Plant Molecular Biology (2005) 59:485–500
DOI 10.1007/s11103-005-0353-0

Discovery of novel conserved peptide domains by ortholog comparison within plant multi-protein families

Ralph Panstruga
Max-Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, D-50829, Köln, Germany (author for correspondence; e-mail panstrug@mpiz-koeln.mpg.de)

Received 19 May 2005; accepted in revised form 28 June 2005

Key words: CONSTANS, MLO, multi-protein family, ortholog, paralog

Abstract

Assigning individual functions to the proteins encoded by the genome of the dicotyledonous reference species *Arabidopsis thaliana* is one of the major challenges in current plant molecular biology. Frequently, Arabidopsis protein families are biocomputationally analyzed by multiple amino acid sequence alignments of the respective family members for detection of conserved peptide motifs that might be of functional relevance. Mere sequence alignment of paralogous sequences may obscure amino acid patches that are highly conserved amongst orthologs and thus potentially relevant for isoform-specific protein function(s). Here I exemplarily illustrate this potential pitfall by amino acid sequence alignments of the heptahelical MLO proteins using either the suite of 15 isoforms (paralogs) encoded by the Arabidopsis genome or a collection of 13 ortholog sequences derived from a set of both monocotyledonous and dicotyledonous plant species. The findings are corroborated by an analogous analysis of the distinct plant multi-protein family of CONSTANS-like transcription regulators. The data reveal that the generally higher sequence similarity of orthologs versus paralogs is not uniformly distributed among the amino acid positions of the orthologs but at least partially clustered in distinct sites/domains, suggesting conservation of isoform-specific functional modules across taxa.

Introduction

At the latest with the completion of the genomic DNA sequence of *Arabidopsis thaliana* in 2000 (The Arabidopsis Genome Initiative, 2000), large-scale functional genomics has become a major research focus in the community of plant scientists. As evidenced by respective community webpages (e.g. http://www.arabidopsis.org/info/genefamily/genefamily.html), a wealth of publications (for recent reviews see Tabata, 2002; Ostergaard and Yanofsky, 2004; Rensink and Buell, 2004) and probably most impressively by major research consortiums like the North American ''Arabidopsis 2010'' project (Ausubel, 2002; http://www.nsf.gov/pubs/2004/nsf04502/nsf04502.htm) or its German complement, the ''Arabidopsis Functional Genomics Network (AFGN; http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/AFGNHome.html)'', the extensive functional analysis of Arabidopsis gene/protein families has entered center stage. A frequent initial step in the analysis of multigene/protein families is a comprehensive biocomputational examination of the respective family members, for example to determine their phylogenetic relationship via multiple sequence alignments (e.g. Eulgem *et al.*, 2000; Andrade *et al.*, 2001; Jiang *et al.*, 2004; Romano *et al.*, 2004). Besides providing the basis for phylogenetic studies, multiple sequence alignments illustrate amino acid conservation across members of a protein family of interest. The identification of

peptide stretches that are either particularly conserved or that exhibit pronounced sequence diversification may provide experimental leads for subsequent functional analyses by targeted site-directed mutational approaches (e.g. Trentmann *et al*., 2000; Shigaki *et al*., 2002; Elliott *et al*., 2005). Another potential application for multiple sequence alignments of protein isoforms is the allocation of putative functionally relevant peptide motifs. Web-based online databases like PROSITE (http://www.expasy.org/prosite/) can be employed for identifying defined peptide domains via short amino acid consensus sequences that may either represent functional entities or correspond to potential sites of posttranslational protein modification(s) (e.g. phosphorylation, glycosylation). However, many of these motifs are so ambiguous or so frequent that generally multiple of the respective sites per protein are detected, leaving the investigator commonly with largely ineffectual information. Thus, there is an urgent need for novel biocomputational approaches to reduce the sheer number of identified peptide domains to a moderate figure of potentially biologically meaningful sites that can be subsequently analyzed by experimentation.

'Paralogy' is defined as the relationship of any two homologous characters (here: proteins) that originate from duplication of the gene encoding that character within a genome (Fitch, 2000; Sonnhammer and Koonin, 2001). Commonly, all isoforms of a gene/protein family present in a given species are considered as paralogs. In contrast, the term 'orthology' describes the relationship of any two homologous characters whose common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained (Fitch, 2000). Depending on the time span that passed since the separation of the two lineages under consideration, orthologs encoded by genes residing in different species may be more closely related among each other than the majority of paralogs within the species since the latter usually result from ancient gene duplication events and were potentially subject to extensive sequence diversification during evolution. This type of paralogs (resulting from gene duplication events predating the lineage split under consideration) is therefore also termed 'out-paralogs' (Sonnhammer and Koonin, 2001). In contrast, in-paralogs represent paralogous characters in a given lineage that evolved by gene duplications following speciation events that separated the given lineage from the other lineage under consideration (Sonnhammer and Koonin, 2001).

MLO proteins comprise a class of plant-specific sequence-diversified proteins that possess seven membrane-spanning domains (Devoto *et al*., 19 99). The Arabidopsis genome encodes 15 isoforms (Devoto *et al*., 2003), whereas in the rice genome 12 family members appear to be present (R. Panstruga, unpublished). Barley *Hv*MLO, the founder of the integral membrane protein family, serves a role as modulator of defence against the phytopathogenic powdery mildew fungus, *Blumeria graminis* f.sp. *hordei* (Büschges *et al*., 1997). Recessively inherited loss-of-function barley *mlo* mutants are fully resistant against all known isolates of the common ascomycete pathogen. It is thought that the fungus corrupts presence of wild-type *Hv*MLO for suppression of a SNARE protein-dependent and possibly vesicle-associated defence mechanism at the cell periphery (Collins *et al*., 2003; Panstruga and Schulze-Lefert, 2003; Schulze-Lefert, 2004; Panstruga, 2005).

In this study I address the question whether a presumably higher sequence similarity of orthologs versus paralogs is uniformly distributed among the amino acid positions of the orthologs or whether conserved residues cluster in particular sites/domains. Using the multi-protein families of heptahelical MLO polypeptides and CONSTANS-like transcription regulators as examples, I demonstrate that several isoform-specific peptide stretches in various regions of the proteins escape attention or are obscured by mere paralog alignments. Conserved amino acid patches identified by ortholog similarity might be instrumental in subsequent functional studies by providing experimental leads for rational structure-function analyses.

Materials and methods

Amino acid sequences of the presumptive *At*MLO2 orthologs *Br*MLO, *Ca*MLO, *Le*MLO, and *Lj*MLO were deduced from full-size cDNA clones *Brassica rapa* E2573 (GenBank accession nos. BG544654 and AY967409), *Capsicum annuum* KS01071D10 (GenBank accession nos. BM06 4796 and AY934528), *Lycopersicon esculentum*

cTOC20K10 and cLEC80N18 (GenBank accession nos. BI931548, BI923467 and AY967408), and *Lotus japonicus* MWM066f10_r (GenBank accession nos. AV426381 and AY967410), respectively. Full-size *At*MLO cDNA sequences for all 15 paralogs were previously obtained (Devoto *et al*., 2003) and used for this study. Likewise, coding sequences of barley *Mlo*, *HvMlo3*, *TaMlo-B1*, all *ZmMlo* genes, *OsMlo1* and *OsMlo2* are predominantly (except *ZmM105* and *ZmM109*) derived from full size cDNAs (Devoto *et al*., 2003), whereas the remaining *Os*MLO amino acid sequences are based on the conceptual translation of genomic bacterial artificial chromosome (BAC) clones from either the subspecies *japonica* or *indica*. Likewise, the *Hv*MLO2 sequence was deduced from a genomic subclone (R. Panstruga, unpublished). Protein sequences of CONSTANS-like proteins are based on the GenBank entries listed in Griffiths *et al*. (2003).

Protein sequence alignments were performed with the CLUSTALW algorithm (http://www.ebi.ac.uk/clustalw/) using standard parameters. CLUSTALW alignments were shaded by means of the Boxshade algorithm (http://www.ch.embnet.org/software/BOX_form.html) choosing 0.8 as the fraction of sequences that must agree for shading.

For phylogenetic analysis of MLO proteins, the Phylip 3.63 software package was used (http://evolution.gs.washington.edu/phylip.html; Felsenstein, 1989). The highly polymorphic N- and C-termini of MLO protein sequences aligned by CLUSTALW were removed before calculating phylogenetic relationships. Thereafter, the Seqboot, Protdist, Neighbor and Consense algorithms were sequentially applied to establish the phylogenetic consensus tree, using 100 replicates each for bootstrap support. ProtML (maximum likelihood inference of protein phylogeny) was used to generate a phylogenetic tree based on the consensus tree calculated by Consense.

Calculations of the ratio of synonymous to nonsynonymous substitutions were performed using either the SNAP (http://www.hiv.lanl.gov/content/hiv-db/SNAP/WEBSNAP/SNAP.html) or yn00 (PAML Version 3.14) (http://abacus.gene.ucl.ac.uk/software/paml.html) software. SNAP (Synonymous/Nonsynonymous Analysis Program) calculates synonymous and nonsynonymous substitution rates based on a set of codon-aligned nucleotide sequences, according to the method of Nei and Gojobori (1986). PAML is a software package for phylogenetic analysis by maximum likelihood. The yn00 program included in the package calculates synonymous and nonsynonymous substitution rates according to the approximate maximum likelihood method of Yang and Nielsen (2000). dN/dS ratios are given as the arithmetic mean ± standard deviation for both procedures.

Results

Monocot and dicot orthologs of barley MLO

MLO proteins are encoded by medium-sized gene families comprising approximately 10–15 members per higher plant species (Devoto *et al*., 2003). While most likely all members encoded by the Arabidopsis and rice genomes are known, only few and mostly incomplete cDNA sequences are available from other plant species. The amino acid sequences of barley *Hv*MLO, barley *Hv*MLO2–*Hv*MLO3, wheat (*Triticum aestivum*) *Ta*MLO-B1, 15 Arabidopsis MLO isoforms (*At*MLO1–*At*MLO15), 9 maize (*Zea mays*) MLO isoforms (*Zm*MLO1–*Zm*MLO9), 11 rice (*Oryza sativa*) MLO isoforms (*Os*MLO1–*Os*MLO11) as well as the dicot sequences *Br*MLO (*Brassica rapa*), *Lj*MLO (*Lotus japonicus*), *Ca*MLO (*Capsicum annuum*), and tomato *Le*MLO (*Lycopersicon esculentum*) were used for phylogenetic analysis as described in Materials and methods. The resulting phylogenetic consensus tree identified two major clades of proteins that represent monophyletic sister lineages of monocot and dicot MLO proteins which both exhibit the highest sequence similarity to *Hv*MLO (Figure 1A, clades 1 and 2). Thus, members of these clades may be considered as orthologs sharing a common ancestor sequence in the last cenancestor of monocot and dicot plants. It should be stressed that this procedure represents an approximation of ortholog identification for those species for which only a single or few isoforms are available. Correct inference of ortholg–paralog relationships would require incorporating all paralogs of a given species. However, this is only possible for species for which the full genome sequence is available (currently Arabidopsis and rice).

Collectively, rice in-paralogs *Os*MLO1 and *Os*MLO2 as well as Arabidopsis in-paralogs *At*MLO2, *At*MLO6, and *At*MLO12 can be considered as co-orthologs of *Hv*MLO (Sonnhammer and Koonin, 2002). Orthologs in different species frequently (though not necessarily) retain

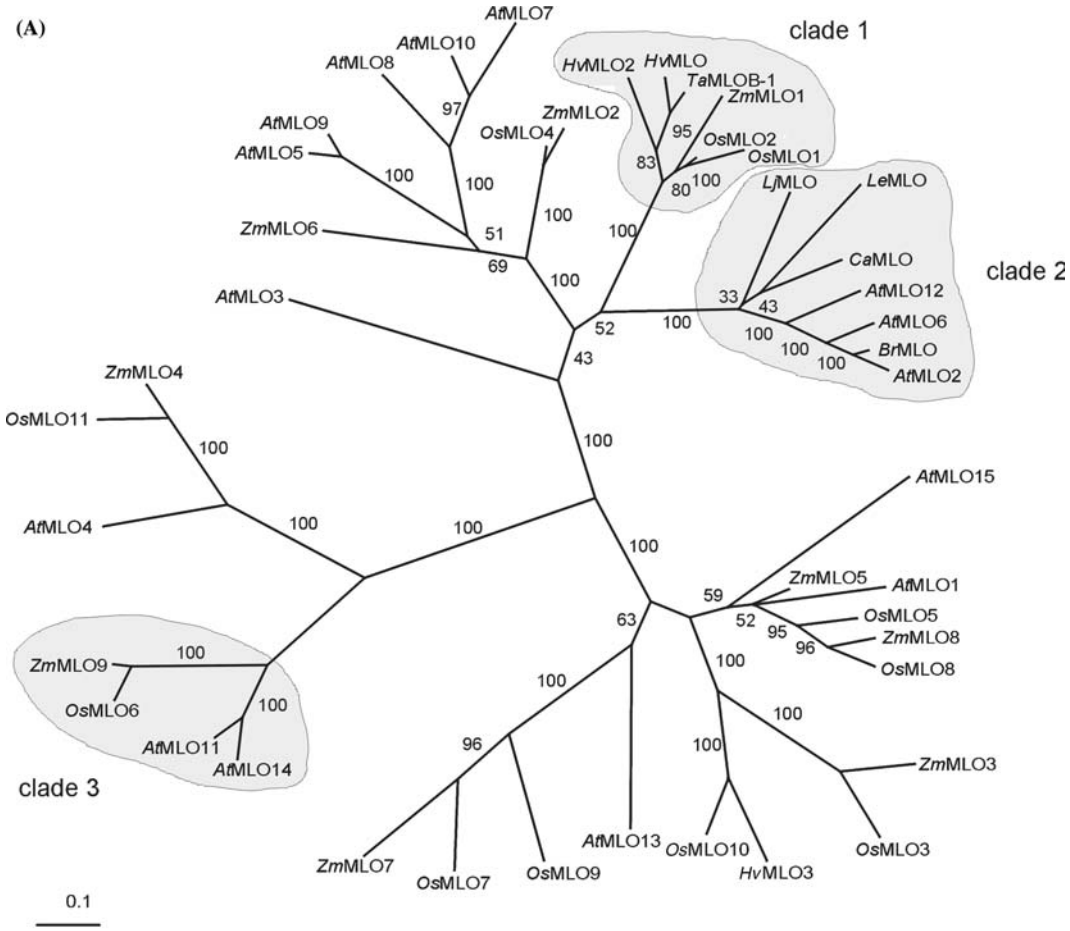

Figure 1. Phylogenetic relationship of MLO isoforms and topology of MLO proteins. (A) Phylogenetic analysis of an amino acid sequence dataset of 43 monocot and dicot MLO family members including 15 Arabidopsis (*At*) MLO isoforms, 11 rice (*Os*) MLO isoforms, 9 maize (*Zm*) MLO isoforms, three barley (*Hv*) MLO isoforms, as well as one wheat (*Ta*), one *Lotus japonicus* (*Lj*), one *Brassica rapa* (*Br*), one *Capsicum annuum* (*Ca*), and one tomato (*Le*) MLO isoform. Apart from *Zm*MLO5 and *Zm*MLO9 all polypeptide sequences are full-size. The phylogenetic tree represents a consensus tree with branch lengths proportional to sequence distance. Numbers indicate bootstrap values (out of 100 replicates) that support the respective branch. The scale (left bottom corner) indicates the number of amino acid substitutions per site (B) Schematic representation of the heptahelical topology of MLO proteins. The serpentine structure depicts the loop domains and the seven transmembrane helices (longitudinal small dark gray boxes) of MLO proteins. Domains investigated in this study (extracellular loop 1, EC1; intracellular loops 2 and 3, IC2 and IC3; C-terminus) are highlighted in black color. The large horizontal light gray box represents the lipid bilayer of the plasma membrane. NH2 and COOH symbolize the amino and carboxyl termini of the protein, respectively.

the same biological function(s) over time. In the case of *HvMlo* it was shown that *OsMlo2*, *ZmMlo1*, and *TaMlo-B1* are able to complement the powdery mildew resistant phenotype of barley *mlo* mutants thus further corroborating the apparent orthologous relationship between these four monocot genes (Elliott *et al.*, 2002 and unpublished data). In contrast, surprisingly neither dicot co-orthologs *AtMLO2*, *AtMLO6*, *AtMLO12* nor the monocot co-ortholog *OsMlo1* are capable of complementing barley *mlo* mutant genotypes (unpublished results), suggesting that an extraordinary degree of sequence similarity to *HvMlo* is required to substitute loss of barley *Mlo* function. However, loss-of-function *Atmlo2* mutants phenocopy the powdery mildew resistance phenotype of barley *mlo* mutants (C. Consonni, M. Humphry, P. Schulze-Lefert, S. Somerville and R. Panstruga, unpublished). An orthologous and/or in-paralogous relationship, respectively, between the Arabidopsis and rice isoforms *At*MLO2, *At*MLO6, *At*MLO12 as well as *Os*MLO1 and *Os*MLO2 is further supported by the 'InParanoid' algorithm. This software detects best-best hits between sequences from two different species based on pairwise similarity scores which are by default calculated with the NCBI BLAST program (http://inparanoid.cgb.ki.se/ehelp.html# how) (Remm *et al.*, 2001).

Multiple sequence alignments of MLO orthologs and paralogs

To unravel the distribution of conserved amino acid residues in *Hv*MLO ortholog and paralog sequences, a comparative sequence analysis was performed. The full size amino acid sequences of the 15 Arabidopsis *At*MLO isoforms were aligned by CLUSTALW (http://www.ebi.ac.uk/clustalw/) using standard parameters. In addition, the amino acid sequences of 12 presumed orthologs of *Hv*MLO (total of 13 sequences including *Hv*MLO; clades 1 and 2 in Figure 1A) were aligned separately. To avoid any subjective bias, none of the sequence alignments was adjusted manually. Subsequently, CLUSTALW alignments of the full protein sequences were chopped and identical residues highlighted by the 'Boxshade' software (http://www.ch.embnet.org/software/BOX_form.html) to illustrate amino acid conservation within particular domains of the polytopic MLO protein. This study is focused on the analysis of the first extracellular loop (EC1), the second and third intracellular loop (IC2 and IC3) as well as the cytoplasmic C-terminus of the seven transmembrane domain proteins (Figure 1B). As outlined below in detail, in most though not all comparative sequence alignments the number of detected conserved or invariant amino acid residues was as expected considerably higher for the 13 aligned monocot and dicot orthologs as compared to the 15 aligned paralogs of *Arabidopsis thaliana*. Surprisingly, however, ortholog sequence conservation was not evenly distributed across the analyzed peptide domains but focused in several cases in particular amino acid positions or short peptide stretches, thereby revealing previously unrecognized sites of potential isoform-specific functional significance.

Intracellular loops 2 and 3

Due to the heptahelical structure, MLO proteins possess three loops that are exposed to the cytoplasm (Figure 1B). These cytoplasmic loops are assumed to play a pivotal role for MLO function(s): All four barley *mlo* loss-of-function mutants known to encode stable protein variants are characterized by defects in either cytoplasmic loop two or three, suggesting that these mutant

variants might be impaired in protein–protein contacts with polypeptide interaction partners (Müller *et al.*, 2005; Panstruga *et al.*, 2005). Moreover, the second cytoplasmic loop hosts 9 of the 25 amino acid residues that appear invariant throughout the MLO protein family (Elliott *et al.*, 2005), suggesting that amino acid preservation in this region might be of particular importance for protein function. In addition, interplay of all cytoplasmic domains was suggested to be required for full MLO functionality (Elliott *et al.*, 2005), possibly by providing a cooperative interface for protein–protein interactions. Thus, identifying conserved amino acids in these loops (for site-directed mutagenesis in the context of structure–function studies) or identification of potential motifs that might be targets for posttranslational modifications is of key interest for future research on this protein family.

The second cytoplasmic loop represents the largest of the three loops of MLO proteins exposed to the cytoplasm (Figure 1B). CLUSTALW-based multiple sequence alignment of the 15 Arabidopsis paralogs identified 15 residues as invariant in the ~100 amino acid stretch of cytoplasmic loop two (Figure 2A). In contrast, 38 residues were found invariant in the mixed monocot and dicot ortholog polypeptide sets (Figure 2B). Notably, the region between amino acids 14 and 40 that appears to be devoid of any recognizable conserved domain in the paralog alignment (boxed in Figure 2A and B) exhibits several highly conserved/invariant amino acids in the ortholog alignment, suggesting that this section may play a role in isoform-specific MLO function(s). Interestingly, however, the stretch of ~10 amino acids juxtaposed downstream of this region is variable both in the paralogs and orthologs, possibly indicating that this region does not require extensive conservation at the primary amino acid sequence level to retain isoform function. Like the proximal half of the cytoplasmic loop two, the central region of cytoplasmic loop three is highly sequence divergent between the 15 Arabidopsis paralogs (Figure 3A, boxed region) but exhibits enhanced sequence conservation amongst the 13 MLO orthologs (Figure 3B).

Extensive sequence diversification might be either the result of relaxed amino acid conservation and the resulting evolutionary drift or, alternatively, the consequence of 'diversifying selection', an evolutionarily conditioned pressure for enhanced amino acid replacements within a particular peptide domain. The latter is commonly found in rapidly evolving genes (proteins) (e.g. Wang *et al.*, 2003). To distinguish between these two formal possibilities, the ratio of nonsynonymous (amino acid changing; dN) to synonymous (silent; dS) substitutions per nonsynonymous and synonymous sites in the respective peptide regions of cytoplasmic loops two and three (see Materials and methods) was calculated. dN/dS ratios < 1 are indicative of conservation (purifying selection), whereas dN/dS ratios > 1 point towards adaptive evolution. In the case of loop two either the coding sequence of the whole loop domain, the indicated boxed domain, or the following variable stretch of ~10 amino acids only was used for calculating dN/dS ratios. Since Arabidopsis paralogs as well as dicot ortholog sequences were due to saturation of synonymous substitutions too divergent ($P_s \geq 0.75$) to determine meaningful results, the analysis was focused on all 15 possible pairwise comparisons of the subset of six monocot ortholog sequences which exhibit moderate rates of synonymous substitutions. Calculations were either performed according to the method described by Nei and Gojobori (1986) or via the approximate maximum likelihood algorithm implemented by Yang and Nielsen (2000) (yn00 program in the PAML software package; see Materials and methods for details). This analysis revealed that, at least when compared among monocot orthologs, the dN/dS values are clearly below 1 without dramatic variation in various regions of the second and third cytoplasmic loop (Table 1). In conclusion, these data suggest that the observed amino acid sequence variation is most likely the long-term result of marginally relaxed evolutionary constraints (though no diversifying selection) in particular sub-domains or even at distinct amino acid positions of the loops.

Extracellular loop 1

We previously reported that the first extracellular loop of MLO proteins represents a region of extraordinary sequence variability that largely prevents proper amino acid sequence alignment and the identification of potentially conserved residues (Devoto *et al.*, 1999, 2003). When performing a multiple sequence alignment with a

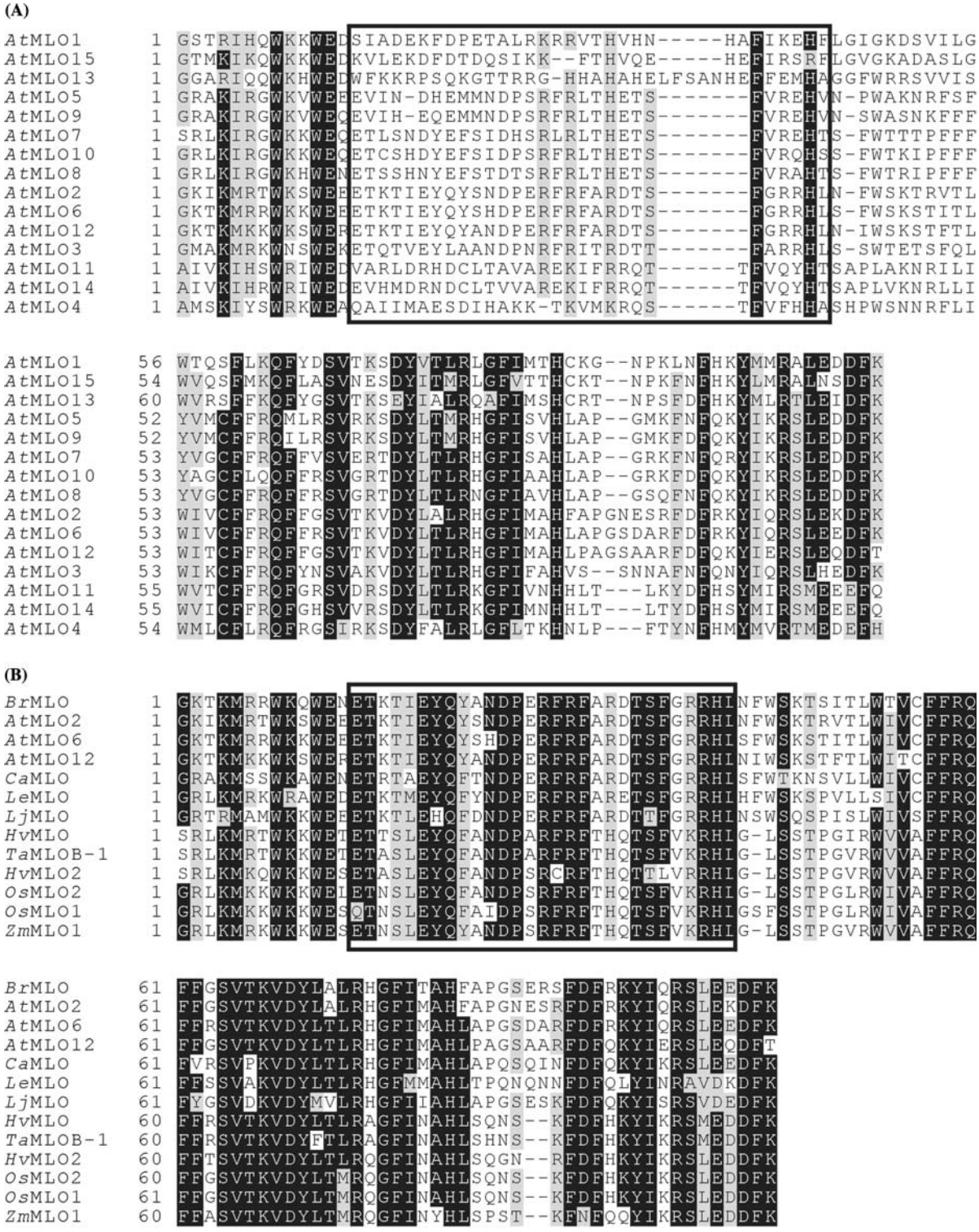

Figure 2. Comparative sequence alignment of the second intracellular loop. Multiple amino acid sequence alignment of the second intracellular loop region of 15 Arabidopsis MLO paralogs (A) and 13 monocot and dicot MLO orthologs (B). The boxed regions in (A) and (B) designate homologous sections in the two protein alignments. Black color indicates invariant residues, gray color marks conservative amino acid exchanges as indicated by the 'Boxshade' algorithm (see Materials and methods).

complex collection of paralogous and orthologous protein sequences, only the three invariant cysteines present in this loop became evident as highly conserved residues (Devoto *et al*., 2003; Elliott *et al*., 2005). However, even this required manual adjustment of the aligned protein sequences. Here, a similar alignment based on the 15 *At*MLO paralog sequences without manual adjustment is presented (Figure 4A). Due to the high sequence variability in this region, the CLUSTALW algorithm fails to accurately align the three invariant cysteines; only the first of the three cysteines is properly associated (see asterisks in Figure 4A). Besides the three cysteines, few amino acids are conserved between the Arabidopsis paralogs. Likewise, when dicot and monocot ortholog MLO sequences were aligned by CLUSTALW, few conserved or invariant amino acid residues flanking each of the three invariant cysteine residues became evident (Figure 4B). In contrast to the boxed regions of cytoplasmic loops two and three that exhibit pronounced sequence diversification between paralogs but considerable conservation among orthologs (see above), the stretch between cysteines two and three remains impossible to align even between ortholog sequences (Figure 4B, boxed region). This raises the intriguing possibility that this area does not directly contribute to protein function but rather serves a role as a scaffold connecting the two peptide domains before and thereafter. This task would be compatible with the relaxed evolutionarily constraints on the region as evidenced by the previously calculated elevated ratio of nonsynonymous to synonymous nucleotide substitutions per site (dN/dS; Devoto *et al*., 2003). Alternatively, this stretch may adopt a higher order peptide fold that is similar amongst orthologs and possibly also paralogs despite significant divergence at the primary amino acid sequence level.

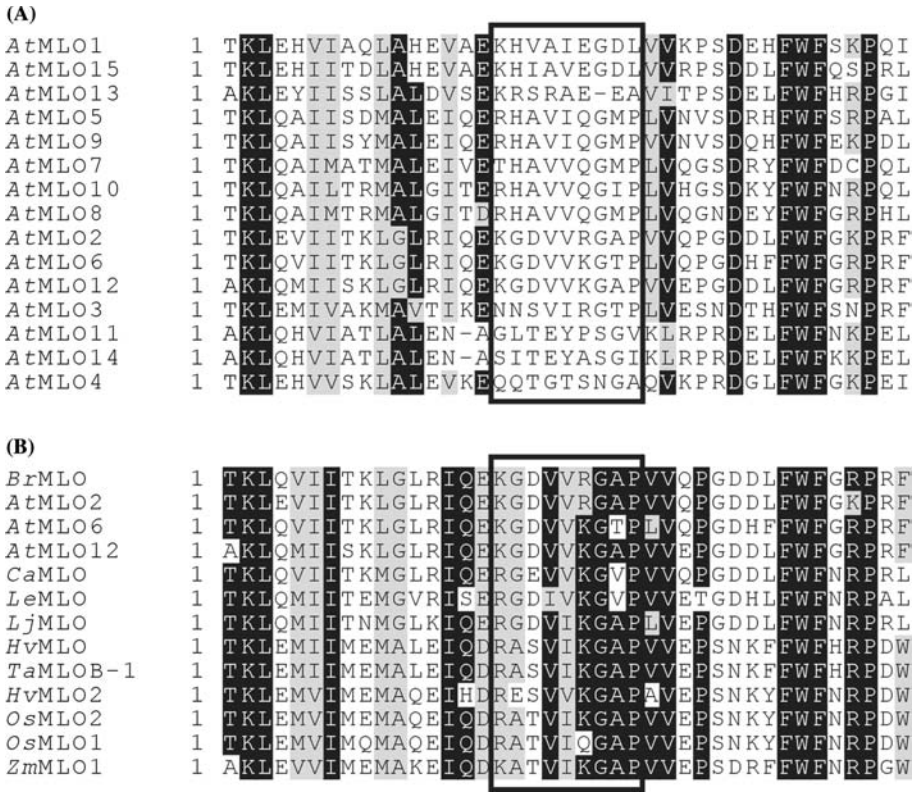

Figure 3. Comparative sequence alignment of the third intracellular loop. Multiple amino acid sequence alignment of the third intracellular loop region of 15 Arabidopsis MLO paralogs (A) and 13 monocot and dicot MLO orthologs (B) The boxed regions in (A) and (B) designate homologous sections in the two protein alignments. Black color indicates invariant residues, gray color marks conservative amino acid exchanges as indicated by the ''Boxshade'' algorithm (see Materials and methods).

Table 1. Ratios of non-synonymous to synonymous substitutions in cytoplasmic loops 2 and 3[a].

Section analyzed	Algorithm used to calculate d_N/d_S values	
	SNAP (Nei and Gojobori, 1986)	PAML (Yang and Nielsen, 2000)
Loop 2 (whole)	0.262 ± 0.072	0.057 ± 0.024
Loop 2 (boxed region)	0.133 ± 0.078	0.074 ± 0.042
Loop 2 (~10 amino acids following boxed region)	0.131 ± 0.115[b]	0.080 ± 0.023[b]
Loop 3 (whole)	0.137 ± 0.118[c]	0.042 ± 0.019[b]
Loop 3 (boxed region)	0.270 ± 0.130[d]	0.094 ± 0.061[e]

[a] Given as mean ± standard deviation of 15 pairwise monocot ortholog comparisons except where indicated differently.
Due to saturation of synonymous substitutions only [b]12, [c]13, [d]7, or [e]14 pairwise comparisons evaluable.

C-terminus

Besides the first extracellular loop, the C-terminus is the second region within MLO proteins that is highly polymorphic (Devoto *et al*., 1999, 2003). In the proximal part, the C-terminus of MLO proteins harbors a region of ~10–15 amino acids with the potential to form an amphiphilic α-helix. This area was previously shown to function as calmodulin binding domain (CaMBD) both *in vitro* and *in vivo* (Figure 5; Kim *et al*., 2002a and b; Bhat *et al*., 2005). Apart from this at the primary sequence level loosely defined domain, no obvious sequence motifs could be previously identified in the C-terminus of aligned amino acid sequences of a mixed set of paralogs and orthologs (Devoto *et al*., 2003). Likewise, a multiple sequence alignment of the 15 Arabidopsis paralogs does not reveal any additional conserved peptide domain(s) downstream of the CaMBD (Figure 5A). In contrast, a multiple sequence alignment of the dicot and monocot orthologs uncovers two further regions of pronounced sequence conservation. The first is located approximately 15–20 residues downstream of the calmodulin-binding domain and is inter alia characterized by presence of conserved serine and threonine residues. This is noteworthy because it has been recently found that the C-termini of two Arabidopsis MLO isoforms (*At*MLO1 and *At*MLO8) are phosphorylated in planta (Nühse *et al*., 2003, 2004). Moreover, detected phosphorylation sites of both isoforms were located in an amino acid stretch corresponding to the boxed region 1 shown in Figure 5A and B. Conserved patches of serines/threonines in this stretch may thus represent candidate phosphorylation sites. The second region of unexpected sequence similarity is located at the distal end of the C-terminus, bearing a peptide domain with the consensus sequence D/E-F-S/T-F (Figure 5B). Remarkably, the relative position of this pattern within the primary amino acid sequence of the C-terminus is not strictly fixed since the distance of the motif to both upstream (CaMBD and motif 1) and downstream anchor points (C-terminal end) appears variable.

To test whether the observed sequence similarity between orthologs also holds true for other clades of the heptahelical MLO protein family, the C-terminus of a further subgroup of MLO proteins (consisting of four presumptive orthologs; clade 3 in Figure 1) was aligned and conserved residues highlighted by the 'Boxshade' algorithm (Materials and methods). This revealed, similarly to the above analyzed family members of clades 1 and 2, additional apparently ortholog-specific amino acid sequence motifs downstream of the CaMBD (Supplementary Figure 1) that are not evident in the Arabidopsis paralog alignment (Figure 5A). This result demonstrates that presence of conserved ortholog-specific peptide stretches is not restricted to a particular phylogenetic clade of the serpentine MLO protein family.

CONSTANS-like proteins

To test the validity of the suggested approach beyond the family of heptahelical MLO proteins, an analogous analysis was performed in the family of CONSTANS-like proteins. The *CONSTANS* gene, originally identified in Arabidopsis, encodes a regulatory element in the photoperiod pathway controlling flowering time (Putterill *et al*. 1995). In Arabidopsis, CONSTANS belongs to a family of

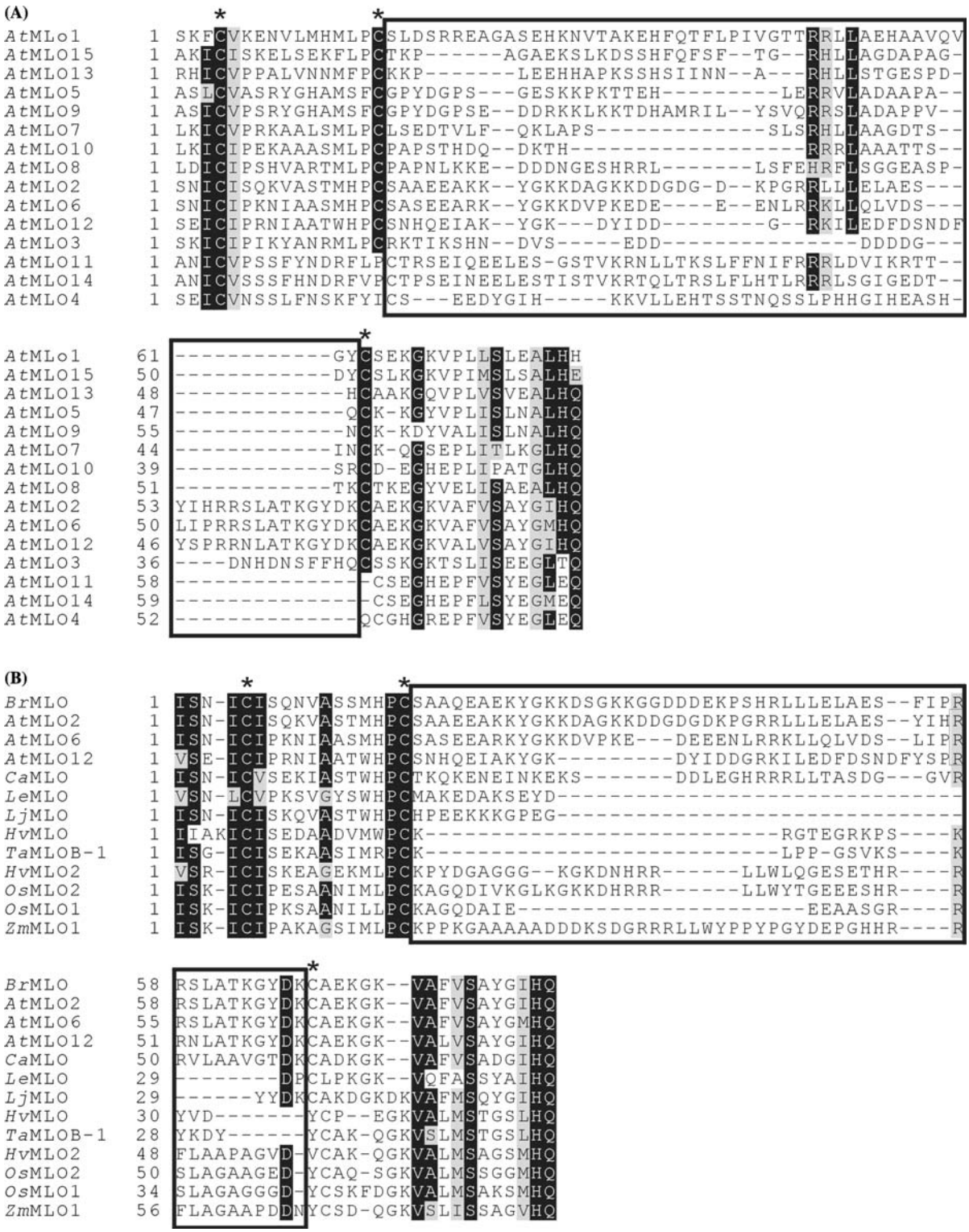
(A)
AtMLo1 1 SKFCVKENVLMHMLPCSLDSRREAGASEHKNVTAKEHFQTFLPIVGTTRRLLAEHAAVQV
AtMLO15 1 AKICISKELSEKFLPCTKP------AGAEKSLKDSSHFQFSF--TG--RHLLAGDAPAG-
AtMLO13 1 RHICVPPALVNNMFPCKKP-------LEEHHAPKSSHSIINN--A---RHLLSTGESPD-
AtMLO5 1 ASLCVASRYGHAMSFCGPYDGPS---GESKKPKTTEH---------LERRVLADAAPA--
AtMLO9 1 ASICVPSRYGHAMSFCGPYDGPSE--DDRKKLKKTDHAMRIL--YSVQRRSLADAPPV--
AtMLO7 1 LKICVPRKAALSMLPCLSEDTVLF--QKLAPS-----------SLSRHLLAAGDTS--
AtMLO10 1 LKICIPEKAAASMLPCPAPSTHDQ--DKTH-------------RRRLAAATTS--
AtMLO8 1 LDICIPSHVARTMLPCPAPNLKKE--DDDNGESHRRL-------LSFEHRFLSGGEASP-
AtMLO2 1 SNICISQKVASTMHPCSAAEEAKK--YGKKDAGKKDDGDG-D--KPGRRLLLELAES---
AtMLO6 1 SNICIPKNIAASMHPCSASEEARK--YGKKDVPKEDE----E--ENLRRKLLQLVDS---
AtMLO12 1 SEICIPRNIAATWHPCSNHQEIAK--YGK---DYIDD--------G--RKILEDFDSNDF
AtMLO3 1 SKICIPIKYANRMLPCRKTIKSHN--DVS-----EDD--------------DDDDG---
AtMLO11 1 ANICVPSSFYNDRFLPCTRSEIQEELES-GSTVKRNLLTKSLFFNIFRRRLDVIKRTT--
AtMLO14 1 ANICVSSSFHNDRFVPCTPSEINEELESTISTVKRTQLTRSLFLHTLRRRLSGIGEDT--
AtMLO4 1 SEICVNSSLFNSKFYICS---EEDYGIH-----KKVLLEHTSSTNQSSLPHHGIHEASH-
AtMLo1 61 ----------GYCSEKGKVPLLSLEALHH
AtMLO15 50 ----------DYCSLKGKVPIMSLSALHE
AtMLO13 48 ----------HCAAKGQVPLVSVEALHQ
AtMLO5 47 ----------QCK-KGYVPLISLNALHQ
AtMLO9 55 ----------NCK-KDYVALISLNALHQ
AtMLO7 44 ----------INCK-QGSEPLITLKGLHQ
AtMLO10 39 ----------SRCD-EGHEPLIPATGLHQ
AtMLO8 51 ----------TKCTKEGYVELISAEALHQ
AtMLO2 53 YIHRRSLATKGYDKCAEKGKVAFVSAYGIHQ
AtMLO6 50 LIPRRSLATKGYDKCAEKGKVAFVSAYGMHQ
AtMLO12 46 YSPRRNLATKGYDKCAEKGKVALVSAYGIHQ
AtMLO3 36 ----DNHDNSFFHQCSSKGKTSLISEEGLTQ
AtMLO11 58 ----------CSEGHEPFVSYEGLEQ
AtMLO14 59 ----------CSEGHEPFLSYEGMEQ
AtMLO4 52 ----------QCGHGREPFVSYEGLEQ
(B)
BrMLO 1 ISN-ICISQNVASSMHPCSAAQEAEKYGKKDSGKKGGDDDEKPSHRLLLELAES--FIPR
AtMLO2 1 ISN-ICISQKVASTMHPCSAAEEAKKYGKKDAGKKDDGDGDKPGRRLLLELAES--YIHR
AtMLO6 1 ISN-ICIPKNIAASMHPCSASEEARKYGKKDVPKE---DEEENLRRKLLQLVDS--LIPR
AtMLO12 1 VSE-ICIPRNIAATWHPCSNHQEIAKYGK---------DYIDDGRKILEDFDSNDFYSPR
CaMLO 1 ISN-ICVSEKIASTWHPCTKQKENEINKEKS-------DDLEGHRRRLLTASDG---GVR
LeMLO 1 VSN-LCVPKSVGYSWHPCMAKEDAKSEYD-------------------------------
LjMLO 1 ISN-ICISKQVASTWHPCHPEEKKKGPEG-------------------------------
HvMLO 1 IIAKICISEDAADVMWPCK----------------------------RGTEGRKPS----K
TaMLOB-1 1 ISG-ICISEKAASIMRPCK----------------------------LPP-GSVKS----K
HvMLO2 1 VSR-ICISKEAGEKMLPCKPYDGAGGG--KGKDNHRR------LLWLQGESETHR----R
OsMLO2 1 ISK-ICIPESAANIMLPCKAGQDIVKGLKGKKDHRRR------LLWYTGEEESHR----R
OsMLO1 1 ISK-ICIPKSAANILLPCKAGQDAIE----------------------EEAASGR----R
ZmMLO1 1 ISK-ICIPAKAGSIMLPCKPPKGAAAAADDDKSDGRRRLLWYPPYPGYDEPGHHR----R
BrMLO 58 RSLATKGYDKCAEKGK--VAFVSAYGIHQ
AtMLO2 58 RSLATKGYDKCAEKGK--VAFVSAYGIHQ
AtMLO6 55 RSLATKGYDKCAEKGK--VAFVSAYGMHQ
AtMLO12 51 RNLATKGYDKCAEKGK--VALVSAYGIHQ
CaMLO 50 RVLAAVGTDKCADKGK--VAFVSADGIHQ
LeMLO 29 --------DPCLPKGK--VQFASSYAIHQ
LjMLO 29 ------YYDKCAKDGKDKVAFMSQYGIHQ
HvMLO 30 YVD-------YCP--EGKVALMSTGSLHQ
TaMLOB-1 28 YKDY------YCAK-QGKVSLMSTGSLHQ
HvMLO2 48 FLAAPAGVD-VCAK-QGKVALMSAGSMHQ
OsMLO2 50 SLAGAAGED-YCAQ-SGKVALMSSGGMHQ
OsMLO1 34 SLAGAGGGD-YCSKFDGKVALMSAKSMHQ
ZmMLO1 56 FLAGAAPDDNYCSD-QGKVSLISSAGVHQ

Figure 4. Comparative sequence alignment of the first extracellular loop. Multiple amino acid sequence alignment of the first extracellular loop region of 15 Arabidopsis MLO paralogs (A) and 13 monocot and dicot MLO orthologs (B). The boxed regions in (A) and (B) designate homologous sections in the two protein alignments. The three asterisks in each panel specify the positions of cysteine residues that are invariant throughout the MLO protein family. Black color indicates invariant residues, gray color marks conservative amino acid exchanges as indicated by the 'Boxshade' algorithm (see Materials and methods).

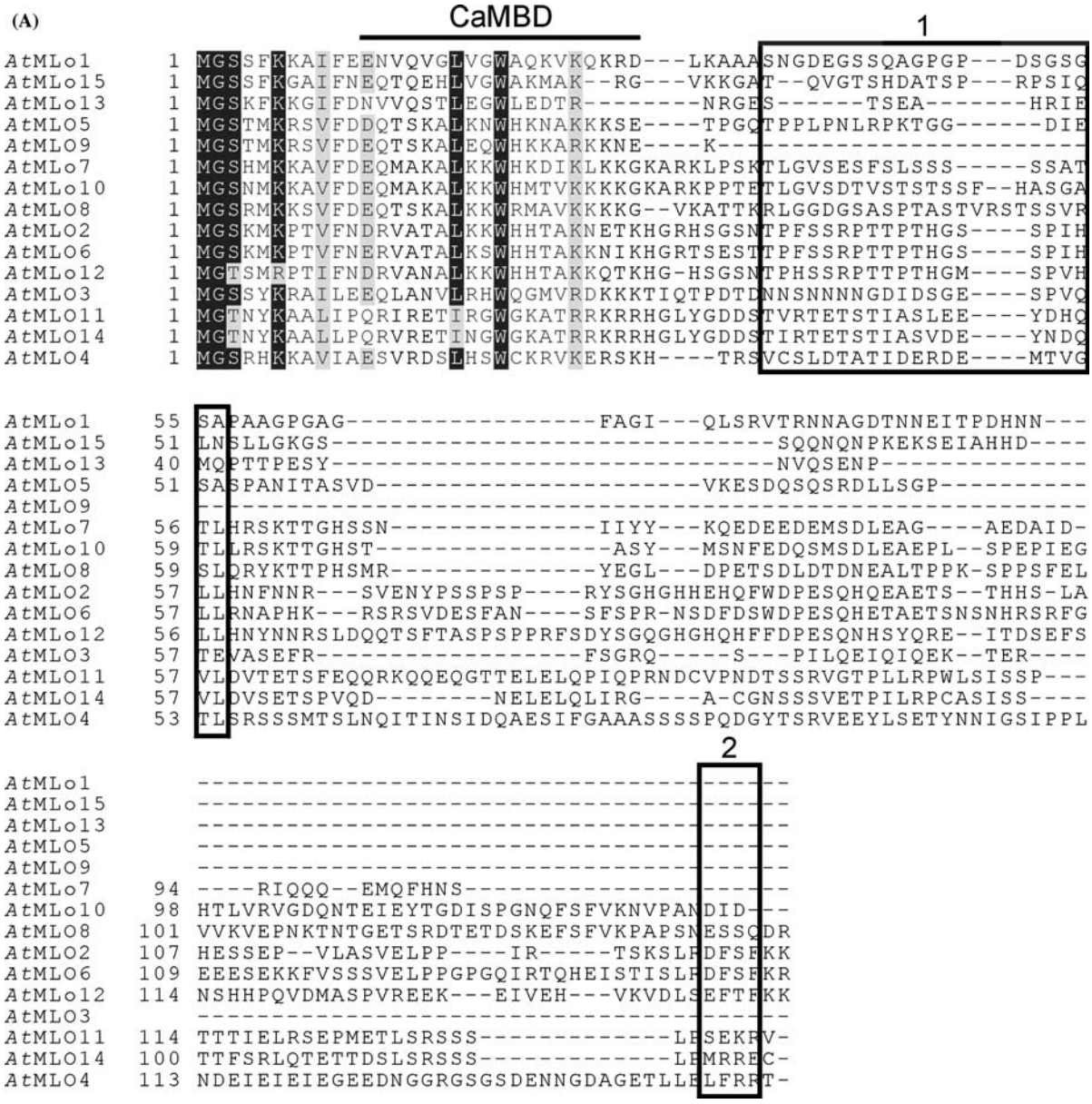

Figure 5. Comparative sequence alignment of the C-terminus. Multiple amino acid sequence alignment of the C-terminus of 15 Arabidopsis MLO paralogs (A) and 13 monocot and dicot MLO orthologs (B). The bar above the sequences in the two panels indicates the approximate position of the calmodulin-binding domain (CaMBD), the numbered boxes designate corresponding regions in the two alignments shown in (A) and (B). Black color indicates invariant residues, gray color marks conservative amino acid exchanges as indicated by the 'Boxshade' algorithm (see Materials and methods).

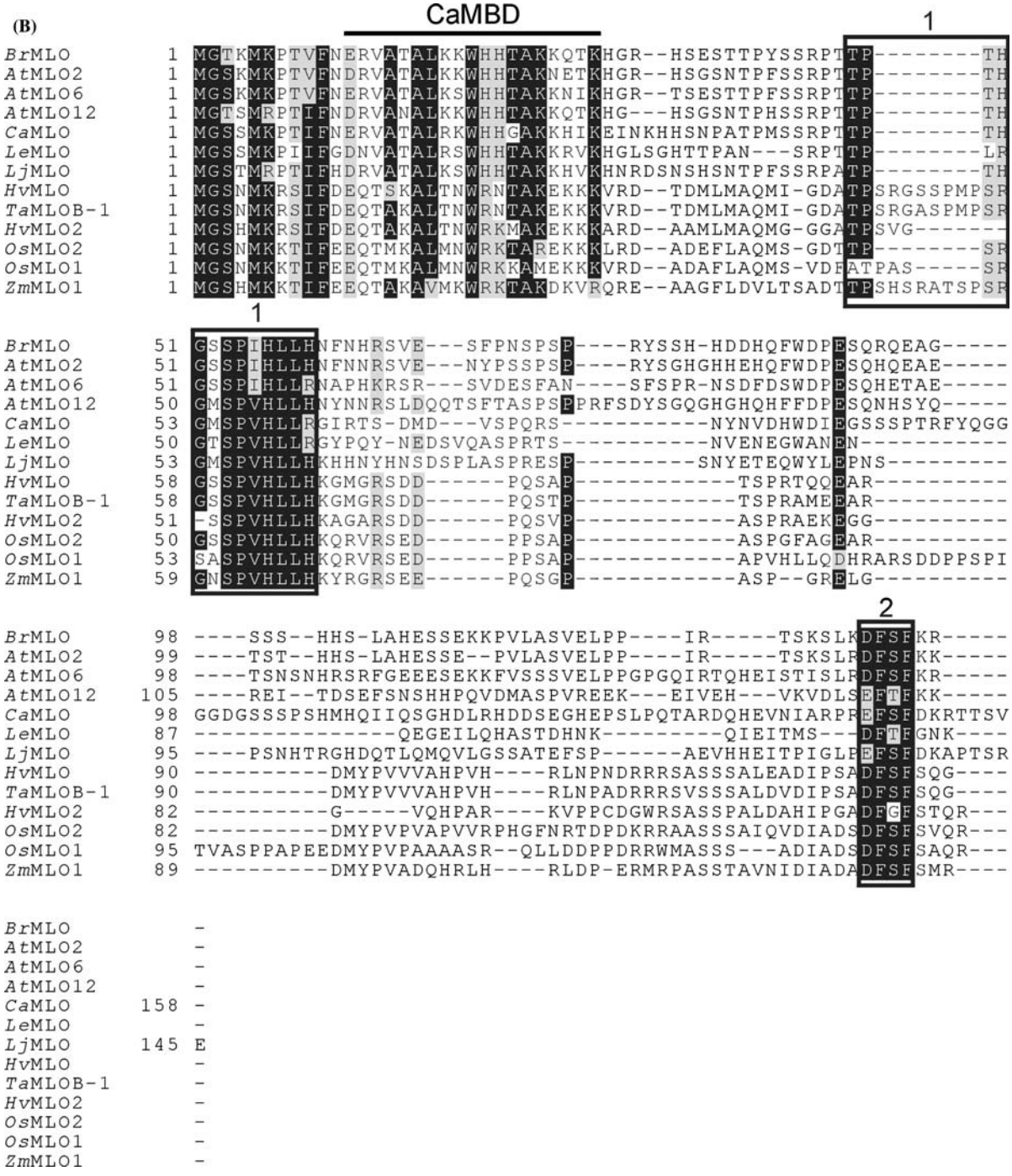

Figure 5. (Continued).

17 putative transcription factors defined by two conserved regions, the so-called B-box and CCT domains. In rice, a minimum of 16 *CONSTANS*-like genes exist (Griffiths *et al*., 2003). Previous phylogenetic analysis of *CONSTANS*-like genes from Arabidopsis, rice and barley revealed several clades comprising putative monocot and dicot orthologs (Griffiths *et al*. 2003). Multiple sequence alignment of either the 17 Arabidopsis isoforms or members of a clade (Ia in Figure 3b in Griffiths

et al., 2003) consisting of nine presumptive (co-)orthologs from Arabidopsis, rice, barley, *Brassica napus* or *Ipomoea nil* (Japanese morning glory) revealed, similar to the MLO protein family, several peptide stretches in various regions of the proteins that are conserved amongst orthologs but variable between paralogs (Supplementary Figure 2). This finding provides first evidence that peptide domains specifically conserved amongst orthologs may occur in several plant multi-protein families.

Discussion

This exemplarily study reveals in two instances, MLO and CONSTANS-like proteins, the potential superiority of multiple amino acid sequence alignments using a set of ortholog sequences versus employing a set of paralog sequences for the discovery of conserved and potentially functionally relevant peptide domains. While the latter yielded a poor resolution with respect to motif detection, several previously unrecognized conserved peptide domains were revealed by the former. This even included a region of the polytopic MLO membrane proteins that was previously thought to be fully devoid of any recognizable sequence motifs, namely the distal end of the highly polymorphic C-terminus (Devoto *et al.*, 2003). It remains to be seen whether these findings can be generalized for proteins encoded by other multigene families. Supporting evidence, however, is provided by a study of Nühse *et al.* (2004) who detected in two instances site-specific conservation of posttranslational phosphorylation sites amongst orthologs but not paralogs (see below). In addition, the occurrence of 'subfamily-specific' sequence motifs was previously noted in case of a few plant multigene families (e.g. Griffiths *et al.*, 2003; Anderson *et al.*, 2004; Tian *et al.*, 2004).

The Arabidopsis genome is thought to have undergone at least two major episodes of duplication, one likely after the monocot-dicot divergence (~200 million years ago; Wolfe *et al.*, 1989) and another more recently, before the Arabidopsis-*Brassica rapa* split and probably during the early emergence of the crucifer family (~24–40 million years ago; Blanc *et al.*, 2003). Thus, apart from individual gene duplications, a substantial set of paralogs present in the Arabidopsis genome are supposed to result from either ancient or recent, local or segmental, gene duplication events, giving rise to either out-paralogs or in-paralogs in relation to other species under consideration. Since their divergence happened comparatively 'recently' it can be assumed that on average monocot and dicot orthologs might be phylogenetically more closely related amongst each other than a considerable fraction of the (out-)paralogs present in the Arabidopsis genome. For example, the wheat–maize divergence is supposed to have occurred ~50–70 million years ago (Wolfe *et al.*, 1989) whereas the Arabidopsis-Medicago split is estimated to have happened ~92 million years ago (Grant *et al.*, 2000). Thus both speciation events happened considerably after the assumed first major duplication event in the progenitor of Arabidopsis (~200 million years ago; see above). Based on this phylogenetic history it was expected to find on average more sequence preservation amongst monocot and dicot orthologs than between Arabidopsis paralogs. However, the concentration of part of this sequence conservation in peptide domains that were previously thought to be generally highly polymorphic was a surprising finding. This result demonstrates that the elevated sequence preservation between orthologs is not evenly distributed but concentrated in particular amino acid positions.

In retrospective, ortholog-specific peptide stretches appear also identifiable upon closer visual inspection of paralog alignments, at first glance suggesting that these motifs might be recognizable without knowledge of further ortholog sequences. This impression is, however, misleading for two reasons. First, the relevant patches are only conserved between respective in-paralogs, e.g. *At*MLO2, 6, and 12, but not amongst out-paralogs. In a considerable number of multi-protein families at least some isoforms are encoded by single copy genes, thus having no in-paralogous but only out-paralogous relatives (for example *At*MLO3 and *At*MLO13; Figure 1A). In these instances conserved residues cannot be inferred from in-paralog comparison. Second, sequence stretches conserved amongst in-paralogs do not have predictive power for conservation of this region between orthologs. This is exemplified by some amino acid sections that are preserved

between in-paralogs but differ significantly amongst the full set of orthologs (Figures 2–5 and Supplementary Figure 2). Vice versa, however, due to the explicit presence of the newly discovered peptide domains in orthologs, these sequence motifs are expected to have predictive power to identify candidate orthologs from phylogenetically uncharacterized sequences, e.g. derived from EST sequencing projects.

Occurrence of ortholog-specific sequence motifs is a likely indicator of isoform specialization, e.g. presence of isoform-specific binding surfaces for protein–protein interactions or isoform-specific sites for posttranslational protein modification(s). Thus, newly discovered domains may be considered for future analysis by targeted structure–function studies. In addition, the multiple sequence alignments of orthologs can be used to estimate the significance of functional peptide domains suggested by motif prediction programs such as PROSITE. Biologically relevant isoform-specific peptide domains ('modules') are expected to be preserved amongst orthologs. A convincing example for this is provided by conservation of the relative position and amino acid sequence environment of posttranslational protein phosphorylation sites between orthologs, but not amongst paralogs, of cellulose synthases and a range of membrane transporters (Nühse *et al*., 2004). In case of *Hv*MLO, in contrast, none of the 16 predicted casein kinase II, protein kinase C, or cAMP or cGMP-dependent protein kinase phosphorylation sites located in the three cytoplasmic domains examined in this study were found to be conserved amongst the set of 13 orthologs and may thus represent a true phosphorylation site (data not shown).

Ortholog alignments might be particularly instructive for plant-specific protein families since in these instances information about functionally relevant peptide domains cannot be deduced from comparison with model species in other kingdoms such as bacteria, yeast or animals. Since approximately one third (~8000) of the genes present in the genome of Arabidopsis appears to be plant-specific (The Arabidopsis Genome Initiative, 2000; Goff *et al*., 2002), motif discovery by ortholog comparison might be a useful strategy for a substantial part of the proteome of this reference species. However, also protein families shared with other kingdoms might have evolved novel plant- and isoform-specific sequence motifs that could be obscured in paralog alignments.

An absolute requirement for multiple sequence alignments of orthologs is the availability of a sufficient number of (full size) DNA/protein sequences. Obtaining full-size cDNAs of (presumed) orthologs from a range of species can be laborious and represents thus a current bottleneck for ortholog comparisons. Many researchers at present therefore focus on biocomputational studies of model organisms like Arabidopsis and rice for which a full genome sequence and comprehensive EST data are readily available. However, these scientists run the risk to miss a lot of important information about their favorite protein families.

Acknowledgements

I wish to thank C.O. Lim (Gyeongyang National University, Chinju, Republic of Korea), the Kazusa DNA Research Institute (Japan), the Clemson University Genomics Institute (SC, U.S.A.), and the Genome Research Center and National Center for Genome Information (Yusong, Republic of Korea) for providing the *Brassica rapa*, *Lotus japonicus*, *Lycopersicon esculentum*, and *Capsicum annuum Mlo* EST clones, respectively. I am grateful to Thomas Nühse and Scott Peck whose initial observations inspired this work. I thankfully acknowledge helpful suggestions by Volker Lipka and Dierk Wanke. Work in my lab is funded by grants of the Max-Planck society and the Deutsche Forschungsgemeinschaft (DFG).

References

Anderson, G.H., Alvarez, N.D.G., Gilman, C., Jeffares, D.C., Trainor, V.C.W., Hanson, M.R. and Veit, B. 2004. Diversification of genes encoding mei2-like RNA binding proteins in plants. Plant Mol. Biol. 54: 653–670.

Andrade, M.A., Gonzalez-Guzman, M., Serrano, R. and Rodriguez, P.L. 2001. A combination of the F-box motif and kelch repeats defines a large Arabidopsis family of F-box proteins. Plant Mol. Biol. 46: 603–614.

Ausubel, F.M. 2002. Summaries of National Science Foundation-sponsored Arabidopsis 2010 projects and National Science Foundation-sponsored plant genome projects that are generating Arabidopsis resources for the community. Plant Physiol. 129: 394–437.

Bhat, R.A., Miklis, M., Schmelzer, E., Schulze-Lefert, P. and Panstruga, R. 2005. Recruitment and interaction dynamics of plant penetration resistance components in a plasma membrane microdomain. Proc. Natl. Acad. Sci. USA 102: 3135–3140.

Blanc, G., Hokamp, K. and Wolfe, K.H. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. 13: 137–144.

Büschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., van Daelen, R., van der Lee, T., Diergaarde, P., Groenendijk, J., Töpsch, S., Vos, P., Salamini, F. and Schulze-Lefert, P. 1997. The barley *Mlo* gene: a novel control element of plant pathogen resistance. Cell 88: 695–705.

Collins, N.C., Thordal-Christensen, H., Lipka, V., Bau, S., Kombrink, E., Qiu, J.L., Hückelhoven, R., Stein, M., Freialdenhoven, A., Somerville, S.C. and Schulze-Lefert, P. 2003. SNARE-protein-mediated disease resistance at the plant cell wall. Nature 425: 973–977.

Devoto, A., Piffanelli, P., Nilsson, I., Wallin, E., Panstruga, R., von Heijne, G. and Schulze-Lefert, P. 1999. Topology, subcellular localization, and sequence diversity of the *Mlo* family in plants. J. Biol. Chem. 274: 34993–35004.

Devoto, A., Hartmann, H.A., Piffanelli, P., Elliott, C., Simmons, C., Taramino, G., Goh, C.S., Cohen, F.E., Emerson, B.C., Schulze-Lefert, P. and Panstruga, R. 2003. Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family. J. Mol. Evol. 56: 77–88.

Elliott, C., Zhou, F.S., Spielmeyer, W., Panstruga, R. and Schulze-Lefert, P. 2002. Functional conservation of wheat and rice *Mlo* orthologs in defense modulation to the powdery mildew fungus. Mol. Plant-Microb. Interact. 15: 1069–1077.

Elliott, C., Müller, J., Miklis, M., Bhatt, R.A., Schulze-Lefertt, P. and Panstruga, R. 2005. Conserved extracellular cysteine residues and cytoplasmic loop-loop interplay are required for functionality of the heptahelical MLO protein. Biochem. J. 385: 243–254.

Eulgem, T., Rushton, P.J., Robatzek, S. and Somssich, I.E. 2000. The WRKY superfamily of plant transcription factors. Trends Plant Sci. 5: 199–206.

Felsenstein, J. 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.

Fitch, W.M. 2000. Homology – a personal view on some of the problems. Trends Genet. 16: 227–231.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R.L., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchinson, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P. Zhong, J.P., Miguel, T., Paszkowski, U., Zhang, S.P., Colbert, M., Sun, W.L., Chen, L.L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y.S., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S. Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science 296: 92–100.

Grant, D., Cregan, P. and Shoemaker, R.C. 2000. Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc. Natl. Acad. Sci. USA 97: 4168–4173.

Griffiths, S., Dunford, R.P., Coupland, G. and Laurie, D.A 2003. The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis. Plant Physiol. 131: 1855–1867.

Jiang, C.Z., Gu, X. and Peterson, T. 2004. Identification of conserved gene structures and carboxy-terminal motifs in the Myb gene family of Arabidopsis and *Oryza* sativa L. ssp *indica*. Genome Biol. 5: R46.

Kim, M.C., Lee, S.H., Kim, J.K., Chun, H.J., Choi, M.S., Chung, W.S., Moon, B.C., Kang, C.H., Park, C.Y., Yoo, J.H., Kang, Y.H., Koo, S.C., Koo, Y.D., Jung, J.C. Kim, S.T., Schulze-Lefert, P., Lee, S.Y. and Cho, M.J. 2002a. Mlo, a modulator of plant defense and cell death, is a novel calmodulin-binding protein – Isolation and characterization of a rice Mlo homologue. J. Biol. Chem. 277: 19304–19314.

Kim, M.C., Panstruga, R., Elliott, C., Müller, J., Devoto, A., Yoon, H.W., Park, H.C., Cho, M.J. and Schulze-Lefert, P. 2002b. Calmodulin interacts with MLO protein to regulate defence against mildew in barley. Nature 416: 447–450.

Müller, J., Piffanelli, P., Devoto, A., Miklis, M., Elliott, C., Ortmann, B., Schulze-Lefert, P. and Panstruga, R. 2005. Conserved ERAD-Like quality control of a plant polytopic membrane protein. Plant Cell 17: 149–163.

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3: 418–426.

Nühse, T.S., Stensballe, A., Jensen, O.N. and Peck, S.C. 2003. Large-scale analysis of *in vivo* phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. Mol. Cell. Proteomics 2: 1234–1243.

Nühse, T.S., Stensballe, A., Jensen, O.N. and Peck, S.C. 2004. Phosphoproteomics of the Arabidopsis plasma membrane and a new phosphorylation site database. Plant Cell 16: 2394–2405.

Ostergaard, L. and Yanofsky, M.F. 2004. Establishing gene function by mutagenesis in *Arabidopsis thaliana*. Plant J. 39: 682–696.

Panstruga, R. and Schulze-Lefert, P. 2003. Corruption of host seven-transmembrane proteins by pathogenic microbes: a common theme in animals and plants?. Microb. Infect. 5: 429–437.

Panstruga, R. 2005. Serpentine MLO proteins as entry portals for powdery mildew fungi. Trans. Biochem. Soc. 33: 389–292.

Panstruga, R., Molina-Cano, L.J., Reinstädler, A. and Müller, J. 2005. Molecular characterization of *mlo* mutants in North American two- and six-rowed malting barley cultivars. Mol. Plant Pathol. 6: 315–320.

Putterill, J., Robson, F., Lee, K., Simon, R. and Coupland, G. 1995. The *CONSTANS* gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. Cell 80: 847–855.

Remm, M., Storm, C.E.V. and Sonnhammer, E.L.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314: 1041–1052.

Rensink, W.A. and Buell, C.R. 2004. Arabidopsis to rice. Applying knowledge from a weed to enhance our understanding of a crop species. Plant Physiol. 135: 622–629.

Romano, P.G.N., Horton, P. and Gray, J.E. 2004. The Arabidopsis cyclophilin gene family. Plant Physiol. 134: 1268–1282.

Schulze-Lefert, P. 2004. Knocking on heaven's wall: pathogenesis of and resistance to biotrophic fungi at the cell wall. Curr. Opin. Plant Biol. 7: 377–383.

Shigaki, T., Sreevidya, C. and Hirschi, K.D. 2002. Analysis of the Ca^{2+} domain in the Arabidopsis H^{+}/Ca^{2+} antiporters CAX1 and CAX3. Plant Mol. Biol. 50: 475–483.

Sonnhammer, E.L.L. and Koonin, E.V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. 18: 619–620.

Tabata, S. 2002. Impact of genomics approaches on plant genetics and physiology. J. Plant Res. 115: 271–275.

The Arabidopsis Genome Initiative., 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Tian, C.G., Wan, P., Sun, S.H., Li, J.Y. and Chen, M.S. 2004. Genome-wide analysis of the GRAS gene family in rice and Arabidopsis. Plant Mol. Biol. 54: 519–532.

Trentmann, O., Decker, G.C., Winkler, H.H. and Neuhaus, H.E. 2000. Charged amino-acid residues in transmembrane domains of the plastidic ATP/ADP transporter from Arabidopsis are important for transport efficiency, substrate specificity, and counter exchange properties. Eur. J. Biochem. 267: 4098–4105.

Wang, H.Y., Tang, H., Shen, C.K.J. and Wu, C.I. 2003. Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. Mol. Biol. Evol. 20: 1795–1804.

Wolfe, K.H., Gouy, M.L., Yang, Y.W., Sharp, P.M. and Li, W.H. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA 86: 6201–6205.

Yang, Z.H. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17: 32–43.