合成基因(包括 SS1、SS2、SE1、SE2、bAS1、bAS2、FPS、CAS、HMGR 和 MDD)、和其他基因(包括 GAPDH 和 Cu/Zn SOD)。以第 1 轴为纵坐标,第 2 轴为横坐标将刺五加各功能基因分布于平面中。其中萜类合成基因集中分布,而其他基因则因功能不同而分散度较大(图 3)。

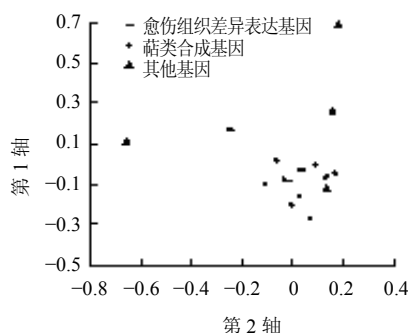


图3 基于 RSCU 的对应性分析

Fig. 3 Corresponding analysis based on RSCU

3.5 最优密码子分析

根据各基因的 ENC 值大小,选择高、低表达基因进行建库,分别计算高、低表达基因库各密码子的 RSCU 值(表 3)。根据两库密码子的 Δ RSCU 值,最终确定 TTT、TTA 和 GTT 等 17 个密码子为最优密码子。

4 讨论

在长期的进化过程中,每种生物体都形成了自

己特有的密码子用法,即密码子的使用具有物种特异性^[1]。中性理论认为密码子的选择仅仅和突变有关,不受自然选择压力影响,即同义密码子突变并不影响生存适合度,在密码子第 3 位上的突变是中性选择的结果;而“选择-突变-漂变”模型则认为,突变的发生是有方向的,同义密码子使用的偏好性反映了对最优密码子的选择和对非同义密码子的“突变-漂变”两方面的作用^[3]。对刺五加已知功能基因的分析结果表明,其密码子 3 个位置的碱基组成不同,第 3 位偏好使用碱基 A 或 T。其与碱基相关的 GC3、GC 指标均与第 1 轴极显著相关,说明碱基差异对刺五加功能基因密码子的偏好性存在较大影响 ($r = 0.786$),同时反应选择作用强度的指标 ENC 与第 1 轴的相关性 ($r = -0.131$) 虽未达显著水平,但 CAI 和 CBI 与第 1 轴确呈极显著相关关系,其相关系数分别达 0.686 和 0.617。不过,在对应性分析过程中,刺五加并没有表现出像葡萄、陆地棉等第 1 轴显著高于其他轴占主导地位的特点,而是第 1 轴和第 2 轴差别较小,两者间差异不显著,且两者同时显著高于其他轴。因此,在对其分析过程中,需要将第 1 轴和第 2 轴的结果进行综合分析,才能全面反映出刺五加功能基因密码子的使用特点。与第 1 轴的分析结果不同,刺五加第 2 轴与 GC3、GC、CAI、CBI 的相关性均不显著,而与 ENC 间呈现极显著的相关关系,但其相关系数 ($r = 0.635$) 略小于第 1 轴与 GC3 的相关系数 0.786,同时第 2 轴所代表的差异比重也略低于第 1 轴。这说明,刺五加功能基因的密码子用法受碱基组成(突变)和选择等多种因素影响。另外,ENC 和 GC3 相关性分析的结果也显示,除部分基因存在于标准曲线下方较远处外,多数位于标准曲线下方较近处,也同样反映出了刺五加功能基因的密码子使用受选择和突变等多种因素影响的特点。

目前,关于五加科药用植物功能基因密码子的用法分析尚未见报道。本研究首次分析了该科重要药用植物刺五加的密码子使用特点。已有的研究表明,亲缘关系相近的物种间,一般表现出相似的密码子使用模式^[1],因此,研究结果对五加科其他药用植物功能基因的体外高效表达、新基因的克隆和功能分析等也具有重要的借鉴意义。但同时应注意到,包括刺五加在内的大部分药用植物的功能基因研究均较少,迄今为止,刺五加中仅克隆出了本实验中涉及到的 17 条功

能基因, 因此, 更加全面的刺五加密码子使用偏好性分析尚有待更多功能基因的克隆及整个基因组测序完成后才能进行。

参考文献

- [1] 武志娟, 钟金城. 密码子偏性及其应用 [J]. 生物学通报, 2012, 47(4): 9-11.
- [2] Ashburner M. Drosophila genomes by the baker's dozen [J]. *Genetics*, 2007, 177(3): 1263-1268.
- [3] 续晨, 蔡小宁, 钱保俐, 等. 葡萄基因组密码子使用偏好模式研究 [J]. 西北植物学报, 2012, 32(2): 409-415.
- [4] 涂正伟, 周渭渭, 单淇, 等. 刺五加的研究进展 [J]. 药物评价研究, 2011, 34(3): 213-216.
- [5] 邢朝斌, 龙月红, 吴鹏, 等. 刺五加环阿屯醇合酶基因的克隆及其表达分析 [J]. 中草药, 2012, 43(7): 1387-1392.
- [6] 邢朝斌, 吴鹏, 陈龙, 等. 刺五加 GAPDH 基因的克隆及序列分析 [J]. 中草药, 2012, 43(1): 155-158.
- [7] 龙月红, 邢朝斌, 梁能松, 等. 刺五加鲨烯合酶基因 2 cDNA 的克隆与蛋白结构分析 [J]. 北方园艺, 2012(2): 117-120.
- [8] 龙月红, 邢朝斌, 王明艳, 等. 刺五加鲨烯合酶基因 cDNA 的克隆与序列分析 [J]. 生物技术通报, 2012(2): 112-116.
- [9] 周晨光. 刺五加胚性细胞团形成差异表达基因的筛选及克隆 [D]. 哈尔滨: 东北林业大学, 2011.
- [10] 尚明照, 刘方, 华金平, 等. 陆地棉叶绿体基因组密码子使用偏性的分析 [J]. 中国农业科学, 2011, 44(2): 245-253.
- [11] Yi D K, Lee H L, Sun B Y, *et al.* The complete chloroplast DNA sequence of *Eleutherococcus senticosus* (Araliaceae); comparative evolutionary analyses with other three asterids [J]. *Mol Cells*, 2012, 33(5): 497-508.