浙 江 农 林 大 学 学 报，2017，**34**(1)：120–128
Journal of Zhejiang A & F University
doi:10.11833/j.issn.2095-0756.2017.01.017

基于全基因组的毛竹同义密码子使用偏好性分析

黄笑宇，许在恩，郭小勤

（浙江农林大学 亚热带森林培育国家重点实验室培育基地，浙江 临安 311300）

摘要：密码子使用偏好性是物种在遗传信息传递过程中的一个重要特点，分析物种的密码子使用偏好性对于了解该物种遗传信息的传递规律具有重要意义。应用 CodonW 软件对毛竹 *Phyllostachys edulis* 基因组中的 26 103 个蛋白质编码基因序列进行了分析，计算了位于密码子 3 个位置的 G+C 含量、有效密码子数、同义密码子的使用频率等，确定了毛竹的最优密码子。结果显示：毛竹密码子第 1 位和第 3 位的 G+C 含量明显高于第 2 位，表现出对以 G 或 C 碱基开头和结尾的密码子发生强烈偏向使用，且确定的 26 种最优密码子均以 G/C 碱基结尾。与模式动植物酵母 *Saccharomys cerevisiae*，大肠埃希菌 *Escherichia coli*，果蝇 *Drosophila melanogaster*，拟南芥 *Arabidopsis thaliana*，烟草 *Nicotiana tabacum*，水稻 *Oryza sativa*，玉米 *Zea mays*，小麦 *Triticum aestivum* 等 8 个代表性物种进行比较，结果显示：毛竹密码子偏好性与模式动植物存在不同程度的差异，与大肠埃希菌、酵母、果蝇、拟南芥、烟草和玉米的差异较大，差异分别为 8 个、11 个、6 个、7 个、6 个和 6 个，而与小麦的差异较小(1 个)，与水稻完全一致。密码子偏好性差异大小在一定程度上反映物种间的进化关系。图 1 表 4 参 29

关键词：植物学；毛竹；密码子偏好性；最优密码子

中图分类号：S722；S795.7　　**文献标志码**：A　　**文章编号**：2095-0756(2017)01-0120-09

Synonymous codon bias of *Phyllostachys edulis*

HUANG Xiaoyu, XU Zaien, GUO Xiaoqin

（The Nurturing Station for the State Key Laboratory of Subtropical Silviculture, Zhejiang A & F University, Lin'an 311300, Zhejiang, China）

Abstract: Analysis of codon usage bias for different species, an important characteristic of genetic information transfer in organisms, is important for understanding the rules of genetic information transfer. To better understand the characteristics of *Phyllostachys edulis*, coding DNA sequences （CDS）of 26 103 proteins in this species were analyzed. The content of G+C at three positions of codons' the Effective number of codons（ENC）and frequency of synonym codon usage for genes were calculated the "optimal codons"were determined by Codon W, CHIPS and CUSP. Then, the frequency of codon usage for *Ph. edulis* with other organisms including five model value species （*Drosophila melanogaster*, *Saccharomys cerevisiae*, *Escherichia coli*, *Arabidopsis thaliana*, and *Nicotiana tabacum*）and three other Gramineae species （*Oryza sativa*, *Zea mays*, and *Triticum aestivum*）were compared. Results showed that the content of G+C at the first and third position of codons was much higher than that of the second positions, and genes preferred codons with C or G in the synonymous position. Meanwhile, 26 codons, ending with G or C, were determined as the "optimal codons". Frequency of codon usage also had fewer differences for the three Gramineae species than it did for the six model value species. To some extent, differences in the size of codon bias reflected the evolutionary relationships between species. These results provided the useful information for understanding the evolution of *Ph. edulis* [Ch,

收稿日期：2016-02-29；修回日期：2016-04-20

基金项目：浙江省自然科学基金资助项目(Y307499，LY13C160011)；国家自然科学基金资助项目(30901155)；浙江省大学生新苗人才计划项目(2015R412020)

作者简介：黄笑宇，从事林学研究。E-mail：593796107@qq.com。通信作者：郭小勤，副教授，博士，从事竹类植物遗传育种等研究。E-mail：xqguo@zafu.edu.cn

1 fig. 4 tab. 29 ref.]
Key words: botany; *Phyllostachys edulis*; codon bias; optimal codon

三联密码子是整个生物王国的核心，作为最基本的编码组分编码特定的氨基酸。除了甲硫氨酸 Met 和色氨酸 Trp 外，同一个氨基酸会由 2~6 个同义密码子编码[1]。根据中心法则，尽管同义突变不会引起蛋白序列的变化，但同义密码子使用偏好性在基因组内和基因组间广泛存在[2-4]。密码子使用偏好性是物种在遗传信息传递过程中的一个重要特点，分析物种的密码子使用偏好性对于了解该物种遗传信息的传递规律具有重要意义。密码子使用偏好性的研究有助于更好地理解分子生物及进化，信使核糖核酸(mRNA)翻译，转基因设计，新基因发现，以及其他生物应用[3-6]。几十年来，大量模式物种如拟南芥 *Arabidopsis thaliana*，水稻 *Oryza sativa*，果蝇 *Drosophila melanogaster*，杨树 *Populus trichocarpa* 等的测序产生了大量的开放阅读框，这些全长编码序列作为密码生物学的基础，为研究密码子使用模式提供了强有力的保障。随着深度测序技术的快速发展，非模式植物也纷纷被测序，产生了大量的序列，有关密码子使用模式的工作也逐渐拉开序幕[7]。毛竹 *Phyllostachys edulis* 为禾本科 Gramineae 多年生木本植物，地上部分可材用，地下部分发育的笋可食用。毛竹基因组序列测定[8]，产生了大量基因组及编码区序列，为研究毛竹密码子的使用情况提供了大量的信息。本研究通过分析毛竹全基因组编码序列数据，了解毛竹基因密码子用法特征，并与不同代表性物种进行比较。这些分析有助于我们理解毛竹的密码子模式，提升植物密码子使用的研究，同时为毛竹基因选择合适的表达系统，优化密码子提高基因表达量等提供重要理论基础。

1　材料与方法

1.1　序列数据

从 http://202.127.18.221/bamboo/index.php 的毛竹基因组注释数据中获取了 31 987 条蛋白质对应的编码基因序列(coding DNA sequence，CDS)，从中挑选出以 ATG 为起始密码子，以 TAA，TAG 或 TGA 为终止密码子的，且 CDS 长度大于 300 bp 的 26 103 个基因作为序列分析样本[9]。这个数据库中已经去除所有假基因的信息。

1.2　序列处理

采用 C 语言编写程序进行序列筛选与处理。

1.3　同义密码子使用偏好性分析

采用 EMBOSS 软件包中的 CHIPS 和 CUSP 程序在线 http://emboss.bioinformatics.nl 及 CodonW1.4.4 (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::codonw) 对毛竹全基因编码序列进行分析，计算有效密码子数(effective number of codons，Enc)，CDS 区的 GC 含量，密码子中第 3 位碱基的 GC 含量(GC3s)，同义密码子相对使用频率 (relative synonymous codon usage，RSCU)及密码子使用概率。

衡量同义密码子使用偏好性参数的含义：①有效密码子数(Enc)。该值被认为是在评价基因整体密码子偏好性用法中最具有参考价值的参数之一，目前被广泛用于评价基因密码子偏好性，其取值范围为 20(每个氨基酸只使用 1 个密码子的极端情况)到 61(各个密码子均被平均使用)[10]。②同义密码子相对使用频率 RSCU。该值的计算方法为某一密码子所使用的频率与其在无偏好使用时预期频率之间的比值，若某一密码子的 RSCU 值等于 1，则表明该密码子的使用没有偏好性；RSCU 值大于 l，表明该密码子的使用频率相对较高，反之亦然。它去除了氨基酸组成对密码子使用的影响，且直观地反映了密码子使用的偏好性[11]。③同义密码子使用的绝对频率(Fract)。该值表示各个密码子在编码该氨基酸的密码子中所占的比例(各比例相加总和为 1)[12]。

1.4　最优密码子的确定

采用 STENICO 等[13]的方法，把密码子使用偏好性强和弱的 2 组基因之间相应密码子出现频率之差达到统计学上显著水平的密码子定义为最优密码子。具体方法如下：通过计算样本中每个基因的有效密码子数，并按该值的大小对基因进行排列，从这一排列的两端各取基因样本总数的 5%，分别组成高、低表达样本组。计算这 2 组基因的相对密码子使用度，并进行卡方检验，确定最优密码子[14]。

1.5 毛竹与其他物种密码子偏好性比较

运用 CUSP 程序计算毛竹基因各密码子的使用频率，并与从 Codon Usage Database(http://www.kazusa.or.jp/codon/)中获得的果蝇，酵母 *Saccharomys cerevisiae*，大肠埃希菌 *Escherichia coli*，拟南芥，烟草 *Nicotiana tabacum*，水稻，玉米 *Zea mays*，小麦 *Triticum aestivum* 等的密码子使用频率进行比较。密码子使用频率若为 0.5~2.0，表明这 2 个物种对该密码子的偏好性较接近，若≥2.0 或≤0.5，则表明偏好性差异较大[15]。

1.6 基于密码子使用偏好性的聚类

利用 SPSS 19.0 对毛竹及其他 9 个物种进行基于密码子使用偏好性的聚类分析，方法参考文献[15]。

2 结果与分析

2.1 有效密码子数与密码子的碱基组成

将毛竹基因组注释数据中获取到的 31 987 条蛋白质对应的编码基因序列进行筛选后，获得 26 103 条有效序列，将这些序列作为一整体，在线计算了其有效密码子数及密码子第 1 位、第 2 位、第 3 位和 3 个位置平均的碱基 GC 百分率，结果见表 1。毛竹基因整体的有效密码子数为 57.88，表明毛竹整体基因的密码子存在一定程度偏好，但偏好性不强。从密码子的 GC 含量来看，3 个位置平均 GC 含量为 0.52。其中，GC_1 含量为 0.56，比 GC_2(0.44)高 0.12，而 GC_3 含量(0.57)比 GC_1 略高 0.01，表明选择压力使得毛竹密码子的第 1 位倾向于选择 G/C，第 2 位倾向于选择 T/A，第 3 位可以有大幅摆动。

表 1 毛竹基因密码子中 3 个位置的 GC 含量及有效密码子数值

Table 1 GC content of different positions and effective number of codons in *Phyllostachys edulis*

基因/个	密码子/个	第 1 位置	第 2 位置	第 3 位置	3 位置平均	有效密码子数
26 103	11 634 464	0.56	0.44	0.57	0.52	57.88

2.2 同义密码子的使用频率

经软件计算的同义密码子使用次数及频率结果见表 2。在 64 个密码子中，GAG 是出现次数最高的密码子，绝对频率为 39.21，是毛竹平均频率的 2.51 倍；紧随其后的是 AAG，达到 33.90；处于第 3 位的是 GAU，为 28.40；CGA 的出现频率最低，仅为 5.34，是毛竹平均频率的 1/3；另有 6 个密码子(UUA，CUA，GUA，ACG，UGU，CGU)的频率小于 10.00(表 2)。

有 34 个密码子的 RSCU 值大于 1，这些密码子为毛竹基因的偏好密码子，其中约 1/3 的密码子以 A/U 结尾，2/3 的密码子以 G/C 结尾。AGG(编码 Arg)，CUC(编码 Leu)和 GUG(编码 Val)的 RSCU 值处于前 3 位，分别为 1.57，1.45 和 1.39。CUG(编码 Cys)以及 AAG(编码 Lys)和 GGC(编码 Gly)相对于其同义密码子的使用频率高，分别为 1.36 和 1.31。这 5 个密码子为本文的高频率密码子。

4 个 NUA 密码子的 RSCU 值最低，AUA 为 0.69，CUA 为 0.52，GUA 为 0.47，UUA 为 0.45，表明这几个是毛竹基因避免使用的密码子。4 个 NCG 的 RSCU 值相对来说接近于平均水平甚至更低，CCG 为 0.99，GCG 为 0.91，ACG 为 0.76，UCG 为 0.75，表明毛竹体内的甲基化水平可能较低或中等，这点从 NCG:NCC 的比值(为 0.82)也可看出。终止密码子 UGA 在毛竹基因中的使用频率较其余 2 个终止密码子高，为 1.02，其次是 UAG，RSCU 值为 0.67，UAA 的使用频率最低，仅为 0.55。

2.3 最优密码子的确定

不仅同义密码子间存在偏好性，且密码子本身的使用也存在偏好性。目前，关于毛竹基因表达的数据偏少，多数转录组测序的数据也基于几个毛竹的特异组织。因此，本研究依据 Enc 值来衡量基因的表达量。表 3 中的结果是通过计算高表达/低表达基因之间同义密码子相对使用频率之差，经卡方测验确定的毛竹中的最优密码子，用 * 号标记，共 26 个。这些密码子均以 G/C 结尾，表明在高表达基因中优先使用这些密码子。这些密码子的使用频率在高表达基因组与低表达基因组之间的差异达到极显著水平。

2.4 毛竹与模式动植物密码子偏好性比较

将毛竹与 3 种模式生物大肠埃希菌、酵母和果蝇密码子使用频率比较，比值 0.5~2.0 表明 2 物种使

表 2　毛竹基因同义密码子的使用频率

Table 2　Frequency of synonymous codons in genes of *Phyllostachys edulis*

氨基酸	密码子	出现次数/次	绝对频率	相对频率（RSCU）	氨基酸	密码子	出现次数/次	绝对频率	相对频率（RSCU）
苯丙氨酸 Phe	UUU	177 259	15.202	0.84	酪氨酸 Tyr	UAU	127 108	10.901	0.85
	UUC	245 828	21.082	1.16		UAC	171 249	14.686	1.15
亮氨酸 Leu	UUA	83 127	7.129	0.45	TER	UAA	6 422	0.551	0.74
	UUG	200 316	17.179	1.08		UAG	7 782	0.667	0.89
	CUU	213 205	18.284	1.14	组氨酸 His	CAU	141 958	12.174	1.00
	CUC	269 853	23.142	1.45		CAC	140 843	12.079	1.00
	CUA	97 208	8.336	0.52	谷氨酰胺 Gln	CAA	167 201	14.339	0.79
	CUG	254 016	21.784	1.36		CAG	254 576	21.832	1.21
异亮氨酸 Ile	AUU	192 698	16.526	1.12	天冬酰胺 Asn	AAU	211 015	18.096	0.99
	AUC	206 620	17.720	1.20		AAC	216 215	18.542	1.01
	AUA	118 527	10.165	0.69	赖氨酸 Lys	AAA	207 746	17.816	0.69
蛋氨酸 Met	AUG	278 037	23.844	1.00		AAG	395 286	33.899	1.31
缬氨酸 Val	GUU	215 712	18.499	1.10	天冬氨酸 Asp	GAU	331 197	28.403	1.05
	GUC	205 738	17.644	1.04		GAC	297 586	25.521	0.95
	GUA	92 064	7.895	0.47	谷氨酸 Glu	GAA	281 313	24.125	0.76
	GUG	274 313	23.525	1.39		GAG	457 181	39.207	1.24
丝氨酸 Ser	UCU	179 360	15.382	1.10	半胱氨酸 Cys	UGU	81 296	6.972	0.74
	UCC	183 357	15.725	1.12		UGC	138 420	11.871	1.26
	UCA	177 440	15.217	1.09	TER	UGA	11 899	1.020	1.37
	UCG	121 924	10.456	0.75	色氨酸 Trp	UGG	152 825	13.106	1.00
脯氨酸 Pro	CCU	165 860	14.224	1.07	精氨酸 Arg	CGU	74 482	6.388	0.62
	CCC	125 023	10.722	0.81		CGC	138 100	11.843	1.14
	CCA	173 732	14.899	1.13		CGA	62 215	5.336	0.51
	CCG	152 555	13.083	0.99		CGG	125 552	10.767	1.04
苏氨酸 Thr	ACU	141 566	12.141	1.02	丝氨酸 Ser	AGU	127 392	10.925	0.78
	ACC	152 386	13.068	1.10		AGC	189 258	16.231	1.16
	ACA	155 105	13.302	1.12	精氨酸 Arg	AGA	134 872	11.567	1.12
	ACG	105 160	9.018	0.76		AGG	189 633	16.263	1.57
丙氨酸 Ala	GCU	258 646	22.181	1.03	苷氨酸 Gly	GGU	192 842	16.538	0.90
	GCC	284 302	24.381	1.13		GGC	280 247	24.034	1.31
	GCA	235 235	20.174	0.94		GGA	187 574	16.086	0.88
	GCG	227 989	19.552	0.91		GGG	195 121	16.733	0.91

说明：下划线表示 RSCU 大于 1 的密码子。

用该密码子的偏好性相似，比值小于 0.5 或大于 2.0，表明该密码子的使用偏好性差异较大。结果显示：毛竹与大肠埃希菌、酵母和果蝇密码子的比值中，分别有 8，11，6 个小于 0.5 或大于 2.0，表明毛竹与这些模式生物之间的密码子偏好性存在一定差异。

表 4 的结果显示：毛竹与双子叶植物的代表种拟南芥和烟草的密码子偏好性差异性较大，比值大于 2.0 或小于 0.5 的分别有 7 个和 6 个，与同科植物相比，与 C_4 植物玉米的密码子偏好性差异也较大，有 6 个，而与 C_3 植物水稻和小麦的偏好性一致。

2.5　不同物种间密码子偏好性的聚类分析

根据各物种编码序列密码子的使用频率，利用 SPSS 19.0 进行聚类分析(图 1)。从图 1 可以看出：双子叶植物拟南芥和烟草密码子使用偏好更相近，禾本科植物毛竹与水稻的最近，其次与小麦和玉米。利用密码子使用频率得出的聚类结果一定程度上反映了各物种间的进化关系。

表 3 毛竹中高/低表达样本的密码子用法

Table 3 Codon usage of high/low expressed genes in *Phyllostachys edulis*

氨基酸	密码子	高		低		氨基酸	密码子	高		低	
		数量/个	RSCU	数量/个	RSCU			数量/个	RSCU	数量/个	RSCU
Phe	UUU	550	0.07	17 107	1.28	Ser	UCU	780	0.17	21 517	1.67
	UUC*	14 471	1.93	9 706	0.72		UCC*	9 894	2.20	7 839	0.61
Leu	UUA	107	0.02	10 064	0.85		UCA	575	0.13	20 029	1.55
	UUG	1 172	0.19	17 142	1.45		UCG*	7 656	1.70	3 859	0.30
	CUU	1 074	0.18	18 765	1.58		AGU	392	0.08	14 622	1.13
	CUC*	19 835	3.28	6 415	0.54		AGC*	7 735	1.72	9 612	0.74
	CUA	605	0.10	8 271	0.70	Pro	CCU	1 054	0.19	15 839	1.63
	CUG*	13 472	2.23	10 464	0.88		CCC*	7 554	1.36	4 437	0.46
Ile	AUU	536	0.12	18 226	1.45		CCA	1 105	0.20	15 688	1.62
	AUC*	11 396	2.73	8 154	0.65		CCG*	12 521	2.25	2 834	0.29
	AUA	611	0.15	11 292	0.90	Thr	ACU	494	0.11	14 804	1.53
Val	GUU	836	0.11	21 662	1.73		ACC*	8 693	1.92	5 894	0.61
	GUC*	13 280	1.77	7 106	0.57		ACA	587	0.13	15 321	1.59
	GUA	493	0.07	9 334	0.75		ACG*	8 327	1.84	2 583	0.27
	GUG*	15 329	2.05	11 844	0.95	Ala	GCU	1 895	0.17	23 625	1.72
Tyr	UAU	316	0.06	12 303	1.31		GCC*	21 775	1.90	7 404	0.54
	UAC	10 241	1.94	6 541	0.69		GCA	1 754	0.15	20 285	1.47
His	CAU	530	0.12	14 620	1.46		GCG*	20 410	1.78	3 769	0.27
	CAC*	8 171	1.88	5 466	0.54	Cys	UGU	223	0.06	8 175	1.15
Gln	CAA	681	0.13	17 457	1.05		UGC*	6 655	1.94	5 992	0.85
	CAG*	9 443	1.87	15 734	0.95	Arg	CGU	707	0.16	5 582	0.82
Asn	AAU	641	0.12	23 601	1.34		CGC*	12 045	2.74	2 958	0.43
	AAC*	9 785	1.88	11 686	0.66		CGA	565	0.13	4 509	0.66
Lys	AAA	650	0.09	22 589	0.98		CGG*	8 353	1.90	3 555	0.52
	AAG*	13 132	1.91	23 377	1.02	Arg	AGA	427	0.10	13 864	2.03
Asp	GAU	1 393	0.14	33 692	1.46		AGG	4 306	0.98	10 524	1.54
	GAC*	18 250	1.86	12 413	0.54	Gly	GGU	1 432	0.18	17 148	1.37
Glu	GAA	984	0.10	30 614	1.13		GGC*	20 195	2.53	8 623	0.69
	GAG*	18 403	1.90	23 754	0.87		GGA	1 551	0.19	15 565	1.25
							GGG*	8 780	1.10	8 577	0.69

说明：经卡方测验确定的毛竹中的最优密码子。

3 讨论

在长期的进化过程中，不同物种对进化环境和选择压力的适应不同，因此，任何一个物种都会形成特定的密码子用法以适应其基因组环境，最终使其宿主适应外界进化环境。由此，不同物种就形成了各自特定的密码子偏好性。若要通过基因工程技术改造某一物种或将某一基因用于体外表达，应先按照宿主的密码子使用偏好性对所导入的基因进行优化和改造。本研究在毛竹全基因组测序的基础上，对编码蛋白基因的密码子偏好性进行了分析，结果表明与很多物种包括人、细菌、酵母、果

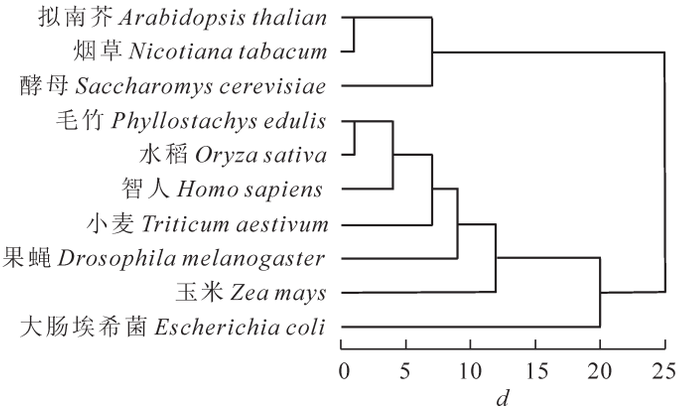

图 1 基于不同物种密码子使用频率的聚类分析

Figure 1 Cluster analysis dendrogram of frenquency of codon usage of different species

表 4　毛竹与模式植物的密码子偏好性比较

Table 4　Comparison of codon preference between bamboo and other model plants

氨基酸	密码子	密码子出现频率/‰						频率比值				
		毛竹	拟南芥	烟草	水稻	玉米	小麦	B/A	B/T	B/R	B/M	B/W
丙氨酸 A	GCA	20.17	17.47	22.93	17.33	13.54	15.69	1.15	0.88	1.16	1.49	1.29
丙氨酸 A	GCC	24.38	10.34	12.65	30.82	41.27	32.36	2.36	1.93	0.79	0.59	0.75
丙氨酸 A	GCG	19.55	9.03	5.94	26.70	37.34	21.82	2.17	3.29	0.73	0.52	0.90
丙氨酸 A	GCT	22.18	28.32	31.74	19.57	15.62	16.36	0.78	0.70	1.13	1.42	1.36
半胱氨酸 C	TGC	11.87	7.16	7.33	12.40	15.46	13.38	1.66	1.62	0.96	0.77	0.89
半胱氨酸 C	TGT	6.97	10.54	9.86	6.20	3.83	5.21	0.66	0.71	1.12	1.82	1.34
天冬氨酸 D	GAC	25.52	17.22	17.07	28.08	38.37	28.73	1.48	1.50	0.91	0.67	0.89
天冬氨酸 D	GAT	28.40	36.65	37.29	25.30	14.58	17.13	0.77	0.76	1.12	1.95	1.66
谷氨酸 E	GAA	24.13	34.34	35.58	21.61	12.79	15.77	0.70	0.68	1.12	1.89	1.53
谷氨酸 E	GAG	39.21	32.24	29.99	38.52	41.97	37.94	1.22	1.31	1.02	0.93	1.03
苯丙氨酸 F	TTC	21.08	20.66	17.95	22.32	28.07	24.65	1.02	1.17	0.94	0.75	0.86
苯丙氨酸 F	TTT	15.20	21.81	24.56	13.05	7.01	12.81	0.70	0.62	1.16	2.17	1.19
苷氨酸 G	GGA	16.09	24.16	23.26	15.91	10.65	15.46	0.67	0.69	1.01	1.51	1.04
苷氨酸 G	GGC	24.03	9.15	11.44	29.54	39.54	30.96	2.63	2.10	0.81	0.61	0.78
苷氨酸 G	GGG	16.73	10.18	10.53	17.13	18.62	18.09	1.64	1.59	0.98	0.90	0.92
苷氨酸 G	GGT	16.54	22.18	22.63	14.83	10.35	13.60	0.75	0.73	1.12	1.60	1.22
组氨酸 H	CAC	12.08	8.72	8.69	13.81	17.61	13.34	1.39	1.39	0.87	0.69	0.91
组氨酸 H	CAT	12.17	13.79	13.24	11.29	7.03	8.56	0.88	0.92	1.08	1.73	1.42
异亮氨酸 I	ATA	10.17	12.60	13.86	8.78	5.39	6.89	0.81	0.73	1.16	1.89	1.48
异亮氨酸 I	ATC	17.72	18.53	13.95	19.35	22.90	24.47	0.96	1.27	0.92	0.77	0.72
异亮氨酸 I	ATT	16.53	21.49	27.51	14.18	7.29	12.04	0.77	0.60	1.17	2.27	1.37
赖氨酸 K	AAA	17.82	30.79	32.22	15.93	8.58	10.62	0.58	0.55	1.12	2.08	1.68
赖氨酸 K	AAG	33.90	32.68	33.92	32.17	32.79	37.98	1.04	1.00	1.05	1.03	0.89
亮氨酸 L	CTA	8.34	9.87	9.20	7.72	5.10	7.54	0.84	0.91	1.08	1.64	1.11
亮氨酸 L	CTC	23.14	16.09	12.44	25.84	32.81	27.48	1.44	1.86	0.90	0.71	0.84
亮氨酸 L	CTG	21.78	9.83	10.28	20.99	33.66	22.26	2.22	2.12	1.04	0.65	0.98
亮氨酸 L	CTT	18.28	24.12	23.97	15.18	9.91	12.68	0.76	0.76	1.20	1.84	1.44
亮氨酸 L	TTA	7.13	12.70	12.82	6.14	2.57	3.90	0.56	0.56	1.16	2.77	1.83
亮氨酸 L	TTG	17.18	20.87	21.90	14.67	8.99	12.40	0.82	0.78	1.17	1.91	1.39
蛋氨酸 M	ATG	23.84	24.53	24.78	23.81	23.76	24.31	0.97	0.96	1.00	1.00	0.98
天冬酰胺 N	AAC	18.54	20.93	18.11	18.49	21.92	21.12	0.89	1.02	1.00	0.85	0.88
天冬酰胺 N	AAT	18.10	22.30	27.54	15.10	7.78	10.38	0.81	0.66	1.20	2.33	1.74
脯氨酸 P	CCA	14.90	16.15	19.40	14.25	10.37	23.49	0.92	0.77	1.05	1.44	0.63
脯氨酸 P	CCC	10.72	5.31	6.63	12.05	15.78	14.64	2.02	1.62	0.89	0.68	0.73
脯氨酸 P	CCG	13.08	8.62	4.99	17.99	24.18	15.63	1.52	2.62	0.73	0.54	0.84
脯氨酸 P	CCT	14.22	18.68	18.39	13.57	9.96	11.83	0.76	0.77	1.05	1.43	1.20
谷氨酰胺 Q	CAA	14.34	19.45	20.70	13.45	7.32	42.52	0.74	0.69	1.07	1.96	0.34
谷氨酰胺 Q	CAG	21.83	15.24	15.26	20.73	25.90	37.85	1.43	1.43	1.05	0.84	0.58
精氨酸 R	AGA	11.57	18.97	15.61	10.52	6.60	6.75	0.61	0.74	1.10	1.75	1.71
精氨酸 R	AGG	16.26	10.96	12.30	15.97	16.44	12.93	1.48	1.32	1.02	0.99	1.26
精氨酸 R	CGA	5.34	6.29	5.27	6.44	3.84	3.06	0.85	1.01	0.83	1.39	1.75
精氨酸 R	CGC	11.84	3.78	3.96	16.18	20.22	12.54	3.13	2.99	0.73	0.59	0.94
精氨酸 R	CGG	10.77	4.87	3.76	13.48	15.65	8.49	2.21	2.86	0.80	0.69	1.27
精氨酸 R	CGT	6.39	9.02	7.69	7.16	4.85	5.63	0.71	0.83	0.89	1.32	1.13
丝氨酸 S	AGC	16.23	11.34	9.89	15.99	21.70	16.36	1.43	1.64	1.02	0.75	0.99
丝氨酸 S	AGT	10.93	14.01	13.23	8.81	5.50	6.62	0.78	0.83	1.24	1.99	1.65

表 4 （续）

Table 4 Continued

氨基酸	密码子	密码子出现频率/‰						频率比值				
		毛竹	拟南芥	烟草	水稻	玉米	小麦	B/A	B/T	B/R	B/M	B/W
丝氨酸 S	TCA	15.22	18.28	17.62	12.46	7.89	10.78	0.83	0.86	1.22	1.93	1.41
丝氨酸 S	TCC	15.73	11.20	10.40	16.36	20.79	17.70	1.40	1.51	0.96	0.76	0.89
丝氨酸 S	TCG	10.46	9.33	5.38	12.35	16.41	10.02	1.12	1.94	0.85	0.64	1.04
丝氨酸 S	TCT	15.38	25.17	20.23	12.72	8.56	10.50	0.61	0.76	1.21	1.80	1.46
苏氨酸 T	ACA	13.30	15.67	17.26	11.58	7.32	9.41	0.85	0.77	1.15	1.82	1.41
苏氨酸 T	ACC	13.07	10.34	9.91	14.87	17.67	18.84	1.26	1.32	0.88	0.74	0.69
苏氨酸 T	ACG	9.02	7.74	4.57	11.37	16.52	9.49	1.17	1.97	0.79	0.55	0.95
苏氨酸 T	ACT	12.14	17.52	20.44	10.63	6.65	9.23	0.69	0.59	1.14	1.83	1.32
缬氨酸 V	GTA	7.90	9.92	11.25	6.78	4.45	5.48	0.80	0.70	1.17	1.78	1.44
缬氨酸 V	GTC	17.64	12.77	11.26	20.09	25.81	21.20	1.38	1.57	0.88	0.68	0.83
缬氨酸 V	GTG	23.53	17.36	16.74	24.27	31.54	24.72	1.36	1.41	0.97	0.75	0.95
缬氨酸 V	GTT	18.50	27.24	26.99	15.51	9.79	14.44	0.68	0.69	1.19	1.89	1.28
色氨酸 W	TGG	13.11	12.48	11.82	13.86	13.36	12.08	1.05	1.11	0.95	0.98	1.09
酪氨酸 Y	TAC	14.69	13.73	13.59	15.10	19.95	20.78	1.07	1.08	0.97	0.74	0.71
酪氨酸 Y	TAT	10.90	14.63	17.62	9.93	4.96	8.39	0.75	0.62	1.10	2.20	1.30
*	TAA	0.55	0.95	1.13	0.66	0.53	0.62	0.58	0.49	0.83	1.04	0.89
*	TAG	0.67	0.53	0.51	0.85	0.91	0.63	1.26	1.31	0.79	0.74	1.06
*	TGA	1.02	1.17	1.01	1.22	1.47	1.46	0.87	1.01	0.84	0.69	0.70

说明：标记有下划线的数据表示比值≥2.0 或≤0.5。字母 B，M，A，T，R 和 W 分别表示毛竹、玉米、拟南芥、烟草、水稻和小麦。

蝇、玉米、草菇*Volvariella volvacea* 等相似[15-16]，毛竹基因密码子偏好使用 G/C 结尾的密码子，这种密码子使用偏好性有利于保证翻译的准确性[17-18]，但与同为植物界的双子叶植物相比，密码子使用偏好性差异较大，双子叶植物偏好使用以 A 或 T 结尾的密码子[14,19-20]。

从 GC 含量上看，很多植物的密码子的 GC_1 含量均比 GC_2 高，两者含量的差异达到 0.096（*Medicago truncatula*）~0.155（*Micromonas pusilla* RCC299）。裸子植物、单子叶植物、绿藻等物种 GC_3 的含量一般来说略高于 GC_1[7]。本研究中的毛竹 GC_1 含量比 GC_2 高 0.12，而 GC_3 含量比 GC_1 高 0.01，表明选择压力对毛竹密码子不同位置的碱基组成影响不同。不同物种中 GC_3 会随着进化不同而发生变化[7,21-22]。一般来说，原始的单细胞或多细胞绿色植物 GC_3 含量会比较高，为 0.690~0.854，苔藓植物为 0.481~0.578，而被子植物 GC_{3s} 的含量变异差异比较大，单子叶植物的变异范围为 0.581~0.609，优等双子叶植物的变异范围为 0.335~0.482。本研究中毛竹 GC_{3s} 为 0.52，超出了单子叶植物的变异范围。这样的特例在其他物种中也有发生，如莱茵衣藻 *Chlamydomonas reinhardtii*，团藻 *Volvox carteri* 和 细小微胞藻 *Micromonas pusilla*[7]。

本研究用 CodonW 软件分析了毛竹同义密码子的 RSCU 值，发现 AGG，CUC，GUG，AAG 和 UGC 5 个密码子为本文的高频率密码子。而 4 个 NUA 密码子 RSCU 值较低，AUA 为 0.69，CUA 为 0.52，GUA 为 0.47，UUA 为 0.45，表明毛竹基因避免使用 UA 密码子，同一现象在其他物种中也发现，可能因为低含量的 UA 抑制了 mRNA 的降解，提高蛋白产物或产量[23]。毛竹中终止密码子的使用以 UGA 的使用频率最高，与大多数植物相吻合[24]。

NCG:NCC 的比值已广泛用于评估 CpG 抑制，反映了编码区甲基化水平，尤其在真双子叶植物。甲基化水平低的物种往往其 NCG:NCC 的比值相对较高，如拟南芥（0.921），深山南芥 *Arabis lyrata*（0.93）；而高甲基化水平的物种，该比值相对较低，如葡萄 *Vitis vinifera*（0.414），杨树（0.463）；甲基化程度中等的物种；该比值中等，如苹果 *Malus×domestica*（0.639），番茄 *Solanum lycopersicon*（0.634）。毛竹中该比值为 0.819 7，表明毛竹为低甲基化水平的物种。由此可以判断：甲基化水平对毛竹的生长发育过程影响有限[7]。

本研究使用同义密码子相对使用频率(RSCU)方法鉴定出 26 个最优密码子，全部以 G/C 结尾，毛竹编码蛋白序列的 GC 含量平均为 52.4%，因此，本研究结果符合一般规律，即富含 GC 碱基的基因组中最优密码子也富含 GC[2, 25]。通过比较某一特定基因与外源表达系统之间的密码子使用偏好性差异，从而分析是否会引起甲基化，导致基因表达量下降或基因沉默[26]，从而改造密码子以提高外源基因在宿主中的表达[27–29]。本研究将毛竹基因组密码子的偏好性与模式动植物大肠埃希菌、酵母、果蝇、拟南芥、烟草、玉米、水稻和小麦待密码子偏好性相比，结果表明：毛竹与不同物种的差异程度不同，其中与大肠埃希菌和酵母的差异最大，而与同科 C_3 植物水稻和小麦的偏好性一致。因此，要将毛竹基因进行体外表达时，需要通过密码子的改造，来提高表达效率。若要将毛竹基因用于水稻和小麦中表达时，可以不用经密码子优化直接进行外源基因表达。本研究的聚类结果表明：密码子偏好性差异大小在一定程度上反映物种间的进化关系，与传统分类有一定的吻合性，但不完全吻合，这与其他物种的基于密码子偏好性聚类的结果类似[4,15]，很可能是因为参数选择单一造成的。该研究结果可为毛竹基因外源表达选择合适的受体提供理论基础，同时，还为将毛竹基因转入模式生物中进行功能验证提供基础资料。

4　结论

本研究对毛竹基因组中的 26 103 个蛋白质编码基因序列进行了分析，根据同义密码子相对使用频率(RSCU 值)确定了毛竹中的最优密码子 26 个，且均以 G/C 结尾。同时与模式动植物 9 个代表性物种进行了比较，毛竹密码子偏好性与水稻完全一致。

5　参考文献

[1] 宋辉，王鹏飞，马登超，等. 蒺藜苜蓿 WRKY 转录因子密码子使用偏好性分析[J]. 农业生物技术学报，2015, **23**(2)：203 – 212.
SONG Hui, WANG Pengfei, MA Dengchao, *et al.* Analysis of codon usage bias of WRKY transcription factors in *Medicago truncatula* [J]. *J Agric Biotechnol*, 2015, **23**(2): 203 – 212.

[2] HERSHBERG R, PETROV D A. Selection on codon bias [J]. *Annu Rev Genet*, 2008, **42**(42): 287 – 299.

[3] 刘汉梅，何瑞，赵耀，等. 玉米密码子用法分析[J]. 核农学报，2008，**22**(2)：141 – 147.
LIU Hanmei, HE Rui, ZHAO Yao, *et al*. Analysis of codon usage in maize [J]. *Acta Agric Nucl Sin*, 2008, **22**(2): 141 – 147.

[4] 晁岳恩，吴政卿，杨会民，等. 11 种植物 *psbA* 基因的密码子偏好性及聚类分析[J]. 核农学报，2011，**25**(5)：927 – 932.
ZHAO Yueen, WU Zhengqing, YANG Huimin, *et al.* Cluster analysis and codon usage bias studies on *psbA* genes from 11 plant species [J]. *Acta Agric Nucl Sin*, 2011, **25**(5): 927 – 932.

[5] DURET L, MOUCHIROUD D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis* [J]. *Proc Nat Acad Sci U S A*, 1999, **96**(8): 4482 – 4487.

[6] SAKAI H, WASHIO T, SAITO R, *et al.* Correlation between sequence conservation of the 5′ untranslated region and codon usage bias in *Mus musculus* genes [J]. *Gene*, 1998, **431**(3): 101 – 105.

[7] FENG Chao, XU Changjie, WANG Yue, *et al.* Codon usage patterns in Chinese bayberry (*Myrica rubra*) based on RNA-Seq data [J]. *BMC Genet*, 2013, **14**(6): 986 – 991.

[8] PENG Zhenhua, LU Ying, LI Lubin, *et al.* The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*) [J]. *Nat Genet*, 2013, **45**(4): 456 – 461.

[9] 刘庆坡，谭军，薛庆中. 籼稻品种 93–11 同义密码子的使用偏好性[J]. 遗传学报，2003，**30**(4)：335 – 340.
LIU Qingpo, TAN Jun, XUE Qingzhong. Synonymous codon usage bias in the rice cultivar 93-11 (*Oryza sativa* L. ssp. *indica*) [J]. *Acta Genet Sin*, 2003, **30**(4):335 – 340.

[10] WRIGHT F. The “effective number of codons” used in a gene [J]. *Gene*, 1990, **87**(1): 23 – 29.

[11] SHARP P M, LI W H. An evolutionary perspective on synonymous codon usage in unicellular organisms [J]. *J Mol Evol*, 1986, **24**(1/2)：28 – 38.

[12] 时慧，王玉，杨路成, 等. 茶树抗寒调控转录因子 ICE1 密码子偏好性分析[J]. 园艺学报，2012, **39**(7)：1341 – 1352.

SHI Hui, WANG Yu, YANG Lucheng, *et al.* Analysis of codon bias of the cold regulated transcription factor ICE1 in tea plant [J]. *Acta Hortic Sin*, 2012, **39**(7): 1341 – 1352.

[13] STENICO M, LLOYD A T, SHARP P M. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases [J]. *Science*, 2002, **296**(5576): 2174 – 2176.

[14] KAWABE A, MIYASHITA N T. Patterns of codon usage bias in three dicot and four monocot plant species [J]. *Genes Genet Syst*, 2003, **78**(5): 343 – 352.

[15] 蒋玮，吕贝贝，何建华，等. 草菇密码子偏好性分析[J]. 生物工程学报，2014，**30**(9)：1424 – 1435.
JIANG Wei, LÜ Beibei, HE Jianhua, *et al*. Codon usage bias in the straw mushroom *Volvariella volvacea* [J]. *Chin J Biotech*, 2014, **30**(9): 1424 – 1435.

[16] 刘汉梅，何瑞，张怀渝，等. 玉米同义密码子偏爱性分析[J]. 农业生物技术学报，2010，**18**(3)：456 – 461.
LIU Hanmei, HE Rui, ZHANG Huaiyu, *et al.* Analysis of synonymous codon bias in maize [J]. *J Agric Biotechnol*, 2010, **18**(3): 456 – 461.

[17] 石秀凡，黄京飞，柳树群，等. 人类基因同义密码子偏好的特征以及与基因 GC 含量的关系[J]. 生物化学与生物物理进展，2002，**29**(3)：411 – 414.
SHI Xiufan, HUANG Jingfei, LIU Shuqun, *et al.* The features of synonymous codon bias and GC content relationship in human genes [J]. *Prog Biochem Biophys*, 2002, **29**(3): 411 – 414.

[18] SHARP P M, COWE E, HIGGINS D G, *et al.* 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 [J]. *Nucl Acid Res*, 1988, **16**(17): 8207 – 8211.

[19] WANG Liangjiang, ROOSSINCK M J. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants [J]. *Plant Mol Biol*, 2006, **61**(4/5): 699 – 710.

[20] WANG Huaichun, HICKEY D A. Rapid divergence of codon usage patterns within the rice genome [J]. *BMC Evolut Biol*, 2007, **7**(S1): 173 – 188.

[21] LYNCH D B, LOGUE M E, BUTLER G, *et al.* Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres [J]. *Gen Biol Evol*, 2010, **2**(1): 572 – 583.

[22] EYRE-WALKER A, HURST LD. The evolution of isochors [J]. *Nat Rev Genet*, 2001, **2**(7): 549 – 555.

[23] MAHER A S, KHABAR K S A. UU/UA dinucleotide frequency reduction in coding regions results in increased mRNA stability and protein expression [J]. *Mol Ther J Am Soc Gene Ther*, 2012, **20**(5): 954 – 959.

[24] SUN Jingchun, CHEN Ming, XU Jinlin, *et al.* Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes [J]. *J Mol Evol*, 2005, **61**(4): 437 – 444.

[25] RAO Yousheng, WU Guozuo, WANG Zhangfeng, *et al.* Mutation bias is the driving force of codon usage in the *Gallus gallus* genome [J]. *J Jpn Veter Med Ass*, 1986, **39**(6): 154 – 158.

[26] 张乐，金龙国，罗玲，等. 大豆基因组和转录组的核基因密码子使用偏好性分析[J]. 作物学报，2011，**37**(6)：965 – 974.
ZHANG Le, JIN Longguo, LUO Ling, *et al.* Analysis of nuclear gene codon bias on soybean genome and transcriptome [J]. *Acta Agron Sin*, 2011, **37**(6): 965 – 974.

[27] SHAO Zhuqing, ZHANG Yanmei, FENG Xueying, *et al.* Synonymous codon ordering: a subtle but prevalent strategy of bacteria to improve translational efficiency [J]. *Plos One*, 2012, **7**(3): e33547. doi:10.1371/ journal.pone.0033547.

[28] QIAN Wenfeng, YANG Jianrong, PEARSON N M, *et al*. Balanced codon usage optimizes eukaryotic translational efficiency [J]. *Plos Genet*, 2012, **8**(3): e1002603. doi: 10.1371/journal.pgen.1002603.

[29] KYOKO H T, MPANJA N, TADAYOSHI H, *et al.* High-level accumulation of recombinant miraculin protein in transgenic tomatoes expressing a synthetic miraculin gene with optimized codon usage terminated by the native miraculin terminator [J]. *Plant Cell Rep*, 2011, **30**(1): 113 – 124.