DOI:10.13275/j.cnki.lykxyj.2021.005.021

杓兰叶绿体基因组密码子偏好性分析

丁　锐[1]，胡　兵[1]，宗小雁[2]，韩辰阳[2]，张丽杰[3]，陈旭辉[2*]

(1. 沈阳农业大学土地与环境学院，辽宁 沈阳　110866；2. 沈阳农业大学生物科学技术学院，辽宁 沈阳　110866；
3. 沈阳农业大学林学院，辽宁 沈阳　110866)

摘要: [**目的**] 分析杓兰叶绿体基因组密码子的使用偏好性，探究影响杓兰叶绿体基因组密码子使用偏好性的主要因素，为兰科叶绿体基因组学研究提供参考。[**方法**] 从 NCBI 数据库中下载完整的杓兰叶绿体基因组序列并进行蛋白编码序列筛选，利用 EMBOSS 在线程序计算各基因及密码子的 GC 含量，利用 CodonW 软件计算各基因的氨基酸长度（Laa）、有效密码子数（ENC）、同义密码子相对使用度（RSCU）、最优密码子使用频率（FOP）及各基因密码子的第 3 核苷酸碱基含量，利用 SPSS 软件分析各指标之间的相关性，利用 Origin 软件绘图。[**结果**] 杓兰叶绿体基因组编码序列的密码子第 3 位碱基富含 A 和 T，GC 含量仅为 29%，ENC 值介于 37.92～61.00 之间，密码子偏好性不强，ENC 与 GC_2 及 GC_3 均呈极显著相关。RSCU 值大于 1 的密码子有 34 个，其中，29 个以 U 或 A 结尾。ENC-plot 分析、PR2-plot 分析及中性绘图分析表明：影响杓兰叶绿体基因组密码子使用偏好性的主要因素为自然选择。对应性分析结果表明：编码光合系统蛋白基因的密码子具有相似的使用模式，编码其它类型基因的密码子则具有不同的使用模式，并最终筛选出最优密码子 16 个。[**结论**] 本研究明确了自然选择是影响杓兰叶绿体基因组密码子使用偏好性的主要因素，并筛选出杓兰叶绿体基因的最优密码子，研究结果能够对兰科系统发育及叶绿体基因组密码子进化研究提供参考。

关键词: 杓兰；叶绿体基因组；密码子偏好性；兰科

中图分类号: Q941.2　　**文献标志码:** A　　**文章编号:** 1001-1498(2021)05-0177-09

密码子是自然界中承载生命信息的基本遗传单位，是蛋白质与核酸的桥梁和纽带，在生物体传递遗传信息的过程中起到重要作用。密码子具备简并性，编码同种氨基酸的不同密码子称为同义密码子；同时，同义密码子在不同物种间的使用频率具有不均一性，这种现象被称为密码子使用偏好性[1]。密码子偏好性是生命体进化的重要特征，在自然界普遍存在并受到自然选择和基因突变等多种因素共同影响，自然选择使得不同物种的基因在选择同义密码子时倾向于使用最优密码子，而基因突变则会使部分非最优密码子存在[2-3]。由于不同物种在进化过程中受到的选择作用和突变压力的影响程度不同，因此，会形成自身独特的密码子使用偏好性[4-5]。研究表明，同一物种或亲缘关系较近的物种中基因一般具有相似的密码子使用模式[6-8]，因此，对密码子使用偏好性进行研究有助于更好地了解物种的进化。

叶绿体是植物进行光合作用的场所，同时也是一种半自主性的细胞器，拥有相对独立的一整套基因组，同时拥有复制、转录及翻译的机制。被子植物的叶绿体基因组一般由 4 部分组成环状双链结构，包括 1 个短单拷贝区（SSC）、1 个长单拷贝区（LSC）及 2 个相同的反向重复区（IR），且 SSC 和 LSC 之间被 2 个 IR 隔开[9]。与体细胞基因

收稿日期：2020-08-08　　修回日期：2021-07-20

基金项目：国家自然科学基金（31670378）；辽宁省“兴辽英才计划”项目（XLYC1807180）；沈阳农业大学大学生创新创业训练计划项目（2020-109）

作者简介: 并列第一作者：丁锐，博士，讲师。主要研究方向：植物生物技术与代谢工程。电话：15040207167。Email：mnidr7@syau.edu.cn；并列第一作者：胡兵，本科。专业：环境工程。电话：13842030986。Email：1735165327@qq.com

* 通讯作者: 陈旭辉，博士，副教授。主要研究方向：兰科菌根生态学。电话：15040126208。Email：xhchen@syau.edu.cn

组相比，叶绿体基因组体量小，基因拷贝数多，进化速率快且保守度高，这些特点使其在研究物种间遗传差异和系统进化关系的过程中充当了理想的工具[10]。自从 1986 年首次公开了烟草[11] 和地钱[12] 的叶绿体基因组序列以来，越来越多的叶绿体基因组信息被 NCBI 数据库收录。

杓兰属（*Cypripedium*）隶属于兰科杓兰亚科，是兰科植物中较为原始的类型，全世界约 50 种，我国有 36 种和 1 变种[13]。杓兰属植物花姿优美、花色丰富，具有较高的观赏价值。然而，随着兰花热的兴起以及生境破碎化的不断加剧，滥采乱挖杓兰属植物的行为日益猖獗，不少种类已近濒危[14]。杓兰（*Cypripedium calceolus* L.）是杓兰属多年生地生植物，主要分布于我国东北、日本、韩国和欧洲地区。目前，该物种已被列入世界自然保护联盟（IUCN）濒危物种红色名录。杓兰的叶绿体基因组序列已被提交至 GenBank 数据库[15]，但目前尚未有关于杓兰叶绿体基因组密码子使用偏好性的研究。本研究通过生物信息学方法分析杓兰叶绿体基因组密码子的使用偏好性，旨在为杓兰的叶绿体基因组学研究提供参考。

1 材料与方法

1.1 基因序列获取

从 NCBI 数据库中下载完整的杓兰叶绿体基因组序列（GenBank 登录号：MN602053.1），序列长度为 175 122 bp，包含 78 条蛋白编码基因。为了避免出现样本误差，移除其中的重复基因序列以及长度小于 300 bp 的编码序列，最终获得 53 条符合分析条件的蛋白编码序列用于后续分析。

1.2 密码子相关参数计算

利用 EMBOSS 程序（http://www.bioinformatics.nl/emboss-explorer）对各基因的 GC 含量进行在线分析，分析结果记为 GC；同时对各基因的密码子第 1、2、3 位核苷酸上的 GC 含量进行在线分析，分析结果分别记为 GC_1、GC_2、GC_3。利用软件 CodonW 对各个基因的密码子在第 3 核苷酸上的 A、G、C、T 含量进行计算，计算结果分别记为 A_3、G_3、C_3、T_3；同时利用该软件对各基因的氨基酸长度（Laa）、有效密码子数（ENC）、同义密码子相对使用度（RSCU）及最优密码子使用频率（FOP）进行计算。

1.3 密码子使用偏好性指标

ENC 及 RSCU 是密码子使用偏好性的重要度量指标[16]。RSCU 意为编码某种特定氨基酸时，某一密码子的实际使用频率与不存在偏好性的状态下其预期使用频率间的比值[17]。RSCU=1 表明该密码子的使用不存在偏好性；RSCU>1 表明该密码子使用频率偏高；RSCU < 1 表明该密码子使用频率偏低。ENC 意为某个基因使用密码子的偏好程度，ENC 值的范围为 20~61。若某基因的 ENC 值为 20，则表明该基因中各种氨基酸只使用特定密码子，同一氨基酸密码子使用无随机性，偏好性高；若某基因的 ENC 值为 61，则表明该基因中各种氨基酸编码时均衡使用其对应的同义密码子，密码子使用随机性高，偏好性低[18-19]。利用 SPSS 软件分析 ENC 与各指标之间的相关性。

1.4 中性绘图分析

取各基因 GC_1 及 GC_2 的平均值，记为 GC_{12}，以各基因的 GC_{12} 为纵坐标、GC_3 为横坐标绘制散点图，并对二者的相关性进行分析。若 GC_3 与 GC_{12} 显著相关，则表明密码子 3 个位点的碱基具有相同的变异模式，突变是密码子使用偏好性的主要影响因素；若 GC_3 与 GC_{12} 相关性不显著，则表明密码子 3 个位点碱基的变异模式差异较大，密码子使用偏好性主要受自然选择影响[20]。

1.5 ENC-plot 绘图分析

取各基因的 ENC 为纵坐标、GC_3 为横坐标绘制散点图。同时，根据公式 $ENC = 2 + GC_3 + 29/[GC_3^2 + (1-GC_3)^2]$ 计算各基因的理论 ENC 值，并以 GC_3 为横坐标、理论 ENC 值为纵坐标绘制标准曲线[21]。标准曲线可以显示出 ENC 及 GC_3 在无选择压力状态下的关联情况，若基因位点在图中分布贴近标准曲线，则突变是密码子使用偏好性的主要影响因素；若基因位点在图中分布远离标准曲线，则密码子偏好性主要受自然选择因素影响。

1.6 PR2-plot 分析

以各基因的 $G_3/(G_3 + C_3)$ 为横坐标、$A_3/(A_3 + T_3)$ 为纵坐标绘制散点图，对密码子第 3 位核苷酸上的碱基组成情况进行分析，从而探讨突变和自然选择对密码子使用偏好性的影响。图中中心点 A = T，G = C，表示某一基因 2 条互补链间不存在任何突变或选择效应上的偏倚，从中心点向其它位点分布的矢量则显示该基因的偏倚程度及方向[22]。

1.7　最优密码子分析

将 53 条基因按 ENC 值由高至低排序，从两端各选出 10% 的基因数作为高、低表达库。根据各基因的 RSCU 值筛选出各库内对应密码子 ΔRSCU>0.08 的密码子作为高表达密码子，并将 ΔRSCU>0.08 且 RSCU>1 的密码子作为最优密码子[23]。

1.8　对应性分析

基于各基因的 RSCU 值，通过软件 CodonW 进行分析，根据分析结果将所有基因在一个 47 维的向量空间进行分布，不同基因在向量空间中的相对分布位置可以表征影响密码子使用偏好性的因素。基因在第 1、2 向量轴（主向量轴）间的分散程度显示出密码子的主要变化趋势，是推断其密码子使用变异的依据[24]。以第 1 轴为横坐标、第 2 轴为纵坐标绘制散点图，根据图中点的分布情况判断基因密码子的使用模式。

2　结果与分析

2.1　密码子的组成特征

杓兰的叶绿体基因组去除长度小于 300 bp 的蛋白编码序列及重复序列后，共剩余 53 条蛋白编码基因，全长 60618 bp，占基因组全长的 35%，基因编码的氨基酸序列长度范围为 100~2310，平均长度 377。对这些基因的密码子组成和偏好性进行统计分析发现，平均 GC 含量为 38%，其中，GC_1（47%）大于 GC_2（39%）大于 GC_3（29%），表明 GC 在密码子 3 个位置上的分布并不均匀，且偏向于以 A 和 T 碱基结尾。各基因的 ENC 值介于 37.92~61.00 之间，平均值为 48.05，ENC 值大于 45 的基因有 38 条，表明杓兰叶绿体基因组密码子的使用偏好性较弱（表 1）。

表 1　杓兰叶绿体基因组的主要参数

Table 1　Main parameters in chloroplast genomics of *Cypripedium calceolus*

基因 Gene	GC_1	GC_2	GC_3	GC	ENC	Laa	基因 Gene	GC_1	GC_2	GC_3	GC	ENC	Laa
accD	0.37	0.36	0.26	0.33	43.81	498	*psbB*	0.54	0.46	0.33	0.44	48.90	508
atpA	0.55	0.40	0.24	0.40	45.08	507	*psbC*	0.54	0.46	0.33	0.45	45.94	473
atpB	0.56	0.41	0.32	0.43	50.39	498	*psbD*	0.52	0.44	0.32	0.43	42.67	353
atpE	0.51	0.40	0.29	0.40	49.10	133	*rbcL*	0.58	0.43	0.31	0.44	48.63	484
atpF	0.49	0.34	0.31	0.38	44.60	184	*rpl14*	0.53	0.37	0.29	0.40	48.78	122
atpI	0.49	0.37	0.27	0.38	45.87	247	*rpl16*	0.51	0.54	0.26	0.44	37.92	137
ccsA	0.31	0.36	0.27	0.32	48.35	327	*rpl2*	0.51	0.49	0.33	0.44	51.58	271
cemA	0.40	0.28	0.32	0.33	52.71	229	*rpl20*	0.36	0.43	0.26	0.35	51.15	136
clpP	0.58	0.36	0.33	0.43	61.00	204	*rpl22*	0.44	0.36	0.19	0.33	41.56	120
matK	0.39	0.30	0.27	0.32	49.06	519	*rpoA*	0.46	0.34	0.29	0.36	51.53	337
ndhA	0.42	0.37	0.23	0.34	44.03	363	*rpoB*	0.50	0.38	0.28	0.38	47.97	1070
ndhB	0.41	0.40	0.32	0.38	47.28	510	*rpoC1*	0.50	0.39	0.30	0.40	49.79	681
ndhC	0.50	0.35	0.29	0.38	51.97	120	*rpoC2*	0.46	0.37	0.28	0.37	49.47	1390
ndhD	0.39	0.37	0.29	0.35	48.05	501	*rps11*	0.55	0.53	0.22	0.43	43.62	138
ndhE	0.42	0.33	0.35	0.37	54.85	101	*rps12*	0.52	0.48	0.27	0.43	46.07	123
ndhF	0.36	0.37	0.24	0.33	46.05	739	*rps14*	0.44	0.49	0.29	0.41	41.06	100
ndhG	0.42	0.35	0.28	0.35	44.20	176	*rps18*	0.37	0.44	0.27	0.36	39.57	101
ndhH	0.48	0.36	0.30	0.38	49.19	393	*rps2*	0.43	0.41	0.32	0.39	52.48	236
ndhI	0.39	0.34	0.21	0.32	40.88	169	*rps3*	0.45	0.33	0.25	0.34	44.70	218
ndhJ	0.48	0.38	0.31	0.39	58.48	158	*rps4*	0.49	0.38	0.29	0.39	51.67	201
ndhK	0.44	0.41	0.30	0.39	52.25	259	*rps7*	0.54	0.46	0.24	0.42	48.50	155
petA	0.54	0.36	0.25	0.38	46.33	320	*rps8*	0.39	0.38	0.23	0.33	43.24	131
petB	0.48	0.42	0.33	0.41	46.93	215	*ycf1*	0.39	0.31	0.33	0.34	54.30	487
petD	0.50	0.37	0.26	0.38	43.69	163	*ycf2*	0.42	0.35	0.37	0.38	52.71	2310
psaA	0.52	0.43	0.34	0.43	52.12	750	*ycf3*	0.47	0.40	0.26	0.38	53.92	168
psaB	0.48	0.43	0.33	0.41	49.95	734	*ycf4*	0.45	0.43	0.35	0.41	50.53	182
psbA	0.49	0.44	0.35	0.43	42.33	353	Average	0.47	0.39	0.29	0.38	48.05	377

基因密码子各参数之间的相关性分析（表 2）结果显示：GC_1 与 GC_2 呈极显著相关，但 GC_1 与 GC_3 及 GC_2 与 GC_3 均不显著相关，表明密码子第 1、2 位碱基组成情况相近，而第 3 位上的碱基组

成随机性较大，与 1、2 位碱基组成具有差异。ENC 与 GC 不显著相关，但与 GC_2 呈显著负相关，与 GC_3 呈极显著正相关，表明密码子第 2、3 位上碱基组成的变化对密码子的使用偏好性影响较大，GC_2 含量越高，密码子的使用偏好性越强；GC_3 含量越低，密码子的使用偏好性越强。ENC 与 Laa 相关不显著，表明基因序列长度并未对密码子使用偏好性造成较大影响。

RSCU 分析（表 3）表明：RSCU>1.00 的密码子共有 34 个，其中，以 A 和 U 结尾的有 29 个，占 85%，表明杓兰叶绿体基因组偏向于使用以 A 或 U 结尾的同义密码子。RSCU < 1.00 的密码子则多以 C 或 G 结尾。

表 2 基因密码子各参数之间的相关性分析

Table 2 Correlation analysis between the indexes of codon use

	GC_1	GC_2	GC_3	GC	ENC	Laa
GC_1	1.000					
GC_2	0.426**	1.000				
GC_3	0.182	0.034	1.000			
GC	0.830**	0.746**	0.463**	1.000		
ENC	0.109	−0.308*	0.528**	0.094	1.000	
Laa	−0.055	−0.153	0.307*	−0.013	0.172	1.000

注：“*”表示显著相关（$p < 0.05$），“**”表示极显著相关（$p < 0.01$）。

Notes: “*”indicates a significant correlation at $p < 0.05$ level, “**” indicates a significant correlation at $p < 0.01$ level.

表 3 杓兰叶绿体基因组各氨基酸的相对同义密码子使用度

Table 3 Relative synonymous codon usage (RSCU) analysis of genes on chloroplast genome in *Cypripedium calceolus*

氨基酸 Amino acid	密码子 Codon	数目 Number	RSCU	氨基酸 Amino acid	密码子 Codon	数目 Number	RSCU	氨基酸 Amino acid	密码子 Codon	数目 Number	RSCU
Phe	UUU	740	1.28	Ser	UCU	425	1.68	TER	UAA	70	1.15
	UUC	418	0.72		UCC	259	1.03		UAG	66	1.09
Leu	UUA	654	1.83		UCA	305	1.21		UGA	46	0.76
	UUG	457	1.28		UCG	128	0.51				
	CUU	437	1.22	Pro	CCU	306	1.48	His	CAU	380	1.51
	CUC	155	0.43		CCC	190	0.92		CAC	124	0.49
	CUA	299	0.83		CCA	239	1.15	Gln	CAA	561	1.50
	CUG	148	0.41		CCG	94	0.45		CAG	185	0.50
Ile	AUU	821	1.45	Thr	ACU	394	1.63	Asn	AAU	681	1.55
	AUC	351	0.62		ACC	175	0.72		AAC	198	0.45
	AUA	531	0.94		ACA	284	1.18	Lys	AAA	701	1.41
Met	AUG	463	1.00		ACG	113	0.47		AAG	291	0.59
Val	GUU	394	1.39	Ala	GCU	479	1.78	Asp	GAU	623	1.59
	GUC	139	0.49		GCC	152	0.57		GAC	161	0.41
	GUA	416	1.47		GCA	328	1.22	Glu	GAA	759	1.44
	GUG	186	0.66		GCG	116	0.43		GAG	298	0.56
Gly	GGU	436	1.25	Arg	CGU	262	1.37	Ser	AGU	302	1.20
	GGC	151	0.43		CGC	69	0.36		AGC	95	0.38
	GGA	553	1.58		CGA	258	1.35	Arg	AGA	352	1.84
	GGG	259	0.74		CGG	84	0.44		AGG	121	0.63
Cys	UGU	183	1.43	Tyr	UAU	585	1.56	Trp	UGG	370	1.00
	UGC	73	0.57		UAC	163	0.44				

2.2 密码子使用的中性绘图分析

中性绘图分析（图 1）表明：GC_{12} 的取值范围为 0.338~0.536，GC_3 的取值范围为 0.190~0.372。所有基因都分布于中线对角线上方，GC_{12} 与 GC_3 的相关系数为 0.147，相关性不显著，表明杓兰叶绿体密码子 3 个位点的碱基组成具有较大差异，杓

兰叶绿体密码子使用偏好性主要受自然选择因素影响。

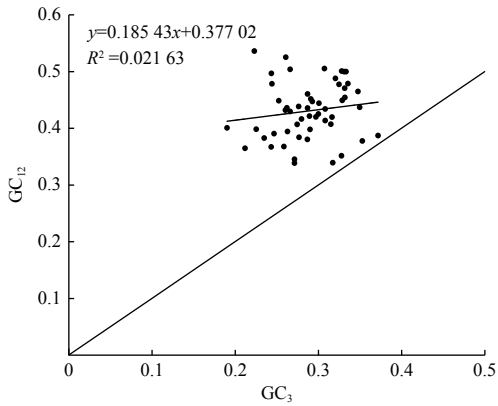

图 1　中性绘图分析
Fig. 1　Analysis of neutrality plot

2.3　ENC-plot 分析

ENC-plot 分析（图 2）发现：大部分基因位点都落在偏离标准曲线的位置，即 ENC 实际值与 ENC 预期值之间差异较大，表明杓兰叶绿体基因组密码子使用偏好性主要是外界自然选择压力等因素作用的结果。

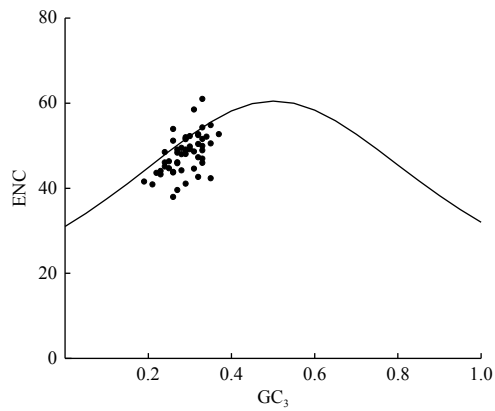

图 2　ENC-plot 分析
Fig. 2　Analysis of ENC-plot

2.4　PR2-plot 分析

PR2-plot 分析结果（图 3）表明：基因位点在平面图的 4 个区域中分布较均匀，其中，右下方区域集中了相对较多的基因位点，表明密码子第 3 位碱基使用 T 频率高于 A，使用 G 的频率高于 C，存在偏好性。因此，可推断杓兰叶绿体基因组密码子的使用模式虽然在一定程度上受到自身突变影响，但该影响作用效果有限，并非密码子偏好性产生的主要因素。

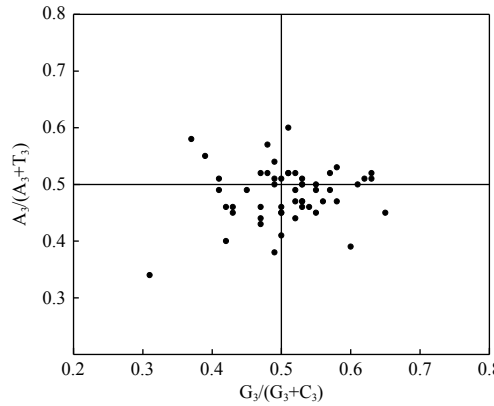

图 3　PR2-plot 分析
Fig. 3　Analysis of PR2-plot

2.5　最优密码子分析

以 ΔRSCU>0.08 为标准共确定 25 个密码子为杓兰叶绿体基因组的高表达密码子，其中，以 A 结尾的有 9 个，U 结尾的有 8 个，C 结尾的有 5 个，G 结尾的有 3 个（表 4）。结合杓兰叶绿体基因的相对同义密码子使用度（表 3），最终分析得出 16 个最优密码子，分别为 GUA、GCA、UCU、UCC、ACU、CCU、CCA、GCU、UAU、UAA、CAU、AAU、CGA、AGU、AGA、GGA，其中，7 个以 A 结尾，8 个以 U 结尾，1 个以 C 结尾（表 4）。

2.6　对应性分析

基于 RSCU 的对应性分析结果显示：第 1、2、3、4 向量轴分别显示了 10.10%、9.00%、8.42% 和 6.89% 的差异，四轴累积差异贡献率为 34.42%，第 1 向量轴是影响密码子使用偏好性的主效因素。从各基因位点在以第 1、2 向量轴为坐标系的平面图（图 4）分布看，编码光合系统蛋白的基因点分布较集中，说明该类基因的密码子具有相似的使用模式；而编码其它蛋白的基因点分布较分散，表明这些基因密码子的使用模式相差较大。

3　讨论

大部分生物体在合成蛋白质时都会偏好性地选择使用同义密码子，这一现象受多种因素共同影响，其中，密码子的碱基组成是最普遍的因素[25-26]。由于密码子第 3 位的碱基改变通常不会引起编码氨基酸的改变，因此，第 3 位的碱基受到的选择压力相对较小，可以作为分析密码子使用偏好性的重要参数[21]。与大部分研究结果相似，杓兰叶绿体基因组各基因密码子的第 3 位碱基 A 和 T 的使用频率

表 4 杓兰叶绿体基因的最优密码子

Table 4 The optimal codons in chloroplast genome of *Cypripedium calceolus*

氨基酸 Amino Acid	密码子 Codon	高表达基因 High expressed gene		低表达基因 Low expressed gene		ΔRSCU
		数目 Number	RSCU	数目 Number	RSCU	
Phe	UUU	48	1.43	24	1.45	−0.02
	UUC	19	0.57	9	0.55	0.02
Leu	UUA	33	1.80	12	2.25	−0.45
	UUG	18	0.98	10	1.88	−0.90
	CUU	18	0.98	7	1.31	−0.33
	CUC**	14	0.76	2	0.38	0.38
	CUA***	13	0.71	1	0.19	0.52
	CUG***	14	0.76	0	0.00	0.76
Ile	AUU	37	1.19	20	1.15	0.04
	AUC	16	0.52	17	0.98	−0.46
	AUA**	40	1.29	15	0.87	0.42
Met	AUG	32	1.00	11	1.00	0.00
Val	GUU	25	1.69	13	2.60	−0.91
	GUC	10	0.68	3	0.60	0.08
	GUA**	17	1.15	4	0.80	0.35
	GUG**	7	0.47	0	0.00	0.47
Ser	**UCU*****	21	1.50	5	0.43	1.07
	UCC**	14	1.00	6	0.52	0.48
	UCA	13	0.93	16	1.39	−0.46
	UCG	11	0.79	15	1.30	−0.51
Pro	**CCU*****	15	1.50	3	0.80	0.70
	CCC	8	0.80	6	1.60	−0.80
	CCA**	13	1.30	3	0.80	0.50
	CCG	4	0.40	3	0.80	−0.40
Thr	**ACU****	17	1.21	10	0.73	0.48
	ACC	14	1.00	13	0.95	0.05
	ACA	17	1.21	19	1.38	−0.17
	ACG	8	0.57	13	0.95	−0.38
Ala	**GCU*****	24	1.71	3	1.20	0.51
	GCC	10	0.71	2	0.80	−0.09
	GCA***	14	1.00	1	0.40	0.60
	GCG	8	0.57	4	1.60	−1.03
Tyr	**UAU***	48	1.81	23	1.53	0.28
	UAC	5	0.19	7	0.47	−0.28
TER	**UAA*****	4	2.40	12	0.92	1.48
	UAG	0	0.00	18	1.38	−1.38
His	**CAU***	16	1.33	6	1.20	0.13
	CAC	8	0.67	4	0.80	−0.13
Gln	CAA	28	1.65	19	1.90	−0.25
	CAG*	6	0.35	1	0.10	0.25
Asn	**AAU***	50	1.56	25	1.32	0.24
	AAC	14	0.44	13	0.68	−0.24
Lys	AAA	40	1.45	33	1.40	0.05
	AAG	15	0.55	14	0.60	−0.05
Asp	GAU	40	1.70	19	1.81	−0.11
	GAC*	7	0.30	2	0.19	0.11

续表 4

氨基酸 Amino Acid	密码子 Codon	高表达基因 High expressed gene		低表达基因 Low expressed gene		ΔRSCU
		数目 Number	RSCU	数目 Number	RSCU	
Glu	GAA	67	1.44	18	1.44	0.00
	GAG	26	0.56	7	0.56	0.00
Cys	UGU	6	1.00	15	1.20	−0.20
	UGC*	6	1.00	10	0.80	0.20
TER	UGA	1	0.60	9	0.69	−0.09
Trp	UGG	18	1.00	11	1.00	0.00
Arg	CGU	9	0.84	5	0.83	0.01
	CGC	5	0.47	3	0.50	−0.03
	CGA**	18	1.69	6	1.00	0.69
	CGG	8	0.75	4	0.67	0.08
Ser	**AGU****	16	1.14	9	0.78	0.36
	AGC	9	0.64	18	1.57	−0.93
Arg	**AGA****	19	1.78	8	1.33	0.45
	AGG	5	0.47	10	1.67	−1.20
Gly	GGU	12	0.84	7	0.85	−0.01
	GGC*	9	0.63	4	0.48	0.15
	GGA**	27	1.89	13	1.58	0.31
	GGG	9	0.63	9	1.09	−0.46

注：加下划线的密码子代表基因组的RSCU＞1，“*”代表ΔRSCU＞0.08，“**”代表ΔRSCU＞0.3，“***”代表ΔRSCU＞0.5，加粗的密码子为最优密码子。

Notes: the underlined codon indicates the genomic RSCU > 1, “*”indicates ΔRSCU > 0.08, “**”indicates ΔRSCU > 0.3, “***”indicates ΔRSCU > 0.5, the bold codons are the optimal codons.

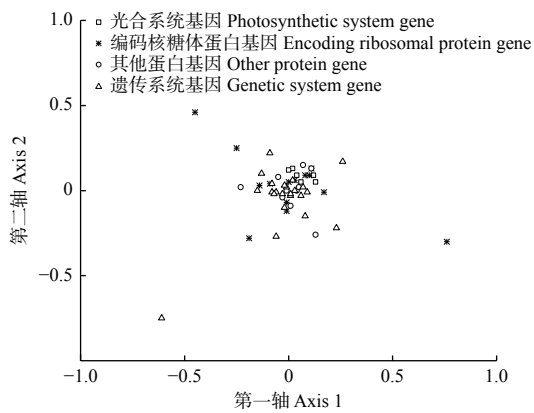

图 4　基于 RSCU 的对应性分析

Fig. 4　Corresponding analysis based on RSCU

高于 G 和 C，存在使用偏好性；同时，T 的使用频率高于 A，G 的使用频率高于 C，这与蒺藜苜蓿[21]和马尾松[27]等植物叶绿体基因的第 3 位密码子偏好性一致，但不同于陆地棉[23]、酸枣[24]及樟树[28]等植物叶绿体基因的分析结果。这表明叶绿体基因组的碱基组成在不同物种中具有各自的特点，密码子使用偏好性存在一定的差异。

突变和自然选择也是影响密码子使用偏好性的主要因素[3]。本研究结合中性绘图分析、ENC-plot 分析和 PR2-plot 分析发现，杓兰叶绿体基因组密码子偏好性主要受到自然选择的影响，突变对密码子的偏好性影响弱于自然选择作用的影响。针对兰科植物的叶绿体基因组密码子偏好性进行分析的研究报道相对较少，研究发现，蝴蝶兰叶绿体密码子产生偏好性的主要原因是碱基差异和自然选择，且碱基组成大于基因表达水平的影响[29]。文心兰叶绿体密码子的使用模式形成过程较复杂，是碱基组成、突变及自然选择等多重因素共同作用的结果[30]。由此可见，不同兰科物种具有不同的叶绿体密码子使用模式，其影响因素并不是单一的。

在突变压力及强正向选择的共同作用下，往往容易形成大量的最优密码子，而突变压力及纯化选择的共同作用，一般会抑制最优密码子的形成[4]。本研究结合杓兰叶绿体高表达密码子分析结果及高频密码子分析结果，共筛选出 16 个最优密码子，且大部分密码子以 U 或 A 结尾。目前，已见报道的绝大多数高等植物和藻类植物叶绿体基因的最优密码子都以 U 或 A 结尾，这一现象与叶绿体基因组进化的相对保守性可能具有相关性[24]。与此同时，最优密码子及其数量在不同物种间又有所不

同，表明不同物种在进化过程中面临的进化压力并不相同。

有研究表明，密码子使用偏好性聚类在较小的分类单元中可能提供较为可靠的分类依据，而当样本量较大时，由于不同基因特殊的密码子偏好性导致这种聚类结果往往不能准确地反映物种亲缘关系[7-8]。本研究基于 RSCU 的兰科聚类呈现杂乱的混合分布（聚类图未列出），不能完全正确地反映兰科植物之间的亲缘关系，因此，基因序列比密码子偏好性更适合于兰科物种分类及系统进化研究。

4 结论

本研究采用生物信息学方法，分析了杓兰叶绿体基因组密码子使用偏好性特点，明确了自然选择是影响杓兰叶绿体基因组密码子使用偏好性的主要因素。筛选出杓兰叶绿体基因的最优密码子，有利于在分子水平上研究兰科植物的进化机制。后续的工作中可以考虑进行同一基因在不同杓兰属植物之间的偏好性对比分析。

参考文献：

[1] 任桂萍, 董璎莹, 党云琨. 密码子中的密码: 密码子偏好性与基因表达的精细调控[J]. 中国科学: 生命科学, 2019, 49（7）：839-847.

[2] Buchan J R, Aucott L S, Stansfield I. tRNA properties help shape codon pair preferences inopen reading frames[J]. Nucleic Acids Research, 2006, 34(3): 1015-1027.

[3] Suzuki Y. Statistical methods for detecting natural selection from genomic data[J]. Genes and Genetic Systems, 2010, 85(6): 359-376.

[4] Hershberg R, Petrov D. Selection on codon bias[J]. Annal Review of Genetics, 2008, 42: 87-99.

[5] Gu W J, Zhou T, Ma J M, *et al*. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales[J]. Virus Research, 2004, 101(2): 155-161.

[6] 赵 森, 邓力华, 陈 芬. 秋茄叶绿体基因组密码子使用偏好性分析[J]. 森林与环境学报, 2020, 40（5）：534-541.

[7] Zhou H, Wang H, Huang L F, *et al*. Heterogeneity in codon usages of sobemovirus genes[J]. Archives of Virology, 2005, 150(8): 1591-1605.

[8] Christianson M. Codon patterns distort phylogenies from or of DNA sequences[J]. American Journal of Botany, 2005, 92(8): 1221-1233.

[9] 张韵洁, 李德铢. 叶绿体系统发育基因组学的研究进展[J]. 植物分类与资源学报, 2011, 33（4）：365-375.

[10] 邢少辰, Liu C J. 叶绿体基因组研究进展[J]. 生物化学与生物物理进展, 2008, 35（1）：21-28.

[11] Shinozaki K, Ohme M, Tanaka M, *et al*. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression[J]. The EMBO Journal, 1986, 5(9): 2043-2049.

[12] Ohyama K, Fukuzawa H, Kohchi T, *et al*. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA[J]. Nature, 1986, 322: 572-574.

[13] 陈丽飞, 刘树英, 江鹏道, 等. 杓兰属植物研究进展[J]. 湖北农业科学, 2012, 51（9）：1733-1735.

[14] 邓 莲, 张 毓, 王苗苗, 等. 濒危兰科植物大花杓兰种子非共生萌发的研究[J]. 种子, 2012, 31（6）：31-34.

[15] Zhang L J, Ding R, Meng W W, *et al*. The complete chloroplast genome sequence of the threatened *Cypripediumcalceolus* (Orchidaceae)[J]. Mitochondrial DNA Part B-Resources, 2019, 4(2): 4220-4222.

[16] 吴宪明, 吴松锋, 任大明, 等. 密码子偏性的分析方法及相关研究进展[J]. 遗传, 2007, 29（4）：420-426.

[17] Sharp P M, Li W H. The codon Adaptation Index-a measureof directional synonymous codonusage bias, and its potential applications[J]. Nucleic Acids Research, 1987, 15(3): 1281-1295.

[18] Fuglsang A. The ‘effective number of codons’ revisited[J]. Biochemical and Biophysical Research Communications, 2004, 317(3): 957-964.

[19] Jiang Y, Deng F, Wang H, *et al*. An extensive analysis on the global codon usage patternof baculoviruses[J]. Archives of Virology, 2008, 153(12): 273-282.

[20] Liu X. A more accurate relationship between‘ effective number of codons’and GC_{3s} under assumptions of no selection[J]. Computational Biology and Chemistry, 2013, 42: 35-39.

[21] 杨国锋, 苏昆龙, 赵怡然, 等. 蒺藜苜蓿叶绿体密码子偏好性分析[J]. 草业学报, 2015, 24（12）：171-179.

[22] Sueoka N. Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses[J]. Journal of Molecular Evolution, 2001, 53(4-5): 469-476.

[23] 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.

[24] 胡晓艳, 许艳秋, 韩有志, 等. 酸枣叶绿体基因组密码子使用偏性分析[J]. 森林与环境学报, 2019, 39（6）：621-628.

[25] Romero H, Zavala A, Musto H. Codon usage inchlamydia trachomatis is the result of strand-specificmutational biases and a complex pattern of selective forces[J]. Nucleic Acids Research, 2000, 28(10): 2084-2090.

[26] Carlini D B, Chen Y, Stephan W. The relationshipbetween third-codon position nucleotide content, codonbias, mRNA secondary structure and gene expression in thedrosophilid alcohol dehydrogenase genes Adh and Adhr[J]. Genetics, 2001, 159(2): 623-633.

[27] 叶友菊, 倪州献, 白天道, 等. 马尾松叶绿体基因组密码子偏好性分析[J]. 基因组学与应用生物学, 2018, 37（10）：4464-4471.

[28] 秦 政, 郑永杰, 桂丽静, 等. 樟树叶绿体基因组密码子偏好性分析[J]. 广西植物, 2018, 38（10）：1346-1355.

[29] 续 晨, 贲爱玲, 蔡晓宁. 蝴蝶兰叶绿体基因组密码子使用的相关分析[J]. 分子植物育种, 2010, 8（5）：945-950.

[30] 李冬梅, 吕复兵, 朱根发, 等. 文心兰叶绿体基因组密码子使用的相关分析[J]. 广东农业科学, 2012（10）：61-65.

Analysis of Codon Usage in the Chloroplast Genome of *Cypripedium calceolus*

DING Rui[1], *HU Bing*[1], *ZONG Xiao-yan*[2], *HAN Chen-yang*[2], *ZHANG Li-jie*[3], *CHEN Xu-hui*[2]

(1. College of Land and Environment, Shenyang Agricultural University, Shenyang　110866, Liaoning, China; 2. College of Bioscience and Biotechnology, Shenyang Agricultural University, Shenyang　110866, Liaoning, China; 3. College of Forestry, Shenyang Agricultural University, Shenyang　110866, Liaoning, China)

Abstract: [**Objective**] To analyze the codon usage bias of *Cypripedium calceolus* chloroplast genome, and identify the main factors influencing codon usage bias of this species in order to provide reference for the chloroplast genomics research of Orchidaceae species. [**Method**] Downloading the complete chloroplast genome sequence of *C. calceolus* and screening the protein coding sequences, the EMBOSS online program was used to calculate the GC content of each gene and codon, and the software CondonW was used to calculate the length of amino acid (LAA), effective number of codon (ENC), relative synonymous codon usage (RSCU), frequency of optimal codons (FOP) and the acid base content of the third nucleoside of each gene codon. The software SPSS was used to analyze the correlation among each index, and software Origin was used to plot. [**Result**] The third codon position of *C. calceolus* chloroplast genome sequence was rich in A and T, and the GC_3 content was only 29%. The ENC values varied from 37.92 to 61.00, indicating a relatively weak codon usage bias. The correlation between the number of effective codons and GC_3 showed an extremely significant level. There were 34 codons with relative synonymous codon usage greater than 1 and 29 codons ending with A and U. Analysis of neutral plot, ENC-plot and PR2-plot showed that the preference of *C. calceolus* chloroplast genome codons was mainly influenced by natural selection. Correspondence analysis showed a similar pattern of codon usage bias of the genes encoding photosynthetic system proteins, while other types of genes were quite different. Sixteen codons were finally determined as the optimal codons. [**Conclusion**] This study confirms that natural selection is the main factor affecting codon usage bias of *C. calceolus* chloroplast genome. The optimal codon of this species is screened. The results can provide a reference for the phylogeny and chloroplast genome codon evolution of Orchidaceae.

Keywords: *Cypripedium calceolus*; chloroplast genome; codon usage bias; Orchidaceae

（责任编辑：张　研）