七种药用植物 *c4h* 基因密码子偏好性及聚类分析

王　晖[1],高德满[2],高妍夏[1],李学军[3],马博辉[4],杨贵明[1]
(1.承德医学院 蚕业研究所,河北省高校特产蚕桑应用技术研发中心,河北 承德 067000;2.锡林郭勒盟第二中学,内蒙古 锡林浩特 026000;3.承德医学院 中药研究所,河北 承德 067000;4.承德医学院 图书馆,河北 承德 067000)

摘要:为了深入研究药用植物基因特性,利用 CodonW 和 SPSS18.0 软件对丹参、黄芪、黄芩、忍冬、桑树、野甘草、长春花共 7 种药用植物 *c4h* 基因密码子进行偏好性和聚类分析。结果表明:7 种药用植物 *c4h* 基因的有效密码子数介于 46.1～55.92,密码子使用偏好性弱,5 种药用植物倾向于 GC 结尾的密码子,2 种倾向于 AT 结尾的密码子。7 种药用植物经聚类分析分为两大类,长春花与黄芪为一大类,另一大类中,丹参与黄芩为一类,野甘草、桑树和忍冬为一类。

关键词:*c4h* 基因;密码子偏好性;聚类分析
中图分类号:Q949.95　**文献标识码**:A　**文章编号**:1002-2767(2015)03-0005-04　DOI:10.11942/j.issn1002-2767.2015.03.0005

肉桂酸羟化酶(cinnamate4-hydroxylase,C4H)是一种 P450 酶,是苯丙烷途径的关键酶,在植物生长、合成多种次生代谢物中起着重要作用[1]。我国是药用植物大国,目前所需的许多化合物主要从药用植物中直接提取,将化合物合成相关基因克隆到表达系统中,用于产生所需要的化合物已成为重要的研究方向。目前美日等发达国家在这方面走在前列,并将一些重要基因注册专利,我国现已完成人参及丹参等的基因组测序。由于药用植物种类多、基因组测序费用昂贵等因素制约了药用植物基因的研究,针对这种现状,应选择用量大且药效显著的药用植物进行重点研究。由于不同的生物物种或同一生物不同的基因具有密码子偏好性,考虑到要将所克隆的基因导入表达系统,因此对所要研究的基因进行密码子偏好性分析很有必要。本研究选择了丹参等 7 种常用药用植物的 *c4h* 基因 CDS(Coding sequence,编码序列)区进行密码子偏好性分析,为进一步将其导入表达系统中进行研究奠定基础。

1　材料与方法

1.1　材料

丹参、黄芪、黄芩、忍冬、桑树、野甘草、长春花的 *c4h* 基因均来自 Genbank,各基因的相关信息见表 1。本研究中用到的软件有密码子分析软件 CodonW、数据分析软件 SPSS18.0。

表 1　7 种药用植物 *c4h* 基因的序列信息
Table 1　Sequence information of *c4h* genes in seven medicinal plants

物种 Species	Genbank 登录号 Accession No. on Genbank	CDS 区位置 Position of CDS
丹参 *Salvia miltiorrhiza*	DQ355979.1	71-1585
长春花 *Catharanthus roseus*	Z32563.1	53-1570
野甘草 *Scoparia dulcis*	KF306081.1	91-1608
黄芩 *Scutellaria baicalensis*	HM062778.1	1-1524
桑树 *Morus alba*	KJ616396.1	1-1518
黄芪 *Astragalus mongholicus*	HQ339960	59-1576
忍冬 *Lonicera japonica*	JX068604.1	1-1518

1.2　方法

从 Genbank 各物种 *c4h* 基因的序列中去除 5'UTR 和 3'UTR,选择 CDS 区,输出 CDS 区序列信息,格式为 fasta,导入到 CodonW 软件包中,对 7 条 *c4h* 基因 CDS 序列进行分析,获得相对同义密码子使用频率(RSCU)、有效密码子数(ENc)、密码子第 3 位 G+C 含量(GC3s)以及 CDS 区 GC 含量,比对 7 种药用植物的试验数据,分析试验结果。利用 SPSS18.0 软件对丹参等药用植物的 *c4h* 基因进行聚类分析,以各物种 *c4h* 基因密码子的 RSCU 值为变量,选择分类中的系统聚类,聚类方法为组间联接,输出分析结果。

2　结果与分析

2.1　*c4h* 基因 GC 含量、ENc、GC3s 的分析

分析结果表明,7 种药用植物的 *c4h* 基因 GC 含量在 50%左右,ENc 均大于 40,接近 60,说明

收稿日期:2014-11-20
第一作者简介:王晖(1985-),男,山西省大同县人,硕士,研究实习员,从事生物学相关研究。E-mail:wanghui861@126.com。

密码子偏好性比较弱[2]，丹参密码子第三位是GC的比例最高，最低为黄芪，说明丹参 *c4h* 基因密码子中的GC主要集中在密码子的第三位，倾向于使用GC结尾的密码子，黄芪 *c4h* 基因的密码子第3位主要是AT，倾向AT结尾的密码子（见表2）。

2.2 密码子使用偏好性分析

分析基因的CDS区，发现丹参、忍冬和野甘草的终止密码子是UGA，黄芪和黄芩的终止密码子是UAA，桑树和长青花的终止密码子是UAG。丹参有27个密码子的RSCU值大于1，长春花有25个，野甘草有26个，黄芩有28个，桑树有26个，黄芪有30个，忍冬有27个。在 *c4h* 基因同一氨基酸的同义密码子中，也存在偏好一种或数种密码子的情况，亮氨酸（Leu）偏好UUG、CUU、CUC三种密码子，异亮氨酸（Ile）偏好AUU、AUC两种密码子，组氨酸（His）偏好CAC，精氨酸（Arg）偏好AGG（见表3）。

表2　7种药用植物 *c4h* 基因密码子与GC含量分析

Table 2　Analysis of codon and GC content in seven medicinal plants

物种 Species	GC	ENc	GC3s
丹参 *Salvia miltiorrhiza*	0.558	46.1	0.752
长春花 *Catharanthus roseus*	0.426	51.2	0.422
野甘草 *Scoparia dulcis*	0.514	55.03	0.638
黄芩 *Scutellaria baicalensis*	0.541	49.21	0.707
桑树 *Morus alba*	0.541	49.93	0.728
黄芪 *Astragalus mongholicus*	0.425	51.61	0.368
忍冬 *Lonicera japonica*	0.494	55.92	0.604

表3　各种有效密码子数量及相对同义密码子使用频率分析

Table 3　Effective number of codons and relative synonymous codon usage frequency

氨基酸 Amino acid	密码子 Codons	丹参 *Salvia miltiorrhiza*	长春花 *Catharanthus roseus*	野甘草 *Scoparia dulcis*	黄芩 *Scutellaria baicalensis*	桑树 *Morus alba*	黄芪 *Astragalus mongholicus*	忍冬 *Lonicera japonica*
Phe	UUU	5/0.37	18/**1.24**	8/0.59	4/0.31	9/0.72	14/**1.04**	13/1.00
	UUC	22/**1.63**	11/0.76	19/**1.41**	22/**1.69**	16/**1.28**	13/0.96	13/1.00
Leu	UUA	1/0.10	6/0.62	2/0.20	1/0.11	0/0.00	3/0.30	5/0.51
	UUG	11/**1.14**	18/**1.86**	11/**1.10**	8/0.84	18/**1.74**	15/**1.50**	17/**1.73**
	CUU	8/0.83	17/**1.76**	12/**1.20**	8/0.84	9/0.87	18/**1.80**	10/1.02
	CUC	23/**2.38**	4/0.41	24/**2.40**	25/**2.63**	21/**2.03**	14/**1.40**	17/**1.73**
	CUA	6/0.62	9/0.93	2/0.20	6/0.63	3/0.39	8/0.80	6/0.61
	CUG	9/0.93	4/0.41	9/0.90	9/0.95	11/**1.06**	2/0.20	4/0.41
Ile	AUU	11/**1.06**	19/**1.68**	17/**1.55**	12/**1.13**	11/**1.03**	20/**1.71**	14/**1.31**
	AUC	17/**1.65**	12/**1.06**	14/**1.27**	15/**1.41**	19/**1.78**	11/0.94	16/**1.50**
	AUA	3/0.29	3/0.26	2/0.18	5/0.47	2/0.19	4/0.34	2/0.19
Met	AUG	13/1.00	12/1.00	11/1.00	14/1.00	12/1.00	11/1.00	15/1.00
Val	GUU	7/0.72	20/**2.22**	9/0.95	4/0.42	6/0.63	24/**2.67**	11/**1.19**
	GUC	11/**1.13**	3/0.33	15/**1.58**	13/**1.37**	13/**1.37**	6/0.67	9/0.97
	GUA	1/0.10	4/0.44	4/0.42	2/0.21	3/0.32	0/0.00	7/0.76
	GUG	20/**2.05**	9/1.00	10/**1.05**	19/**2.00**	16/**1.68**	6/0.67	10/**1.08**
Ser	UCU	0/0.00	5/**1.50**	1/0.29	0/0.00	1/0.30	5/**1.43**	1/0.38
	UCC	1/0.33	2/0.60	6/**1.71**	2/0.63	3/0.90	1/0.29	4/1.50
	UCA	2/0.67	2/0.60	2/0.57	4/**1.26**	2/0.60	4/**1.14**	1/0.38
	UCG	8/**2.67**	2/0.60	3/0.86	2/0.63	5/**1.50**	0/0.00	1/0.38
	AGU	1/0.33	6/**1.80**	2/0.57	5/**1.58**	1/0.30	6/**1.71**	5/**1.88**
	AGC	6/**2.00**	3/0.90	7/**2.00**	6/**1.09**	8/**2.40**	5/**1.43**	4/**1.50**
Pro	CCU	6/0.86	12/**1.71**	9/**1.24**	4/0.57	9/**1.29**	11/**1.57**	3/0.44
	CCC	6/0.86	2/0.29	11/**1.52**	8/**1.14**	8/**1.14**	2/0.29	7/**1.04**
	CCA	3/0.43	8/**1.14**	4/0.55	7/1.00	6/0.86	15/**2.14**	6/0.89
	CCG	13/**1.86**	6/0.86	5/0.69	9/**1.29**	5/0.71	0/0.00	11/**1.63**
Thr	ACU	3/0.63	11/**2.32**	4/0.80	2/0.38	5/**1.11**	6/**1.26**	5/0.87
	ACC	6/**1.26**	3/0.63	6/**1.20**	10/**1.90**	9/**2.00**	7/**1.47**	10/**1.74**
	ACA	4/0.84	5/**1.05**	6/**1.20**	5/0.95	2/0.44	5/**1.05**	4/0.70
	ACG	6/**1.26**	0/0.00	4/0.80	4/0.76	2/0.44	1/0.21	4/0.70
Ala	GCU	5/0.71	5/0.87	4/0.62	4/0.52	4/0.62	12/**1.71**	6/0.86
	GCC	11/**1.57**	11/0.91	10/**1.54**	16/**2.06**	13/**2.00**	4/0.57	10/**1.43**
	GCA	2/0.29	5/0.87	5/0.77	4/0.52	4/0.62	8/**1.14**	8/**1.14**
	GCG	10/**1.43**	2/0.35	7/**1.08**	7/0.90	5/0.77	4/0.57	4/0.57

续表 3 **Continuing Table** 3

氨基酸 Amino acid	密码子 Codons	丹参 *Salvia miltiorrhiza*	长春花 *Catharanthus roseus*	野甘草 *Scoparia dulcis*	黄芩 *Scutellaria baicalensis*	桑树 *Morus alba*	黄芪 *Astragalus mongholicus*	忍冬 *Lonicera japonica*
Tyr	UAU	2/0.33	7/0.93	5/0.77	4/0.67	2/0.31	7/**1.27**	7/**1.08**
	UAC	10/**1.67**	8/**1.07**	8/**1.23**	8/**1.33**	11/**1.69**	4/0.73	6/0.92
His	CAU	1/0.17	3/0.60	4/0.73	1/0.18	3/0.67	5/0.83	3/0.60
	CAC	11/**1.83**	7/**1.40**	7/**1.27**	10/**1.82**	6/**1.33**	7/**1.17**	7/**1.40**
Gln	CAA	1/0.11	6/0.75	7/0.74	3/0.33	3/0.33	9/**1.06**	8/0.94
	CAG	18/**1.89**	10/**1.25**	12/**1.26**	15/**1.67**	15/**1.67**	8/0.94	9/**1.06**
Asn	AAU	6/0.46	16/**1.23**	9/0.72	7/0.52	6/0.43	14/**1.12**	11/0.79
	AAC	20/**1.54**	10/0.77	16/**1.28**	20/**1.48**	22/**1.57**	11/0.88	17/**1.21**
Lys	AAA	6/0.35	14/0.76	11/0.63	5/0.27	3/0.17	21/**1.20**	8/0.44
	AAG	28/**1.65**	23/**1.24**	24/**1.37**	32/**1.73**	33/**1.83**	14/0.80	28/**1.56**
Asp	GAU	11/0.88	16/**1.45**	13/1.00	12/**1.04**	7/0.64	19/**1.41**	16/**1.23**
	GAC	14/**1.12**	6/0.55	13/1.00	11/0.96	15/**1.36**	8/0.59	10/0.77
Glu	GAA	4/0.23	19/0.93	9/0.51	8/0.44	8/0.42	21/**1.20**	8/0.44
	GAG	31/**1.77**	22/**1.07**	26/**1.49**	28/**1.56**	30/**1.58**	14/0.80	28/**1.56**
Cys	UGU	0/0.00	2/**1.33**	2/**1.33**	1/0.50	1/0.67	2/**1.33**	1/0.67
	UGC	3/2.00	1/0.67	1/0.67	3/**1.50**	2/**1.33**	1/0.67	2/**1.33**
Trp	UGG	8/1.00	7/1.00	8/1.00	8/1.00	8/1.00	8/1.00	8/1.00
Arg	CGU	3/0.53	4/0.69	4/0.73	1/0.18	3/0.51	6/**1.06**	2/0.35
	CGC	8/**1.41**	0/0.00	3/0.55	7/**1.27**	5/0.86	4/0.71	3/0.53
	CGA	4/0.71	4/0.69	3/0.55	2/0.36	3/0.51	3/0.53	2/0.35
	CGG	5/0.88	0/0.00	5/0.91	7/**1.27**	7/**1.20**	1/0.18	8/**1.41**
	AGA	7/**1.24**	15/**2.57**	4/0.73	10/**1.82**	5/0.86	10/**1.76**	5/0.88
	AGG	7/**1.24**	12/**2.06**	14/**2.55**	6/**1.09**	12/**2.06**	10/**1.76**	14/**2.47**
Gly	GGU	4/0.46	9/**1.06**	3/0.38	1/0.13	3/0.35	15/**1.82**	4/0.52
	GGC	18/**2.06**	10/**1.18**	13/**1.63**	12/**1.50**	15/**1.76**	6/0.73	6/0.77
	GGA	3/0.34	11/**1.29**	9/**1.13**	10/**1.25**	8/0.94	12/**1.45**	9/**1.16**
	GGG	10/**1.14**	4/0.47	7/0.88	9/**1.13**	8/0.94	0/0.00	12/**1.55**

"/"前的数字表示各物种 *c4h* 基因有效密码子的数量,"/"后的数字表示 RSCU 值,加粗数字表示 RSCU 值大于 1。

The number before"/"means the offective codon number of *c4h* gene, the number after"/"means RSCU, the coarse number means RSCU greater than 1.

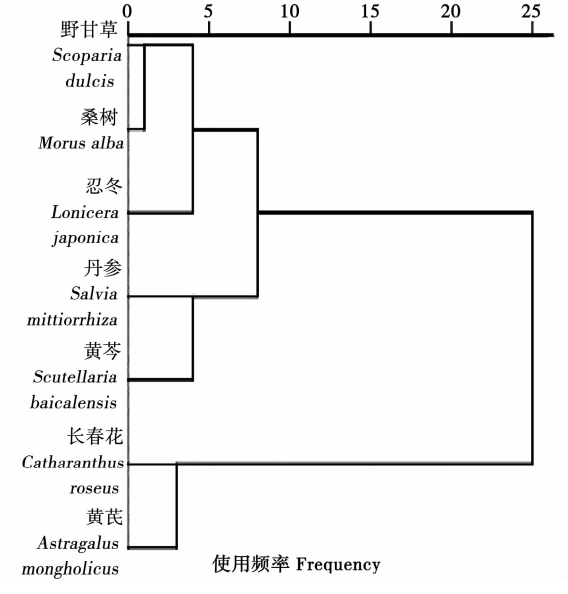

图 1　基于 *c4h* 基因相对同义密码子使用频率的聚类分析

Fig. 1　Cluster analysis of RSCU for *c4h* genes

2.3　药用植物 *c4h* 基因密码子偏好性聚类分析

从 SPSS 软件分析结果可知,7 种药用植物分为两大类,长春花与黄芪为一大类,另一大类中,丹参与黄芩为一类,野甘草、桑树和忍冬为一类(见图 1),在植物分类学中,忍冬、丹参、黄芩和长春花均属于合瓣花亚纲,丹参与黄芩又都属于管状花目唇形科。聚类分析与传统分类结果有一定的差距,可能是由于本研究的实验材料或基因样本数量不够大。

3　结论与讨论

生物体的 20 种氨基酸是由 61 种密码子编码,除了甲硫氨酸和色氨酸之外的氨基酸都有2～6 种同义密码子,在不同的物种和基因组甚至同一基因组的不同基因间,同义密码子的使用频率并不相同,主要表现为密码子使用的偏好性,ENc 被广泛地用来表示密码子的偏好性[3]。密码子的第 1、2 位大多由自然选择决定,而第 3 位则受突

变压力的影响，病毒等原核生物密码子的第3位偏好A和T，双子叶植物和哺乳动物偏好G和C[4]，基因的GC和GC3值越高，突变偏性和翻译选择在密码子偏好性选择上起的作用越大[5]，研究表明突变和密码子偏好性决定人致癌基因的表达水平，高表达的致癌基因含有高GC含量和高度密码子偏好性。此外，密码子偏好性也受如基因表达水平、RNA结构、基因长度、进化压力、蛋白质亲水性和蛋白质二级结构等因素的影响[6]。

RSCU值大于1表示对应的密码子使用频率较高，小于1则表示较低。本研究中RSCU值最高的是黄芩亮氨酸CUC，最低的是丹参缬氨酸GUA和亮氨酸UUA。在进行转基因研究时，应结合目的基因密码子偏好程度，选择与表达载体偏好密码子匹配程度最好的密码子。在7种药用植物中，丹参*c4h*基因的密码子偏好性最强，忍冬最弱。同义密码子使用的选择可以为基因预测、分类、进化和设计高表达的基因及载体提供有价值的参考，本研究首次对丹参等药用植物*c4h*基因的密码子偏好性进行了分析，获得了*c4h*基因的ENc、RSCU等试验数值，尝试利用RSCU值构建丹参等的聚类分析图，并结合传统分类学进行有意义的探讨，为试验材料*c4h*基因在基因工程技术应用研究奠定了基础。

参考文献：

[1] 王树斌，全雪丽，具红光，等. 膜荚黄芪肉桂酸-4-羟化酶(*c4h*)基因克隆与序列分析[J]. 延边大学农学学报，2012，34(4)：277-281.

[2] Hu C，Chen J，Ye L，et al. Codon usage bias in human cytomegalovirus and its biological implication[J]. Gene，2014，545(1)：5-14.

[3] Mazumder T H，Chakraborty S，Paul P. A cross talk between codon usage bias in human oncogenes[J]. Bioinformation，2014，10(5)：256-262.

[4] Chen H，Sun S，Norenburg J L，et al. Mutation and selection cause codon usage and bias in mitochondrial genomes of ribbon worms(Nemertea)[J]. PLOS ONE，2014，9(1)：e85631.

[5] Mirsafian H，Mat Ripen A，Singh A，et al. A comparative analysis of synonymous codon usage bias pattern in human albumin superfamily[J]. The Scientific World Journal，2013，2014：639682.

[6] Behura S K，Severson D W. Codon usage bias：causative factors，quantification methods and genome-wide patterns：with emphasis on insect genomes[J]. Biological Reviews，2013，88(1)：49-61.

Codon Usage Bias and Cluster Analysis of *c4h* Genes in Seven Medicinal Plants

WANG Hui[1]，GAO De-man[2]，GAO Yan-xia[1]，LI Xue-jun[3]，MA Bo-hui[4]，YANG Gui-ming[1]

(1. The Sericultural Research Institute，Hebei Universities Technology Research and Development Center of Sericulture，Chengde Medical University，Chengde，Hebei 067000；2. The Second Middle School of Xilingol League，Xilinhot，Inner Mongolia 026000；3. Institute of Chinese Traditional Medicine，Chengde Medical University，Chengde，Hebei 067000；4. Library of Chengde Medical University，Chengde，Hebei 067000)

Abstract：In order to further study genetic features of medicinal plants，coding usage of *c4h* genes in medicial plants including *Salvia miltiorrhiza*，*Catharanthus roseus*，*Scoparia dulcis*，*Scutellaria baicalensis*，*Morus alba*，*Astragalus mongholicus*，*Lonicera japonica* were analyzed by CodonW. Cluster analysis of *c4h* genes codon bias were carried out by SPSS18.0. The result showed that *c4h* genes ENc were between 46.1 and 55.92 in seven medical plants，they had a weak codon bias，5 medicinal plants tended to use codon ending with GC，2 medicinal plants tends to use codon ending with AT. Seven medicinal plants were divided into two parts，one part included *Catharanthus roseus* and *Astragalus mongholicus*. In the other part，*Salvia miltiorrhiza* and *Scutellaria baicalensis* were in a group，*Scoparia dulcis*，*Morus alba* and *Lonicera japonica* were in another group.

Keywords：*c4h* genes；codon bias；cluster analysis