西北农业学报 2018,27(1):52-61
Acta Agriculturae Boreali-occidentalis Sinica doi: 10.7606/j.issn.1004-1389.2018.01.008

网络出版日期:2018-01-12
网络出版地址:http://kns.cnki.net/kcms/detail/61.1220.S.20180112.0934.018.html

谷子类甜蛋白基因家族的鉴定与密码子偏性分析

刘 潮,韩利红,王海波,唐利洲

(曲靖师范学院 云南高原生物资源保护与利用研究中心,生物资源与食品工程学院,云南省高校云贵高原动植物多样性及生态适应性进化重点实验室,云南曲靖 655011)

摘 要 类甜蛋白在植物的生长发育和抵御胁迫过程中发挥作用。利用生物信息学方法对谷子类甜蛋白家族基因组成、结构、启动子顺式作用元件、密码子使用偏性等进行分析。结果表明,谷子类甜蛋白家族包含43个成员,分布在9条染色体上,分为3种基因结构类型,其中16个成员仅包含1个外显子,14个包含3个外显子成员的内含子相位均为1-2型。根据系统发育分析归为12个聚类组,聚类组5中的基因主要来自Ⅲ号染色体,聚类组6中的基因主要来自Ⅰ号染色体,其内含子相位多为1-2型,这些基因可能来自同一祖先基因,与进化过程中的染色体重组事件有关。88.4%的基因参与应激反应,多个基因启动子区含有激素和胁迫响应元件,说明该家族基因在植物抵御胁迫过程中发挥作用。多数基因有效密码子数(*ENC*)较小,密码子的第3位的G+C含量(*GC3s*)值分布集中,包含10个最优密码子,均以G或C结尾,表明谷子类甜蛋白家族基因密码子使用偏性强,基因表达潜力高,进化过程中主要受自然选择压力影响。

关键词 谷子;类甜蛋白;聚类分析;密码子偏性

中图分类号 S515 **文献标志码** A **文章编号** 1004-1389(2018)01-0052-10

病程相关蛋白(Pathogenesis related protein,PR)是植物受病原物侵染或非生物因子刺激后产生的一类水溶性蛋白。目前,发现至少17个PR家族[1]。类甜蛋白(Thaumatin like protein,TLP)属于PR5家族,因与热带植物西非竹竿(*Thaumatococcus danielli* Benth.)果实中分离到的甜蛋白(Thaumatin)氨基酸序列有很高的同源性而得名,广泛分布于多种植物、动物及微生物中[2-3]。典型的TLP由16个半胱氨酸残基对形成8个二硫键,不仅稳定了分子结构,也保证蛋白的正确折叠,能够抵抗热变性、酸、碱和蛋白酶降解作用[4-5]。大多数TLP均具有索玛甜家族标签和5个保守的氨基酸残基[6],后者参与蛋白维持适当的拓扑结构和酸裂周围的表面静电势,对TLPs抗真菌活性必不可少[7]。TLP家族蛋白进化分析发现,动物TLP单独分在一支,并以单一祖先序列的形式来自于植物[3],陆生植物进化过程中TLP基因含量和多样性显著增加[8],而水稻和拟南芥TLP分布于多个支系,并存在染色体内和染色体间的复制[3]。单子叶和双子叶植物进化上发生分离后,*TLP*基因在进化枝上发生不对称的增加[3]。Liu等[9]认为*TLP*基因来自于大约10亿年前的植物、动物和真菌的共同祖先。在有些植物中,同一染色体上甚至同一位点存在TLP基因簇,说明串联重复是TLP超家族不对称扩张的重要机制[9]。研究表明,TLP具有抗真菌活性[10-11]、葡聚糖酶活性[12]、致敏原活性[13]等,在植物的生长发育和抵御胁迫过程中发挥作用。

密码子具有简并性,在物种的稳定上起着重要的作用。同义密码子在不同物种不同基因间的使用频率不同,特定物种或基因家族在长期进化中形成了适应自身基因组环境的密码子使用偏性。研究表明,同义密码子的选择使用对基因的表达起着重要的调节作用,有利于翻译的准确性和效率[14]。密码子偏性分析有助于预测基因的表达水平[15]、基因功能分析[16]、选择基因异源表达最适宿主和优化密码子以提高异源表达水平等[17]。

收稿日期:2017-09-11 **修回日期**:2017-10-09

基金项目:国家自然科学基金(31460179);云南省高校科技创新团队项目[云教科(2014)14号]。

第一作者:刘 潮,男,博士,讲师,研究方向为分子植物病理学。E-mail:liuchao@mail.qjnu.edu.cn

通信作者:唐利洲,男,博士,教授,研究方向为分子谱系地理学。E-mail:tanglizhou@163.com

谷子(*Setaria italica*)为禾本科粮食作物,富含蛋白质、脂肪和维生素,广泛栽培于欧亚大陆的温带和热带地区,其主产区多为干旱少雨地区[18],中国主要集中在黄河中上游地区,是北方地区的主要粮食之一。栽培过程中,谷瘟病、干旱等生物和非生物胁迫是谷子高产的严重障碍,抗病和抗胁迫相关基因的研究将为谷子品种的选育提供借鉴。目前,谷子全基因组数据已公布[18],这为基因功能和进化研究提供了条件。本试验从蛋白氨基酸和基因的碱基组成出发,对谷子 TLP 家族蛋白聚类关系、基因选择性、编码序列(Coding sequence,CDS)密码子的组成及使用偏性进行分析,旨在阐明 TLP 家族基因特征,为进一步利用 TLP 家族基因培育优良谷子品种奠定基础。

1　材料与方法

1.1　谷子 TLP 基因的鉴定与分析

以拟南芥 TLP 序列为探针,搜索谷子基因组数据库(http://www.plantgdb.org/SiGDB/)和 GenBank 谷子蛋白数据库,候选蛋白序列在 SMART 数据库(http://smart.embl-heidelberg.de/)中对蛋白功能域进行确认。所有蛋白序列生理生化特征通过 Expasy(http://www.expasy.org/tools/)预测。

1.2　基因结构分析

谷子 TLP 对应的基因序列和 CDS 序列从 GenBank 数据库中下载。使用基因结构显示系统(http://gsds.cbi.pku.edu.cn/index.php)绘制基因结构示意图。

1.3　蛋白基因本体和聚类分析

利用基因本体(Gene ontology)数据库(http://amigo1.geneontology.org/cgi-bin/amigo/blast.cgi)查询 TLP 功能分类。利用 WEGO 在线软件(http://wego.genomics.org.cn/)对蛋白 Gene ontology(GO)富集度进行计算。应用 MEGA 5.0 软件,采用邻接法(Neighbor-Joining,NJ)构建系统发育树。NJ 进化树分析步长值为 1 000,采用泊松校验(Poisson correction)的方法计算距离,其余参数取默认值。

1.4　启动子特征分析

通过 GenBank 数据库获取谷子 *TLP* 基因转录起始位点上游 1 kb 序列,通过 PlantCARE(http://bioinformatics.psb.ugent.be/webtools/plantcare/html/)数据库进行基因启动子区顺式作用元件分析。

1.5　密码子偏性分析

使用软件 CodonW 对谷子 *TLP* 基因 CDS 序列密码子使用性参数进行分析。参数包括:密码子适应指数(Codon adaptation index,*CAI*)、有效密码子数(Effective number of codons,*ENC*)、密码子的第 3 位的 G+C 含量(*GC3s*)。以 *GC3s* 为横坐标,*ENC* 为纵坐标,绘制 *ENC* 与 *GC3s* 的关联分布图[19]。图中曲线为密码子偏性仅受碱基突变影响时的 *ENC* 预期值的位置,计算公式为:$ENC = 2 + GC3s + 29/[GC3s^2 + (1 - GC3s)^2]$。使用 EMBOSS explorer 网站(http://emboss.toulouse.inra.fr/)在线软件对同义密码子相对使用度(Relative synonymous codon usage,*RSCU*)进行分析。根据 *ENC* 偏性对最优密码子(Optimal codon)进行分析,选择 *ENC* 值前后各 10%作为低表达和高表达基因,分别计算 2 组基因中 TLP 基因密码子的 *RSCU*。当 $\triangle RSCU > 0.3$,且在高表达组中 $RSCU > 1$,在低表达组中 $RSCU < 1$,可确定该密码子为最优密码子[20]。

2　结果与分析

2.1　谷子 *TLP* 基因家族的鉴定

以拟南芥 TLP 序列为探针,从谷子基因组数据库中共搜索并鉴定到 43 个 TLP 家族成员(表 1)。通过 SMART 在线数据库对谷子 TLP 进行结构分析发现,均含有典型索玛甜(Thaumatin,THN)结构域。谷子 *TLP* 基因在所有 9 条染色体上均有分布,其中Ⅸ和Ⅲ号染色体上基因成员较多(基因数均为 9),其次为Ⅱ、Ⅰ和Ⅴ号染色体(基因数分别为 7、5 和 5),Ⅳ、Ⅵ、Ⅶ和Ⅷ号染色体较少(基因数分别为 2、3、1 和 2)。生理生化分析显示,蛋白氨基酸数为 160～666,其中氨基酸数较多的 Si003953m 和 Si004228m 均含有蛋白激酶功能域,可能在蛋白的磷酸化过程中发挥作用。蛋白等电点为 4.42～9.17,其中酸性蛋白占 76.7%。疏水性与蛋白结构域形成和高级结构的稳定性有重要关系,谷子 TLP 中疏水性蛋白占 60.5%,具有较强亲水和疏水活性的蛋白均为酸性蛋白。

表 1 谷子中 TLP 家族信息

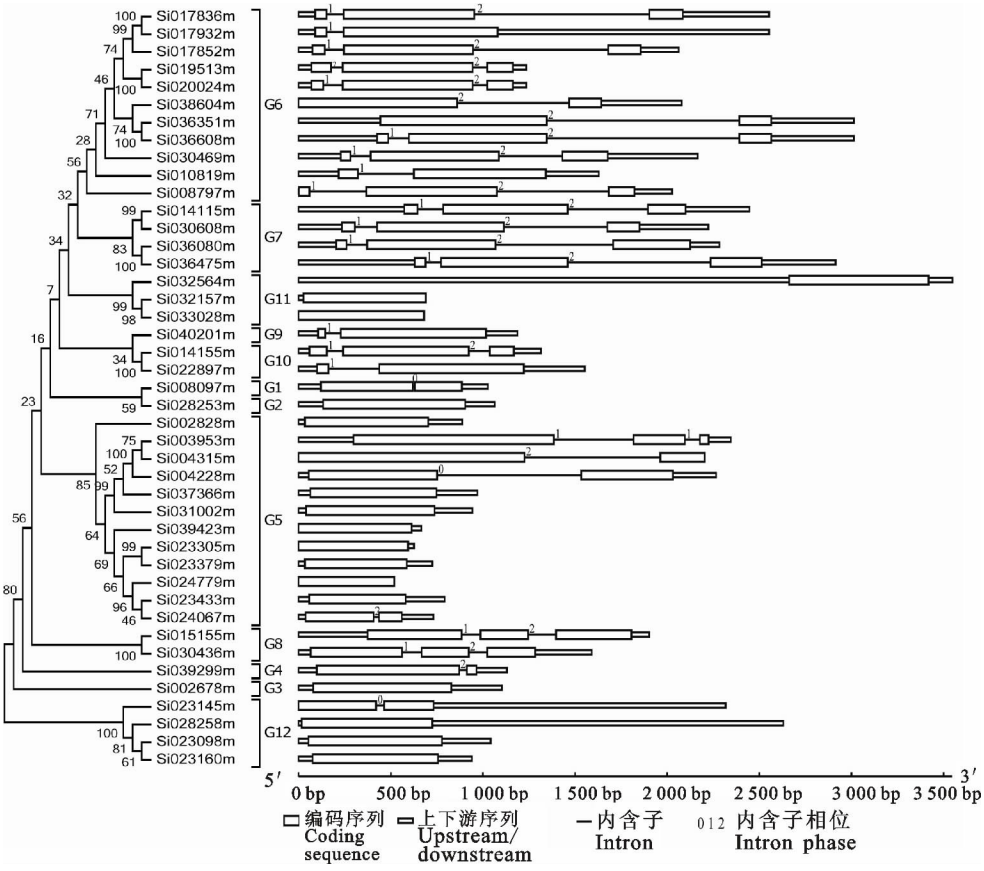
Table 1 Information of TLP family in foxtail millet

登录号 Accession	基因编号 Gene symbol	染色体位置 Chromosome location	氨基酸长度/aa Sequence length	蛋白分子质量/ku Molecular mass	等电点 pI	平均疏水指数 GRAVY
Si002678m	*LOC101755851*	chⅤ:41043640-41044742	249	26 220.90	7.49	−0.02
Si002828m	*LOC101779500*	chⅤ:39666880-39667767	222	22 896.83	4.72	−0.02
Si003953m	*LOC101770711*	chⅤ:6630774-6633118	661	73 264.38	6.53	−0.31
Si004228m	*LOC101770305*	chⅤ:6638602-6640866	666	73 118.47	6.11	−0.25
Si004315m	*LOC101758963*	chⅤ:6643287-6645489	413	44 748.77	6.86	−0.18
Si008097m	*LOC101754594*	chⅣ:37288744-37289770	254	25 811.41	8.69	0.09
Si008797m	*LOC101783927*	chⅣ:39626269-39628294	303	29 701.53	4.42	0.41
Si010819m	*LOC101774920*	chⅧ:3711314-3712942	273	27 327.61	5.46	0.22
Si014115m	*LOC101786770*	chⅥ:35012429-35014873	316	33 509.46	5.15	−0.17
Si014155m	*LOC101778574*	chⅥ:31681772-31683086	301	32 083.67	8.66	0.04
Si015155m	*LOC101777763*	chⅥ:33103296-33105197	392	39 697.05	4.55	−0.13
Si017836m	*LOC101773689*	chⅠ:17906345-17908897	316	31 642.59	4.94	0.20
Si017852m	*LOC101774354*	chⅠ:18040087-18042149	313	31 530.32	4.91	0.12
Si017932m	*LOC101773689*	chⅠ:17906345-17908897	298	30 639.35	5.30	0.06
Si019513m	*LOC101775841*	chⅠ:18364157-18365392	302	31 000.77	6.33	0.07
Si020024m	*LOC101775841*	chⅠ:18364157-18365392	309	31 776.63	6.13	0.05
Si022897m	*LOC101760244*	chⅢ:188185-189736	281	29 905.10	8.21	0.14
Si023098m	*LOC101778229*	chⅢ:41258347-41259388	241	24 506.98	8.58	0.15
Si023145m	*LOC101778779*	chⅢ:41241292-41243608	160	16 466.87	6.71	0.25
Si023160m	*LOC101779705*	chⅢ:41312188-41313127	226	22 848.97	5.70	0.30
Si023305m	*LOC101762279*	chⅢ:49895100-49895725	197	20 381.93	5.09	−0.05
Si023379m	*LOC101762680*	chⅢ:49903698-49904422	183	18 535.06	5.49	0.22
Si023433m	*LOC101764711*	chⅢ:49971842-49972633	173	17 591.55	4.60	0.07
Si024067m	*LOC101764306*	chⅢ:49960855-49961585	173	18 195.42	8.36	−0.12
Si024779m	*LOC101763494*	chⅢ:49947110-49947628	172	18 261.21	4.60	−0.08
Si028253m	*LOC101774025*	chⅧ:40429634-40430697	256	26 574.32	8.30	−0.01
Si028258m	*LOC101761990*	chⅧ:40279516-40282145	443	46 268.25	5.86	−0.08
Si030436m	*LOC101754007*	chⅡ:36280064-36281654	336	35 104.64	5.90	0.02
Si030469m	*LOC101774559*	chⅡ:38463784-38465948	331	32 708.56	4.86	0.19
Si030608m	*LOC101777928*	chⅡ:38474979-38477201	310	31 446.08	4.66	0.07
Si031002m	*LOC101757778*	chⅡ:44087991-44088932	231	24 199.40	8.13	0.03
Si032157m	*LOC101753062*	chⅡ:11053196-11053885	229	23 572.23	5.74	−0.17
Si032564m	*LOC101757622*	chⅡ:11233294-11236840	277	28 411.68	6.84	−0.05
Si033028m	*LOC101754138*	chⅡ:11046609-11047289	226	23 512.10	4.67	−0.16
Si036080m	*LOC101771079*	chⅨ:32359211-32361493	390	38 595.48	4.70	0.18
Si036351m	*LOC101774059*	chⅨ:51405438-51408450	357	35 631.06	5.03	0.33
Si036475m	*LOC101774730*	chⅨ:51424102-51427015	340	34 190.57	5.53	0.07
Si036608m	*LOC101774059*	chⅨ:51405438-51408450	326	32 376.37	5.21	0.24
Si037366m	*LOC101785089*	chⅨ:9644816-9645785	227	23 706.35	4.42	−0.20
Si038604m	*LOC101770679*	chⅨ:32257972-32260049	317	31 085.74	4.96	0.31
Si039299m	*LOC101779988*	chⅨ:21573122-21574251	288	29 399.79	8.55	0.20
Si039423m	*LOC101769623*	chⅨ:9647662-9648327	221	22 674.64	9.17	−0.14
Si040201m	*LOC101781881*	chⅨ:52346479-52347666	275	27 173.42	5.51	0.19

2.2　基因和蛋白结构及聚类分析

通过基因结构显示系统对谷子 *TLP* 基因结构、内含子组成与相位进行分析(图 1)。43 个谷子 *TLP* 基因分为 3 种结构类型,其中含有 1、2 和 3 个外显子的基因数目分别为 16、12 和 15 个。1-2 型内含子相位类型的基因数目最多(14 个),其次为 1 型和 2 型相位类型(均为 5 个),同一聚类组中的多数基因外显子数和内含子相位类型一致(图 1)。

参考拟南芥和水稻等的研究[3,21]对谷子 TLP 家族进行聚类分析。谷子 TLP 家族归为 12 个聚类组,各聚类组中基因数不一致,聚类组 5 和 6 中基因数较多(分别为 12 和 11),其他聚类组中基因数相对较少(图 1)。聚类组 5 中的基因 Si023305m、Si023379m、Si024779m、Si023433m 均来自Ⅲ号染色体,基因 Si037366m、Si039423m 来自Ⅸ号染色体并且均只有 1 个外显子,聚类组 6 中的基因 Si017836m、Si017932m、Si017852m、Si019513m、Si020024m 均位于Ⅰ号染色体,并且其位置临近,内含子相位均为 1-2 型。



TLP 进化组编号参考 Shatters 等[3],Zhao 等[21]　Evolution group numbers according to the results of the evolutionary analysis Shatters, *et al*.[3], Zhao, *et al*[21]

图 1　谷子 TLP 家族进化及基因结构

Fig. 1　Evolution and gene structure of TLP family in foxtail millet

2.3　蛋白 GO 功能分析

分析 TLP 的 GO 组成和功能分类,对了解其在植物生命进程中的功能具有重要意义。在 43 条谷子 TLPs 中,发现 12 条不同的 GO 注释子条目(图 2)。4 类为细胞结构组分,其中参与细胞(cell,GO:0005623)和细胞组分(cell part,GO:0044464)结构组成的均占 58.1%,参与胞外区作用的(extracellular region,GO:0005576)占 86.0%,参与共质体构成的(symplast,GO:0055044)占 16.3%。3 类在分子功能中起作用,具有结合功能(binding,GO:0005488)和催化活性(catalytic activity,GO:0003824)的蛋白均占 4.7%,分子转导活性(molecular transducer activity,GO:0060089)的蛋白占 2.3%。5 类参与

了生物学过程，其中参与胞内进程（cellular process，GO：0009987）和代谢进程（metabolic process，GO：0008152）的只占 4.7%，参与多器官进程（multi\-organism process，GO：0051704）和响应应激（response to stimulus，GO：0050896）的均占88.4%，参与免疫系统进程（immune system process，GO：0002376）的占 20.9%。

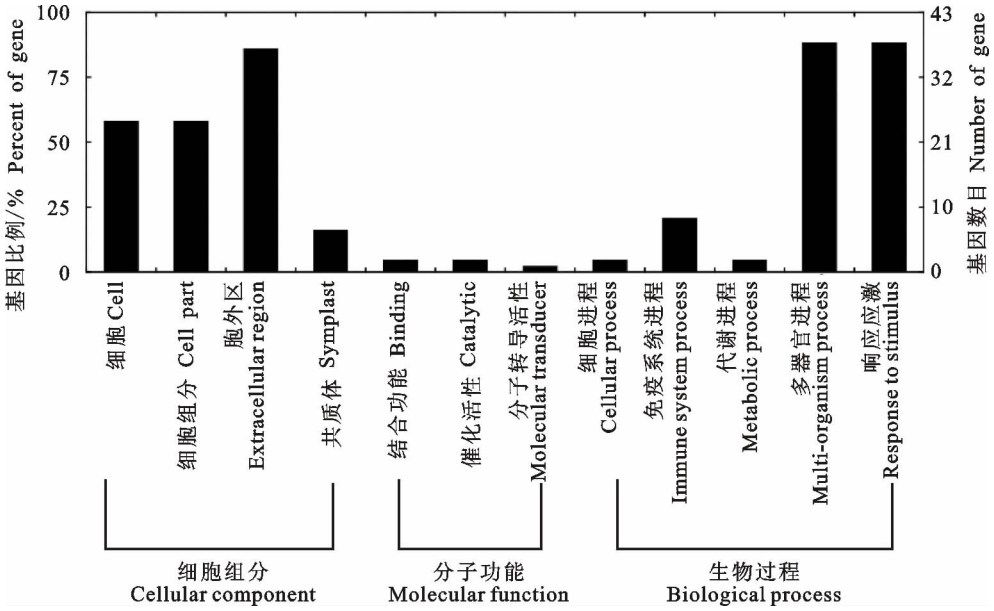


图 2 谷子 TLP 蛋白 GO 功能分类

Fig. 2 GO classification of TLP in foxtail millet

2.4 启动子特征

通过 PLACE 数据库对谷子 *TLP* 基因上游 1 kb 启动子区顺式作用元件进行分析（表 2），发现所有基因启动子区均含有多个 TATA-box 和 CAAT-box，部分基因含有 Py-rich；所有基因启动子区均含有 1 个到多个激素响应元件，包括脱落酸响应元件 ABRE，茉莉酸响应元件 CGTCA-motif，赤霉素响应元件 GARE-motif 和 P-box，水杨酸响应元件 TCA-element 等；每个基因启动子区均含有多个胁迫响应元件，包括热激响应元件 HSE，低温反应顺式作用元件 LTR，干旱响应 MYB 结合位点 MBS，防御和胁迫相关响应元件 TC-rich，受伤和真菌激发子响应元件W-box。基因含有的激素或胁迫响应元件数量和类型不同，可能是不同的基因在不同的信号通路中发挥作用，也说明该家族基因功能的多样性和复杂性。

2.5 密码子使用偏性

利用 CodonW 软件和 EMBOSS explorer 数据库对基因密码子使用偏性进行分析（表 3），发现谷子 *TLP* 基因的 *CAI* 值平均为 0.282，60.5%的基因 *ENC* 值为 28.12～35.00，*ENC* 值反映基因编码对密码子选择性强弱，一般 *ENC* 值低于 35 表示基因表达对密码子的使用偏性较强。93.0%的基因 *GC3s* 为 0.834～0.973，分布较集中，*GC3s* 分布反映了植物所受的选择压力，*GC3s* 分布范围越小，表明密码子使用偏性受自然选择压力影响越大[22]。以上结果表明，谷子 TLP 家族基因密码子使用偏性较强，多数基因具有较高的表达潜力，基因在进化过程中主要受到自然选择压力影响。

ENC 与 *GC3s* 关联分析显示，谷子 *TLP* 基因分布在标准曲线下方，多数 *ENC* 较小，*ENC* 和 *GC3s* 分布相对集中（图 3），说明不同的基因密码子偏性较强，多数基因进化过程中主要受到自然选择压力影响。

谷子 *TLP* 基因密码子 *RSCU* 分析显示（表 4），*RSCU*＞1 的密码子均以 G 或 C 结尾，发现 10 个最优密码子，△*RSCU*＞1 的分别为编码丙氨酸（Ala，GCG）、谷氨酸（Glu，GAG）、异亮氨酸（Leu，CTG）、脯胺酸（Pro，CCG）、精氨酸（Arg，CGC）、苏氨酸（Thr，ACG）（表 4），表明谷子 TLP 家族基因偏好使用 G 或 C 结尾的密码子。

表 2　谷子 *TLP* 基因启动子区顺式作用元件信息

Table 2　Information about putative cis-acting elements in the 1 kb upstream promoter region of TLP genes in foxtail millet

登录号 Accession	TATA-box	Py-rich	CAAT-box	ABRE	CGTCA-motif	GARE-motif	P-box	TCA-element	HSE	LTR	MBS	TC-rich	W-box
Si002678m	13		19		2		1			1	1		2
Si002828m	26	2	14	6	3			1			2	1	2
Si003953m	23	1	21	1	2		1			2			
Si004228m	25		29	1	1	1	2	1			1	2	
Si004315m	34		31		3	1					2	1	
Si008097m	18	2	26	1	2			1		1	3		
Si008797m	15		28		1	2		1			2		
Si010819m	20		31	3	1	2	1		1	1		1	1
Si014115m	3	1	15	3	7						2	1	1
Si014155m	16		5	6	6			1					
Si015155m	20	1	33		1	1		1			1	2	
Si017836m	32	3	14	1		3		3			1	1	3
Si017852m	25		29	2	5			2	1	1	1		1
Si017932m	32	3	14	1		3		3			1	1	3
Si019513m	32	3	14	1		3		3			1	1	3
Si020024m	32		19	3			1				1	2	1
Si022897m	25	1	29	2	2		1	1		1	2	1	
Si023098m	17		15		4			1			1	1	1
Si023145m	50		21	3		1					3		1
Si023160m	19	1	11	5	2		1				1		1
Si023305m	121		25	1		1					1	2	
Si023379m	22	2	31			2	1	2			3		1
Si023433m	53	1	23	3	1	1	2		1		2		
Si024067m	36							1	2		1		1
Si024779m	49		27	2	1				1			2	
Si028253m	36		27		1			2			3		1
Si028258m	10		8		2	1			1		1	1	
Si030436m	22	1	28	1	2		1	1		1	1	2	
Si030469m	24	1	17	2	3						2		2
Si030608m	23		16	2	2		1		1		1		
Si031002m	29		39	5	3			2				1	
Si032157m	60	1	44	3	1				1		1		1
Si032564m	15		23	3	1	2				1		1	
Si033028m	26		24	3	3	1			3		1	3	1
Si036080m	26		28	3	1	2					6	2	
Si036351m	11		8	1	4			1		2		1	1
Si036475m	11		26										
Si036608m	11		8	2	4			1		2		1	1
Si037366m	44		24	6	1			1	2		2	2	2
Si038604m	13	1	15	7	1	1					1	1	3
Si039299m	29		29		2						1	1	
Si039423m	47		31	6	2	1			1		3	2	3
Si040201m	13		18	2	2	2		1	1	1		1	

注:TATA-box 转录起始区－30 bp 核心启动子元件; Py-rich 高转录水平相关顺式作用元件;CAAT-box 启动增强元件;ABRE 脱落酸响应元件;CGTCA-motif 茉莉酸响应元件;GARE-motif 赤霉素响应元件;P-box 赤霉素响应元件;TCA-element 水杨酸响应元件;HSE 热激响应元件;LTR 参与低温反应的顺式作用元件;MBS 干旱响应 MYB 结合位点;TC-rich 防御和胁迫相关响应元件;W-box 受伤和真菌激发子响应元件。

Note:TATA-box core promoter element around －30 of transcription start;Py-rich cis-acting element conferring high transcription levels; CAAT-box common cis-acting element in promoter and enhancer regions;ABRE cis-acting element involved in the abscisic acid responsiveness;CGTCA-motif cis-acting regulatory element involved in the MeJA-responsiveness;GARE-motif gibberellin-responsive element; P-box gibberellin-responsive element;TCA-element cis-acting element involved in salicylic acid responsiveness;HSE cis-acting element involved in heat stress responsiveness;LTR cis-acting element involved in low-temperature responsiveness;MBS MYB binding site involved in drought-inducibility;TC-rich cis-acting element involved in defense and stress responsiveness;W-box fungal elicitor responsive element.

表 3 谷子 *TLP* 家族基因密码子使用特性

Table 3 Characterization of codon usage of *TLP* genes in foxtail millet

登录号 Accession	*CAI*	*ENC*	*GC3s*	登录号 Accession	*CAI*	*ENC*	*GC3s*	登录号 Accession	*CAI*	*ENC*	*GC3s*
Si017836m	0.316	33.81	0.899	Si024067m	0.33	32.45	0.925	Si036608m	0.306	33.04	0.930
Si017932m	0.308	40.05	0.855	Si023433m	0.348	31.35	0.946	Si036475m	0.262	36.14	0.913
Si017852m	0.304	38.26	0.846	Si008097m	0.265	30.40	0.971	Si040201m	0.243	36.33	0.933
Si020024m	0.233	41.58	0.840	Si004228m	0.229	54.50	0.556	Si008797m	0.265	36.56	0.899
Si033028m	0.321	34.25	0.930	Si003953m	0.225	58.24	0.523	Si004315m	0.194	56.32	0.590
Si032564m	0.345	29.53	0.952	Si002828m	0.294	33.58	0.940	Si019513m	0.231	41.13	0.834
Si030436m	0.269	41.06	0.855	Si002678m	0.342	29.58	0.959	Si039299m	0.312	33.18	0.940
Si030608m	0.256	33.43	0.914	Si014155m	0.236	36.59	0.881	Si039423m	0.258	34.56	0.960
Si031002m	0.269	30.19	0.960	Si015155m	0.254	39.15	0.840	Si036351m	0.310	33.99	0.910
Si032157m	0.334	36.42	0.912	Si014115m	0.284	37.13	0.874	Si023098m	0.280	29.32	0.966
Si030469m	0.281	30.26	0.957	Si010819m	0.290	31.38	0.914	Si023145m	0.216	36.66	0.910
Si022897m	0.285	34.24	0.923	Si028253m	0.282	34.83	0.911	Si023160m	0.280	28.12	0.968
Si023305m	0.291	33.04	0.958	Si037366m	0.318	32.14	0.973	Si028258m	0.236	32.42	0.952
Si023379m	0.302	32.03	0.899	Si038604m	0.272	32.26	0.932				
Si024779m	0.379	35.31	0.918	Si036080m	0.248	32.60	0.944				

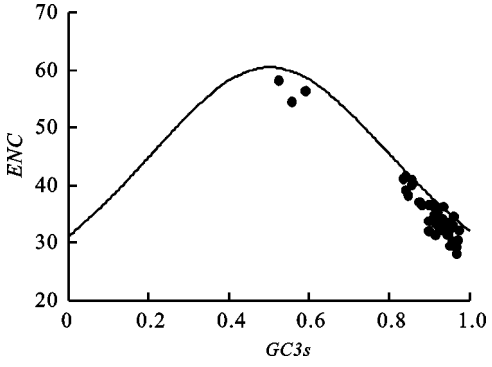


图 3 谷子 TLP 家族基因 *ENC* 与 *GC3s* 的关系

Fig. 3 Correlative analysis of *ENC* and *GC3s* of TLP genes in foxtail millet

3 讨 论

谷子是目前种植第二广泛的粟类作物，属于典型的 C_4 植物，比 C_3 植物具有更高的水分利用效率和光合效率，特别能适应干旱环境条件，是非生物胁迫抗性，特别是抗旱研究的模式植物[23]。Zhang 等[18]研究发现，谷子在进化过程中发生了 3 次染色体重组事件，其中 2 次发生在谷子从水稻分化之后，1 次发生在谷子从高粱分化之后。这些事件导致部分基因家族大量扩张，并为谷子具有强的抗旱能力奠定了基础。TLP 属于多基因家族，不同物种基因组中 TLP 数量有很大差异，胡萝卜（*Daucus carota*）、黄瓜（*Cucumis sativus*）等蔬菜中大约 30 个，水稻（*Oryza sativa*）、高粱（*Sorghum bicolor*）等作物中 50～60 个，火炬松（*Pinus taeda*）中大于 80 个[24]。研究表明，植物 TLP 家族蛋白在植物的生长发育和抵御胁迫过程中发挥作用[6-7]。

本研究通过生物信息学方法从谷子基因组中共发现 43 个 *TLP* 基因，多数为酸性蛋白。有 3 种基因结构类型，其中仅有 1 个外显子的基因有 16 个，1-2 型内含子相位的基因有 14 个，主要来自聚类组 5 和 6 中。这些基因多数位于同一染色体上，说明这些基因可能来源于同一祖先基因，是染色体内和染色体间复制的结果。蛋白 GO 分类分析显示 86.0%的蛋白参与胞外作用，88.4%的蛋白参与多器官发育进程和响应应激反应，说明该家族蛋白在信号肽的引导下在胞外空间参与器官发育和应对胁迫等进程，在植物的生长发育和抵御环境胁迫过程中发挥重要作用。启动子是基因表达调控的重要元件，通过分析和预测启动子区顺式作用元件可以为基因表达和功能研究奠定基础。谷子 TLP 家族启动子区富含植物激素和病原响应元件，表明这些基因参与植物多种生命进程。

在进化过程中，植物基因组密码子偏性主要受碱基突变和自然选择压力的影响，不同物种以

表 4　谷子 *TLP* 基因同义密码子使用情况

Table 4　Usage of synonymous codon of *TLP* genes in foxtail millet

氨基酸 Amino acid	密码子 Codon	*RSCU*	氨基酸 Amino acid	密码子 Codon	*RSCU*	氨基酸 Amino acid	密码子 Codon	*RSCU*	氨基酸 Amino acid	密码子 Codon	*RSCU*
Ala	GCA	0.23	His	CAC	1.64	Gln	CAA	0.30	Thr	ACA	0.32
Ala	GCC	1.73	His	CAT	0.36	Gln	CAG	1.70	Thr	ACC	1.75
Ala	GCG***	1.75	Lys	AAA	0.37	Arg	AGA	0.44	Thr	ACG***	1.75
Ala	GCT	0.28	Lys	AAG	1.63	Arg	AGG	1.15	Thr	ACT	0.19
Cys	TGC	1.92	Leu	CTA	0.27	Arg	CGA	0.22	Val	GTA	0.17
Cys	TGT	0.08	Leu	CTC	2.65	Arg	CGC***	2.38	Val	GTC**	1.41
Asp	GAC	1.78	Leu	CTG***	2.04	Arg	CGG*	1.58	Val	GTG	2.16
Asp	GAT	0.22	Leu	CTT	0.43	Arg	CGT	0.23	Val	GTT	0.26
Glu	GAA	0.36	Leu	TTA	0.17	Ser	AGC	2.05	Ile	ATA	0.27
Glu	GAG***	1.64	Leu	TTG	0.44	Ser	AGT	0.20	Ile	ATC	2.33
Phe	TTC	1.85	Asn	AAC	1.81	Ser	TCA	0.37	Ile	ATT	0.40
Phe	TTT	0.15	Asn	AAT	0.19	Ser	TCC	1.66			
Gly	GGA	0.22	Pro	CCA	0.38	Ser	TCG**	1.51			
Gly	GGC	2.48	Pro	CCC	1.33	Ser	TCT	0.20			
Gly	GGG	1.08	Pro	CCG***	2.01	Tyr	TAC**	1.88			
Gly	GGT	0.22	Pro	CCT	0.29	Tyr	TAT	0.12			

注:最优密码子标注如下 *.△*RSCU*>0.3;** △*RSCU*>0.6;*** △*RSCU*>1。

Note:Optimal codon marked as * △*RSCU*>0.3;** △*RSCU*>0.6;*** △*RSCU*>1.

及同一物种的不同基因受到的两种压力强度也不同。密码子偏性与物种进化和生存环境有关,密码子偏性越强,基因在进化过程中受到的选择压力越强。谷子 *TLP* 家族多数基因 *ENC* 低于 35,*GC3s* 分布集中,表明基因密码子偏性强,基因具有较高的表达潜力,进化过程中主要受自然选择压力影响。单子叶植物植物偏好 G 或 C 结尾的密码子[25]。基因 *RSCU*>1 的密码子均为 G 或 C 结尾,10 个最优密码子,其中 7 个以 G 结尾,3 个以 C 结尾,表明谷子偏好使用 G 或 C 结尾的密码子,与玉米[25]、水稻和高粱[26]等密码子使用偏性一致。物种受到正向选择时会形成大量最优密码子[27],进一步说明谷子 *TLP* 基因在进化过程中受到较强的正向选择。进行基因工程操作时,基因表达往往受到宿主密码子偏性影响,通过修改外源基因密码子,使之与宿主密码子偏性一致,可实现外源基因的高效表达[17,28]。本研究为谷子 *TLP* 基因的开发利用和转基因研究提供了借鉴。

参考文献　Reference:

[1] VAN LOON L C,REP M,PIETERSE C M. Significance of inducible defense-related proteins in infected plants[J]. *Annual Review of Phytopathology*,2006,44:135-162.

[2] ABAD L R,D'URZO M P,LIU D,*et al*. Antifungal activity of tobacco osmotin has specificity and involves plasma membrane permeabilization[J]. *Plant Science*,1996,118(1):11-23.

[3] SHATTERS RG,BOYKIN L M,LAPOINTE S L,*et al*. Phylogenetic and structural relationships of the *PR5* gene family reveal an ancient multigene family conserved in plants and select animal taxa[J]. *Journal of Molecular Evolution*,2006,63(1):12-29.

[4] SMOLE U,BUBLIN M,RADAUER C,*et al*. Mal d 2,the thaumatin-like allergen from apple,is highly resistant to gastrointestinal digestion and thermal processing[J]. *International Archives of Allergy and Immunology*,2008,147(4):289-298.

[5] FIERENS E,ROMBOUTS S,GEBRUERS K,*et al*. TLXI,a novel type of xylanase inhibitor from wheat (*Triticum aestivum*) belonging to the thaumatin family[J]. *Biochemical Journal*,2007,403(3):583-591.

[6] JAMI S K,ANURADHA T S,GURUPRASAD L,*et al*. Molecular,biochemical and structural characterization of osmotin-like protein from black nightshade (*Solanum nigrum*)[J]. *Journal of Plant Physiology*,2007,164(3),238-252.

[7] LIU D,HE X,LI W,*et al*. Molecular cloning of a thaumatin-like protein gene from *Pyrus pyrifolia* and overexpression of this gene in tobacco increased resistance to pathogenic fungi[J]. *Plant Cell,Tissue and Organ Culture* (*PC-*

TOC),2012,111(1):29-39.

[8] PETRE B, MAJOR I, ROUHIER N, *et al*. Genome-wide analysis of eukaryote thaumatin-like proteins (PR5s) with an emphasis on poplar[J]. *Plant Biology*, 2011, 11(1):33.

[9] LIU J J, STURROCK R, EKRAMODDOULLAH A K M. The superfamily of thaumatin-like proteins: its origin, evolution, and expression towards biological function[J]. *Plant Cell Reports*, 2010, 29(5):419-436.

[10] RAMOS M V, DE OLIVEIRA R S B, PEREIRA H M, *et al*. Crystal structure of an antifungal osmotin-like protein from *Calotropis procera* and its effects on *Fusarium solani* spores, as revealed by atomic force microscopy: insights into the mechanism of action[J]. *Phytochemistry*, 2015, 119:5-18.

[11] LIU C, CHENG F, SUN Y, *et al*. Structure—function relationship of a novel PR-5 protein with antimicrobial activity from soy hulls[J]. *Journal of Agricultural and Food Chemistry*, 2016, 64(4):948-959.

[12] VAN DAMME E J, CHARELS D, MENU-BOUAOU ICHE L, *et al*. Biochemical, molecular and structural analysis of multiple thaumatin-like proteins from the elderberry tree (*Sambucus nigra* L.)[J]. *Planta*, 2002, 214(6): 853-862.

[13] BREITENEDER H. Thaumatin-like proteins-a new family of pollen and fruit allergens[J]. *Allergy*, 2004, 59(5):479-481.

[14] CARLINI D B, CHEN Y, STEPHAN W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr [J]. *Genetics*, 2001, 159(2):623-633.

[15] PEK H B, KLEMENT M, ANG K S, *et al*. Exploring codon context bias for synthetic gene design of a thermostable invertase in *Escherichia coli*[J]. *Enzyme and Microbial Technology*, 2015, 75:57-63.

[16] PAN L L, WANG Y, HU J H, *et al*. Analysis of codon use features of stearoyl-acyl carrier protein desaturase gene in *Camellia sinensis*[J]. *Journal of Theoretical Biology*, 2013, 334:80-86.

[17] ZHOU M, WANG T, FU J, *et al*. Nonoptimal codon usage influences protein structure in intrinsically disordered regions[J]. *Molecular Microbiology*, 2015, 97(5):974-987.

[18] ZHANG G, LIU X, QUAN Z, *et al*. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential[J]. *Nature Biotechnology*, 2012, 30(6):549-554.

[19] WRIGHT F. The 'effective number of codons' used in a gene[J]. *Gene*, 1990, 87(1):23-29.

[20] DURET L, MOUCHIROUD D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*[J]. *Proceedings of the National Academy of Sciences*, 1999, 96(8):4482-4487.

[21] ZHAO J P, SU X H. Patterns of molecular evolution and predicted function in thaumatin-like proteins of *Populus trichocarpa*[J]. *Planta*, 2010, 232(4):949-962.

[22] KAWABE A, MIYASHITA N T. Patterns of codon usage bias in three dicot and four monocot plant species[J]. *Genes & Genetic Systems*, 2003, 78(5):343-352.

[23] LI P, BRUTNELL T P. Setaria viridis and *Setaria italica*, model genetic systems for the Panicoid grasses[J]. *Journal of Experimental Botany*, 2011, 62(9), 3031-3037.

[24] 刘 潮,韩利红,王海波,等. 胡萝卜类甜蛋白家族鉴定与生物信息学分析[J]. 中国蔬菜,2017(2):38-44.
LIU CH, HAN L H, WANG H B, *et al*. Identification and bioinformatics analysis of thaumatin-like protein family in *Daucus carota*[J]. *China Vegetables*, 2017(2):38-44.

[25] LIU H, HE R, ZHANG H, *et al*. Analysis of synonymous codon usage in *Zea mays*[J]. *Molecular Biology Reports*, 2010, 37(2):677.

[26] TATARINOVA T V, ALEXANDROV N N, BOUCK J B, *et al*. GC 3 biology in corn, rice, sorghum and other grasses[J]. *BMC Genomics*, 2010, 11(1):308.

[27] HERSHBERG R, PETROV D A. Selection on codon bias [J]. *Annual Review of Genetics*, 2008, 42:287-299.

[28] ZELASKO S, PALARIA A, DAS A. Optimizations to achieve high-level expression of cytochrome P450 proteins using *Escherichia coli* expression systems[J]. *Protein Expression and Purification*, 2013, 92(1):77-87.

Identification and Codon Bias Analysis of Thaumatin-like Protein Gene Family in Foxtail Millet

LIU Chao, HAN Lihong, WANG Haibo and TANG Lizhou

(Center for Yunnan Plateau Biological Resources Protection and Utilization / Key Laboratory of Yunnan Province Universities of the Diversity and Ecological Adaptive Evolution for Animals and Plants on Yungui Plateau, College of Biological Resource and Food Engineering, Qujing Normal University, Qujing Yunnan 655011, China)

Abstract The thaumatin-like proteins(TLP) play a role in plant growth, development and stress resistance. The composition, structure, cis-acting element and codon usage bias of TLP family of foxtail millet were analyzed by bioinformatics. The results showed that the foxtail millet TLP family consists of 43 members, distributing on nine chromosomes and being divided into three kinds of gene structure types. Sixteen of 43 members contains merely one exon, and introns phase of 14 members possessing three exons belong to 1-2 type. Phylogenetic analysis classifies into twelve clusters , genes of group five and six are mainly from chromosome Ⅲ and Ⅰ, respectively, and most of the intron phase is 1-2 type. These genes may be originated from the same ancestor, and correlated to the evolution events of chromosome recombination. 88.4% genes are involved in stress responses, and multiple gene promoter regions contain hormones and stress responsive elements, indicating that *TLP* genes play a role in stress resistance. *ENC* values of most genes are smaller, the distribution of *GC3s* value is centralized, and 10 optimal codons ended in G or C are found. It is suggested that codon usage in *TLP* family of foxtail millet is biased, gene expression of potential is high, and is mainly influenced by natural selection pressure during evolution.

Key words *Setaria italica*; Thaumatin-like protein; Cluster analysis; Codon bias

Received 2017-09-11　　**Returned** 2017-10-09

Foundation item National Natural Science Foundation of China(No. 31460179); Yunnan Province University Science and Technology Innovation Team Project[Yunnan Education Science(2014)14].

First author LIU Chao, male, Ph. D, lecturer. Research area: molecular plant pathology. E-mail: liuchao@mail.qjnu.edu.cn

Corresponding author TANG Lizhou, male, Ph. D, professor. Research area: molecular lineage geography. E-mail: tanglizhou@163.com

(责任编辑:史亚歌　**Responsible editor:SHI Yage**)