GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Positive selection sites in tertiary structure of Leguminosae *Chalcone isomerase 1*

R.K. Wang[1], S.F. Zhan[1], T.J. Zhao[2], X.L. Zhou[1] and C.E. Wang[1]

[1]Jiujiang Laboratory for Seed Quality of Soybean National Center for
Soybean Improvement, Research Center for Soybean,
College of Life Science, Jiujiang University, Jiangxi Province, China
[2]National Key Laboratory for Crop Genetics and Germplasm Enhancement,
Soybean Research Institute, National Center for Soybean Improvement,
Nanjing Agricultural University, Jiangsu Province, China

Corresponding author: C.E. Wang
E-mail: soybeanw@163.com

Genet. Mol. Res. 14 (1): 1957-1967 (2015)
Received January 9, 2014
Accepted March 26, 2014
Published March 20, 2015
DOI http://dx.doi.org/10.4238/2015.March.20.5

ABSTRACT. Isoflavonoids and the related synthesis enzyme, chalcone isomerase 1 (CHI1), are unique in the Leguminosae, with diverse biological functions. Among the Leguminosae, the soybean is an important oil, protein crop, and model plant. In this study, we aimed to detect the generation pattern of Leguminosae *CHI1*. Genome-wide sequence analysis of *CHI* in 3 Leguminosae and 3 other closely related model plants was performed; the expression levels of soybean chalcone isomerases were also analyzed. By comparing positively selected sites and their protein structures, we retrieved the evolution patterns for Leguminosae CHI1. A total of 28 *CHI* and 7 *FAP3* (*CHI4*) genes were identified and separated into 4 clades: *CHI1*, *CHI2, CHI3*, and *FAP3*. Soybean genes belonging to the same chalcone isomerase subfamily had similar expression patterns. *CHI1*, the unique chalcone isomerase subfamily in Leguminosae, showed signs of significant positive selection as well as special expression characteristics, indicating an accelerated evolution throughout its divergence. Eight sites were identified as undergoing positive selection with high confidence. When mapped onto the tertiary structure of CHI1, these 8 sites were observed surrounding the enzyme substrate only; some of them connected to the catalytic core of CHI. Thus, we inferred that the generation of Leguminosae CHI1 is dependent on the positively selected amino acids surrounding its catalytic substrate. In other words, the evolution of CHI1 was driven by specific selection or processing conditions within the substrate.

Key words: Leguminosae; Soybean; Chalcone isomerase;
Positive selection; Protein structure

INTRODUCTION

Isoflavonoids are secondary metabolites that mediate diverse biological functions and have significant ecological impacts. Plants use isoflavonoids and their derivatives as phytoalexin compounds targeting disease-causing pathogenic fungi and other microbes. In addition, the soybean uses isoflavonoids to stimulate soil-microbe rhizobium to form nitrogen-fixing root nodules (Verma, 1992). In the plant kingdom, approximately 95% of isoflavonoids are found in the family Leguminosae, and 60% of Leguminosae flavonoids are 5-deoxyflavonoids (Hegnauer and Gpayer-Barkmeijer, 1993). The soybean is an important oil and protein crop, as well as a model plant of the Leguminosae. Soybean isoflavonoids have direct but complex effects on human health. In particular, they can prevent many hormone-dependent cancers and improve a woman's health (Ferrer et al., 2008).

Isoflavonoids are derived from the phenylpropanoid pathway, which is conserved in all plant species. In most species, chalcone synthase (CHS) and chalcone isomerase (CHI) are important in the production of isoflavonoids and flavonoids. Some CHIs isolated from *Medicago* and *Arabidopsis* have been studied structurally and mechanistically (Bednar and Hadcock 1988; Jez et al., 2000; Hur et al., 2004; Ngaki et al., 2012). These structure-function analyses suggest that the formation of a hydrogen bond network between the active site of CHI and its substrates is crucial for the catalytic activity of an enzyme (Jez and Noel, 2002; Hur et al., 2004). CHIs were derived from fatty-acid-binding proteins (FAP). Three *CHI* clades and one *FAP3* clade, with distinctive phylogenic lineages, coexist in the soybean; they are *CHI1*, *CHI2*, *CHI3*, and *FAP3* (previously termed *CHI4* in some articles) (Ralston et al., 2005). CHI2 is ubiquitous in plants and converts 6'-deoxychalcone into (2S)-5-deoxychalcone. In contrast, CHI1 appears to be legume specific and possesses additional catalytic activity, which allows it to convert chalcone into (2S)-naringenin. (2S)-5-deoxychalcone and (2S)-naringenin are the precursors of many flavonoids and isoflavonoids. *In vitro* CHI3 has the similar produced as CHI1 and CHI2, but the function of CHI3 in plants has not been ascertained (Ralston, 2005). The *FAP3* gene is highly homologous with *CHI*; however, their ligands in plants may be composed of different fatty acids. *FAP3* knockout *Arabidopsis thaliana* plants show elevated α-linolenic acid levels and marked reproductive defects, including aberrant seed formation (Ngaki et al., 2012).

Evolutionary biologists have conducted numerous studies on the factors affecting the evolutionary rates of protein-coding regions over the years (Ridout et al., 2010). Evolutionary rates vary not only between proteins but also between different sites within a single protein (Larracuente et al., 2008). The likelihood that positive selection alters an amino acid at a given site may depend on several factors such as the physical and chemical nature of the amino acid, the functional importance of the site, the surrounding environment, the physical properties of the structure, and the folding properties of the structure (Kosiol et al., 2008). Protein secondary structure, the physical arrangement of the amino acid chain, is produced mainly by the amino acid sequence and is another factor that may contribute to varying evolution rates at different amino acid positions. The amino acid order directly affects protein folding, and, therefore, tertiary structure and function; it is also highly conserved between homologous proteins (Petersen et al., 2007; Ridout et al., 2010).

The isoflavonoid and related *CHI1* gene family are only found in the Leguminosae. In this paper, the positively selected sites of *CHI* and *FAP3* sequences were compared among 3 Leguminosae and 3 other closely related plants to detect whether there are different evolution patterns among the Leguminosae *CHI1* genes. This will provide a new insight into the specific genetic system of isoflavonoids in the Leguminosae and will be beneficial for soybean breeding.

MATERIAL AND METHODS

Gene retrieval and identification

Genome data for 6 model plants (i.e., *Glycine max*, *Medicago truncatula*, *Phaseolus vulgaris*, *Arabidopsis thaliana*, *Ricinus communis*, and *Cucumis sativus*) were used for sequence analyses. CHI proteins were searched on the Phytozome website (http://www.phytozome.net/) via BLASTP (E-value <1e^{-50}). The sequences collected were used for subsequent BLASTP searches to identify any missing sequences. Finally, all of the sequences identified were submitted to the Pfam database (http://pfam.sanger.ac.uk/search) (Punta et al., 2012) to check whether they contained a complete chalcone domain (the Pfam number of the chalcone isomerase is PF02431). Some genes were adjusted manually. All information for the *CHI* genes, including accession Nos., chromosomal locations, open reading frame lengths, and number of exons and introns, were retrieved from the Phytozome database. Structures of the genes and regions encoding the Pfam domains were created using the Gene Structure Display Server (http://gsds.cbi.pku.edu.cn/) (Guo et al., 2007).

Phylogenetic and expansion pattern research

The *CHI* candidate genes collected were used for phylogenetic analyses. Amino acid sequences were aligned using the CLUSTAL program of Molecular Evolutionary Genetics Analysis 4.0 (Tamura et al., 2007). The alignment results guided the neighbor-joining tree construction using the DNA sequences, and any alignment gaps were deleted manually. The parameters were set as follows: model, p-distance; bootstrap, 1000 replicates; and gap/missing data, pairwise deletion.

Expressed sequence tag (EST) resources have increased significantly for the soybean and may be used to research gene transcripts. The EST data can be used to detect the expansion pattern of a given gene family. In soybean, the identification of ESTs has proceeded rapidly, with approximately 1.5 million ESTs now available in GenBank (as of July 12, 2012). The EST data can be searched from GenBank (http://www.ncbi.nlm.nih.gov/dbEST/index.html) and well analyzed via web SoyKB (http://soykb.org/).

Molecular evolution and tertiary structure rebuilding

We used a maximum likelihood approach to investigate positive selection, with a Codeml procedure via Phylogenetic Analysis by Maximum Likelihood 4.0 under branch-site model situations. Sites under positive selection will show a non-synonymous/synonymous rate ratio (dN/dS, denoted as ω) that differs from that of the other sites. The branch-site models allow ω to vary both between sites in the sequences and across branches on the tree. It assumes that there are 4 site classes in the sequence of a tree. The first class is highly conserved in all branches, with a small ω ratio (i.e., ω_0). The second class includes neutral or weakly constrained sites with ω_1, where ω_1 is near 1. In the third and fourth classes, the given branch is defined as foreground branches. The background lineages have ω_0 or ω_1, but the foreground branches have ω_2, which may be >1. Its purpose is to detect positive selection that affects a few sites along given lineages (i.e., foreground branches). In the likelihood ratio test, the null hypothesis fixes $\omega_2 = 1$ (i.e., neutral selection), and the alternative hypothesis constrains $\omega_2 \geq 1$ (i.e., positive selection). In the presence of positive selection, the posterior probabilities for the sites with positive selection were calculated using the Bayes empirical Bayes (BEB) method (Yang, 2007).

The predicted secondary structures for CHI were obtained using the UCL web site tools (http://bioinf.cs.ucl.ac.uk/psipred/). *MtCHI1* is well studied (Jez et al., 2000; Ngaki et al., 2012), and it was acquired and analyzed from the protein data bank (http://www.rcsb.org/pdb/home/home.do).

RESULTS

Identification and organization of *CHI* genes

Twenty-eight *CHI* and 7 *FAP3* genes were identified by BLAST from the 6 model plants. Notably, leguminous plants have more *CHI* members than nonleguminous plants. There were always 6-8 genes encoding CHI in leguminous plants, while only 3-4 genes were found in the other plants. This is because the Leguminosae have additional copies of *CHI1* (Table 1). Phylogenetic relationships between these genes were similar according to the neighbor-joining and Bayesian methods (data not shown). In agreement with previous studies, these genes could be divided into 4 clades: *CHI1*, *CHI2*, *CHI3*, and *FAP3* (also known as *CHI4* in some articles). The *CHI3* and *FAP3* genes were found in *M. truncatula*, possibly because of its low sequencing throughput.

The information from the phylogenetic tree and the exon/intron structure permitted exploration of the evolutionary relationships of the gene families. *CHI1* is a special isoflavonoid related subfamily only found in the Leguminosae, with 3, 2, or 5 members in *G. max*, *P. vulgaris*, and *M.* truncatula, respectively. The leguminous *CHI1* genes always contained 4 exons. *MtCHI1B*, *MtCHI1C*, and *MtCHI1D* were divergent from the other *CHI1s* and contained 5, 2, or 3 exons, respectively. Moreover, they had no complete *CHI* regions, possibly indicating a process of pseudogenization. A few copies were found as *CHI2* in these plants. *CHI1* and *CHI2* were identified as highly homologous genes, with similar exon/intron structures, untranslated regions, and CHI regions (Figure 1). Furthermore, genome sequence analysis revealed that *CHI1* and *CHI2* form a tandem cluster in the Leguminosae (Table 1). We also analyzed the sequence of *FAP3*, which is the ancestor gene subfamily of *CHI* (Ngaki et al., 2012). The results of phylogenetic analyses were similar to those of Lyle Ralston (Ralston et al., 2005), which defined *FAP3* as *CHI4*.

Table 1. Candidate genes for *CHI* and *FAP3*.

Gene symbol	Local tag in phytozome	Gene length (bp)	E value in Pfam*
AthCHI 2A	AT3G55120	915	8.00E-55
AthCHI 2B	AT5G66220	2051	4.60E-39
AthCHI 3	AT1G53520	523	6.40E-54
AthFAP 3	AT5G05270	1540	8.70E-50
CsaCHI 2	Cucsa.182640	1356	7.40E-54
CsaCHI 3	Cucsa.058090	1687	4.60E-55
CsaFAP 3	Cucsa.143940	997	1.20E-47
GmaCHI 1A	Glyma20g38560	4930	1.30E-52
GmaCHI 1B1	Glyma20g38570	3243	8.90E-58
GmaCHI 1B2	Glyma10g43850	2900	1.20E-56
GmaCHI 2	Glyma20g38580	1964	4.90E-60
GmaCHI 3A	Glyma13g33730	1916	6.50E-53
GmaCHI 3B	Glyma15g39050	2630	1.70E-53
GmaFAP 3A	Glyma06g14820	4118	5.20E-48
GmaFAP 3B	Glyma04g40030	4101	6.10E-47
MtrCHI 1A1	Medtr1g146040	2437	8.90E-51
MtrCHI 1A2	Medtr1g146050	2071	2.50E-53
MtrCHI 1B	Medtr1g146060	1323	2.00E-06
MtrCHI 1C	Medtr1g146210	984	4.20E-02
*MtrCHI 1D**	Medtr1g146080	2932	1.70E-51
MtrCHI 2A	Medtr1g146030	1688	3.10E-59
MtrCHI 2B	Medtr1g146220	1846	5.50E-59
PtrCHI 2	POPTR_0010s21980	1390	2.30E-62
PtrCHI 3A	POPTR_0011s10330	777	5.30E-56
PtrCHI 3B	POPTR_0001s38970	1799	7.80E-40
PtrFAP 3	POPTR_0019s08610	1514	5.00E-46
PvuCHI 1A	Phvulv091023708m.g	2308	9.80E-53
PvuCHI 1B	Phvulv091005737m.g	2296	1.40E-52
PvuCHI 2A	Phvulv091003077m.g	1796	3.20E-61
PvuCHI 2B	Phvulv091005738m.g	1867	5.80E-61
PvuCHI 3	Phvulv091004776m.g	1744	2.20E-50
PvuFAP 3	Phvulv091026497m.g	1721	1.60E-46
RcoCHI 2	29740.t000017	2722	1.90E-44
RcoCHI 3	29989.t000018	4477	2.60E-29
RcoFAP 3	29729.t000101	1857	6.10E-48

*The protein family identified is PF02431.10.

Expression patterns of soybean chalcone isomerase

EST databases contain libraries organized by organism, tissue type, and developmental stage that can provide valuable information for gene expression research. Thus, EST mining was used to analyze transcript levels of soybean chalcone isomerase in the leaf, flower, root, root nodule, pod, and seed. From the different tissue types and developmental stages, expression of all the soybean chalcone isomerase genes was detected. Specialized enzymes biosynthesize chemicals of given functions, often sharing a pedigree with primary metabolic enzymes (Hartmann, 2007). Genes belonging to a given isomerase subfamily always have similar expression patterns. The 3 copies of *GmCHI1* were highly expressed in the root. Thus, we suggest that *GmCHI1* plays an important role in isoflavonoid synthesis in the root. Moreover, *GmCHI1A* was expressed at a higher level than that of *GmCHI1B1* and *GmCHIB2* in all tissues, suggesting its functional importance. Despite their close relationship, *GmCHI1B1* was typically expressed in the seed, while *GmCHI1B2* was expressed in the pod shell. *GmCHI2* and *GmCHI3* were barely detected in the root or root nodule, and they shared similar tissue expression characteristics. This was different from the 2 *GmFAP3* genes, which had high transcription in the seed and root. Thus, phylogenetically divergent genes also exhibited different expression patterns.

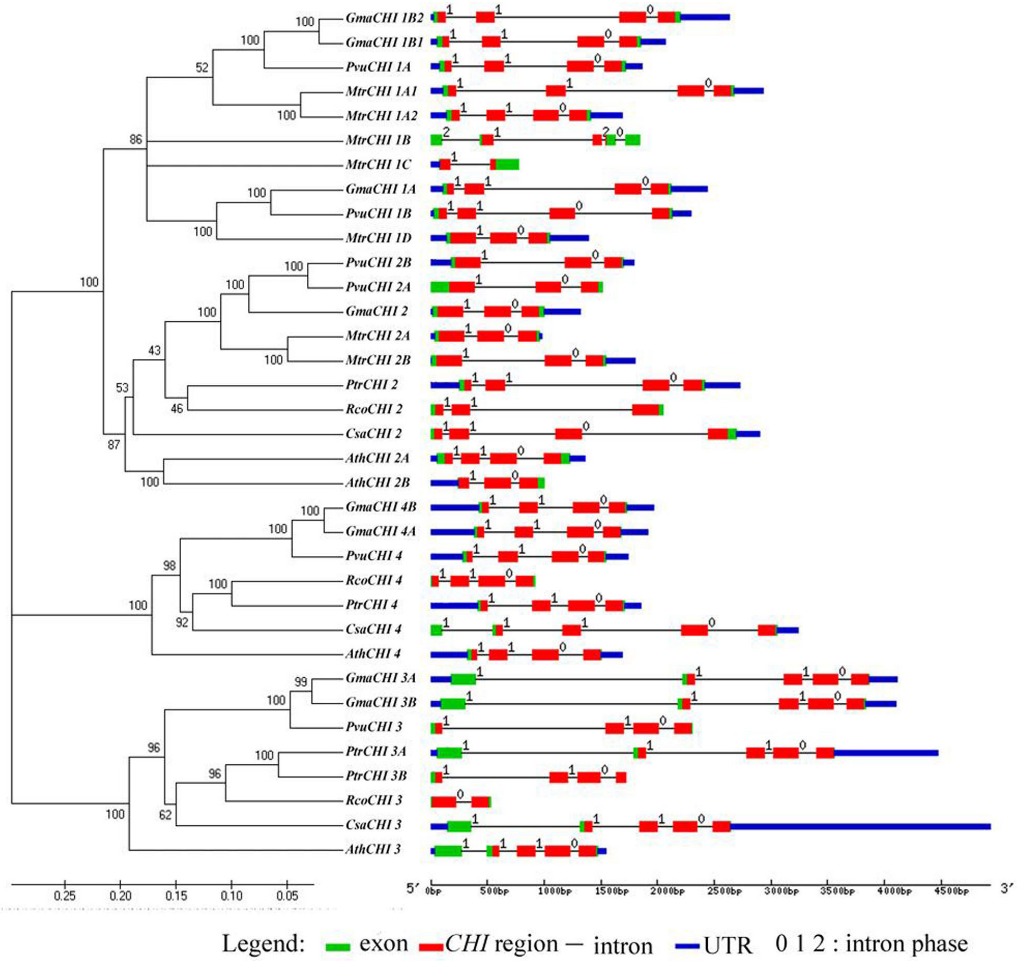

Figure 1. Neighbor-joining tree and gene structure of *CHI* and *FAP3*. Sequences are shown by gene symbols. Bootstrap values from neighbor-joining analyses are listed to the left of each node. Values >50 are shown. Exons and introns are shown by filled boxes and single lines, respectively. The *CHI* or *FAP3* region was identified from the Pfam database.

Positively selected site detection in Leguminosae CHI

To determine which sites differ in their evolutionary patterns, the ω values for sites were calculated using a branch-specific model. The 4 main clades of *CHI* and *FAP3* were independently defined as the foreground branches. When the *CHI1*, *CHI2*, and *FAP3* branches were defined as the foreground branches, the null hypothesis was rejected ($P < 0.01$), and the estimated parameters of the alternative hypotheses indicated that an average of 8-39% of the sites on these branches were under positive selection, with ω values of $\omega_2 > 1$ (Table 2). In the FAP3 branch ($\omega_2 = 999$), some sites undergoing synonymous mutation did not contain nonsynonymous mutations, and the value of dN/dS could not be calculated. A BEB analysis was used, rather than Naive Empirical Bayes, to identify putative codons under positive selection;

Naive Empirical Bayes is less conservative and can be more prone to error in smaller data sets (Casola and Hahn., 2009).

In the branch-site model leading to *CHI1* or *FAP3*, 28 or 39% of sites were under positive selection, respectively; and 8 or 10 sites were found in the BEB analysis, respectively, with a posterior probability of >95% (Table 2). This suggests that many sites in *CHI1* and *FAP3* have undergone adaptive evolution. Considering that *FAP3* is functionally different from *CHI*, its abundant number of positively selected sites is not surprising. As a special protein existing only in Leguminosae, *CHI1* also has many positively selected sites indicating functional conversion. On the other hand, no sites were detected by BEB analysis in branch CHI2, and only 8% of sites were under positive selection in the estimated parameters of the branch-site model. Meanwhile, the alternative hypothesis of *CHI3* could not reject the null hypothesis. This suggests that *CHI2* and *CHI3* have not undergone significant positive selection during their divergence from the other *CHI* genes.

Table 2. Summary statistics for the detection of selection using branch-site models.

Foreground	Null hypothesis		Alternative hypothesis	
branch	Estimated parameters	*ln L*	Estimated parameters	*ln L*
CHI1	P_0=0.40, P_1=0.05, P_2+P_3=0.56 ω_0=0.16, ω_1=ω_2=1	-10667.13	P_0=0.64, P_1=0.08, P_2+P_3=0.28 ω_0=0.16, ω_1=1, ω_2=14.69	-10663.67**
CHI2	P_0=0.08, P_1=0.0_1, P_2+P_3=0.91 ω_0=0.17, ω_1=ω_2=1	-10674.08	P_0=0.81, P_1=0.11, P_2+P_3=0.08 ω_0=0.16, ω_1=1, ω_2=999	-10667.46**
CHI3	P_0=0.58, P_1=0.07, P_2+P_3=0.35 ω_0=0.17, ω_1=ω2=1	-10673.77	P_0=0.56, P_1=0.07, P_2+P_3=0.37 ω_0=0.17, ω_1=1, ω_2=3.91	-10672.79
FAP3	P_0=0.58, P_1=0.07, P_2+P_3=0.35 ω_0=0.17, ω_1=ω_2=1	-10671.51	P_0=0.54, P_1=0.07, P_2+P_3=0.39 ω_0=0.17, ω_1=1, ω_2=999	-10663.33**

MtrCHI 1B, MtrCHI 1C, RcoCHI 2, AthCHI 2B, PtrCHI 3B, RcoCHI 3, were not included in the branch-specific model because its long gaps; 197 codons were used. The test statistic *2Δl* is compared to a χ^2 distribution with 1 degree of freedom; **significant results.

Secondary and tertiary structures of CHI1

Understanding the structure of CHI and FAP3 is important for clarifying the roles of these proteins and providing insight into potential evolutionary pathways of CHIs. Highly similar α-helix and β-sheet structures were observed for all of the CHI and FAP3 proteins examined, which always included 7 α-helixes and 7 β-sheets in the same order (Figure 2a). Near these structural domains, conserved amino acid residues were found, implying their functional importance. Figure 2b shows the structure of MtCHI1 from the Protein Data Bank, which was derived from X-ray crystallography by Jez et al. (2000). The overall structure of MtCHI resembles an upside-down bouquet that adopts an open-faced β-sandwich fold; a large β-sheet (β3a-β3f) and a layer of α-helices (α1-α7) comprise the core structure, with short β-strands (β1, β2) on the opposite side of the large β-sheet (Jez et al., 2000; Ngaki et al., 2012).

We have detected the primary, positively selected sites in *CHI1* and *FAP3* (in the branch-site model, *CHI1*, *CHI2*, and *FAP3* were under positive selection but no sites were identified via BEB analysis of *CHI2*). Note that these sites were calculated based on the branches of all *CHI1* and *FAP3* proteins rather than a single sequence. In the branch leading to *CHI1*, 8 selection sites were found (95% level of BEB), which were always in the regions of β3a, β3c, β3e, α7, and, especially, α4, indicating functionally accelerated evolution of these secondary structures. FAP3, β3a, β3d, β3f, and α6 were found to contain positively selected sites. Domains containing positively selected sites usually have a significant impact on protein evolution.

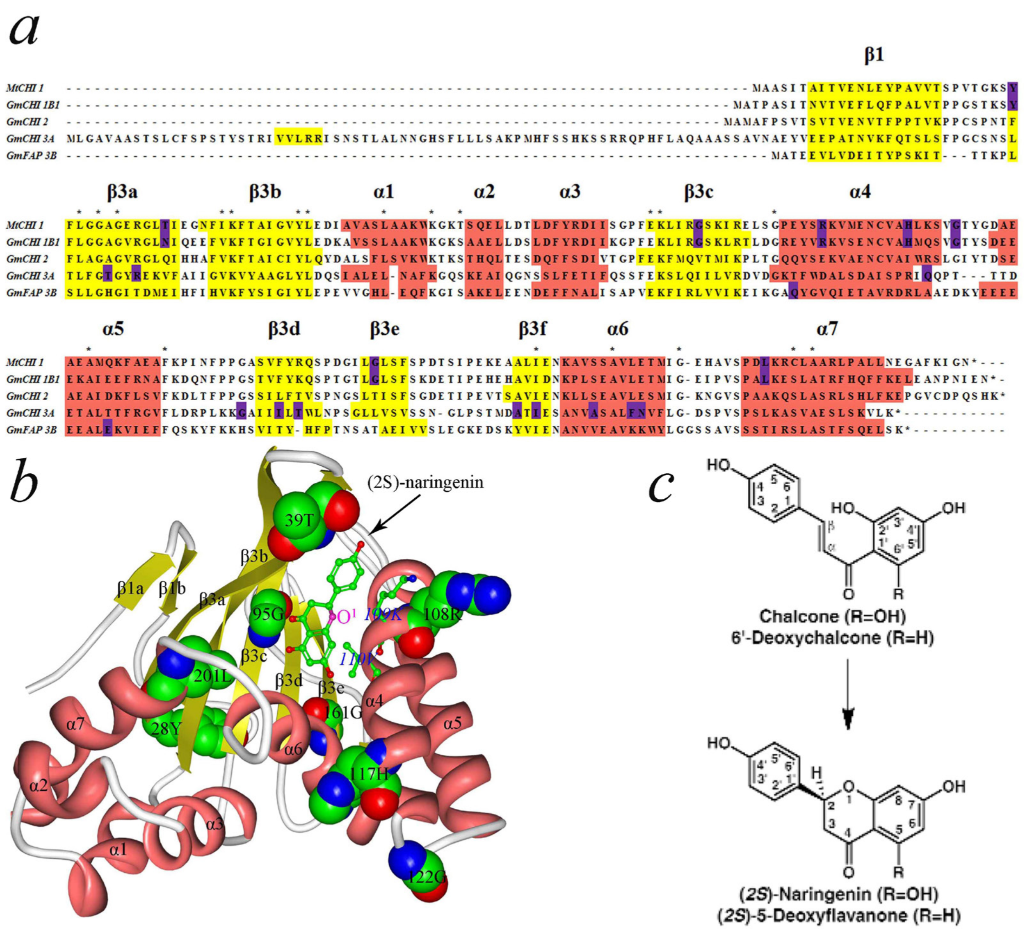
a
β1
MtCHI 1 --MAASITAITVENLEYPAVVTSPVTGKSY
GmCHI 1B1 --MATPASITNVTVEFLQFPALVTPPGSTKSY
GmCHI 2 ---MAMAFPSVTSVTVENVTFPPTVKPPCSPNTF
GmCHI 3A MLGAVAASTSLCFSPSTYSTRIVVLRRISNSTLALNNGHSFLLLSAKPMHFSSHKSSRRQPHFLAQAAASSAVNAEYVEEPATNVKFQTSLSFPGCSNSL
GmFAP 3B --MATEEVLVDEITYPSKIT---TTKPL
β3a
β3b
α1
α2
α3
β3c
α4
MtCHI 1 FLGGAGERGLTIEGNFIKFTAIGVYLEDIAVASLAAKWKGKTSQELLDTLDFYRDIISGPFEKLIRGSKIRELSGPEYSRKVMENCVAHLKSVGTYGDAE
GmCHI 1B1 FLGGAGVRGLNIQEEFVKFTGIGVYLEDKAVSSLAAKWKGKSAAELLDSLDFYRDIIKGPFEKLIRGSKLRTLDGREYVRKVSENCVAHMQSVGTYSDEE
GmCHI 2 FLAGAGVRGLQIHHAFVKFTAICIYLQYDALSFLSVKWKTKSTHQLTESDQFFSDIVTGPFEKFMQVTMIKPLTGQQYSEKVAENCVAIWRSLGIYTDSE
GmCHI 3A TLFGTGYREKVFAIIGVKVYAAGLYLDQSIALEL-NAFKGQSKEAIQGNSSLFETIFQSSFEKSLQIILVRDVDGKTFWDALSDAISPRIQQPT---TTD
GmFAP 3B SLLGHGITDMEIHFIHVKFYSIGIYLEPEVVGHL-EQFKGISAKELEENDEFFNALISAPVEKFIRLVVIKEIKGAQYGVQIETAVRDRLAAEDKYEEEE
α5
β3d
β3e
β3f
α6
α7
MtCHI 1 AEAMQKFAEAFKPINFPPGASVFYRQSPDGILGLSFSPDTSIPEKEAALIENKAVSSAVLETMIG-EHAVSPDLKRCLAARLPALLNEGAFKIGN*---
GmCHI 1B1 EKAIEEFRNAFKDQNFPPGSTVFYKQSPTGTLGLSFSKDETIPEHEHAVIDNKPLSEAVLETMIG-EIPVSPALKESLATRFHQFFKELEANPNIEN*-
GmCHI 2 AEAIDKFLSVFKDLTFPPGSSILFTVSPNGSLTISFSGDETIPEVTSAVIENKLLSEAVLESMIG-KNGVSPAAKQSLASRLSHLFKEPGVCDPQSHK*
GmCHI 3A ETALTTFRGVFLDRPLKKGAIIILTWLNPSGLLVSVSSN-GLPSTMDATIESANVASALFNVFLG-DSPVSPSLKASVAESLSKVLK*----------
GmFAP 3B EEALEKVIEFFQSKYFKKHSVITY-HFPTNSATAEIVVSLEGKEDSKYVIENANVVEAVKKWYLGGSSAVSSSTIRSLASTFSQELSK*----------
b
(2S)-naringenin
39T
β3b
β1a
β1b
β3a
95G
O¹
109K
108R
110V
201L
β3c
β3d
β3e
161G
α4
α7
28Y
α6
α5
117H
α2
α3
α1
122G
c
HO
OH
HO
R
O
Chalcone (R=OH)
6'-Deoxychalcone (R=H)
HO
H
OH
O
O
R
(2S)-Naringenin (R=OH)
(2S)-5-Deoxyflavanone (R=H)

Figure 2. Structure and positively selected sites of CHI1. **a.** Chalcone and naringenin substrates of CHI1 are shown with their numbering systems. **b.** Structure of MtCHI from the Protein Data Bank that was submitted by Jez et al. (2000). The labels of α-helices (pale red) and β-sheets (yellow) are in accordance with Jez et al. (2000). Eight positively selected sites (blue), with their molecular structures, are shown on the protein, as well as the position of (2S)-naringenin and (**c**) the important catalytic core constituted by 109K and 110V. Positively selected sites are in purple, and conservative amino acid residues are marked by and asterisk (*).

DISCUSSION

The *CHI* genes only exist in plants, which are involved in isoflavonoid and flavonoid synthesis; and *CHI*s are derived from *FAP3* (Ngaki et al., 2012). The topologies of their phylogenetic trees, intron structure, and Pfam domains (Punta et al., 2012) were carefully considered to reveal the evolution of the *CHI* gene family. There are 3 types of *CHI*s; 2 members were identified as *CHI2* and *CHI3* and could be found in all of the 6 plants studied. All *CHI* and

FAP3 proteins have similar Pfam domains, and secondary and tertiary structures. The chalcone domain identified in Pfam nearly covers all of the coding regions of *CHI*, except *CHI3*.

Leguminous isoflavonoids are secondary metabolites that are beneficial for plant resistance to diseases and insect pests (Hegnauer and Gpayer-Barkmeijer, 1993). CHI1 is a special protein that only exists in the Leguminosae and plays an important role in the generation of Leguminosae isoflavonoids; *CHI1* is derived from *CHI2* and *CHI3* (Shimada et al., 2003; Ralston et al., 2005). *GmCHI1* is highly expressed in the root and differed from the parental *CHI2* and *CHI3* genes, which were barely detectable in the roots (Table 3). Otherwise, the roots of the soybean can secrete isoflavonoids to stimulate soil-microbe rhizobium and form nitrogen-fixing root nodules (Verma, 1992). Thus, we suggest that CHI1 diverged from the parental CHIs; thus, CHI1 is expressed in the roots, where it has adapted to its novel role in root isoflavonoid synthesis. In correlation with the unique characteristics of *CHI1*s in the Leguminosae, significant positive selection in the branch-site model was detected along its branch. Moreover, 8 sites with a posterior probability of >95% were found in the BEB analysis, indicating that these amino acid of CHI1 experienced natural positive selection. In positive selection, the given sites have a larger nonsynonymous substitution rate than the synonymous substitution rate. It most often occurs under environmental changes or when plants migrate to new areas with different environmental pressures (Ridout et al., 2010). Thus, the positive selection of Leguminosae CHI1 offers a probable explanation as to why the *CHI1* genes encode proteins important to Leguminosae pathogen resistance. The positively selected sites may be a primary cause of the divergence between *CHI1* and the other *CHI*s.

Table 3. *In silico* expression analysis of *GmCHI* genes.

Gene symbol	Young leaf	Flower	Root	Root nodule	1-cm pod	Pod shell 10 DAF	Pod shell 14 DAF	Seed 10 DAF	Seed 14 DAF	Seed 21 DAF	Seed 35 DAF
GmaCHI 1A	51	35	523	35	57	35	21	3	9	8	16
GmaCHI 1B1	7	5	105	19	1	1	0	0	1	4	15
GmaCHI 1B2	3	16	53	6	11	10	3	0	0	0	0
GmaCHI 2	34	85	1	1	25	17	13	7	11	15	2
GmaCHI 3A	12	6	1	0	11	10	3	1	2	8	8
GmaCHI 3B	11	5	0	0	7	10	4	1	0	2	1
GmaFAP 3A	18	10	98	13	17	10	9	20	23	27	37
GmaFAP 3B	4	8	30	7	5	4	4	13	10	15	7
Total	140	170	811	81	134	97	57	45	56	140	170

Data are the number of EST sequences matching a given DNA, with the following parameters: maximum identity >95%, length >200 bp and E value $<10^{-10}$.

The catalytic chemistry core and substrate binding play central roles during the evolution of new functions (Petsko et al., 1993; Babbitt and Gerlt, 1997; Zhang et al., 2010). The evolution of *CHI1* agrees with this theory. We have found that the positively selected codons of *CHI1* formed a bag surrounding the substrate (2S)-naringenin (Figure 2b), by the regions of β3a, β3c, α4, β3e, and α7. CHI catalyzes on the O^1 of (2S)-naringenin (Figure 2c, 2b), which is connected to 109K and 110V (Jez et al., 2000; Jez and Noel, 2002); interestingly the nearby site 108R is a positively selected site, indicating that it may modify the catalytic core of CHI1. In the tertiary structure, the α4 domain adjoins (2S)-naringenin, and it included 4 positively selected sites such as 108R, 117H, 109K, and 110V, as well as 122G (located nearby). Moreover, the side chains of some positively selected amino acids (39T, 95G, 161G, and 201L) point forward towards (2S)-naringenin. Thus, we infer that the positively selected codons, especially site 108R, play important roles in the function of CHI1. In addition, the evolution of CHI1 may depend on structural changes surrounding its substrate; during this process, the enzyme began to catalyze (2S)-naringenin rather than fatty acids. Indeed, (2S)-naringenin lies in the binding cleft of *MtrCHI1* in the catalytic reaction, and the function of *MtrCHI1* relies on the binding of the residue to the substrate (Jez et al., 2000; Jez and Noel, 2002).

ACKNOWLEDGMENTS

Research supported by the National Science Foundation of China (#31360339), National Key Laboratory Program of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University (#ZW2009005), Jiangxi Provincial Key Technology R&D Program (#20111BBF60005), Science Foundation of Jiangxi Province (#20114BAB204022), Foundation of Jiangxi Province Educational Committee (#GJJ12612), and Foundation of JiuJiang University (#2012KJ09).

REFERENCES

Babbitt PC and Gerlt JA (1997). Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* 272: 30591-30594.

Bednar RA and Hadcock JR (1988). Purification and characterization of chalcone isomerase from soybeans. *J. Biol. Chem.* 263: 9582-9588.

Casola C and Hahn MW (2009) Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J. Mol. Evol.* 68: 679-687.

Ferrer JL, Austin MB, Stewart C Jr and Noel JP (2008). Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. *Plant Physiol. Biochem.* 46: 356-370.

Guo AY, Zhu QH, Chen X and Luo JC (2007). GSDS: a gene structure display server. *Yi Chuan* 29: 1023-1026.

Hartmann T (2007). From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry* 68: 2831-2846.

Hegnauer R and Gpayer-Barkmeijer RJ (1993). Relevance of seed polysaccharides and flavonoids for the classification of the Leguminosae: a chemotaxonomic approach. *Phytochemistry* 34: 3-16.

Hur S, Newby ZE and Bruice TC (2004). Transition state stabilization by general acid catalysis, water expulsion, and enzyme reorganization in *Medicago savita* chalcone isomerase. *Proc. Natl. Acad. Sci. U S A* 101: 2730-2735.

Jez JM and Noel JP (2002). Reaction mechanism of chalcone isomerase: pH dependence, diffusion control, and product binding differences. *J. Biol. Chem.* 277: 1361-1369.

Jez JM, Bowman ME, Dixon RA and Noel JP (2000). Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase. *Nature Struct. Biol.* 7: 786-791.

Kosiol C, Vinar T, Fonseca RR, Hubisz MJ, et al. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* 4: e1000144.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, et al. (2008). Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* 24: 114-123.

Ngaki MN, Louie GV, Philippe RN, Manning G, et al. (2012). Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis. *Nature* 485: 530-533.

Petersen L, Bollback JP, Dimmic M, Hubisz M, et al. (2007). Genes under positive selection in *Escherichia coli*. *Genome Res.* 17: 1336-1343.

Petsko GA, Kenyon GL, Gerlt JA, Ringe D, et al. (1993). On the origin of enzymatic species. *Trends Biochem. Sci.* 18: 372-376.

Punta M, Coggill PC, Eberhardt RY, Mistry J, et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40: D290-D301.

Ralston L, Subramanian S, Matsuno M and Yu O (2005). Partial reconstruction of flavonoid and isoflavonoid biosynthesis in yeast using soybean type I and type II chalcone isomerases. *Plant Physiol.* 137: 1375-1388.

Ridout KE, Dixon CJ and Filatov DA (2010). Positive selection differs between protein secondary structure elements in *Drosophila*. *Genome Biol. Evol.* 2: 166-179.

Shimada N, Aoki T, Sato S, Nakamura Y, et al. (2003). A cluster of genes encodes the two types of chalcone isomerase involved in the biosynthesis of general flavonoids and legume-specific 5-deoxy(iso) flavonoids in *Lotus japonicas*. *Plant Physiol.* 3: 941-951.

Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599.

Verma D (1992). Signals in root nodule organogenesis and endocytosis of rhizobium. *Plant cell* 4: 373-382.

Yang ZH (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586-1591.

Zhang J, Yang H, Long M, Li L, et al. (2010). Evolution of enzymatic activities of testis-specific short-chain dehydrogenase/reductase in *Drosophila*. *J. Mol. Evol.* 71: 241-249.