江苏农业学报(*Jiangsu J.of Agr.Sci.*),2017,33(5):998~1006
http://www.jsnyxb.com

刘 潮,韩利红,宋培兵,等. 桑树类甜蛋白家族鉴定与生物信息学分析[J].江苏农业学报,2017,33(5):998-1006.
doi:10.3969/j.issn.1000-4440.2017.05.007

桑树类甜蛋白家族鉴定与生物信息学分析

刘 潮, 韩利红, 宋培兵, 王德琴, 袁国芳, 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: 为了全面分析桑树类甜蛋白家族基因特征,从桑树全基因组中鉴定类甜蛋白家族(Thaumatin-like protein,TLP)基因,并对基因结构和组成、蛋白质结构域和系统进化以及密码子使用特性进行分析。通过研究,筛选并鉴定了25个桑树TLP家族成员。多数具有典型的索马甜家族标签和保守氨基酸残基。基因结构分析结果表明,12个*TLP*基因家族成员含有1个内含子,10个成员含有2个内含子,且内含子相位较保守。系统进化分析结果显示,桑树*TLP*分为8个聚类组,组内序列一致性较高。密码子偏性分析结果显示,桑树*TLP*家族基因密码子第3位的核苷酸使用偏性较弱,多数基因的进化主要是由碱基随机突变造成的,属于低表达基因,少数基因受自然选择压力影响,基因表达水平较高,在植物响应环境胁迫中发挥作用。

关键词: 桑树;类甜蛋白;系统进化;密码子偏性
中图分类号: S888.71 **文献标识码:** A **文章编号:** 1000-4440(2017)05-0998-09

Identification and bioinformatics analysis of thaumatin-like protein family in *Morus notabilis*

LIU Chao, HAN Li-hong, SONG Pei-bing, WANG De-qin, YUAN Guo-fang, WANG Hai-bo, TANG Li-zhou
(*Center for Yunnan Plateau Biological Resources Protection and Utilization/College of Biological Resource and Food Engineering/Key Laboratory of Yunnan Province Universities of the Diversity and Ecological Adaptive Evolution for Animals and Plants on Yungui Plateau, Qujing Normal University, Qujing* 655011, *China*)

Abstract: In order to comprehensively understand the genetic characteristics of thaumatin-like protein (TLP) family, the *TLP* genes of mulberry whole genome were identified, and the gene structure, protein domain, phylogeny and codon usage bias were analyzed. A total of 25 TLP sequences were identified, most of which have the typical thaumatin sequence and conserved amino acid residues. Each of 12 TLP family members contains one conservative intron, and each of 10 family members contains two conservative introns. The phylogenetic analysis revealed that the *TLPs* were divided into eight groups, and the sequence similarity was relatively high in each group. Codon usage bias of the third codon of TLP family genes was weak, indicative of the low expression. Base mutation pressure are responsible for the evolution of most of mulberry *TLP* family genes. Natural selection pressure contributes to a few genes evolution, suggestive of high expression which may play an important role in plant in response to environment stress.

Key words: *Morus notabilis*; thaumatin-like protein; phylogeny analysis; codon usage bias

收稿日期:2017-03-14
基金项目:国家自然科学基金项目(31460179);云南省高校科技创新团队项目
作者简介:刘 潮(1980-),男,河北景县人,博士,讲师,主要从事分子植物病理学研究,(E-mail)liuchao@mail.qjnu.edu.cn
通讯作者:唐利洲,(E-mail)tanglizhou@163.com

病程相关蛋白分为17个家族,在防御植物病原中发挥作用[1]。植物类甜蛋白与甜蛋白(Thaumatin)氨基酸序列有很高的同源性,因此被称作类甜

蛋白(Thaumatin-like protein,TLP),属于病程相关蛋白第5家族。*TLP*基因家族是一个具有多个成员的高度复杂的基因家族,参与寄主防御和多种发育进程[2-3]。多数TLP具有N端信号肽,在保证蛋白运输到特定区域中起作用。超过20个来自不同物种的TLP具有抗真菌活性[4],有些TLP具有葡聚糖酶活性[5],结构域I和II之间的酸性氨基酸是TLP具有β-1, 3葡聚糖酶活性所必需的[6]。TLP在多种水果和作物中具有过敏原活性[7]。TLP蛋白也提高了植物对多种非生物胁迫的耐受性[8]。除了参与植物的胁迫反应之外,TLP蛋白还参与生长发育的多项进程。

密码子具有简并性,编码同种氨基酸的不同密码子称为同义密码子,在物种的稳定上起着重要作用。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特性,此现象称为密码子偏好性(Codon bias)。同义密码子的偏好使用可能与基因的G、C含量有关,同义密码子的选择使用影响了基因的表达,能提高翻译的准确性和效率,改变编码蛋白的疏水性[9]。密码子偏好性的分析有助于预测基因的表达水平,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[10]。

目前,陆续从多种物种中发现了TLP蛋白[11-12],然而对桑树中*TLP*家族基因及其编码的蛋白研究较少。桑树是一种常见的落叶树种,其叶片多汁是桑蚕的主要饲料,在中国多个省份均有栽培;桑椹可供食用,桑树皮可用于造纸,桑叶和根皮可入药,具有较高的经济价值。桑树病虫害种类多、危害广,在一定程度上影响了蚕桑生产的发展。鉴于川桑(*Morus notabilis*)的基因组测序工作已完成,本研究利用生物信息学方法对川桑基因组[13]中*TLP*家族基因的结构和组成、系统进化以及基因密码子使用偏好性进行分析,为进一步揭示TLP家族生物学功能提供借鉴。

1　材料与方法

1.1　桑树TLP家族氨基酸序列获取

蛋白1Z3Q是小果野蕉(*Musa acuminate*)中典型的TLP蛋白,常被用作TLP蛋白结构和功能研究的模板。以蛋白1Z3Q氨基酸序列为探针序列,利用GenBank数据在线分析软件BLAST,搜索桑树(*M. notabilis*)蛋白质数据库,同时以蛋白质标签结构域为查询关键词在Pfam数据库(http://pfam.sanger.ac.uk/)进行确认。

1.2　桑树TLP蛋白理化特性分析

通过Expasy(http://www.expasy.org/tools/)对蛋白质生理生化特征进行预测。使用SMART(http://smart.embl-heidelberg.de/)对蛋白质结构域进行预测。使用CBS数据库(http://www.cbs.dtu.dk)中的在线软件预测蛋白质信号肽和跨膜结构域。

1.3　桑树*TLP*基因碱基序列的获取与分析

从GenBank数据库中获得TLP蛋白对应的基因组序列和基因编码序列(Coding sequence,CDS)。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。通过MEME SUITE在线工具(http://meme-suite.org/tools/meme)预测桑树TLP转录因子蛋白质氨基酸序列的保守基序(Motif)。使用ClustalX进行蛋白质氨基酸序列比对,应用MEGA5.0软件,采用NJ(Neighbor-Joining)法构建系统进化树。

1.4　桑树*TLP*家族基因密码子偏性分析

使用软件CodonW对桑树TLP家族基因CDS序列密码子使用特性进行分析。使用参数包括:基因的同义密码子在第3位上碱基的出现频率,密码子适应指数(*CAI*),密码子偏好指数(*CBI*),有效密码子数(*ENC*),密码子的第3位的G、C含量(*GC3s*),基因的G、C含量(*GC*),同义氨基酸数(*Lsym*)等参数。以*GC3s*为横坐标,*ENC*为纵坐标,绘制*ENC*与*GC3s*的关联分布图[14]。图中曲线为密码子偏性仅受碱基突变影响时的*ENC*预期值位置,计算公式为:$ENC = 2 + GC3s + 29/[GC3s^2 + (1 - GC3s)^2]$。分布点越靠近标准曲线表示密码子偏性受碱基突变影响越大,越远离标准曲线表示密码子偏性受到自然选择影响越大。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。

2　结果与分析

2.1　桑树TLP家族蛋白鉴定

以典型TLP蛋白1Z3Q氨基酸序列为探针,利用生物信息学方法,从桑树蛋白质组中搜索到25个TLP家族成员。利用PFAM和SMART数据库进行蛋白质结构分析,发现所有序列均含有典型索玛甜

功能域(Thaumatin,THN),15 个成员在 N 端存在潜在的信号肽序列(表 1)。大多数 TLP 蛋白均具有索玛甜家族标签 G-X-[GF]-X-C-X-T-[GA]-D-C-X-(1,2)-G-X-(2,3)-C[15]和 5 个保守的 REDDD(1 个精氨酸,1 个谷氨酸,3 个天门冬氨酸)残基,后者参与蛋白质维持适当的拓扑结构和酸裂周围的表面静电势,对 TLP 抗真菌活性必不可少[16]。本研究选取 20 个序列一致性较强的桑树 TLP 蛋白进行序列比对,发现均具有典型的索马甜家族标签和保守氨基酸残基。

2.2 桑树 TLP 家族蛋白氨基酸序列分析

通过 ExPASy 对桑树 25 条 TLP 蛋白氨基酸序列进行分析,发现 19 条 TLP 蛋白分子量介于 2.4×10^4~3.8×10^4,酸性氨基酸数量略多于碱性氨基酸。鉴定的 25 条 TLP 蛋白氨基酸序列中,有 21 条与蛋白 1Z3Q 相似性高于 40%,*E* 值小于 1×10^{-20}(表 1)。

表 1 桑树 TLP 家族信息

Table 1 Information of TLP family identified in mulberry

基因编号	氨基酸长度(aa)	等电点	分子量	*E* 值	相似度(%)	结构域
L484_001577	304	5.18	32 019	4.00×10^{-50}	43	信号肽+THN 结构域
L484_024982	253	8.31	27 360	6.00×10^{-48}	41	信号肽+THN 结构域
L484_024983	365	4.44	36 947	8.00×10^{-47}	44	信号肽+THN 结构域
L484_020153	288	8.24	31 007	3.00×10^{-51}	47	信号肽+THN 结构域
L484_010482	184	9.34	19 324	4.00×10^{-6}	35	THN 结构域
L484_023752	97	8.97	10 437	2.00×10^{-23}	62	信号肽+THN 结构域
L484_023753	134	8.06	15 104	8.00×10^{-19}	54	THN 结构域+COP 结构域
L484_023755	558	5.29	61 544	2.00×10^{-69}	54	信号肽+THN 结构域+水解酶结构域
L484_025349	246	5.25	26 240	3.00×10^{-38}	39	信号肽+THN 结构域
L484_017379	249	9.07	26 350	1.00×10^{-48}	41	信号肽+THN 结构域
L484_022589	244	4.68	25 195	3.00×10^{-48}	43	信号肽+THN 结构域
L484_022590	244	8.9	26 202	4.00×10^{-44}	40	信号肽+THN 结构域
L484_022591	240	8.87	25 049	5.00×10^{-26}	41	THN 结构域
L484_022592	244	5.01	26 088	5.00×10^{-45}	41	信号肽+THN 结构域
L484_004031	621	8.98	70 059	2.00×10^{-15}	42	THN 结构域
L484_010920	228	7.36	24 163	5.00×10^{-49}	43	THN 结构域
L484_024191	259	5.84	28 035	4.00×10^{-52}	45	信号肽+THN 结构域
L484_020224	508	5.26	54 848	7.00×10^{-70}	55	信号肽+THN 结构域
L484_021218	251	4.53	25 726	1.00×10^{-47}	43	信号肽+THN 结构域
L484_009461	264	6.84	28 455	5.00×10^{-42}	38	THN 结构域
L484_021878	335	5.71	35 884	3.00×10^{-52}	44	THN 结构域
L484_021879	341	4.55	35 778	3.00×10^{-48}	42	THN 结构域
L484_026587	343	6.44	37 922	3.00×10^{-70}	54	信号肽+THN 结构域
L484_026644	333	8.21	36 101	5.00×10^{-18}	32	THN 结构域
L484_026669	347	5.92	36 726	2.00×10^{-51}	45	THN 结构域

E 值和相似度表示与小果野蕉 1Z3Q 蛋白比对的数据。

同时从 NCBI 中获得桑树 TLP 蛋白对应的基因组序列和 CDS 序列。通过对基因结构、内含子位置和相位进行分析,发现 4 种基因结构类型(图 1)。根据剪接位置的不同,内含子分为 3 种相位类型,0 型内含子位于两个密码子之间,1 型位于密码子的第 1 和第 2 碱基之间,2 型位于密码子的第 2 和第 3 碱基之间[17]。相位的改变会导致后续阅读框发生改变,因此内含子的相位通常是比较保守的。12 个

含 1 个内含子的桑树 *TLP* 基因中,10 个均为 1 型相位;10 个含 2 个内含子的桑树 *TLP* 基因中,7 个属于 1 型、2 型相位。含有相同内含子相位的基因可能来源于同一祖先,而含有不同内含子相位的基因可能发生了独立的内含子获取或丢失,属于较古老的基因。

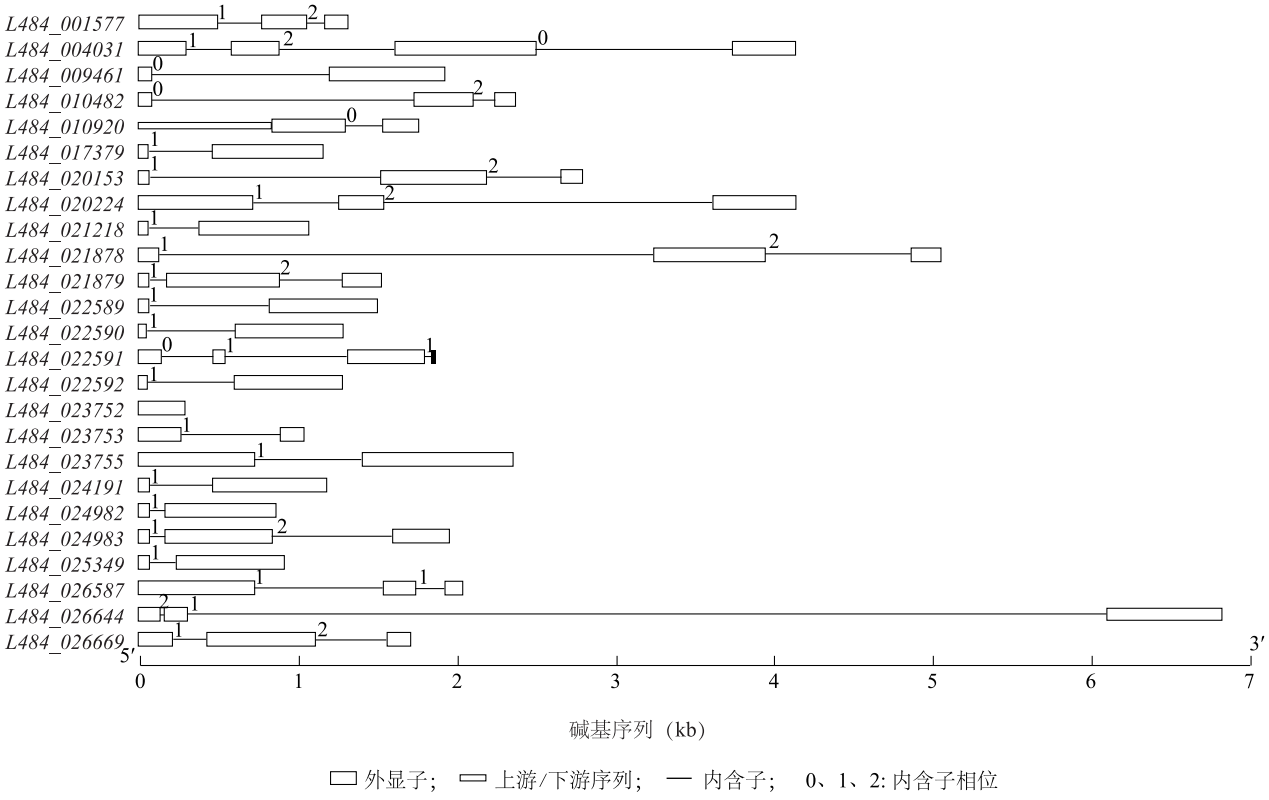


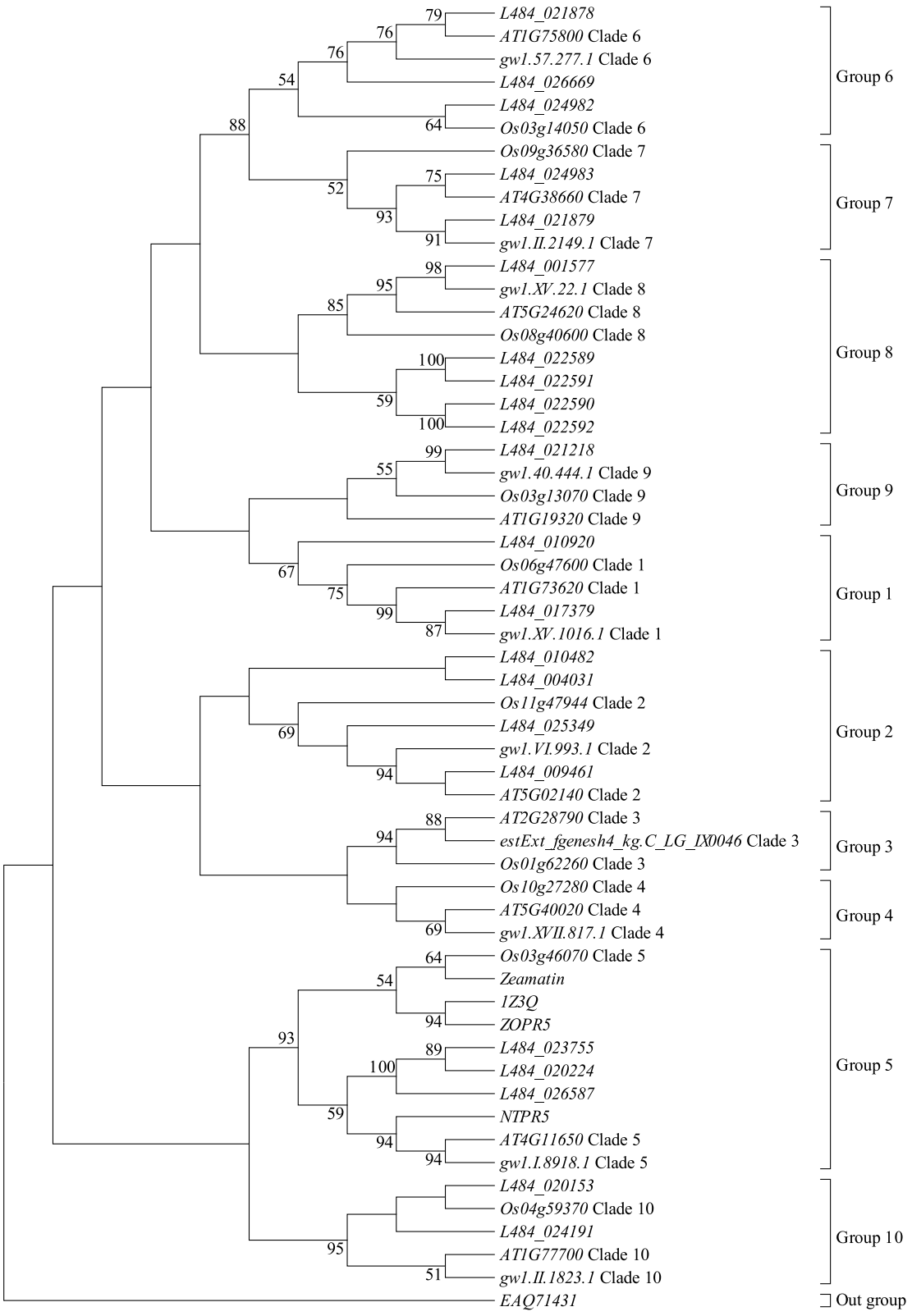
图 1　桑树 *TLP* 家族基因结构

Fig.1　The gene structure of *TLP* family in mulberry

使用 MEME SUITE 在线软件对桑树 TLP 家族蛋白保守氨基酸 Motif 进行分析,搜索参数最小宽度 6 个氨基酸,最大宽度为 50 个氨基酸。发现 5 类 Motif 的保守性较强,分别为基序 1(SGQDFYDVSLVDGFNLPVSVAPSGGSGGC)、基序 2(PDTCKPTVYSRIFKAACPRAYSYAYDD)、基序 3(GSDAATFTIKNNCKYTVWPGILSGAGKPQ)、基序 4(TGFELPPGZSRSLTVPPGWSGRFWGRTGC)、基序 5(CKSACEAFNTPEYCC)。23 条序列中包含 Motif 1 基序,22 条序列中含有 Motif 2、3 和 5 基序,18 条序列中含有 Motif 4 基序。

2.3　桑树 *TLP* 家族基因系统进化分析

Zhao 等[18]通过对杨树(*Populus trichocarpa*)的 50 个 *TLP*、拟南芥(*Arabidopsis thaliana*)的 10 个 *TLP*、水稻(*Oryza sativa*)的 10 个 *TLP* 进行系统进化分析,发现杨树与拟南芥和水稻类似,*TLP* 也聚类为 10 个聚类组。本研究中去除短链和分离序列,选取 22 个桑树 *TLP*、10 个拟南芥 *TLP*、10 个水稻 *TLP* 和 10 个杨树 *TLP* 构建 NJ 系统进化树(图 2),以稻瘟病菌(*Magnaporthe oryzae*) *TLP*(NCBI 登录号 EAQ71431)作为外群。与前人研究结果[12,18]类似,拟南芥、水稻和杨树的 10 个 *TLP* 分别属于 10 个聚类组,桑树 22 个 *TLP* 分属于其中的 8 个聚类组,且各个组中 *TLP* 数量很不均衡。其中,最大的聚类组 8 包含 5 个成员,聚类组 2 包含 4 个成员,聚类组 5 和聚类组 6 均包含 3 个成员,聚类组 1、7 和 10 均包含 2 个成员,聚类组 9 包含 1 个成员,聚类组 3 和聚类组 4 不含桑树 *TLP*。多项研究结果表明,聚类组 5 中的其他物种成员能响应病原或环境胁迫[6,19]。具有抗真菌活性的 TLP 往往也具有 β-1,3 葡聚糖酶活性[5]。葡聚糖酶活性使 TLP 能结合并降解真菌细胞壁的主要组分 β-1, 3 葡聚糖,为 TLP 进一步破坏真菌细胞膜奠定了基础[20]。聚类组 5 中的 *L484_023755* 编码蛋白含有 THN 和水解酶结构域,可能具有抑菌或葡糖苷酶活性,具体活性有待深入研究。



TLP 后缀的 Clade 序号参照文献[12]、[18]。

图 2 桑树 *TLP* 家族基因系统进化树

Fig.2 The phylogenetic tree of *TLP* family in mulberry

2.4　桑树 *TLP* 家族基因密码子使用特性

为全面了解桑树 *TLP* 家族基因密码子使用特性,对基因密码子使用的多个参数进行分析。结果(表 2)显示,桑树 *TLP* 家族基因密码子第 3 位的核苷酸使用偏性不强,仅有 10 个基因 *GC3s* 数值高于 60%,说明这些基因偏好 GC 作为第 3 位密码子。*ENC* 值反映基因编码密码子选择性强弱,其数值一般为 20~61,高表达基因密码子 *ENC* 值越靠近 20,表示对某些同义密码子偏性越大,基因表达对密码子的使用偏性越强[14]。*CAI* 值表示密码子适应指数,其值越高表示基因表达水平越强。分析发现,桑树 *TLP* 家族基因 *ENC* 值介于 42.4 至 61.0 之间,68%的基因 *ENC* 值大于 50.0,*CAI* 值介于 17.4 至 35.6 之间,只有 7 个基因的 *CAI* 值高于 25.0,说明桑树 *TLP* 家族基因密码子使用偏好性不强,多数属于低表达基因,响应植物生长或外界刺激信号的能力较弱,仅有少数基因属于高表达基因,这部分基因可能在植物响应环境胁迫中发挥作用。80%的桑树 *TLP* 家族基因 *Gravy* 值为负值,表示多数蛋白属于亲水性蛋白。

表 2　桑树 *TLP* 家族基因密码子使用特性

Table 2　Characterization of codon usage of *TLP* family genes in mulberry

基因	*T3s*	*C3s*	*A3s*	*G3s*	*CAI*	*CBI*	*GC3s*	*GC*	*Gravy*	*ENC*	*L-sym*	*L-aa*
LA484_001577	30.3	33.0	32.1	26.3	19.7	−5.4	48.3	51.0	10.3	61.0	294	304
LA484_004031	33.4	33.0	28.3	32.5	17.4	−3.3	50.7	46.5	−19.1	56.8	594	621
LA484_009461	30.6	35.2	28.3	33.9	21.2	−3.6	53.1	49.2	−3.1	58.5	254	264
LA484_010482	26.4	35.0	23.0	33.3	17.4	0.1	57.5	59.2	−33.7	60.7	181	184
LA484_010920	37.7	31.2	27.7	24.7	20.1	2.3	45.7	51.5	−26.6	59.3	219	228
LA484_017379	21.4	59.4	7.6	28.2	27.0	22.4	74.6	62.4	−1.6	42.4	244	249
LA484_020153	34.7	33.1	29.3	25.0	20.6	−2.5	47.1	48.4	9.4	60.6	278	288
LA484_020224	34.3	39.3	26.2	23.6	24.7	4.8	50.7	49.1	−22.0	58.9	491	508
LA484_021218	23.5	47.4	18.0	30.1	25.4	7.1	65.3	58.4	−0.9	46.5	245	251
LA484_021878	14.8	53.5	20.6	33.2	25.7	21.0	71.3	59.4	−46.4	46.7	327	335
LA484_021879	24.8	41.0	23.5	30.3	22.5	6.4	59.4	54.4	−4.2	57.0	335	341
LA484_022589	19.1	53.5	23.0	24.9	28.4	24.6	65.7	58.2	4.1	49.3	239	244
LA484_022590	16.1	56.9	13.1	37.5	35.6	30.1	76.2	58.7	−23.5	44.1	240	244
LA484_022591	14.2	52.9	28.4	24.4	22.3	14.7	64.5	57.2	2.8	51.7	234	240
LA484_022592	16.0	56.6	12.6	38.6	35.2	30.8	76.7	59.8	−23.2	45.7	240	244
LA484_023752	23.2	47.6	23.3	26.1	17.8	−0.3	61.3	56.7	−12.1	44.5	93	97
LA484_023753	34.5	39.5	25.8	26.5	18.7	−0.7	51.9	46.5	−49.9	61.0	133	134
LA484_023755	30.2	41.0	26.8	28.5	21.6	3.5	54.4	48.9	−28.6	57.1	544	558
LA484_024191	39.2	32.8	30.3	19.8	22.2	0.0	42.9	46.8	−10.5	58.0	252	259
LA484_024982	16.1	50.9	14.3	38.5	21.8	14.1	74.5	60.7	−7.9	48.8	247	253
LA484_024983	24.5	37.6	23.5	29.8	20.1	5.9	58.4	57.6	−5.0	58.1	356	365
LA484_025349	25.9	47.6	19.4	29.0	23.1	10.3	62.5	55.7	8.0	51.8	240	246
LA484_026587	32.2	40.3	28.8	23.1	24.1	8.7	51.1	49.5	−36.6	61.0	333	343
LA484_026644	32.3	39.7	21.2	30.9	26.3	7.8	56.4	51.2	−36.6	61.0	321	333
LA484_026669	27.6	40.3	22.1	29.6	22.8	3.4	58.1	55.5	−2.6	56.3	339	347

T3s、*C3s*、*A3s*、*G3s*:同义密码子第 3 位上 T、C、A、G 的出现频率;*CAI*:密码子适应指数;*CBI*:密码子偏好指数;*GC3s*:密码子的第 3 位的 G+C 含量;*GC*:基因的 G、C 含量;*Gravy*:平均亲水性值;*ENC*:有效密码子数;*L-sym*:同义氨基酸数;*L-aa*:氨基酸数。

ENC 值与 *GC3s* 关联分布图中的 *GC3s* 分布反映了植物所受的选择压力,*GC3s* 分布越广泛,表明密码子使用偏好性受碱基突变压力越大,*GC3s* 分布范围越小,表明密码子使用偏好性受自然选择压力影响越大[21]。在 *ENC* 值与 *GC*3*s* 关联分析中,基因分布越靠近标准曲线表明密码子使用偏好性越不受自然选择压力的影响,基因分布在标准曲线下方或较远区域,表明该基因受到选择压力或其他因素的

影响较大[22]。桑树 *TLP* 家族多数基因 *ENC* 值靠近标准曲线,且 *GC3s* 分布较广泛,说明基因进化主要受碱基突变压力影响,只有少数 *ENC* 值远离标准曲线的基因受到较强的自然选择压力影响(图 3)。

同义密码子相对使用度(*RSCU*)是指同义密码子实际使用数与理论使用数的比值。若 *RSCU*=1,表示密码子使用无偏性;若 *RSCU*>1,表示密码子使用频率高;若 *RSCU*<1,表示密码子使用频率低。使用 EMBOSS explorer 网站 cusp 软件分析并计算 *TLP* 家族基因 *RSCU* 值(表 3)。结果表明,桑树不同基因编码蛋白的密码子 *RSCU* 值不同,说明桑树 *TLP* 家族基因密码子存在偏性,但多数 *RSCU* 值偏离中心数值 1 的强度较弱,又说明密码子使用偏性不强。

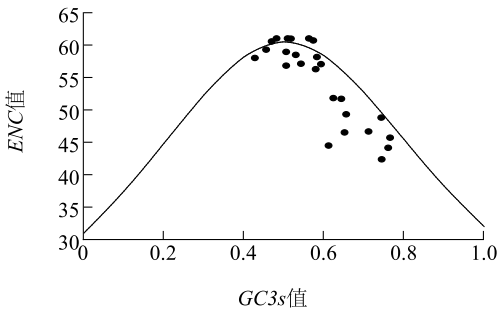


图 3 桑树 *TLP* 家族基因有效密码子数(*ENC*)与同义密码子第 3 位 G、C 含量(*GC3s*)的关系

Fig.3 Correlative analysis of effective number of codon (*ENC*) and G/C content on the 3rd site of synonymous codon (*GC3s*) of *TLP* family genes in mulberry

表 3 桑树 *TLP* 家族基因同义密码子使用情况

Table 3 Usage of synonymous codon of *TLP* family genes in mulberry

氨基酸	密码子	*RSCU* 值	氨基酸	密码子	*RSCU* 值	氨基酸	密码子	*RSCU* 值	氨基酸	密码子	*RSCU* 值
Ala	GCA	0.74	His	CAC	1.16	Gln	CAA	1.01	Thr	ACA	0.86
Ala	GCC	1.40	His	CAT	0.84	Gln	CAG	0.99	Thr	ACC	1.31
Ala	GCG	0.91	Lys	AAA	0.81	Arg	AGA	1.37	Thr	ACG	0.82
Ala	GCT	0.95	Lys	AAG	1.19	Arg	AGG	1.27	Thr	ACT	1.01
Cys	TGC	1.42	Leu	CTA	0.75	Arg	CGA	0.87	Val	GTA	0.51
Cys	TGT	0.58	Leu	CTC	1.92	Arg	CGC	0.65	Val	GTC	1.37
Asp	GAC	1.05	Leu	CTG	0.58	Arg	CGG	1.27	Val	GTG	1.06
Asp	GAT	0.95	Leu	CTT	0.95	Arg	CGT	0.57	Val	GTT	1.06
Glu	GAA	0.83	Leu	TTA	0.64	Ser	AGC	1.50	Ile	ATA	0.67
Glu	GAG	1.17	Leu	TTG	1.17	Ser	AGT	0.63	Ile	ATC	1.34
Phe	TTC	1.26	Asn	AAC	1.22	Ser	TCA	0.88	Ile	ATT	0.99
Phe	TTT	0.74	Asn	AAT	0.78	Ser	TCC	1.15			
Gly	GGA	0.77	Pro	CCA	0.87	Ser	TCG	0.84			
Gly	GGC	1.63	Pro	CCC	0.98	Ser	TCT	1.01			
Gly	GGG	0.88	Pro	CCG	1.21	Tyr	TAC	1.11			
Gly	GGT	0.72	Pro	CCT	0.94	Tyr	TAT	0.89			

甲硫氨酸和色氨酸为单密码子编码,3 个终止密码子均不存在偏性,未列出。

3 讨 论

TLP 属于多基因家族,是一类重要的病程相关蛋白,研究发现来自多个物种的 TLP 具有抗真菌活性[4]。目前已鉴定 28 个拟南芥 TLP、31 个水稻 TLP 和 55 个杨树 TLP[4,18]。本研究通过生物信息学方法获得 25 个桑树 TLP 蛋白氨基酸序列,均具有典型的 THN 结构域,多数具有信号肽,属于胞外分泌蛋白,具有较高的亲水性,说明桑树 TLP 蛋白在胞外基质中能较好地发挥功能,是否参与了将胞内胁迫信号传递给周围细胞有待深入研究。

Shatters 等[12]通过对昆虫、线虫、水稻、拟南芥 TLP 家族蛋白氨基酸序列进化分析发现,动物 TLP 蛋白单独分在一支,并可能是以单一祖先序列的形式来自于植物,而水稻和拟南芥 TLP 蛋白分布于多个支系,并存在染色体内和染色体间的复制。单子

叶和双子叶植物进化上发生分离后,*TLP* 基因在 10 个进化枝上发生了不对称的增加[10]。植物、动物和真菌等真核生物 *TLP* 碱基序列进化分析结果显示,陆生植物进化过程中 *TLP* 基因含量和多样性显著增加[23]。Liu 等[4]将获得的 118 个 *TLP* 碱基序列分为 9 组,植物 *TLP* 主要属于 5 个组(Ⅳ、Ⅵ、Ⅶ、Ⅷ、Ⅸ),认为 *TLP* 基因来自于大约 1.0×10^9 年前的植物、动物和真菌的共同祖先。对从 18 个植物基因组序列获取的 *TLP* 多家族基因比较发现,从莱茵衣藻(*Chlamydomonas reinhardtii*)到杨树(*P. trichocarpa*)*TLP* 发生了多样性的进化[23]。与之前的研究结果[10,18]类似,桑树 *TLP* 主要包含在 8 个聚类组中,且组内序列一致性较高,这是基因进化过程中基因发生复制的结果,是基因家族扩张的重要事件。有 3 个桑树 *TLP* 分布在第 5 聚类组中,其编码的蛋白质三级结构与抗菌蛋白 1Z3Q 有较高的一致性,说明这些蛋白质在进化上比较保守,该组的其他物种 TLP 具有抗真菌活性[6,19],说明该组的 3 个成员可能在植物响应真菌病原胁迫过程中发挥重要作用,值得深入研究。

密码子是蛋白质编码的基本结构单位,通过分析密码子参数的变化可以多角度了解进化事件。氨基酸密码子第 3 位碱基的改变通常不会引起编码氨基酸的改变,因此多数氨基酸往往由不止一个三联密码子编码。通常用 G、C 含量来反映突变的整体趋势,因此 *GC3s* 常被用作密码子使用偏好性的重要依据。桑树 *TLP* 家族基因偏好以 G 或 C 作为第 3 位密码子。双子叶植物偏爱 A/T 结尾的密码子,单子叶植物偏爱 G/C 结尾的密码子[22,24]。高水平的 *GC3s* 表示基因表达受 DNA 甲基化或基因改变导致的加速进化的影响[25]。在我们的研究中,双子叶植物桑树 *TLP* 家族基因的密码子第 3 位的核苷酸使用偏性并不强,且有部分基因偏好 G、C 作为第 3 位密码子;桑树 *TLP* 家族基因 *ENC* 值、*CAI* 值以及 *ENC*-plot 分析结果显示,该家族基因主要受碱基突变选择压力影响,部分基因也受到自然选择压力的影响。因该家族多数基因主要受内因碱基突变影响,这部分基因在植物生长发育或应对胁迫过程中的作用可能较弱,而另外少数 *GC3s* 值较高的基因主要受外因自然选择压力的影响,这部分基因 *ENC* 值也相对较小,说明这些基因表达水平较高,在植物应对环境胁迫中可能发挥了作用,这些基因值得进一步深入研究。

参考文献:

[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] 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: 289-298.

[3] 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: 583-591.

[4] 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.

[5] LAURENCE M B, VRIET C, PEUMANS W J, et al. A molecular basis for the endo-b-1, 3-glucanase activity of the thaumatin-like proteins from edible fruits[J]. Biochimie, 2003, 85: 123-131.

[6] LEONE P, MENU-BOUAOUICHE L, PEUMANS W J, et al. Resolution of the structure of the allergenic and antifungal banana fruit thaumatin-like protein at 1.7-A[J]. Biochimie, 2006, 88(1): 45-52.

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

[8] RAJAM M V, CHANDOLA N, SAIPRASAD GOUD P, et al. Thaumatin gene confers resistance to fungal pathogen as well as tolerance to abiotic stresses in transgenic tobacco plants[J]. Biologia Plantarum, 2007, 51: 135-141.

[9] 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.

[10] 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.

[11] 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: 11-23.

[12] SHATTERS R G J, 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.

[13] HE N, ZHANG C, QI X, et al. Draft genome sequence of the mulberry tree *Morus notabilis*[J]. Nature Communications, 2013, 4: 2445.

[14] WRIGHT F. The 'effective number of codons' used in a gene[J].

Gene,1990,87(1):23-29.

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

[16] 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,2012,111:29-39.

[17] SHARP P A. Speculations on RNA splicing (minireview)[J]. Cell,1981,23(643):621.

[18] 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.

[19] SINGH N K,KUMAR K R R,Kumar D,et al. Characterization of a pathogen induced thaumatin-like protein gene *AdTLP* from *Arachis diogoi*,a wild peanut[J].PLoS One,2013,8(12):e83963.

[20] GRENIER J,POTVIN C,TRUDEL J,et al. Some thaumatin-like proteins hydrolyse polymeric beta-1,3-glucans[J].Plant Journal,1999,19:473-480.

[21] 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.

[22] ZHANG W J,ZHOU J,LI Z F,et al. Comparative analysis of codon usage patterns among mitochondrion,chloroplast and nuclear genes in *Triticum aestivum* L.[J].Journal of Integrative Plant Biology,2007,49(2):246-254.

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

[24] 刘汉梅,赵 耀,顾 勇,等.几种植物 *waxy* 基因的密码子用法特性分析[J].核农学报,2010,24(3):476-481.

[25] 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.

(责任编辑:张震林)