DOI: 10.11964/jfc.20220713597

鲤科鱼类 *SP1* 密码子的偏好性与进化

何志敏， 朱肖霞， 余清婷， 曾子豪， 肖 阳， 钟高德，
李 当， 唐建洲， 刘 臻*
长沙学院生物与化学工程学院，水生动物营养与品质调控湖南省重点实验室，湖南 长沙 410003

摘要：
【目的】特异性蛋白 1 (SP1) 是 Sp/KLF 蛋白家族成员之一，是最早被发现的转录因子，参与基因的转录调控。本研究目的是为了解鲤科鱼类 *SP* 1 在进化过程中形成的密码子使用模式及其与不同物种之间的亲缘关系，为实现鲤科鱼类异源高效表达提供理论依据。
【方法】本研究运用 Codon W、Clustal X、MEGA 4.0 及 SPSS 软件，对 4 种鲤科鱼类与其他 10 个不同物种的 *SP* 1 序列进行密码子偏好性和进化分析。
【结果】4 种鲤科鱼类 *SP* 1 编码亮氨酸和异亮氨酸时分别对 CUG 和 AUC 密码子具有较强的偏好性。14 个物种 *SP* 1 的 ENC 平均值为 50.57，CAI 值介于 0.184~0.379，均远小于 1，表明该基因密码子偏好性较弱。4 个鲤科鱼类 *SP* 1 均表现出相似的密码子偏好性。ENC-plot 分析发现 *SP* 1 密码子偏好性主要受自然选择的影响。基于 *SP* 1 序列的系统进化分析与相对同义密码子使用度 RSCU 聚类分析结果差异较小。大肠杆菌是草鱼 *SP* 1 最佳的外源表达系统，模式动物斑马鱼、小鼠均可作为草鱼 *SP* 1 的遗传转化受体。
【结论】CUG 和 AUC 为 4 个鲤科鱼类 *SP* 1 最优密码子，物种之间密码子偏好性存在差异，自然选择是导致 14 个物种 *SP* 1 密码子偏好性的主要影响因素。本研究为鲤科鱼类 *SP*1 的分类、演化与表达提供了理论依据。

关键词: 鲤科；*SP*1；密码子偏好性；分子进化；聚类分析

转录因子 SP1 是一种序列特异性的 DNA 结合蛋白，可调控某些启动子中富含 GC/GT 序列基因的转录过程，参与多种生理和病理过程的调控。在肿瘤的生长和转移过程中，*SP*1 可以通过调控癌基因、抑癌基因、细胞周期调控分子和生长相关信号转导通路、血管生成相关因子和细胞的凋亡过程，对各型肿瘤细胞产生重要影响[1]。目前已有许多动物 *SP*1 的 cDNA 被克隆和表征，包括小尾寒羊 (*Ovis aries*)[2]、猪 (Suidae)[3] 等。

某一物种或某一基因通常倾向于使用一种或几种特定的同义密码子，这些密码子被称为最优密码子 (optimal codon)，此现象被称为

第一作者：何志敏，从事水产动物分子营养研究，E-mail：Z20180831@ccsu.edu.cn

通信作者：刘臻，教授，博导，水生动物营养与品质调控湖南省重点实验室主任，湖南省领军人才，行业岗位专家，湖南省121创新工程人才，从事水生动物分子营养与饲料研究，主持包括国家面上、区域联合基金、国家重点研发等30项课题，E-mail：liuzhen_2015@sina.com

资助项目：国家自然科学基金 (31902345，U21A20267)

收稿日期：2022-07-11
修回日期：2022-08-22

文章编号：
1000-0615(2025)02-029103-12
中图分类号：S 917.4
文献标志码：A

作者声明本文无利益冲突

密码子偏性 (codon bias)[4]。根据以往研究，密码子偏性与诸多因素有关，如氨基酸组分[5]、mRNA 的二级结构[6]、翻译起始效应、G+C 含量[7]、基因长度[8-9]、tRNA 的丰度[10]、蛋白质的结构[11] 等。研究发现，密码子的使用存在着不均等现象，即使是同一物种，其不同蛋白质中相同氨基酸对密码子的选用频率也不同，即密码子的使用具有偏好性，使用频率高的密码子为高频优越密码子。密码子偏好性即指在编码氨基酸合成蛋白时，往往优先使用某一种或几种密码子，这一现象广泛存在于绝大部分的生物类群中[12-13]。

鲤科 (Cyprinidae)，属鲤形目 (Cypriniformes)，全世界共有鲤科鱼类 210 属 3 700 种以上。鲤科鱼类是鲤形目中分布最广、种类最多的一群[14-15]。本研究通过利用 Condon W 软件和 EMBOSS 在线程序分析比较 14 个不同物种的 *SP*1 密码子使用偏好性特点以及进化亲缘关系的聚类分析，为研究鲤科鱼类 *SP*1 进化过程中的使用模式提供一定的参考，为进一步研究鲤科鱼类分子系统进化、提高外源基因表达及遗传转化等奠定了基础。

1　材料与方法

1.1　*SP*1 序列数据来源

本研究选用的 14 个物种 *SP*1 序列全部来自 GenBank 数据库 (https://www.ncbi.nlm.nih.gov/nucleotide/)，登录序号见表 1。

1.2　*SP*1 的密码子偏好性分析

在 NCBI 下载各物种的 *SP*1 序列，去除 5′UTR 和 3′ UTR 序列，筛选出 CDS 序列，在 txt 文档中建立 fasta 格式，导入 codon W 软件中，对 14 条 NCBI 中不同物种的 *SP*1 序列进行分析，获得包括相对密码子使用频率 (RSCU，relative synonymouscodon usage)、有效密码子数 (ENC，effective number of codon)、密码子第 3 位各碱基的含量等密码子偏好性相关参数的数据。根据获得的 RSCU 值绘制热能图，对 RSCU 值进行图形化展示。RSCU 是指对于某一特定的密码子在编码对应氨基酸的同义密码子间的相对使用概率，若 RSCU 值为 1，表明其无偏好性；若其值大于 1，表明该密码子相对使用频率较高；若其值小于 1，则说明该密码子相对使用频率较低[16]。Fop (frequencyof optimal codons) 是密码子偏好性常用的指标，数值范围为 0.36~1.00，越接近 0.36 代表偏好性越弱[17-18]。GRAVY (grand Average of hydrophobicity) 是蛋白疏水水平，反映蛋白质的疏水性对密码子使用偏好的影响。CAI (codon adaptation index) 的值在 0~1 之间, 数值越高则表明该基因的密码子使用偏好性越强。

表 1　14 个物种 *SP*1 序列登录号

Tab. 1　Accession number of *SP*1 of 14 species

物种中文名称 Chinese name of species	物种拉丁名 Latin name of species	登录号 accession no.
草鱼	*Ctenopharyngodon idella*	KY081668.1
斑马鱼	*Danio rerio*	BC067713.1
鲫	*Carassius auratus*	KJ095609.1
齐口裂腹鱼	*Schizothorax prenanti*	MN428455.1
球形芽孢杆菌	*Bacillus sphaericus*	AF081278.1
小鼠	*Mus musculus*	AF022363.1
褐家鼠	*Rattus norvegicus*	D12768.1
小麦	*Triticum aestivum*	MN296511.1
斜纹夜蛾	*Spodoptera litura*	JN232200.1
山羊	*Capra hircus*	HM236311.1
西伯利亚花栗鼠	*Tamias sibiricus*	LC388393.1
智人	*Homo sapiens*	BC062539.1
热带爪蟾	*Xenopus tropicalis*	BC061414.1
福寿螺	*Siganus canaliculatus*	MK572810.1

1.3　ENc-plot 绘图

ENc 值是描述密码子使用偏离随机选择程度的参数，它能反映密码子家族中同义密码子非均衡使用的偏好程度，其值范围为 20~61，数值越接近于 20，说明偏好性越强。ENc-plot 分析图是以 GC3s 为变量，以 ENc 为因变量，按照 $ENc=2+GC3s+29/[GC3s^2+(1-GC3s)^2]$ 建立期望曲线，将所得的各序列以 (GC3s, ENc) 绘制散点图。各点与期望曲线的相对位置可以反映出密码子偏好性的形成是由于碱基突变还是自然选择[19]。若某一基因的密码子偏好性受突变影响较大时，其 ENc-GC3s 点将分布于期望曲线附近；若其受自然选择影响较大时，则会分布在偏离期望曲线较远的位置[20-21]。

1.4　*SP*1 的进化分析及 RSCU 值的相关聚值分析

通过 MEGA 软件构造 *SP*1 的系统进化树，分析其在进化上的亲缘关系。建树方法选用邻接法 (Neighbor-joining)，检验方法设为 Bootstrp method，检验次数设为 1 000。利用 SPSS 软件

组间联结法对各基因的 RSCU 值进行系统聚类分析。通过二者对比，分析 *SP*1 密码子偏好性与物种进化之间的关系。

2　结果

2.1　*SP*1 同义密码子相对使用度分析

Codon W 分析后获得 14 个物种 *SP*1 序列的 RSCU 值，将 RSCU 值标准化后，获得基于RSCU值的热图 (图 1)，发现不同物种 *SP*1 的密码子相对使用度有一定差异。对 14 个物种的 59 个密码子的 RSCU 值进行计算，若 RSCU 值大于 1.6，说明密码子偏好性较强。RSCU 值等于 1 的密码子有 34 个，如草鱼 *SP*1 编码脯氨酸的 CCA；RSCU 值大于 1 的密码子共有 389 个，小鼠、褐家鼠、山羊和福寿螺均有 31 个；偏好性较强的密码子有斑马鱼 *SP*1 编码亮氨酸的 CUG 密码子、球形芽孢杆菌 *SP*1 编码亮氨酸的 UUA 密码子、鲫 *SP*1 编码异亮氨酸的 AUC 等 108 个，4 个鲤科鱼类的 CUG 和 AUC 密码子

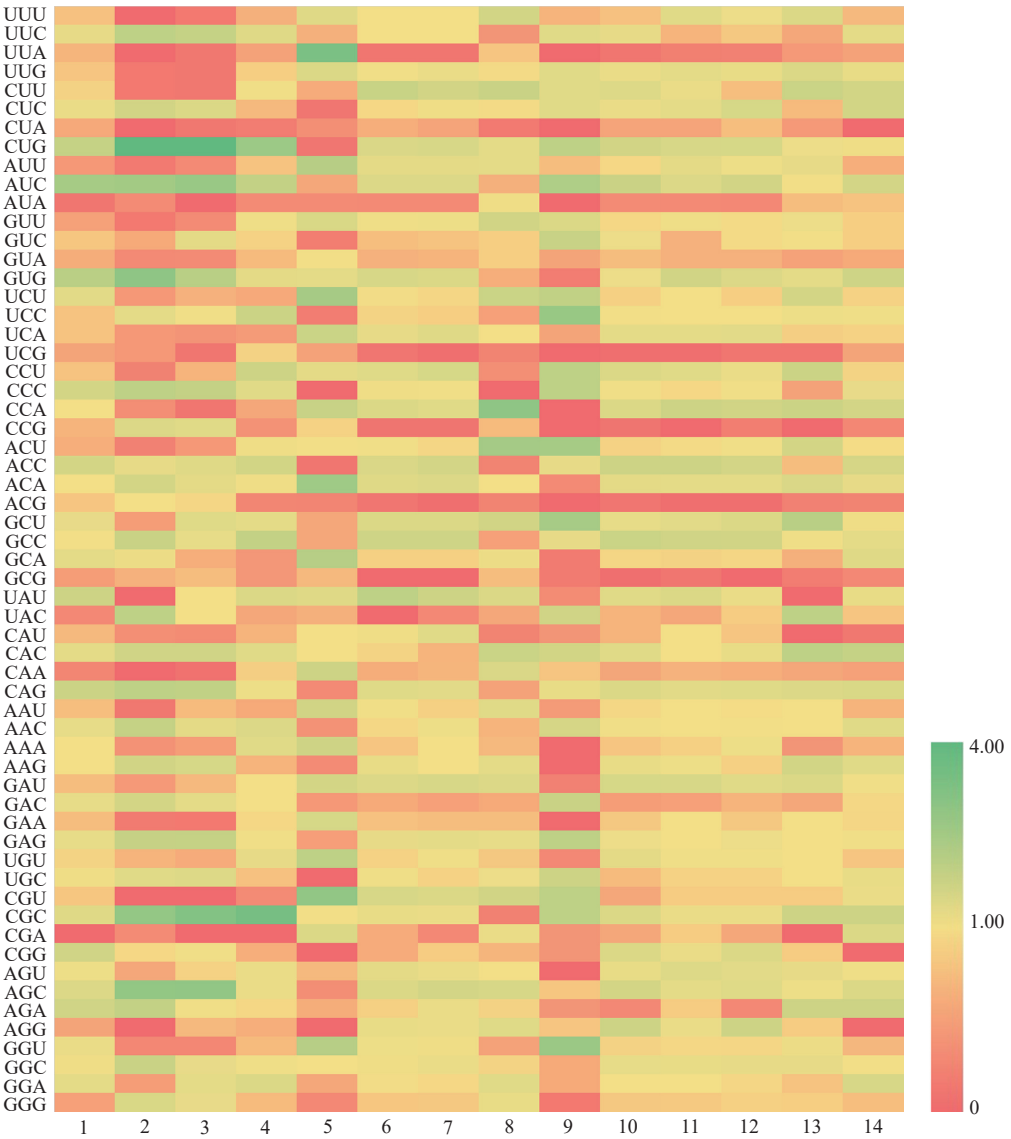

图 1　14 个物种 *SP*1 的 RSCU 热图

1. 草鱼，2. 斑马鱼，3. 鲫，4. 齐口裂腹鱼，5. 球形芽孢杆菌，6. 小鼠，7. 褐家鼠，8. 小麦，9. 斜纹夜蛾，10. 山羊，11. 西伯利亚花栗鼠，12. 智人，13. 热带爪蟾，14. 福寿螺，下同。

Fig. 1　Heat map of RSCU values of *SP*1 from 14 species

1. *C. idella*, 2. *D. rerio*, 3. *C. auratus*, 4. *S. prenanti*, 5. *B. sphaericus*, 6. *M. musculus*, 7. *R. norvegicus*, 8. *T. aestivum* , 9. *S. litura* , 10. *C. hircus*, 11. *T. sibiricus*, 12. *H. sapiens*, 13. *X. tropicalis*, 14. *S. canaliculatus*, the same below.

RSCU 均值分别为 3.150和 2.395，表明这两个密码子偏好性很强，CUG 和 AUC 为 4 个鲤科鱼类 *SP*1 最优密码子 (表 2)。另外，这 4 种鱼类中，斑马鱼和鲫在 UUC、CUG、AUC、GUG、CCC、CAC、CAG、CGC、AGC 这 9 个密码子中的偏好性均较强且趋势相近，说明斑马鱼和鲫在密码子的偏好性上具有一定相似性，在进化过程中亲缘关系更近。此外 4 个鲤科鱼类 *SP*1 都对 UUU、UUA、UUG、CUU 等 38 个密码子偏好性较弱，说明这 4 个鲤科鱼类部分密码子存在偏好性一致的现象。其余 10 个物种中，斜纹夜蛾 CUG、AUC、UCC 等 16 个密码子偏好性较强，球形芽孢杆菌 UUA、AUU、UCU 等 13 个密码子偏好性较强，小鼠等 8 个物种 *SP*1 的绝大多数密码子 RSCU 值均小于 1，偏好性较弱。10 个物种同一个密码子的 RSCU 值差异较大，但也存在偏好性一致的密码子，包括 UUU、CUA、UCG、ACG 等密码子，14 个物种 *SP*1 约二分之一的密码子偏好性均较弱。

表 2　14 种物种 *SP*1 同义密码子的相对使用度

Tab. 2　Relative synonymous codon usage of *SP*1 in 14 species

编码氨基酸 encoded amino acid	密码子 codon	草鱼 *C. idella*	斑马鱼 *D. rerio*	鲫 *C. auratus*	齐口裂腹鱼 *S. prenanti*	球形芽孢杆菌 *B. sphaericus*	小鼠 *M. musculus*	褐家鼠 *R. norvegicus*	小麦 *T. aestivum*	斜纹夜蛾 *S. litura*	山羊 *C. hircus*	西伯利亚花栗鼠 *T. sibiricus*	智人 *H. sapiens*	热带爪蟾 *X. tropicalis*	福寿螺 *S. canaliculatus*
苯丙氨酸 Phe	UUU	0.77	0.00	0.15	0.64	1.36	1.00	1.00	1.60	0.67	0.77	1.33	1.17	1.43	0.71
	UUC	1.23	2.00	1.85	1.36	0.64	1.00	1.00	0.40	1.33	1.23	0.67	0.83	0.57	1.29
亮氨酸 Leu	UUA	0.69	0.00	0.14	0.53	3.36	0.11	0.11	0.81	0.00	0.11	0.21	0.22	0.44	0.53
	UUG	0.81	0.15	0.14	0.88	1.44	1.05	1.18	0.97	1.33	1.18	1.29	1.20	1.46	1.20
	CUU	0.92	0.15	0.14	1.06	0.60	1.79	1.61	1.78	1.33	1.39	1.18	0.76	1.76	1.60
	CUC	1.15	1.61	1.43	0.71	0.12	0.95	1.07	0.97	1.33	1.18	1.29	1.53	0.73	1.60
	CUA	0.58	0.00	0.14	0.18	0.36	0.63	0.54	0.16	0.00	0.54	0.54	0.76	0.44	0.00
	CUG	1.85	4.10	4.00	2.65	0.12	1.47	1.50	1.30	2.00	1.61	1.50	1.53	1.17	1.07
异亮氨酸 Ile	AUU	0.43	0.16	0.30	0.79	2.11	1.27	1.27	1.29	0.75	0.95	1.27	1.10	1.22	0.63
	AUC	2.46	2.53	2.70	1.89	0.57	1.42	1.42	0.64	2.25	1.74	1.42	1.61	1.03	1.58
	AUA	0.11	0.32	0.00	0.32	0.32	0.30	0.30	1.07	0.00	0.32	0.30	0.29	0.75	0.79
缬氨酸 Val	GUU	0.51	0.15	0.32	1.09	1.49	1.08	1.11	1.63	1.45	0.95	1.08	0.97	1.16	0.88
	GUC	0.82	0.59	1.28	0.91	0.19	0.76	0.78	0.88	1.82	1.16	0.65	0.97	1.03	0.88
	GUA	0.62	0.30	0.32	0.73	1.02	0.65	0.67	0.88	0.55	0.74	0.65	0.65	0.52	0.59
	GUG	2.05	2.96	2.08	1.27	1.30	1.51	1.44	0.63	0.18	1.16	1.62	1.41	1.29	1.66
丝氨酸 Ser	UCU	1.34	0.43	0.65	0.58	2.47	0.99	0.94	1.75	1.91	0.90	1.01	0.88	1.56	0.91
	UCC	0.79	1.29	1.04	1.73	0.18	0.93	0.88	0.50	2.73	1.02	1.01	1.00	1.11	1.09
	UCA	0.79	0.43	0.39	0.46	1.76	1.22	1.35	1.00	0.55	1.26	1.26	1.31	0.89	0.91
	UCG	0.55	0.43	0.13	0.92	0.53	0.12	0.06	0.25	0.00	0.06	0.06	0.13	0.11	0.55
脯氨酸 Pro	CCU	0.78	0.24	0.67	1.71	1.27	1.37	1.52	0.36	2.00	1.43	1.33	1.14	1.73	0.93
	CCC	1.56	2.00	1.87	1.33	0.00	1.07	1.05	0.00	2.00	1.05	0.95	1.05	0.53	1.21
	CCA	1.00	0.35	0.13	0.57	1.82	1.46	1.33	2.91	0.00	1.43	1.71	1.62	1.73	1.58
	CCG	0.67	1.41	1.33	0.38	0.91	0.10	0.10	0.73	0.00	0.10	0.00	0.19	0.00	0.28
苏氨酸 Thr	ACU	0.63	0.22	0.43	1.07	1.03	1.06	0.94	2.5	2.46	0.91	0.97	1.06	1.56	0.98
	ACC	1.56	1.22	1.36	1.60	0.13	1.47	1.59	0.25	1.23	1.70	1.71	1.59	0.74	1.54
	ACA	1.00	1.56	1.28	1.07	2.58	1.35	1.41	1.00	0.31	1.27	1.26	1.29	1.48	1.23
	ACG	0.81	1.00	0.94	0.27	0.26	0.12	0.06	0.25	0.00	0.12	0.06	0.06	0.22	0.25

·续表 2·

编码氨基酸 encoded amino acid	密码子 codon	草鱼 *C. idella*	斑马鱼 *D. rerio*	鲫 *C. auratus*	齐口裂腹鱼 *S. prenanti*	球形芽孢杆菌 *B. sphaericus*	小鼠 *M. musculus*	褐家鼠 *R. norvegicus*	小麦 *T. aestivum*	斜纹夜蛾 *S. litura*	山羊 *C. hircus*	西伯利亚花栗鼠 *T. sibiricus*	智人 *H. sapiens*	热带爪蟾 *X. tropicalis*	福寿螺 *S. canaliculatus*
丙氨酸 Ala	GCU	1.21	0.47	1.38	1.26	0.57	1.45	1.45	1.63	2.43	1.25	1.31	1.44	2.09	1.07
	GCC	1.02	1.77	1.25	1.89	0.57	1.66	1.66	0.50	1.22	1.75	1.64	1.63	1.09	1.29
	GCA	1.30	1.12	0.63	0.42	2.14	0.90	0.90	1.13	0.17	0.94	0.92	0.94	0.64	1.36
	GCG	0.47	0.65	0.75	0.42	0.71	0.00	0.00	0.75	0.17	0.06	0.13	0.00	0.18	0.29
酪氨酸 Tyr	UAU	1.71	0.00	1.00	1.43	1.36	2.00	1.71	1.43	0.33	1.33	1.43	1.14	0.00	1.20
	UAC	0.29	2.00	1.00	0.57	0.64	0.00	0.29	0.57	1.67	0.67	0.57	0.86	2.00	0.80
组氨酸 His	CAU	0.71	0.36	0.33	0.67	1.00	1.07	1.33	0.25	0.40	0.67	1.00	0.8	0.00	0.14
	CAC	1.29	1.64	1.67	1.33	1.00	0.93	0.67	1.75	1.60	1.33	1.00	1.20	2.00	1.86
谷氨酰胺 Gln	CAA	0.27	0.00	0.08	0.89	1.70	0.62	0.69	1.47	0.80	0.56	0.66	0.63	0.57	0.51
	CAG	1.73	2.00	1.92	1.11	0.30	1.38	1.31	0.53	1.20	1.44	1.34	1.37	1.42	1.49
天冬酰胺 Asn	AAU	0.76	0.14	0.73	0.59	1.62	1.04	0.90	1.33	0.46	0.96	1.00	0.98	1.00	0.67
	AAC	1.24	1.86	1.27	1.41	0.38	0.96	1.10	0.67	1.54	1.04	1.00	1.02	1.00	1.33
赖氨酸 Lys	AAA	1.00	0.38	0.47	1.33	1.68	0.80	1.00	0.71	0.00	0.80	0.90	1.10	0.4	0.67
	AAG	1.00	1.63	1.53	0.67	0.32	1.20	1.00	1.29	0.00	1.20	1.10	0.90	1.60	1.33
天冬氨酸 Asp	GAU	0.75	0.44	0.71	1.00	1.57	1.41	1.50	1.41	0.22	1.53	1.50	1.33	1.43	1.04
	GAC	1.25	1.56	1.29	1.00	0.43	0.59	0.5	0.59	1.78	0.47	0.50	0.67	0.57	0.96
谷氨酸 Glu	GAA	0.75	0.17	0.15	0.95	1.52	0.77	0.74	0.75	0.00	0.83	1.00	0.83	1.00	0.94
	GAG	1.25	1.83	1.85	1.05	0.48	1.23	1.26	1.25	2.00	1.17	1.00	1.17	1.00	1.06
半胱氨酸 Cys	UGU	0.93	0.67	0.60	1.23	2.00	0.91	1.09	0.83	0.29	1.27	1.09	1.09	1.00	0.80
	UGC	1.07	1.33	1.40	0.77	0.00	1.09	0.91	1.17	1.71	0.73	0.91	0.91	1.00	1.20
精氨酸 Arg	CGU	0.82	0.00	0.00	0.32	2.90	1.50	1.43	1.62	2.00	0.57	0.86	0.86	0.86	1.14
	CGC	1.36	2.84	3.18	3.47	1.03	1.20	1.14	0.23	2.00	1.43	1.14	1.14	1.71	1.71
	CGA	0.00	0.32	0.00	0.00	1.45	0.60	0.29	1.15	0.40	0.57	0.86	0.57	0.00	1.43
	CGG	1.64	0.95	1.06	0.63	0.00	0.60	0.86	0.69	0.40	1.43	1.14	1.43	0.86	0.00
丝氨酸 Ser	AGU	1.11	0.57	0.91	1.15	0.71	1.28	1.18	1.00	0.00	1.20	1.39	1.31	1.22	1.09
	AGC	1.42	2.86	2.87	1.15	0.35	1.46	1.59	1.50	0.82	1.56	1.26	1.38	1.11	1.45
精氨酸 Arg	AGA	1.64	1.89	1.06	0.95	0.62	0.9	1.14	0.92	0.40	0.29	0.86	0.29	1.71	1.71
	AGG	0.55	0.00	0.71	0.63	0.00	1.20	1.14	1.38	0.80	1.71	1.14	1.71	0.86	0.00
甘氨酸 Gly	GGU	1.14	0.28	0.28	0.73	2.14	1.11	1.07	0.53	2.67	0.92	0.96	0.94	1.13	0.70
	GGC	1.07	1.77	1.21	1.09	1.00	1.07	1.15	0.93	0.59	1.25	1.20	1.22	1.22	1.03
	GGA	1.29	0.47	1.30	1.45	0.57	0.99	0.95	1.33	0.59	1.00	1.00	0.94	0.78	1.51
	GGG	0.50	1.49	1.21	0.73	0.29	0.82	0.83	1.20	0.15	0.83	0.84	0.90	0.87	0.76

2.2 ENC-plot 分析

ENC-plot 图显示，14 个物种 *SP*1 位点均分布在标准曲线下方，实际 ENC 值与理论 ENC 值存在差异，且大部分离标准曲线较远，说明大部分物种 *SP*1 密码子偏好性形成受自然选择的作用较大，14 个物种的 ENC 平均值为 50.57，说明这 14 个物种 *SP*1 密码子的偏好性均较弱。其中，西伯利亚花栗鼠和智人这两个物种位于

期望曲线附近，表明这两个物种的 *SP*1 密码子偏好性的形成主要受到突变影响。相反地，褐家鼠和斜纹夜蛾这两个物种分布于离期望曲线较远的位置，表明自然选择主导这两个物种的 *SP*1 密码子偏好性 (图 2)。

2.3　*SP*1 密码子偏好性相关参数分析

14 个物种的 CAI 值范围为 0.184~0.379，数值差异不大，且均远小于 1，说明密码子的偏好性水平较弱 (表 3)。其中，14 个物种的 CBI (codon bias index) 值存在一定的差异。4 个鱼种的 FOP (frequency of optimal codons) 值均大于 0.5，相对来说密码子使用强度较强，其中球形芽孢杆菌（*Bacillus sphaericus*）的 FOP 值最低，密码子的使用强度最弱。14 个物种的 L-sym 和 L-aa 数值相差较大。相比之下，小鼠、山羊和智人的 L-sym 值最大，斜纹夜蛾的 L-sym 值最低，不同物种间的 L-aa 值大小趋势和 L-sym 值一致，且同一物种的 L-aa 均比 L-sym 值大。除斜纹夜蛾外，其余物种的 GRAVY 值均小于 0，表明氨基酸的亲水性对 14 个物种中大部分物种的密码子使用偏好存在一定影响。Aromo 值反映芳香族蛋白质对密码子使用偏好性的影响，14 个物种的 Aromo 值均小于 0.13，足以说明芳香族蛋白质对密码子使用偏好性影响不大(表 3)。其中，斑马鱼和鲫的 CAI、CBI、FOP 相差不大，说明这两类鱼种在 *SP*1 表达水平上有高度的相似性。

2.4　*SP*1 系统进化及 RSCU 值聚类分析

通过邻接法构建进化树可以判断各物种 *SP*1 在进化上的亲缘关系。从总体上看，14 个物种的 *SP*1 被分为三大类。4 个鲤科鱼类有 3 个聚为一支，草鱼与福寿螺等聚为一支，且同为鼠类的褐家鼠与小鼠聚类在不同的分支上 (图 3)。这种基因的聚类和物种分类存在冲突的现象，在动植物中已经被广泛发现，其可能原因包括基因的倍增和重组、水平的基因转移等。

图 4 为基于各物种 *SP*1 RSCU 值的聚类分析结果。其聚类结果与基于 CDS 序列的进化树

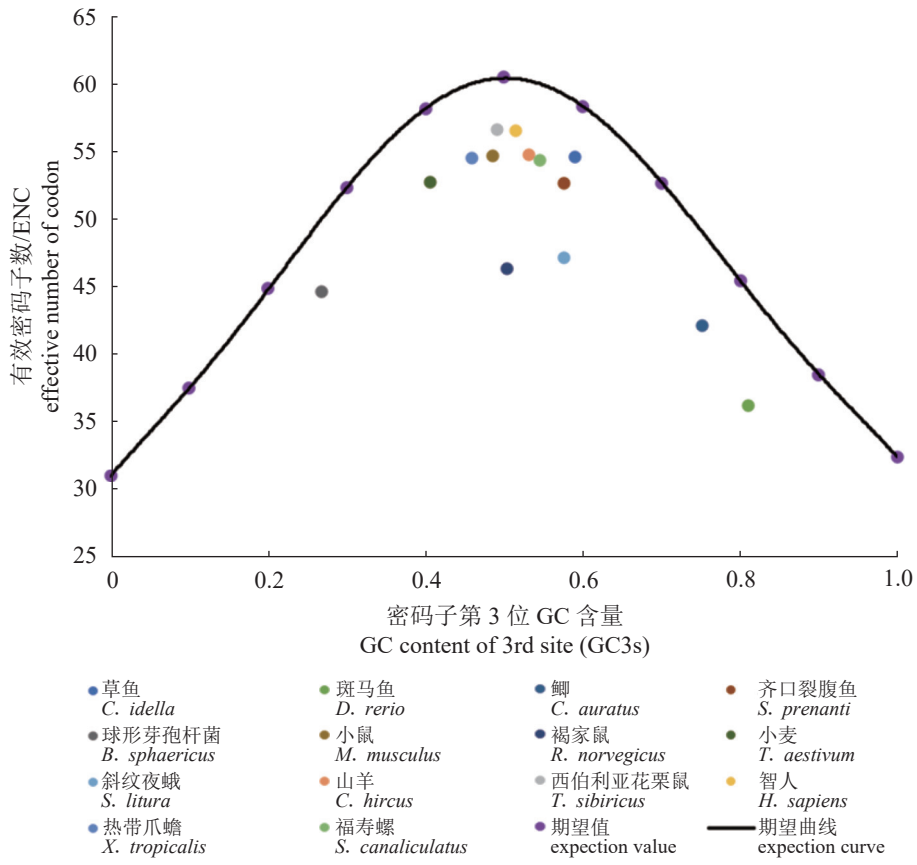

图 2　不同物种 *SP*1 的 ENC-plot 分析

Fig. 2　Analysis of ENc-Plot of *SP*1 in different species

表 3　14 种物种 *SP*1 密码子偏好性相关参数

Tab. 3　Codon bias parameters of *SP*1 in 14 species

物种 species	密码子适应指数 CAI	密码子偏好性指数 CBI	最优密码子使用频率 Fop	同义氨基酸数 L_sym	氨基酸数 L_aa	氨基酸疏水性平均值 GRAVY	芳香度 Aromo
草鱼　*C. idella*	0.267	0.156	0.519	634	650	−0.432 154	0.036 923
斑马鱼　*D. rerio*	0.294	0.250	0.566	447	457	−0.268 271	0.039 387
鲫　*C. auratus*	0.292	0.263	0.577	461	470	−0.424 468	0.042 553
齐口裂腹鱼　*S. prenanti*	0.258	0.147	0.504	349	359	−0.093 593	0.083 565
球形芽孢杆菌　*B. sphaericus*	0.201	−0.160	0.320	272	285	−0.507 368	0.108 772
小鼠　*M. musculus*	0.235	0.046	0.463	767	784	−0.452 041	0.029 337
褐家鼠　*R.norvegicus*	0.184	−0.043	0.411	711	759	−0.902 767	0.126 482
小麦　*T. aestivum*	0.196	−0.022	0.397	330	341	−0.010 850	0.061 584
斜纹夜蛾　*S. litura*	0.379	0.436	0.679	224	232	0.036 207	0.090 517
山羊　*C. hircus*	0.237	0.069	0.476	767	786	−0.440 458	0.029 262
西伯利亚花栗鼠　*T. sibiricus*	0.231	0.043	0.461	764	781	−0.437 132	0.029 449
智人　*H. sapiens*	0.235	0.062	0.472	767	785	−0.435 924	0.029 299
热带爪蟾　*X. tropicalis*	0.245	0.098	0.485	497	509	−0.278 978	0.025 540
福寿螺　*S. canaliculatus*	0.236	0.089	0.483	679	701	−0.419 971	0.032 810

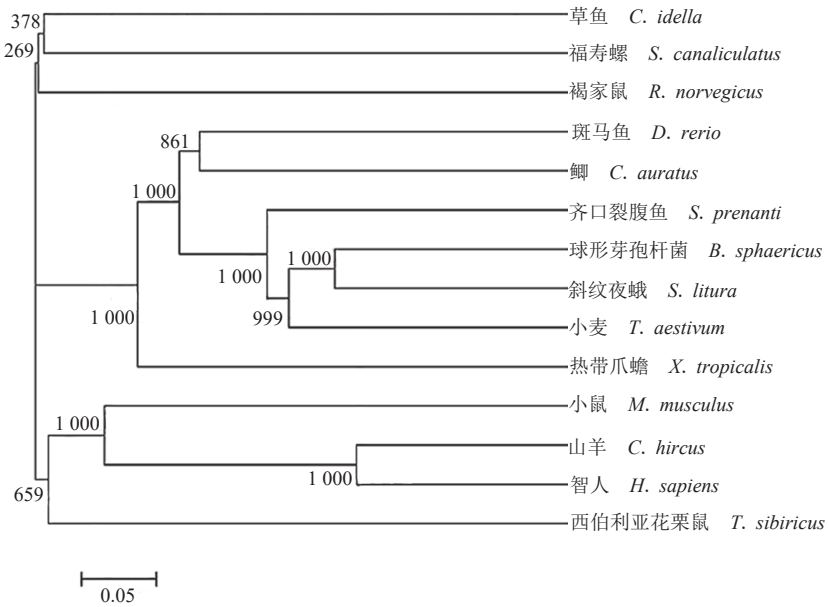

图 3　*SP*1 系统进化分析

Fig. 3　Phylogenetic analysis of *SP*1

的分析结果基本一致，但是存在一定的差异。在 RSCU 聚类分析结果中，可以明显看出原核生物和真核生物分别聚类。14 个物种中，原核生物球形芽孢杆菌单独聚为一支，其余 13 种真核生物又聚为两个分支，分别是鲤科鱼类与非鲤科鱼类，而草鱼的聚类情况与系统进化树分析结果一致，其与福寿螺等聚为一支。同科属动物的 *SP*1 被聚类到不同的分支上，说明鲤科鱼类的 *SP*1 具有密码子偏好性差异。

2.5　*SP*1 密码子成分相关分析

表 4 列出了 14 个物种的 *SP*1 编码区核苷酸序列中 T3s、C3s、A3s、G3s、Nc、GC3s、GC 的含量。除球形芽孢杆菌、小麦、小鼠和热带爪蟾外，其余物种均是 C3s 值最大，说明这 14 个物种中的大部分物种在第 3 位的核苷酸

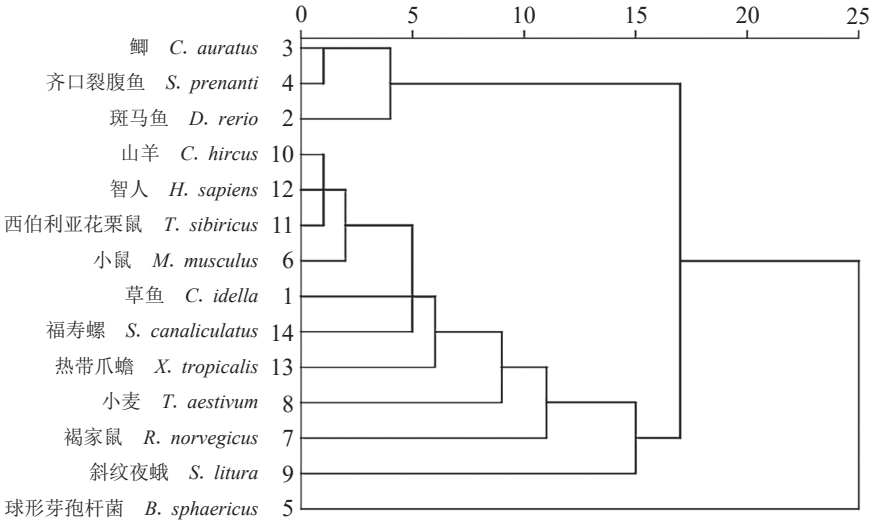

图 4　*SP*1 的 RSCU 值聚类分析

Fig. 4　RSCU cluster analysis of *SP*1

表 4　*SP*1 密码子碱基组成特性分析

Tab. 4　Characteristic analysis of base composition of *SP*1 codon

物种 species	密码子第3位上的T含量 T3s	密码子第3位上的C含量 C3s	密码子第3位上的A含量 A3s	密码子第3位上的G含量 G3s	密码子数 Nc	密码子第3位上的GC含量 GC3s	密码子GC含量 GC
草鱼 *C. idella*	0.271 5	0.384 3	0.223 9	0.346 7	54.64	0.590	0.544
斑马鱼 *D. rerio*	0.088 2	0.518 7	0.137 9	0.469 3	36.19	0.810	0.649
鲫 *C. auratus*	0.165 8	0.484 2	0.135 8	0.446 3	42.14	0.751	0.609
齐口裂腹鱼 *S. prenanti*	0.301 0	0.433 7	0.209 1	0.274 6	52.63	0.576	0.513
球形芽孢杆菌 *B. sphaericus*	0.461 9	0.133 3	0.492 8	0.239 4	44.62	0.268	0.360
小鼠 *M. musculus*	0.366 1	0.337 6	0.249 6	0.257 7	54.66	0.485	0.523
褐家鼠 *R. norvegicus*	0.260 8	0.373 8	0.371 2	0.259 8	46.31	0.504	0.538
小麦 *T. aestivum*	0.393 6	0.219 9	0.314 8	0.281 3	52.77	0.406	0.492
斜纹夜蛾 *S. litura*	0.363 2	0.509 4	0.113 2	0.147 9	47.15	0.576	0.533
山羊 *C. hircus*	0.317 5	0.383 9	0.242 7	0.265 8	54.78	0.531	0.540
西伯利亚花栗鼠 *T. sibiricus*	0.343 4	0.346 5	0.263 0	0.254 1	56.67	0.491	0.527
智人 *H. sapiens*	0.321 8	0.369 1	0.258 0	0.260 6	56.54	0.514	0.535
热带爪蟾 *X. tropicalis*	0.406 4	0.290 6	0.241 9	0.276 4	54.53	0.459	0.500
福寿螺 *S. canaliculatus*	0.265 7	0.393 2	0.282 0	0.283 3	54.35	0.545	0.539

多为 C，即碱基 C 的使用频率更高，表明 *SP*1 密码子使用优先选择 C 末端密码子，即 C 末端统一密码子优先用于 *SP*1 编码区[21]。这 14 个物种的 Nc 值为 36.19~56.67，均高于偏好性显著阈值 35，而接近于无偏好性的情况，说明这些物种的密码子偏好性较弱，在编码蛋白质过程中对密码子的选择更倾向于随机。球形芽孢杆菌的 GC3s 为 0.268，斑马鱼的 GC3s 为 0.81，鲫的 GC3s 为 0.751，其余物种的 GC3s 值范围为 0.459~0.590，差异较小。

2.6　草鱼与模式生物密码子偏好性比较

草鱼与酵母 (*Saccharomyces cerevisiae*)、小鼠、斑马鱼、大肠杆菌 (*Escherichia coli*) 这 4 种物种密码子使用频率进行比较 (表 5)，其中 *Ci/Sc*、*Ci/Mm*、*Ci/Dr*、*Ci/Ec* 分别表示草鱼 *SP*1 与大肠杆菌、小鼠、斑马鱼、酵母基因组的每种密码子使用频率比值，比值为 0.5~2.0，表明这两个物种对该密码子的偏好性较接近，

表 5 草鱼 *SP*1 与模式生物基因组密码子使用频率的比较

Tab. 5 Comparison of codon usage bias between *C. idella SP*1 and genome of pattern organism

氨基酸 amino acid	密码子 codon	草鱼/大肠杆菌 *Ci/Ec*	草鱼/酵母 *Ci/Sc*	草鱼/小鼠 *Ci/Mm*	草鱼/斑马鱼 *Ci/Dr*	氨基酸 amino acid	密码子 codon	草鱼/大肠杆菌 *Ci/Ec*	草鱼/酵母 *Ci/Sc*	草鱼/小鼠 *Ci/Mm*	草鱼/斑马鱼 *Ci/Dr*
苯丙氨酸 Phe	TTT	0.216	0.192	0.291	0.275	丙氨酸 Ala	GCT	0.834	0.613	0.650	0.622
	TTC	0.475	0.435	0.367	0.385		GCC	0.426	0.873	0.423	0.564
亮氨酸 Leu	TTA	0.432	0.229	0.896	0.857		GCA	0.678	0.864	0.886	0.843
	TTG	0.500	0.257	0.522	0.569		GCG	0.157	0.806	0.781	0.581
	CTT	0.685	0.650	0.597	0.630	酪氨酸 Tyr	TAT	0.363	0.319	0.492	0.476
	CTC	0.907	1.852	0.495	0.588		TAC	0.083	0.068	0.062	0.059
	CTA	1.254	0.373	0.617	0.806	组氨酸 His	CAT	0.369	0.368	0.472	0.459
	CTG	0.318	1.524	0.405	0.426		CAC	0.924	1.154	0.588	0.608
异亮氨酸 Ile	ATT	0.133	0.133	0.260	0.242	谷氨酰胺 Gln	CAA	0.666	0.366	0.833	0.847
	ATC	0.951	1.337	1.022	0.970		CAG	2.124	5.207	1.848	1.881
	ATA	0.185	0.056	0.135	0.130	天冬酰胺 Asn	AAT	0.701	0.364	0.833	0.798
蛋氨酸 Met	ATG	0.444	0.574	0.526	0.471		AAC	0.984	0.847	1.034	0.871
缬氨酸 Val	GTT	0.270	0.226	0.467	0.355	赖氨酸 Lys	AAA	0.333	0.263	0.502	0.375
	GTC	0.533	0.678	0.519	0.541	天冬氨酸 Asp	AAG	1.025	0.357	0.327	0.358
	GTA	0.541	0.508	0.811	0.896		GAT	0.281	0.239	0.429	0.363
	GTG	0.786	1.852	0.704	0.707		GAC	0.806	0.743	0.577	0.540
丝氨酸 Ser	TCT	1.954	0.723	1.049	1.006	谷氨酸 Glu	GAA	0.158	0.132	0.222	0.246
	TCC	1.124	0.704	0.552	0.658		GAG	0.563	0.521	0.254	0.234
	TCA	1.282	0.535	0.847	0.758	半胱氨酸 Cys	TGT	1.270	0.864	0.614	0.619
	TCG	0.805	0.814	1.667	1.250		TGC	1.160	1.667	0.650	0.714
	AGT	1.476	0.986	1.102	1.061	精氨酸 Arg	CGT	0.145	0.469	0.638	0.435
	AGC	1.125	1.837	0.914	0.978		CGA	0.000	0.000	0.000	0.000
脯氨酸 Pro	CCT	0.959	0.519	0.380	0.422		CGG	0.950	3.529	0.588	0.909
	CCC	2.415	2.059	0.769	1.102		AGA	2.069	0.282	0.496	0.420
	CCA	1.059	0.492	0.520	0.573		AGG	1.053	0.217	0.164	0.196
	CCG	0.276	1.132	0.968	0.732	甘氨酸 Gly	GGT	0.655	0.669	1.404	1.168
苏氨酸 Thr	ACT	1.101	0.493	0.730	0.690		GGC	0.545	1.531	0.708	0.872
	ACC	1.084	1.969	1.316	1.543		GGA	1.997	1.651	1.071	0.837
	ACA	1.950	0.899	1.000	0.941		GGG	0.621	1.167	0.461	0.700
色氨酸 Trp	TGG	0.263	0.385	0.320	0.345						

若≥2.0 或≤0.5，则表明偏好性差异较大。通过比较密码子使用频率，分析频率差异性较大(小于 0.5 大于 2 的频率比值) 的密码子个数以确定最佳的外源表达系统和遗传转化受体。草鱼 *SP*1 与酵母、大肠杆菌差异较大的密码子个数分别为 27 和 23，说明大肠杆菌是草鱼 *SP*1 的最佳外源表达系统。与小鼠、斑马鱼基因组使用频率比较，差异较大的密码子个数均为 21，说明小鼠与斑马鱼均可作为草鱼 *SP*1 的遗传转化受体。

3 讨论

本研究通过对 14 个不同物种 *SP*1 密码子的偏好性分析，发现密码子 CUG 为 4 个鲤科鱼类 *SP*1 的最优密码子。通过分析不同物种 RSCU 大于 2 的偏好性密码子发现，鲫、球形芽孢杆菌、斑马鱼、斜纹夜蛾中分别有 5、7、8 和 11 个 RSCU 大于 2 的密码子，热带爪蟾、草鱼、齐口裂腹鱼和小麦分别有 3、2、2 和 2 个，其余物种只有 1 个或没有，表明 *SP*1 对

优势密码子的使用偏好性较弱。进一步 ENC 和 Fop 的分析发现，14 个物种的 ENC 平均值为 50.57，Fop 均值比较接近 0.50，表明所选样本 *SP*1 的密码子偏好性都普遍较弱。综上所述，ENC 与 RSCU 分析结果均指明 *SP*1 密码子偏好性较弱。密码子适应指数 CAI 与基因的表达水平通常呈正相关，而 ENC 则相反。本研究的 14 个物种 *SP*1 序列的 CAI 值均小于 0.3，而 ENC 均值大于 50，呈现低 CAI 值、高 ENC 值的趋势，表明 14 个物种 *SP*1 的表达水平较低，而如果基因表达水平低，往往其对密码子的偏好性也较低，这一结果与前面的分析结论一致。另外，*SP*1 的较低相对表达水平也在草鱼中被验证[22]。

密码子的偏好性在形成中的影响因素主要是突变和自然选择。在本研究中，ENC-plot 显示，14 个不同物种中大部分离标准曲线较远，说明 *SP*1 密码子使用偏好性在形成中的原因主要是自然选择，如褐家鼠和斜纹夜蛾的 *SP*1 密码子偏好性主要受到自然选择或其他因素的影响，可能是 tRNA 丰度、基因长度或结构等[23-24]，而西伯利亚花栗鼠和智人的 *SP*1 密码子偏好性的形成主要受到突变影响。对 14 个不同物种 *SP*1 密码子的相关参数分析发现，不同物种的 L-aa 值与 L-sym 值有相同的大小趋势，同一物种的 L-aa 值大于 L-sym 值；GRAVY 值体现蛋白质的亲水性对大部分物种的密码子使用偏好存在一定影响；Aromo 值均小于 0.13，说明芳香族蛋白质对密码子使用偏好性影响不大。在对 *SP*1 密码子成分相关分析后，发现大部分物种的 *SP*1 密码子优先选择 C 末端密码子，而 14 个物种的 Nc 值均高于显著阈值，所以在编码蛋白质过程中对密码子的选择倾向于随机。

亲缘关系相近的物种在密码子使用偏好性方面会有一定的相似性[25]。基于 *SP*1 序列的进化分析与传统的动物分类结果基本一致，表明 *SP*1 与动物的进化过程密切相关。同为鼠科 (Muridae) 的不同种属有与其他物种混合交错聚为一支的现象，其原因可能与 *SP*1 序列在不同物种中的高同源性有关。在对 14 个物种进行基于 *SP*1 CDS 聚类分析与基于 RSCU 聚类分析后发现，鲫、斑马鱼和齐口裂腹鱼，草鱼和福寿螺的聚类结果在两种聚类分析中保持一致。两种聚类分析结果不完全相同，如球形芽孢杆菌在 RSCU 聚类分析中单独聚为一支，但在系统进化分析中发现其与斜纹夜蛾等聚为一大支，这可能是由于单一基因密码子使用偏好性聚类分析不能准确细化地反映物种间的系统进化关系，或是各个物种对 *SP*1 密码子的使用模式存在差异，还需综合其他方面的研究。RSCU 聚类在一定程度上能够反映出物种进化的关系，但容易受到更多因素的干扰，比较适合在较小的分类单元中进行细化分析时提供更多研究依据，因此相较于不同物种间比对，基于序列的系统进化树分析具有更高的准确度[25]。但是从本研究结果中可以看出，基于 RSCU 的聚类分析结果相比基于 CDS 的进化分析结果更符合传统物种分类学结果。

通过对比分析草鱼与模式生物的密码子使用频率，发现大肠杆菌为草鱼 *SP*1 的最佳外源表达系统，但是为达到较高水平的蛋白表达需要对密码子进行优化，而斑马鱼与小鼠均可作为其遗传转化受体。本研究为鲤科鱼类 *SP*1 进化过程中的使用模式提供一定的参考，为鲤科鱼类 *SP*1 的分类与演化、表达调控及遗传育种等研究提供了科学依据。

参考文献 (References):

[1] 白雪, 邓红. 转录因子 Sp1 与肿瘤关系研究的新进展 [J]. 浙江大学学报 (医学版), 2010, 39(2): 215-220.
Bai X, Deng H. Research progress on relationship between transcription factor Sp1 and tumor[J]. Journal of Zhejiang University (Medical Sciences), 2010, 39(2): 215-220 (in Chinese).

[2] 王凤, 赵弼时, 刘旭莹, 等. 小尾寒羊 *SP*1 基因克隆及其对前体脂肪细胞分化的影响 [J]. 中国畜牧兽医, 2021, 48(5): 1544-1557.
Wang F, Zhao B S, Liu X Y, *et al*. Cloning of *SP*1 gene and its effect on preadipocytes differentiation in small-tailed Han Sheep[J]. China Animal Husbandry & Veterinary Medicine, 2021, 48(5): 1544-1557 (in Chinese).

[3] 张广杰, 崔悦悦, 邱庆庆, 等. 广西巴马小型猪 *SP1* 基因克隆测序及其真核表达载体的构建 [J]. 南方农业学报, 2018, 49(2): 360-366.
Zhang G J, Cui Y Y, Qiu Q Q, *et al*. Clone sequencing of gene *SP1* in Guangxi Bama mini pig and construction of its eukaryotic expression vector[J]. Journal of Southern Agriculture, 2018, 49(2): 360-366 (in Chinese).

[4] 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.

Wu X M, Wu S F, Ren D M, *et al*. The analysis method and progress in the study of codon bias[J]. Hereditas (Beijing), 2007, 29(4): 420-426 (in Chinese).

[5] D'Onofrio G, Mouchiroud D, Aïssani B, *et al*. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins[J]. Journal of Molecular Evolution, 1991, 32(6): 504-510.

[6] Mita K, Ichimura S, Zama M, *et al*. Specifie codon usage pattern and its implications on the secondary structure of silk fibroin mRNA[J]. Journal of Molecular Biology, 1988, 203(4): 917-925.

[7] Knight R D, Freeland S J, Landweber L F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes[J]. Genome Biology, 2001, 2(4): research0010.

[8] Eyre-Walker A. Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy?[J]. Molecular Biology and Evolution, 1996, 13(6): 864-872.

[9] Moriyama E N, Powell J R. Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli*[J]. Nucleic Acids Research, 1998, 26(13): 3188-3193.

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

[11] Gupta S K, Majumdar S, Bhattacharya T K, *et al*. Studies on the relationships between the synonymous codon usage and protein secondary structural units[J]. Biochemical and Biophysical Research Communications, 2000, 269(3): 692-696.

[12] 原晓龙, 郝佳波, 王毅, 等. 铁核桃叶绿体基因组密码子偏好性分析 [J]. 分子植物育种, 2020, 18(20): 6671-6677.
Yuan X L, Hao J B, Wang Y, *et al*. Codon usage bias analysis of chloroplast genome in Juglans sigillata[J]. Molecular Plant Breeding, 2020, 18(20): 6671-6677 (in Chinese).

[13] 李若愚, 张小丹, 马昕怡, 等. 桃基因密码子使用模式及其偏好性分析 [J]. 分子植物育种, 2021, 19(3): 799-807.
Li R Y, Zhang X D, Ma X Y, *et al*. Analysis of codon usage patterns and codon usage bias in peach (*Prunus persica*)[J]. Molecular Plant Breeding, 2021, 19(3): 799-807 (in Chinese).

[14] Gibson R N. Fishes of the world, 3rd edition[J]. Journal of Experimental Marine Biology and Ecology, 1995, 186(2): 291-292.

[15] 何舜平, 刘焕章, 陈宜瑜, 等. 基于细胞色素 b 基因序列的鲤科鱼类系统发育研究 (鱼纲:鲤形目)[J]. 中国科学 C 辑:生命科学, 2004, (1): 96-104.
He S P, Liu H Z, Chen Y Y, *et al*. Phylogeny of cyprinid fishes based on cytochrome b gene sequence (Ichthyoid: Cypriniformes)[J]. Science in China Series C: Life Sciences, 2004, (1): 96-104(in Chinese).

[16] 王英, 张太奎, 张孟伟, 等. 石榴等陆生植物 *Actin1* 基因密码子偏向性与进化分析 [J]. 分子植物育种, 2020, 18(6): 1799-1807.
Wang Y, Zhang T K, Zhang M W, *et al*. Codon usage bias and evolution analyses of *Actin1* genes from *Punica granatum* and other land plants[J]. Molecular Plant Breeding, 2020, 18(6): 1799-1807 (in Chinese).

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

[18] Fuglsang A. The effective number of codons for individual amino acids: some codons are more optimal than others[J]. Gene, 2003, 320: 185-190.

[19] Zhao Y C, Zheng H, Xu A Y, *et al*. Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution[J]. BMC Genomics, 2016, 17(1): 677-686.

[20] Lal D, Verma M, Behura S K, *et al*. Codon usage bias in phylum *Actinobacteria*: relevance to environmental adaptation and host pathogenicity[J]. Research in Microbiology, 2016, 167(8): 669-677.

[21] Li G, Ren Y M, Pan H X, *et al*. Comprehensive analysis and comparison on the codon usage pattern of whole *Mycobacterium tuberculosis* coding genome from different area[J]. BioMed Research International, 2018, 2018: 3574976.

[22] He Z M, Cai Y Y, Yang M, *et al*. Transcription factor CDX2 directly regulates the expression of *Ctenopharyngodon idellus* intestinal PepT1 to mediate the transportation of oligopeptide[J]. Aquaculture Reports, 2022, 24: 101148.

[23] Bera B C, Virmani N, Kumar N, *et al*. Genetic and codon usage bias analyses of polymerase genes of equine influenza virus and its relation to evolution[J]. BMC Genomics, 2017, 18(1): 652.

[24] 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.

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

Codon preference and evolutionary analysis of *SP*1 in Cyprinidae

HE Zhimin , ZHU Xiaoxia , YU Qingting , ZENG Zihao , XIAO Yang ,
ZHONG Gaode , LI Dang , TANG Jianzhou , LIU Zhen *

Hunan Provincial Key Laboratory of Nutrition and Quality Control of Aquatic Animals,
Department of Biological and Chemical Engineering, *Changsha University*, *Changsha* 410003, *China*

Abstract: Specificity Protein 1 (SP1), a member of the Sp/KLF protein family, is one of the earliest identified transcription factors and participates in the transcriptional regulation of many genes. This study aimed to explore the codon usage pattern in evolution process and the phylogenetic relationship among different species of *SP*1 and also provide references for its high heterologous expression. The software such as Codon W, Clustal X, MEGA 4.0, and SPSS were used to analyze the codon bias and evolutionary of *SP*1 of 4 cyprinidae fish species and 10 other species. The results showed that the Cyprinidae fish *SP*1 hithly preferred the codons of CUG and AUC. The mean of an effective number of codons (ENC) of *SP*1 was 50.57 and the value of the codon adaptation index (CAI) was between 0.184 and 0.379, which was far less than 1. The above two indexes illustrated that *SP*1 in different species showed codon bias. Moreover, the *SP*1 of four cyprinid fish showed similar codon preference. The ENC-plot analysis revealed that natural selection was the main reason for the *SP*1 codon usage bias. In cluster analysis, there was little difference between the phylogenetic analysis based on the CDS sequences of *SP*1 and the clustering analysis based on RSCU. *Escherichia coli* was the most suitable heterologous expression system for *Ctenopharyngodon idella SP*1，and the model animals *Danio rerio* and *Mus musculus* could both be used as genetic transformation receptors for *C. idella SP*1. This study showed that CUG and AUC were the optimal codons of *SP*1 in 4 cyprinids, there were differences in codon preference between species, and natural selection was the main influencing factor leading to *SP*1 codon preference in the 14 species. This study provides a theoretical basis for the classification, evolution and expression of Cyprinidae *SP*1.

Key words: Cyprinidae; *SP*1; codon preference; molecular evolution; clustering analysis

Corresponding author: LIU Zhen. E-mail: liuzhen_2015@sina.com

Funding projects: National Natural Science Foundation of China (31902345, U21A20267)