JSE Journal of Systematics and Evolution

Journal of Systematics and Evolution 50 (6): 502–511 (2012) doi: 10.1111/j.1759-6831.2012.00211.x

Research Article

Evolutionary divergence of the APETALA1 and CAULIFLOWER proteins

[1,2]Bin WANG [1,3]Ning ZHANG [1,2]Chun-Ce GUO [1]Gui-Xia XU [1]Hong-Zhi KONG [1]Hong-Yan SHAN*

[1](*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences,* Beijing 100093, China)
[2](*Graduate University of Chinese Academy of Sciences,* Beijing 100049, China)
[3](*State Key Laboratory of Genetic Engineering, Institute of Plant Biology, Center for Evolutionary Biology, School of Life Sciences, Fudan University,* Shanghai 200433, China)

Abstract *APETALA1* (*AP1*) and *CAULIFLOWER* (*CAL*) are a pair of paralogous genes that were generated through the pre-Brassicaceae whole-genome duplication event. *AP1* and *CAL* have both partially redundant and unique functions. Previous studies have shown that the K and C regions of their proteins are essential for the functional divergence. However, which differences in these regions are the major contributors and how the differences were accumulated remain unknown. In the present study, we compared the sequences of the two proteins and identified five gaps and 55 amino acid replacements between them. Investigation of genomic sequences further indicated that the differences in the proteins were caused by non-synonymous substitutions and changes in exon–intron structures. Reconstruction of three-dimensional structures revealed that the sequence divergence of AP1 and CAL has resulted in differences between the two in terms of the number, length, position and orientation of α-helices, especially in the K and C regions. Comparisons of sequences and three-dimensional structures of ancestral proteins with AP1 and CAL suggest that the ancestral AP1 protein experienced fewer changes, whereas the ancestral CAL protein accumulated more changes shortly after gene duplication, relative to their common ancestor. Thereafter, AP1-like proteins experienced few mutations, whereas CAL-like proteins were not conserved until the diversification of the Brassicaceae lineage I. This indicates that AP1- and CAL-like proteins evolved asymmetrically after gene duplication. These findings provide new insights into the functional divergence of *AP1* and *CAL* genes.

Key words ancestral sequence inference, *APETALA1*, *CAULIFLOWER*, coding-sequence divergence, gene duplication, three-dimensional structure.

Gene duplication provides raw genetic material for biological evolution and phenotypic innovation (Ohno, 1970). Studies on the divergence of duplicate genes are of great importance for understanding the mechanisms underlying organismal evolution. Divergence between duplicate genes can occur in coding regions or regulatory regions. Gain, loss, or mutation of transcription factor-binding sites can result in divergence in both regulatory regions and expression patterns (Papp et al., 2003; Li et al., 2005; Ganko et al., 2007). Divergences in coding regions can be achieved by non-synonymous substitutions and changes in exon–intron structures that can occur as a result of exon/intron gain/loss, exonization/pseudoexonization, and insertion/deletion (Zhang et al., 2002; Vandenbussche et al., 2003; Xu et al., 2012). The evolutionary mechanisms underlying duplicate gene divergence in coding regions can be revealed by comparing the ancestral sequences estimated by several methods with extant gene sequences (Zhang et al., 1998; Thornton, 2001; Thornton et al., 2003; Zhang, 2006). Thus, studies on coding sequence divergence of duplicate genes are not only easy to perform, but can also provide us with more information on important evolutionary steps that have led to the functional divergence of duplicate genes.

The *Arabidopsis thaliana APETALA1* (*AP1*) and *CAULIFLOWER* (*CAL*) genes are famous paralogous genes that were created by a gene duplication event prior to the origin of the Brassicaceae (Lawton-Rauh et al., 1999). Both play important roles in the floral regulatory network (Liu et al., 2011). Previous studies have indicated that *AP1* and *CAL* have both partially redundant and unique functions. *In situ* hybridization results

Received: 15 May 2012 Accepted: 3 June 2012
* Author for correspondence. E-mail: shanhongyan@ibcas.ac.cn. Tel.: 86-10-62836736. Fax: 86-10-62590843.

show that both genes are expressed in floral meristems, but in sepals and petals from Stage 4 at different intensities, in addition to the specific expression of *CAL* in vascular bundles of inflorescence stems (Mandel et al., 1992; Kempin et al., 1995). The phenotypes of *ap1* and *cal* single mutants, as well as *ap1 cal* double mutants, suggest that *AP1* plays key roles in the formation of floral meristems and the specification of sepal and petal identities, whereas *CAL* functions only as a floral meristem identity gene and is a positive regulator of *AP1* (Bowman et al., 1993). In the past decade, much has been learned about the mechanisms underlying essential divergence in the coding regions between the two genes. For example, molecular evolutionary studies have shown that the evolutionary rate of *CAL* was approximately twofold greater than that of *AP1*, suggesting that *CAL* has evolved under relaxed selection after gene duplication (Lawton-Rauh et al., 1999; Liu et al., 2011). Domain-swapping experiments have shown that the K domain of AP1 is important for the formation of floral meristem identity, both the K and C regions are essential for the establishment of sepal identity, and either the K domain or the C region is indispensable for the formation of petal identity (Alvarez-Buylla et al., 2006). These findings emphasize the importance of the K and C regions to the functional divergence of AP1 and CAL (Alvarez-Buylla et al., 2006). However, which differences in these regions are the major contributors and how the differences were accumulated remain unclear.

To reveal the mechanisms underlying the divergence of AP1 and CAL proteins, we carefully compared their sequences and predicted three-dimensional (3D) structures, and traced their evolutionary processes since gene duplication. We identified five gaps and 55 amino acid replacements in the alignment of AP1 and CAL. Investigation of their genomic sequences further indicated that both non-synonymous substitutions and changes in exon–intron structures, most of which occurred in the I, K, and C regions, have contributed to their sequence divergence. The differences in proteins have resulted in the divergence of AP1 and CAL in terms of their 3D structures, including the number, length, position, and orientation of α-helices. Step-by-step comparisons of ancestral amino acid sequences and 3D structures with AP1 and CAL suggest that the CAL-like proteins have accumulated more mutations than the AP1-like proteins not only before the diversification of the Brassicaceae, but also prior to the origin of the Brassicaceae lineage I, suggestive of an asymmetric evolutionary pattern.

1 Material and methods

1.1 Plant material and gene isolation

To explore the evolutionary divergence of the AP1 and CAL proteins, we sampled six Brassicaceae species (*Arabidopsis thaliana*, *A. lyrata*, *Capsella rubella*, *Lepidium apetalum*, *Rorippa indica*, and *Thellungiella parvula*) and three non-Brassicaceae species from Brassicales (*Cleome spinosa* of Cleomaceae, *Carica papaya* of Caricaceae, and *Tropaeolum majus* of Tropaeolaceae). This enabled us to infer the evolutionary processes of AP1 and CAL via step-by-step comparisons. The sequences of *AP1*- and *CAL*-like genes of *A. thaliana*, *A. lyrata*, *C. rubella*, *T. parvula*, and *C. papaya* were retrieved from publicly available databases by blast searches (see Table S1 available as Supplementary Material for this paper). Others were isolated from young floral buds and inflorescences by 3′ rapid amplification of cDNA ends (RACE). Tissues of *L. apetalum*, *R. indica*, *C. spinosa*, and *T. majus* were collected from the Botanical Garden of the Institute of Botany, Chinese Academy of Sciences, or the Beijing Botanical Garden, China.

Extraction of total RNA, purification of mRNA, and synthesis of first-strand cDNA were performed as described previously (Shan et al., 2007). For isolation of *AP1* and *CAL* homologs from sampled species, two rounds of polymerase chain reaction (PCR) amplification were performed under the following conditions: initiation denaturalization at 94 °C for 4 min, followed by 35 cycles of 94 °C for 30 s, 52 °C for 30 s, and 72 °C for 1 min, with a final extension at 72 °C for 10 min. All primers used in the present study are listed in Table S2, and the primer combinations for each round of PCR amplification are given in Table S3. The PCR products of expected length were gel purified using an AxyPrep DNA Gel Extraction Kit (Axygen, Union City, CA, USA) and cloned into the pGEM-T Easy Vector (Promega, Madison, WI, USA). For each transformation, at least 10 clones were sequenced.

To obtain the exon–intron structure of *C. spinosa CspAP1*, genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method and then amplified using the primer combinations listed in Table S3. The PCR conditions were as follows: initiation denaturalization at 94 °C for 5 min, followed by 40 cycles of 94 °C for 30 s, 52 °C for 30 s, and 72 °C for 3 min, with a final extension at 72 °C for 10 min. The amplified fragments were then cloned and sequenced as described above.

1.2 Sequence alignment and phylogenetic analysis

All putative protein sequences coded by *AP1*- and *CAL*-like genes were aligned with CLUSTALX 1.83 (Thompson et al., 1997) and then adjusted manually using GeneDoc (Nicholas et al., 1997). Finally, two protein matrices were generated, named "Bra_AP1_P", and "Bra_AP1_P_good". "Bra_AP1_P" included all sites, whereas "Bra_AP1_P_good" included residues with >12 quality scores that were calculated by CLUSTALX 1.83 (Thompson et al., 1997). The corresponding DNA matrices were generated using aa2dna software (https://homes.bio.psu.edu/people/faculty/nei/software.htm, accessed 15 May 2012) and named "Bra_AP1_D", and "Bra_AP1_D_good", respectively. Maximum likelihood (ML) estimates of phylogenetic relationships were performed with DNA matrices in PhyML version 2.4.3 (Guindon & Gascuel, 2003). The general time reversible (GTR) substitution model was used, with optimization of the proportion of invariable sites, nucleotide frequencies and the gamma shape parameter. Bootstrap analyses were performed for 1000 replicates.

1.3 Ancestral sequence inference

To trace the evolutionary processes of the AP1 and CAL proteins since gene duplication, we inferred the ancestral sequences of all the interior nodes in the phylogenetic tree. Considering that the species relationships within *AP1* and *CAL* lineages in our phylogenetic trees were not perfectly consistent with the species tree of the Brassicaceae (Fig. S1), which may affect the reliability of ancestral sequence inference, a constraint tree was used (Fig. 2:A). Here, the phylogenetic relationships among the Brassicaceae species were determined based on the recently published phylogeny of the Brassicaceae (Couvreur et al., 2010). Sequences from *C. spinosa* of Cleomaceae, *C. papaya* of Caricaceae, and *T. majus* of Tropaeolaceae were used as outgroups. Amino acid sequences at all interior nodes were inferred by using the distance-based and likelihood-based Bayesian methods, which were performed using the ANCESTOR software (Zhang & Nei, 1997) and the CODEML program in PAML 4.3 (Yang, 2007), respectively.

1.4 Protein structure prediction

To determine which differences at the protein level have caused changes in structures, we performed 3D structure modeling using the I-TASSER method (http://zhanglab.ccmb.med.umich.edu/I-TASSER/, accessed 15 May 2012; Zhang, 2008) for AP1 and CAL, as well as for all the ancestral proteins. We first excluded the models that were inconsistent with human myocyte-specific enhancer factor 2A (MEF2A) in the MADS domain, because its crystal structure has been determined (Protein Data Bank ID: 1EGW; Huang et al., 2000; Santelli & Richmond, 2000). We then compared the C-score values of the rest of the protein models and selected those with the highest scores for further analyses. The C-score value is a confidence score for estimating the qualities of predicted models by I-TASSER. Further comparisons of the selected structures were performed using SPDBV v.4.0.1 software (http://www.expasy.ch/spdbv/, accessed 15 May 2012; Guex & Peitsch, 1997). Similarities between the structures were evaluated by Z-score values using DaliLite (http://www.ebi.ac.uk/Tools/dalilite/index.html, accessed 15 May 2012; Holm & Park, 2000). The higher the Z-score value, the more similar the structures are. As a general rule, a Z-score >20 means that the two structures are definitely homologous, whereas scores between 8 and 20 mean that the two are probably homologous, scores between 2 and 8 are inconclusive, and scores <2 mean that the two are not significantly similar (Holm & Park, 2000).

2 Results

2.1 Sequence differences of the AP1 and CAL proteins in *Arabidopsis thaliana*

The *AP1* and *CAL* genes encode 256 and 255 amino acids, respectively. Pairwise sequence alignment of their proteins indicated that there were 55 sites showing amino acid differences, including seven in the MADS domain, three in the I region, 16 in the K domain, and 29 in the C region (Fig. 1: A). Based on the amino acid properties, we divided them into four types: (i) differences between hydrophobic and hydrophilic amino acids (Type I); (ii) differences between uncharged and charged amino acids (Type II); (iii) differences between negatively and positively charged amino acids (Type III); and (iv) differences that do not result in differences in amino acid properties (Type IV). Generally, the first three types of differences can cause changes in amino acid properties. According to this criterion, of the 55 sites showing amino acid differences, 18 (32.7%) were considered Type I, 10 (18.2%) were considered Type II, two (3.6%) were considered Type III, and 25 (45.5%) were considered Type IV (Fig. 1: A). In addition, five gaps, including two in the I region and three in the C region, were observed (Fig. 1: A).

To determine whether the five gaps were created by intra-exonic insertion/deletion or exonization/pseudoexonization at the DNA level, we further compared the genomic sequences of *AP1* and *CAL*. Both have eight exons and seven introns, but the length of the

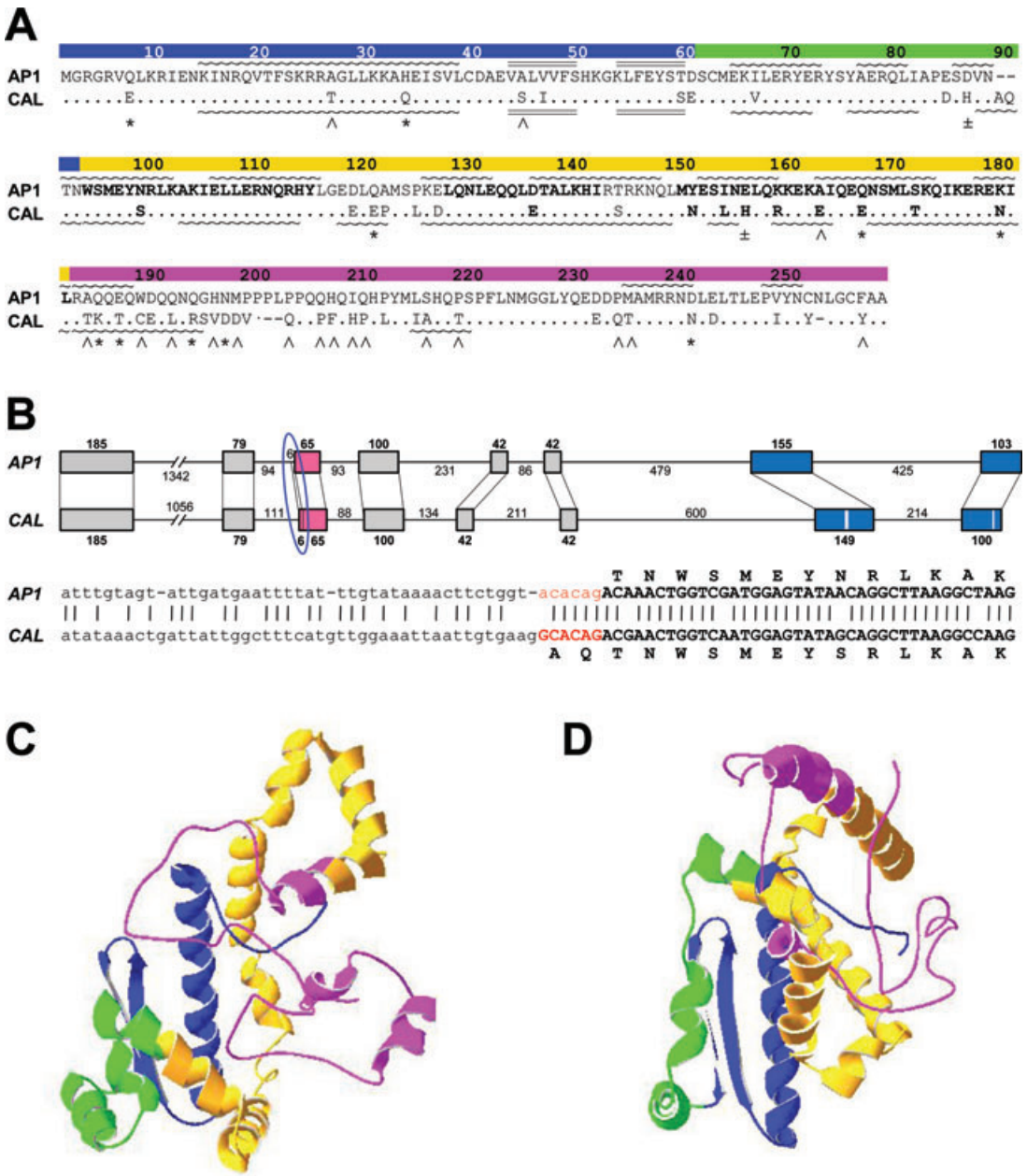


Fig. 1. Comparisons of protein sequences and three-dimensional (3D) structures of AP1 and CAL, and their exon-intron structures at the genomic level. **A,** Pairwise alignment of AP1 and CAL. The MADS domain, I region, K domain, and C region are shown in blue, green, yellow and pink, respectively. The K1, K2, and K3 subdomains in the K domain, defined according to Yang & Jack (2004), are highlighted in bold. Amino acids with Type I, II, and III differences are indicated by the $\wedge$, $*$, and $\pm$ symbols, respectively. The α-helices predicted in the 3D structures are denoted by curved lines, whereas the β-sheets are indicated by double lines. **B,** Schematic representation of divergence of exon–intron structures. Exons that have experienced exonization/pseudoexonization are highlighted in pink, whereas insertions/deletions are in blue and exons that do not differ structurally are shown in gray. The numbers represent the lengths of the exons and introns, which are largely proportional to the real lengths. Regions (especially exons) that match each other are connected with thin lines. Intra-exonic insertions/deletions are indicated by white bars. The region encircled by the oval indicates the position where exonization/pseudoexonization occurred, with the detailed genomic alignment for this region is given below. Uppercase letters denote exon sequences, whereas lowercase letters denote intron sequences. Vertical lines indicate identical nucleotides between the two sequences. Amino acid sequences are given above and below the exons. The nucleotides involved in exonization/pseudoexonization are highlighted in red. **C,** The 3D structure of AP1. The four regions are distinguished by different colors as shown for **A**. **D,** The 3D structure of CAL.

third, seventh, and eighth exons differs (Fig. 1: B). Genomic sequence comparison suggested that the two gaps at sites 200 and 201 in the protein alignment were located in the seventh exon, whereas the gap at site 252 was in the eighth exon, and all were caused by intra-exonic insertions/deletions (Fig. 1: B). In contrast, the gaps at sites 89 and 90 in the protein alignment were at the beginning of the third exon of *AP1* and *CAL* (Fig. 1: B). Interestingly, the last six nucleotides of the second intron of *AP1* (acacag) matched very well with the first six nucleotides of the third exon of *CAL* (GCACAG). This indicated an exonization/pseudoexonization event that was actually caused by the selection of a different splice acceptor site "AG" (Fig. 1: B).

2.2 Structural differences of the AP1 and CAL proteins in *Arabidopsis thaliana*

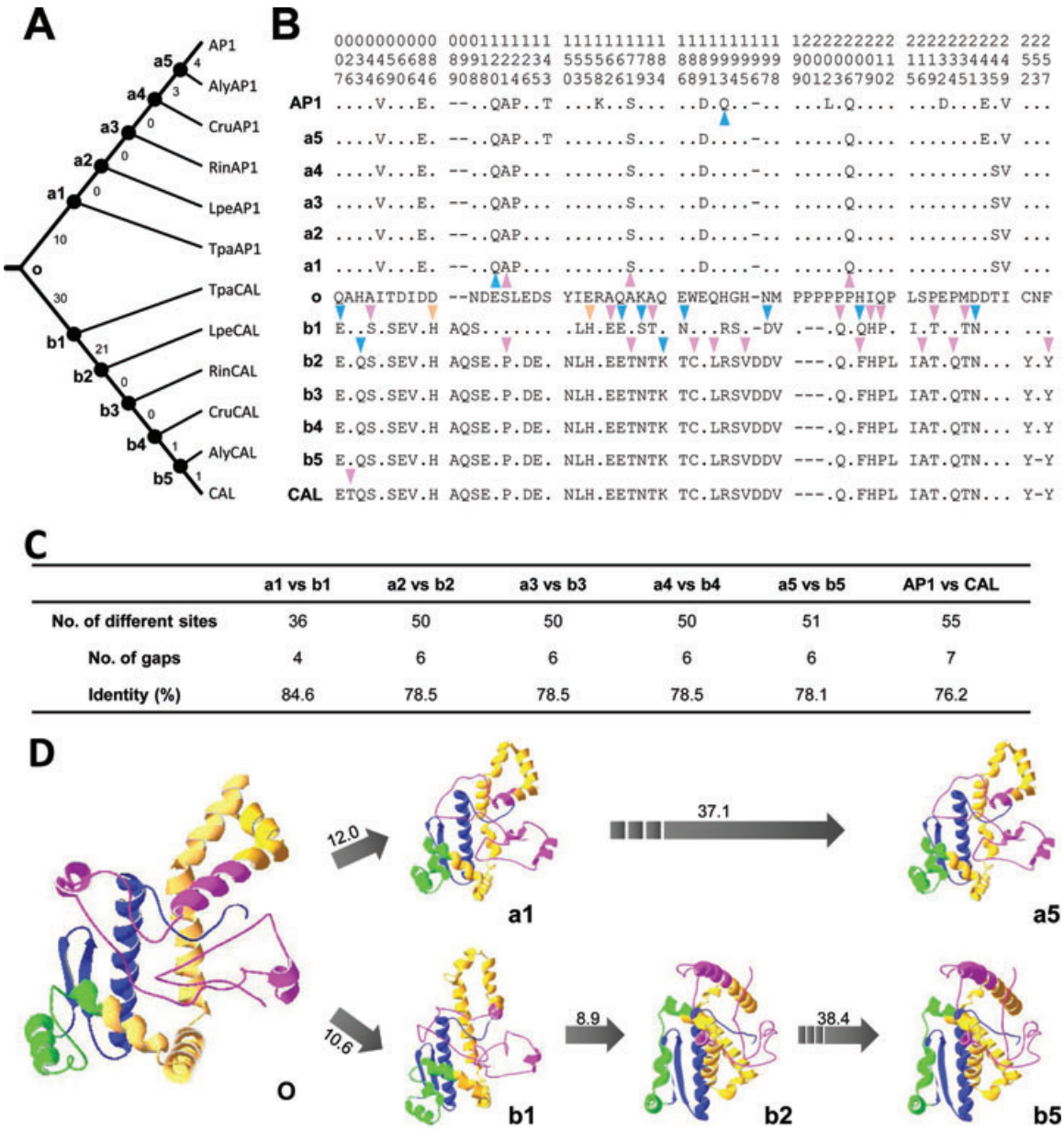
To determine whether the divergence of AP1 and CAL could lead to differences in their 3D structure, we performed structure modeling of AP1 and CAL using the I-TASSER server (Zhang, 2008). The Z-score value was 7.9, suggesting significant divergence of the two protein structures. By comparing their structures, we found that they shared the same conformation in the MADS domain, including one α-helix (sites 14–38) and two β-sheets (sites 43–49 and 53–59; Fig. 1: C, D; Table S4). In the I region, they possessed three α-helices, the position and orientation of which were all obviously different (Fig. 1: C, D; Table S4). In the K domain, both had seven α-helices. Their first two covered the K1 subdomain (93–114), the third and fourth of AP1 and the fourth of CAL covered the K2 subdomain (127–141), and the fifth–seventh nearly covered the K3 subdomain (149–181). The third helix of CAL was specific to itself (Fig. 1: A, D). Despite the similarities in the number and position of most helices, the topologies have diverged dramatically due to differences in the orientations of the helices (Fig. 1: C, D). There were three and two α-helices in the C regions of AP1 and CAL, respectively. Although the first one of CAL corresponded to that of AP1 in terms of position, the length and orientation of the α-helices showed obvious divergence (Fig. 1: C, D, Table S4). The second helix and the last two helices were specific to AP1 and CAL, respectively (Fig. 1: C, D). Thus, the conformations of AP1 and CAL have diverged in terms of the length, position, number, and orientation of the α-helices in the I, K, and C regions.

2.3 Sequence evolution of the AP1 and CAL proteins

To investigate the evolutionary processes of AP1 and CAL, we first performed phylogenetic analysis using sequences from the present study and publicly available databases (Table S1). Both phylogenetic trees reconstructed with different matrices indicated that the *AP1*- and *CAL*-like genes from the Brassicaceae were grouped into two clades with high bootstrap support. The *AP1* homologous genes from outgroup species were at the base of the Brassicaceae *AP1* and *CAL* lineages (Fig. S1). This implies that the gene duplication event that produced *AP1*- and *CAL*-like genes happened before the origin of the Brassicaceae, but after the divergence of the Brassicaceae and the Cleomaceae. We subsequently inferred the ancestral protein sequences of all interior nodes in the constraint gene tree with distance-based and likelihood-based Bayesian methods. For the 11 interior nodes, the average accuracies and posterior probabilities of the ancestral amino acid inference were in the range 0.961–0.980 for the distance-based method and 0.971–0.998 for the likelihood-based method, respectively. The inferred ancestral proteins are shown in Fig. S2.

By comparing the ancestral protein sequences, we found that amino acid changes have occurred at 10 and 26 sites of the ancestral AP1 (Node a1) and CAL (Node b1) proteins, respectively, relative to their common ancestor (Node o). In addition to amino acid replacements, Node b1 has gained alanine (A) and glutamine (Q) at sites 89 and 90, but has lost two prolines (P) at sites 200 and 201 (Fig. 2: B). These differences led to an 84.6% protein identity between Nodes a1 and b1 (Fig. 2: C). In particular, we found that the amino acid properties have changed at some sites during evolution (Fig. 2: B). For example, from Nodes o to a1, three Type I changes and one Type II change were observed (Fig. 2: B), whereas from Nodes o to b1, eight Type I changes, seven Type II changes, and two Type III changes were found (Fig. 2: B).

During the following evolution, Node b2 accumulated 19 additional amino acid replacements, including nine Type I changes, two Type II changes, and eight Type IV changes (Fig. 2: B), whereas Node a2 remained unchanged (Fig. 2: B). As a result, the identity of these two proteins has decreased to 78.5% (Fig. 2: C). Then, the AP1- and CAL-like proteins were unchanged until Nodes a5 and b5 (Fig. 2: B). From Node a4 to Node a5, only three Type IV amino acid changes were found, but a reverse mutation occurred at site 245 (Fig. 2: B). From Node b4 to Node b5, a loss of asparagine (N) at site 253 was detected. These changes have led to 78.1% protein identity between Nodes a5 and b5 (Fig. 2: C). From Node a5 to AP1, four amino acid replacements, including one Type II change and three Type IV changes, were observed, whereas from Node b5 to CAL, only one Type I amino acid substitution was detected. These new changes further decreased protein identity



	a1 vs b1	a2 vs b2	a3 vs b3	a4 vs b4	a5 vs b5	AP1 vs CAL
No. of different sites	36	50	50	50	51	55
No. of gaps	4	6	6	6	6	7
Identity (%)	84.6	78.5	78.5	78.5	78.1	76.2

Fig. 2. Comparisons of protein sequences and three-dimensional (3D) structures of ancestral sequences with AP1 and CAL. **A,** The constraint tree of the AP1- and CAL-like proteins. The ancestral nodes are indicated with dots. The gene duplication event occurred at Node o. The total number of differences, including amino acid replacements and gaps, between nodes are shown beside each branch. **B,** Amino acid residues showing differences in the alignment of ancestral and present-day AP1- and CAL-like proteins. The residue positions in the alignment are shown at the top. The amino acids are represented by single-letter codes. Dots show the same amino acids as those of the sequences at Node o and dashes indicate gaps. Pink, blue, and orange triangles represent Type I, II, and III amino acid changes, respectively. **C,** Comparisons of protein sequences of ancestral paralogs with those of AP1 and CAL. **D,** The evolutionary processes of 3D structures of AP1- and CAL-like proteins. The *Z*-score values between proteins are shown beside the arrows.

between AP1 and CAL to 76.2% in *A. thaliana* (Fig. 2: C). Together, these results indicate that the ancestral AP1 protein underwent fewer mutations, whereas the ancestral CAL protein accumulated more shortly after gene duplication. Thereafter, the AP1-like proteins experienced few changes, whereas the CAL-like proteins were not conserved until the diversification of the Brassicaceae lineage I.

As mentioned above, the amino acids A and Q at sites 89 and 90 of CAL were actually generated by an exonization/pseudoexonization event. Furthermore, by comparing the ancestral sequences of Nodes o, a1 and

b1, we found that the CAL-like proteins have gained A and Q shortly after gene duplication. To determine whether the gain of A and Q in the ancestral CAL protein was also caused by exonization/pseudoexonization, we compared the genomic sequences of the *AP1*- and *CAL*-like genes from *A. thaliana*, *A. lyrata*, *Capsella rubella*, and *Thellungiella parvula* with the outgroups *CspAP1* from *C. spinosa* and *CpaAP1* from *C. papaya*. The results indicate that A and Q at sites 89 and 90 of the CAL-like proteins were caused by exonizations/pseudoexonizations predating the diversification of the Brassicaceae (Fig. S3).

2.4 Structural evolution of the AP1 and CAL proteins

To understand the evolutionary processes of protein structures of AP1 and CAL, we modeled 3D structures of ancestral proteins with the I-TASSER server (Zhang, 2008). Before gene duplication, the ancestor of AP1- and CAL-like proteins (Node o) had one α-helix and two β-sheets in the MADS domain, two α-helices in the I region, six α-helices in the K domain, and three α-helices in the C region (Fig. 2: D; Table S4). In contrast with Node o, the ancestral AP1 protein (Node a1) gained one α-helix in the I region and one in the C region, and lost one α-helix in the C region. In the K domain of Node a1, the total number of α-helices was seven, but the third and fourth helices corresponded to the third helix of Node o. For the remaining 12 comparable α-helices and β-sheets, 10 have changed in length (Fig. 2: D; Table S4). In the case of the ancestral CAL protein (Node b1), it gained one α-helix in the I region and lost one in the C region (Table S4). Of the 13 comparable α-helices and β-sheets, seven showed different lengths from Node o (Fig. 2: D; Table S4). To evaluate the divergent degree of the structures of Nodes a1 and b1 relative to Node o, we calculated the Z-score values of Nodes o and a1, as well as Nodes o and b1. Our results indicate that the overall topology of Node a1 was more similar to that of Node o than Node b1, as evidenced from the Z-score values (12.0 for Node o vs. Node a1; 10.6 for Node o vs. Node b1; Fig. 2: D).

During the following evolution, Node b2 independently gained one α-helix and lost one in the C region, with the number of α-helices increasing to eight in the K domain (Table S4). The big divergence of Nodes b1 and b2 was also confirmed by the lower Z-score value (8.9) between them (Fig. 2: D). Thereafter, the topologies of the 3D structures of the remaining nodes did not change, although the number of α-helices in the K domain differed between ancestral CAL-like proteins and the present-day CAL protein (Fig. 2: D; Table S4). In contrast with the CAL-like proteins, the structures of the ancestral AP1 protein were maintained nearly conserved during subsequent evolution (Fig. 2: D; Table S4). These results suggest that the evolutionary patterns of the 3D structures of the AP1 and CAL proteins were essentially consistent with those of their protein sequences.

3 Discussion

3.1 Divergent mechanisms of the AP1 and CAL proteins and their functional implications

In recent decades, considerable progress has been made in exploring the patterns, mechanisms, and consequences of the coding-sequence divergence of duplicate genes (Lynch & Conery, 2000; Moore & Purugganan, 2005; Innan & Kondrashov, 2010; Xu et al., 2012). In particular, it has been found that non-synonymous substitutions that could lead to changes in amino acid properties, as well as changes in exon–intron structures, generally resulted in the distinct divergence of protein sequences at key sites, protein length and domain organization (Hanzawa et al., 2005; Xu et al., 2012). In the present study, we found that non-synonymous substitutions and changes in exon–intron structures that were produced by intra-exonic insertions/deletions and exonizations/pseudoexonizations have contributed to the divergence of *AP1* and *CAL*. Furthermore, the relative contributions were by no means the same. The marked sequence divergence at the DNA level has led to 76.2% identity between the AP1 and CAL proteins, suggestive of potential functional divergence between the two.

A previous study has shown that the functional divergence of *AP1* and *CAL* was caused by differences in their coding regions, especially the K and C regions (Alvarez-Buylla et al., 2006). By searching the BioGRID database (Stark et al., 2006), we found that AP1 could form a homodimer with itself and heterodimers with another 21 proteins, whereas CAL could only form heterodimers with five proteins. It has been proposed that the K domain of the MADS-box proteins is responsible for protein–protein interactions by forming amphipathic α-helices, and the C region is required for forming higher-order protein complexes (Cho et al., 1999; Egea-Cortines et al., 1999; Yang et al., 2003). By comparing the K and C regions of AP1 and CAL, we observed five and 20 sites showing differences in amino acid properties, respectively. Moreover, three intra-exonic insertions/deletions in the C regions have led to the divergence in protein length. These differences altogether have resulted in significant divergence in the number, length, position, and orientation of the

α-helices in the 3D structure. Therefore, it is highly likely that the distinct structures of AP1 and CAL have led to their different protein–protein interaction capabilities, and even functions.

3.2 Asymmetric evolution of AP1 and CAL proteins

In the regulatory network for floral development, *AP1* and *CAL*, *SHP1* (*SHATTERPROOF 1*) and *SHP2* (*SHATTERPROOF 2*), as well as *SEP1* (*SEPALLATA1*) and *SEP2* (*SEPALLATA2*), were generated by the gene duplication event before the origin of the Brassicaceae. However, only *AP1* and *CAL* have both partially redundant and unique functions, whereas the other two gene pairs play redundant roles in the floral development of *Arabidopsis thaliana* (Pelaz et al., 2000; Pinyopich et al., 2003). Previous work indicates that *AP1*, *SHP1*, *SHP2*, *SEP1*, and *SEP2* evolved under strong purifying selection, whereas *CAL* was subject to relaxed purifying selection (Lawton-Rauh et al., 1999; Liu et al., 2011). In the present study, we found that the CAL-like proteins have accumulated more mutations than the AP1-like proteins not only before the diversification of the Brassicaceae, but also prior to the origin of the Brassicaceae lineage I. Overall, the AP1-like proteins evolved conservatively, but the CAL-like proteins showed significant differences from their ancestral proteins. Our results, in combination with the functional and molecular evolutionary data, suggest that *AP1* and *CAL* evolved in an asymmetric pattern after gene duplication.

A considerable number of studies has demonstrated the prevalence of asymmetric divergence between duplicate genes, which could have been caused by changes in the regulatory and/or coding regions (Conant & Wagner, 2003; Braasch et al., 2006; Zou et al., 2009). The asymmetric divergence of duplicate genes in coding regions could be reflected in sequences, evolutionary rates, 3D structures, protein–protein interactions, and functions, as we have seen for *AP1* and *CAL*, as well as other duplicate genes (Wagner, 2002; Yang et al., 2005; Lin et al., 2010). Compared with their contemporaneous duplicated gene pairs (*SHP1*/*SHP2* and *SEP1*/*SEP2*), the asymmetric evolutionary pattern of *AP1* and *CAL* seems to be coupled with their functions. In the present study, we have found that the *AP1*-like genes resemble the preduplicated ancestor in terms of protein sequence, 3D structure, and possibly function. If *AP1* and *CAL* did not diverge after gene duplication, the dosage of their proteins will be very high. The plants may flower very early and show abnormal sepals and petals, as the phenotype in transgenic plants that overexpress *AP1* (Liljegren et al., 1999). Early flowering may be detrimental to the plant under normal conditions and, finally, it may lead to a decline in fitness. Therefore, to avoid early flowering, the ancestors of *AP1*- and *CAL*-like genes diverged in a short evolutionary time after gene duplication in an asymmetric pattern by accumulating more mutations via different mechanisms in the *CAL* lineage. As a result, *AP1* plays major roles in the formation of floral meristems, sepals, and petals, whereas *CAL* regulates *AP1* and plays minor roles in determining floral meristem identity.

In the present study, we revealed the mechanisms underlying the divergence of the AP1 and CAL proteins, and their evolutionary processes and patterns, providing new insights into the functional divergence of *AP1* and *CAL*. However, as mentioned in the Introduction, *AP1* and *CAL* have diverged in terms of both expression patterns and functions, which implies that the regulatory regions may also be involved in their divergence, in addition to the coding regions. Therefore, to comprehensively understand the mechanisms underlying the divergence of *AP1* and *CAL*, their regulatory regions also need to be studied.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Grant Nos. 30800065, 30970210, and 31070202).

References

Alvarez-Buylla ER, Garcia-Ponce B, Garay-Arroyo A. 2006. Unique and redundant functional domains of *APETALA1* and *CAULIFLOWER*, two recently duplicated *Arabidopsis thaliana* floral MADS-box genes. Journal of Experimental Botany 57: 3099–3107.

Bowman JL, Alvarez J, Weigel D, Meyerowitz EM, Smyth DR. 1993. Control of flower development in *Arabidopsis thaliana* by *APETALA1* and interacting genes. Development 119: 721–743.

Braasch I, Salzburger W, Meyer A. 2006. Asymmetric evolution in two fish-specifically duplicated receptor tyrosine kinase paralogons involved in teleost coloration. Molecular Biology and Evolution 23: 1192–1202.

Cho S, Jang S, Chae S, Chung KM, Moon YH, An G, Jang SK. 1999. Analysis of the C-terminal region of *Arabidopsis thaliana APETALA1* as a transcription activation domain. Plant Molecular Biology 40: 419–429.

Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. Genome Research 13: 2052–2058.

Couvreur TL, Franzke A, Al-Shehbaz IA, Bakker FT, Koch MA, Mummenhoff K. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Molecular Biology and Evolution 27: 55–71.

Egea-Cortines M, Saedler H, Sommer H. 1999. Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in *Antirrhinum majus*. The EMBO Journal 18: 5370–5379.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. Molecular Biology and Evolution 24: 2298–2309.

Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18: 2714–2723.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.

Hanzawa Y, Money T, Bradley D. 2005. A single amino acid converts a repressor to an activator of flowering. Proceedings of the National Academy of Sciences USA 102: 7748–7753.

Holm L, Park J. 2000. DaliLite workbench for protein structure comparison. Bioinformatics 16: 566–567.

Huang K, Louis JM, Donaldson L, Lim FL, Sharrocks AD, Clore GM. 2000. Solution structure of the MEF2A-DNA complex: Structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. The EMBO Journal 19: 2615–2628.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: Classifying and distinguishing between models. Nature Reviews Genetics 11: 97–108.

Kempin SA, Savidge B, Yanofsky MF. 1995. Molecular basis of the *cauliflower* phenotype in *Arabidopsis*. Science 267: 522–525.

Lawton-Rauh AL, Buckler ESt, Purugganan MD. 1999. Patterns of molecular evolution among paralogous floral homeotic genes. Molecular Biology and Evolution 16: 1037–1045.

Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. Trends in Genetics 21: 602–607.

Liljegren SJ, Gustafson-Brown C, Pinyopich A, Ditta GS, Yanofsky MF. 1999. Interactions among *APETALA1*, *LEAFY*, and *TERMINAL FLOWER1* specify meristem fate. The Plant Cell 11: 1007–1018.

Lin JY, Stupar RM, Hans C, Hyten DL, Jackson SA. 2010. Structural and functional divergence of a 1-Mb duplicated region in the soybean (*Glycine max*) genome and comparison to an orthologous region from *Phaseolus vulgaris*. The Plant Cell 22: 2545–2561.

Liu Y, Guo C, Xu G, Shan H, Kong H. 2011. Evolutionary pattern of the regulatory network for flower development: Insights gained from a comparison of two *Arabidopsis* species. Journal of Systematics and Evolution 49: 528–538.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

Mandel MA, Gustafson-Brown C, Savidge B, Yanofsky MF. 1992. Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1*. Nature 360: 273–277.

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. Current Opinion in Plant Biology 8: 122–128.

Nicholas KB, Nicholas HB Jr, Deerfield DW II. 1997. GeneDoc: Analysis and visualization of genetic variation. EMBnet. NEWS 4: 1–4.

Ohno S. 1970. Evolution by gene duplication. New York: Springer.

Papp B, Pal C, Hurst LD. 2003. Evolution of *cis*-regulatory elements in duplicated genes of yeast. Trends in Genetics 19: 417–422.

Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF. 2000. B and C floral organ identity functions require *SEPALLATA* MADS-box genes. Nature 405: 200–203.

Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, Wisman E, Yanofsky MF. 2003. Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature 424: 85–88.

Santelli E, Richmond TJ. 2000. Crystal structure of MEF2A core bound to DNA at 1.5 Å resolution. Journal of Molecular Biology 297: 437–449.

Shan H, Zhang N, Liu C, Xu G, Zhang J, Chen Z, Kong H. 2007. Patterns of gene duplication and functional diversification during the evolution of the *AP1*/*SQUA* subfamily of plant MADS-box genes. Molecular Phylogenetics and Evolution 44: 26–41.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: A general repository for interaction datasets. Nucleic Acids Research 34: D535–D539.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25: 4876–4882.

Thornton JW. 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. Proceedings of the National Academy of Sciences USA 98: 5671–5676.

Thornton JW, Need E, Crews D. 2003. Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signaling. Science 301: 1714–1717.

Vandenbussche M, Theissen G, Van de Peer Y, Gerats T. 2003. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. Nucleic Acids Research 31: 4401–4409.

Wagner A. 2002. Asymmetric functional divergence of duplicate genes in yeast. Molecular Biology and Evolution 19: 1760–1768.

Xu G, Guo C, Shan H, Kong H. 2012. Divergence of duplicate genes in exon–intron structure. Proceedings of the National Academy of Sciences USA 109: 1187–1192.

Yang J, Xie Z, Glover BJ. 2005. Asymmetric evolution of duplicate genes encoding the CCAAT-binding factor NF-Y in plant genomes. The New Phytologist 165: 623–631.

Yang Y, Jack T. 2004. Defining subdomains of the K domain important for protein–protein interactions of plant MADS proteins. Plant Molecular Biology 55: 45–59.

Yang Y, Fanning L, Jack T. 2003. The K domain mediates heterodimerization of the *Arabidopsis* floral organ identity proteins, APETALA3 and PISTILLATA. The Plant Journal 33: 47–59.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586–1591.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nature Genetics 38: 819–823.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. Journal of Molecular Evolution 44: S139–S146.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proceedings of the National Academy of Sciences USA 95: 3708–3713.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nature Genetics 30: 411–415.
Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9: 40.
Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. 2009. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. PLoS Genetics 5: e1000581.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/j.1759-6831.2012.00211.x/suppinfo:

Table S1. *AP1* and *CAL* homologs included in the present study.
Table S2. Primers used in the present study.
Table S3. Primer combinations for the isolation of *AP1* and *CAL* homologs.
Table S4. Information for the α-helices and β-sheets in the predicted three-dimensional structures of ancestral and present-day AP1- and CAL-like proteins.
Fig. S1. Maximum-likelihood trees of the *AP1*- and *CAL*-like genes, which were reconstructed based on the nucleotide matrices ‘‘Bra_AP1_D_good’’ (**A**) and “Bra_AP1_D” (**B**), respectively. Higher than 50% bootstrap supports are indicated for each node.
Fig. S2. Sequence alignment of ancestral and present-day AP1- and CAL-like proteins. Dots show the same amino acids as those of the ancestral protein at Node o and gaps are indicated by dashes.
Fig. S3. Evolution of exon–intron structures of the *AP1*- and *CAL*-like genes. The tree on the left illustrates the phylogenetic relationships of the genes. The star indicates the gene duplication event. Exons that have experienced exonization/pseudoexonization are highlighted in pink and insertions/deletions are blue; those without structural differences are gray. The regions that have experienced exonization/pseudoexonization are shaded in yellow, suggesting that the exonization/pseudoexonization event happened before the origin of the Brassicaceae.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.