Genome-Wide Screening for Lectin Motifs in *Arabidopsis thaliana*

Lore Eggermont, Bruno Verstraeten, and Els J.M. Van Damme*

Abstract
For more than three decades, *Arabidopsis thaliana* served as a model for plant biology research. At present only a few protein families have been studied in detail in *Arabidopsis*. This study focused on all sequences with lectin motifs in the genome of *Arabidopsis*. Based on amino acid sequence similarity (BLASTp searches), 217 putative lectin genes were retrieved belonging to 9 out of 12 different lectin families. The domain organization and genomic distribution for each lectin family were analyzed. Domain architecture analysis revealed that most of these lectin gene sequences are linked to other domains, often belonging to protein families with catalytic activity. Many protein domains identified are known to play a role in stress signaling and defense, suggesting a major contribution of the putative lectins in development and plant defense. This genome-wide screen for different lectin motifs will help to unravel the functional characteristics of lectins. In addition, phylogenetic trees and WebLogos were created and showed that most lectin sequences that share the same domain architecture evolved together. Furthermore, the amino acids responsible for carbohydrate binding are largely conserved. Our results provide information about the evolutionary relationships and functional divergence of the lectin motifs in *A. thaliana*.

Core Ideas
- More than 200 lectin genes were identified in *Arabidopsis thaliana*.
- The biological role(s) of proteins containing lectin motifs was discussed.
- Evolutionary relationships of several lectin motifs were investigated.

Proteins are key molecules that fulfill a range of biological roles in a cell. Protein domains are distinct parts in the protein sequence that can fold and function separately (Nasir et al., 2014). A protein domain can be defined from different perspectives. The structural viewpoint defines a protein domain as an independent protein fold. From an evolutionary point of view, protein domains are defined as conserved parts of the sequence. Moreover, these domains typically have a particular recurring function. Altogether a protein domain represents...
a conserved part of the sequence with a specific fold and function (Kelley and Sternberg, 2015; Moore et al., 2008).

The domain architecture of a protein contains all the information of the domains that build the protein and can be determined by scanning the protein sequence through a domain database, such as Pfam. Several domains are found at the root of the species tree, indicating that these are common to most species and are used to create domain architectures by modular rearrangements (Moore et al., 2008). Between 5.6 and 12.4% of all currently found domain architectures have been generated more than once throughout evolutionary history. The only reason that this could have happened is that the same domain architectures were formed in different branches of the tree of life as a consequence of selection (Forslund et al., 2008).

Most proteins consist of multiple domains (Rentzsch and Orengo, 2013). Single-domain architectures are common for the major groups of organisms (prokaryotes, eukaryotes, and viruses), whereas multidomain architectures are usually unique to a species and explain the diversification thereof (Levitt, 2009). Indeed, during the evolution of proteins both domain gains and losses have occurred (Nasir et al., 2014).

Lectins are proteins in which the domain architecture contains at least one lectin domain. A lectin domain can bind reversibly to specific carbohydrate structures, either free carbohydrates or glycans from glycoproteins and glycolipids (Peumans and Van Damme, 1995). According to their carbohydrate binding domain, plant lectins can be divided into 12 families: the Agaricus bisporus (Lange) Imbach agglutinin family, the amaranth family, the homologs of class V chitinases (CRA), the cyanovirin family, the Euonymus europaeus L. lectin (EUL) family, the Galanthus nivalis L. agglutinin (GNA) family, the hevein family, the jacalin-related lectin (JRL) family, the legume lectin family, the Lysin Motif (LysM) domain family, the Nicotiana tabacum L. agglutinin (Nictaba) family, and the ricin B lectin family (Van Damme et al., 2008). Most lectins contain, in addition to their carbohydrate binding domain, at least one other functional protein domain (Van Damme, 2014).

For more than three decades, Arabidopsis thaliana (L.) Heynh served as a model for plant biology research. The domain architecture of a protein contains all the information of the domains that build the protein and can be determined by scanning the protein sequence through a domain database, such as Pfam. Several domains are found at the root of the species tree, indicating that these are common to most species and are used to create domain architectures by modular rearrangements (Moore et al., 2008). Between 5.6 and 12.4% of all currently found domain architectures have been generated more than once throughout evolutionary history. The only reason that this could have happened is that the same domain architectures were formed in different branches of the tree of life as a consequence of selection (Forslund et al., 2008).

Most proteins consist of multiple domains (Rentzsch and Orengo, 2013). Single-domain architectures are common for the major groups of organisms (prokaryotes, eukaryotes, and viruses), whereas multidomain architectures are usually unique to a species and explain the diversification thereof (Levitt, 2009). Indeed, during the evolution of proteins both domain gains and losses have occurred (Nasir et al., 2014).

Lectins are proteins in which the domain architecture contains at least one lectin domain. A lectin domain can bind reversibly to specific carbohydrate structures, either free carbohydrates or glycans from glycoproteins and glycolipids (Peumans and Van Damme, 1995). According to their carbohydrate binding domain, plant lectins can be divided into 12 families: the Agaricus bisporus (Lange) Imbach agglutinin family, the amaranth family, the homologs of class V chitinases (CRA), the cyanovirin family, the Euonymus europaeus L. lectin (EUL) family, the Galanthus nivalis L. agglutinin (GNA) family, the hevein family, the jacalin-related lectin (JRL) family, the legume lectin family, the Lysin Motif (LysM) domain family, the Nicotiana tabacum L. agglutinin (Nictaba) family, and the ricin B lectin family (Van Damme et al., 2008). Most lectins contain, in addition to their carbohydrate binding domain, at least one other functional protein domain (Van Damme, 2014).

For more than three decades, Arabidopsis thaliana (L.) Heynh served as a model for plant biology research. At present several reports are accessible on individual proteins containing a lectin domain, but only a few families of proteins have been studied in detail in Arabidopsis. This study aimed to make an inventory of all sequences with lectin motifs in the genome of Arabidopsis. All the lectin sequences from A. thaliana were identified, and their domain architectures were determined. Known functions from the literature for several members of each family were discussed. Phylogenetic trees and WebLogos for three lectin families (the JRL family, the LysM domain family, and the Nictaba domain family) yielded new insights into the phylogenetic relationships of these lectins in plants. Our study provides information about the evolutionary relationships and functional divergence of the lectin motifs in A. thaliana.

Materials and Methods

Identification of the Putative Lectin Genes in Arabidopsis thaliana

Sequences encoding the putative lectin genes were searched for in the A. thaliana genome on the Phytozome v10.3 website (https://phytozome.jgi.doe.gov/pz/portal.html) (Altschul et al., 1990; Goodstein et al., 2012). Protein sequences encoding model proteins from each lectin family were used as a query for BLASTp searches (Basic Local Alignment Search Tool; comparison matrix BLOSUM62 and word length 3). Each model sequence represents the first lectin sequence described for a particular lectin family (Supplemental Table S1). The top hits resulting from these BLASTp searches were again used as a query. All sequences (E-value <10) were downloaded with the BioMart application from Phytozome v10.3 (Smedley et al., 2015).

Alternatively, sequences of putative lectin genes in A. thaliana were retrieved by using Pfam identification numbers for each lectin domain (Supplemental Table S1). However, Pfam identification numbers are not available for the lectin domains of the EUL and CRA family.

Finally, Pfam domain names (Supplemental Table S1) were used to search for potential lectins with the SMART (Simple Modular Architecture Research Tool) database (http://smart.embl-heidelberg.de/) (Letunic et al., 2009). Because the SMART database does not use Arabidopsis Gene Initiative codes, the Arabidopsis Gene Initiative codes of the resulting protein sequences were found in the UniProtKB database (http://www.uniprot.org/uniprot/).

Identical protein sequences retrieved from the three methods described above were deduplicated.

Annotating the Protein Domains of Putative Lectins

All sequences for the putative lectin genes from A. thaliana were checked for the presence of a lectin domain with InterProScan5 (http://www.ebi.ac.uk/interpro). InterProScan5 scans protein sequences on conserved protein domains and combines data from multiple databases: HAMAP, PANTHER, PfamA, PIRSF, ProDom, PRINTS, Prosite-Profiles, SMART, TIGRFAM, Prosite-Patterns, Gene3d, and SUPERFAMILY (Jones et al., 2014). InterProScan5 version 5.7–48.0 was installed and run on the local server. Proteins with at least one lectin domain were considered as a putative lectin. Protein domains other than lectin domains were identified. The start and end position of each protein domain was determined and used to draw the domain architecture for each putative lectin on scale using the DomainDraw software (http://domaindraw.imb.uq.edu.au/) (Fink and Hamilton, 2007).

Because the lectin domains of the EUL and CRA family have no Pfam identification number, the lectin domain sequences for the model sequences (Supplemental Table S1) were aligned with the protein sequences encoding the putative lectins from the EUL and CRA family with Clustal Omega (http://www.ebi.ac.uk/Tools/
msa/clustalo/) (Sievers et al., 2011). InterProScan5 was used to identify the domains other than lectin domains in these EUL and CRA family lectins.

**Determining Signal Sequences and/or Transmembrane Domains**

Each potential lectin sequence was checked for the presence of a signal sequence and/or a transmembrane domain with Phobius (http://phobius.sbc.su.se/index.html). Phobius combines the models from the SignalP and TMHMM server in a slightly modified way (Käll et al., 2004). The start and end positions for the signal sequences and/or transmembrane domains were determined and used to draw the schematic domain architectures.

**Mapping the Putative Lectin Genes on the Chromosomes of Arabidopsis and Analysis of Tandem Duplications**

Using the BioMart application of Phytozome v10.3, the transcription start position (base pairs) of each putative lectin was retrieved and used to map the lectin genes on the chromosomes of Arabidopsis (Voorrips, 2002). Only primary transcripts were mapped on the chromosomes. The positions of the centromeres are according to Feraru et al. (2012). The chromosomes were drawn to scale using their golden path lengths (TAIR, https://www.Arabidopsis.org/index.jsp). Tandem duplicated genes were defined as two lectin genes from the same family located on the same chromosome separated by a maximum of 10 other (not lectin) genes.

**Phylogenetic Analysis**

Phylogenetic trees were created using the lectin domain sequences identified for each family. For putative lectin sequences that contain more than one lectin domain, each lectin domain was used as a separate entry. Alignment of these sequences was performed with Multiple Alignment using Fast Fourier Transform (http://www.ebi.ac.uk/Tools/msa/mafft/) using the default parameters (Katoh and Standley, 2013). The alignments were trimmed using the automated1 option of trimAl, which was installed locally (Capella-Gutiérrez et al., 2009). Phylogenetic trees were created with RAxML v8.2.4 using the GAMMA model for rate heterogeneity and automatic determination of the best amino acid substitution model (i.e., the model with the highest likelihood score on the starting tree) (Stamatakis, 2014). Bootstrap analysis was performed using the rapid bootstrap algorithm of RAxML (Stamatakis et al., 2008). The number of bootstraps was determined using the frequency criterion up to a maximum number of 1000. Visualization of the phylogenetic tree was done with the FigTree v1.4.3 software (http://tree.bio.ed.ac.uk/software/figtree/).

**Analysis of Amino Acids Responsible for Carbohydrate Binding**

The untrimmed sequence alignments used to generate the phylogenetic trees were used in WebLogo3 to make a graphical representation of the amino acid conservation at each position of the sequence (http://weblogo.berkeley.edu/logo.cgi) (Crooks et al., 2004). A comparative analysis with model sequences and the amino acids known to be important for carbohydrate binding activity allowed us to check if amino acids essential for interaction with carbohydrate structures are conserved.

**Results and Discussion**

**Identification and Distribution of the Genes with a Lectin Domain in Arabidopsis thaliana**

BLASTp searches against the A. thaliana genome retrieved 217 putative lectin sequences that could be classified in 9 of the 12 plant lectin families as defined in Van Damme et al. (2008) (Supplemental Table S2). Sequences with lectin domains homologous to the Agaricus bisporus agglutinin, amaranthin, and cyanovirin families were not found (Table 1). The legume lectins represent the most abundant lectin family in A. thaliana with 54 putative lectin genes (24.9%), followed by the JRL family (50 genes; 23.0%) and the GNA lectin family (49 genes; 22.6%).

Mapping of the transcript start positions on the five chromosomes revealed that putative lectin genes are present throughout the A. thaliana genome (Table 1; Fig. 1). Genes for some lectin families (e.g., LysM domain family) are present on each chromosome, whereas genes from other lectin families (e.g., hevein family) are only present on one or a few chromosomes. The nine putative lectin genes from the CRA lectin family are present in one tandem duplication cluster on chromosome 4. Chromosome 4 is the smallest chromosome and shows...
the lowest lectin gene density (1.3 genes per Mbp). Chromosome 1 is the largest chromosome and has the highest lectin density (2.4 genes per Mbp).

Tandem duplications of lectin sequences are spread throughout the genome and are found for putative lectin genes from six families (Fig. 1). Chromosome 1 has the most tandem duplication clusters ($n = 10$), and chromosome 3 has the least ($n = 4$). The legume lectin homologs are the only family with tandem duplication clusters on each chromosome. The GNA family has four tandem duplication clusters on chromosome 1. This represents the highest number of tandem duplication clusters on one chromosome.

## Domain Architecture and Importance of the Putative Lectins in *Arabidopsis thaliana*

All sequences with a putative lectin domain were searched for the presence of other protein domains with a known function. Different (combinations of) protein domains can give information on the possible functions of the protein. Different domain architectures for the putative lectins from each family are shown in Fig. 2 through 5. All sequences were checked for the presence of a signal peptide and/or a transmembrane domain to give information about the localization of the putative lectins in the plant cell (Table 2). If available, literature reporting on the importance of the lectin sequences for *Arabidopsis* growth and development is briefly discussed. Supplemental Table S3 gives an overview of the publications discussing the biological role(s) of the putative lectins in *A. thaliana*.

### Class V Chitinase-Related Agglutinin Homologs

Nine proteins with a CRA domain were found in the *A. thaliana* genome (Table 1). All these putative lectin genes are located in one tandem duplication cluster on chromosome 4 (Fig. 1). The size of the chitinase-related domain of these sequences varies between 210 and

---

**Table 2. Putative lectin sequences with a signal peptide and transmembrane region.**

<table>
<thead>
<tr>
<th>Lectin family†</th>
<th>Signal peptide</th>
<th>Transmembrane domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRA</td>
<td>2/9</td>
<td>1/9</td>
</tr>
<tr>
<td>EUL</td>
<td>0/1</td>
<td>0/1</td>
</tr>
<tr>
<td>GNA</td>
<td>47/49</td>
<td>33/49</td>
</tr>
<tr>
<td>Hevein</td>
<td>10/10</td>
<td>0/10</td>
</tr>
<tr>
<td>Jacalin</td>
<td>3/50</td>
<td>0/50</td>
</tr>
<tr>
<td>Legume lectin</td>
<td>52/54</td>
<td>46/54</td>
</tr>
<tr>
<td>LysM</td>
<td>10/12</td>
<td>7/12</td>
</tr>
<tr>
<td>Nictaba</td>
<td>0/30</td>
<td>0/30</td>
</tr>
<tr>
<td>Ricin B</td>
<td>1/2</td>
<td>0/2</td>
</tr>
</tbody>
</table>

† CRA, class V chitinase-related agglutinin; EUL, *Euonymus europaeus* lectin; GNA, *Galanthus nivalis* agglutinin; LysM, Lysin Motif; Nictaba, *Nicotiana tabacum* agglutinin.
349 amino acids, with sequence identities to the CRA domain of the model sequence (Supplemental Table S1) varying between 41.29 and 55.06%. Next to their CRA domain, seven out of nine proteins also contain a chitinase insertion domain (CID) (Li and Greene, 2010), a domain typical for chitinases that belong to the glycosyl hydrolase (GH) family 18 (CAZy database). Therefore, all nine proteins are classified in the GH family 18 according to this CAZy database. Two out of nine sequences with a CRA domain encode a signal peptide, and one of the two proteins contains a trans-membrane domain (Table 2; Fig. 2). Transcript levels for one homolog AtChiC (AT4G19810, containing a signal peptide) were slightly upregulated after abscisic acid (ABA), jasmonic acid, and salt treatment (Ohnuma et al., 2011). The recombinant AtChiC showed enzymatic activity. In particular, it hydrolyzed N-acetylglucosamine oligomers (chitinase activity). By definition, a lectin domain should not exert any catalytic activity; therefore, this protein cannot be referred to as a lectin.

**Euonymus europaeus Lectin Homologs**

Only one EUL homolog was retrieved from the *A. thaliana* genome, referred to as ArathEULS3 (AT2G39050) (Fouquaert et al., 2009). The EUL domain of 154 amino acids is preceded by an N-terminal domain of unknown function (Fig. 2). The sequence does not contain a signal peptide or a trans-membrane domain (Table 2; Fig. 2). Microscopic analysis of an EGFP-fusion protein revealed that ArathEULS3 is located in the cytoplasm and the nucleus (Van Hove et al., 2011). More recent experiments showed an elevated expression of ArathEULS3 after treatments of *Arabidopsis* seedlings with glutathione, ABA, methyl jasmonate, and salt (Hacham et al., 2014; Van Hove et al., 2014). Furthermore, Van Hove et al. (2015) revealed
increased levels of the lectin transcripts after infection of wild-type Arabidopsis plants with Pseudomonas syringae. It was suggested that ArathEULS3 plays a role in the ABA-induced stomatal closure (Van Hove et al., 2015).

**Galanthus nivalis Agglutinin Homologs**

The GNA homologs represent one of the three largest lectin families in A. thaliana (Fig. 2). Only 6 out of 49 GNA homologs contain only a GNA domain; all other protein sequences have a chimeric domain architecture. The GNA domain is combined with an S-locus glycoprotein domain and/or a Pan/Apple domain and/or a protein kinase domain. In addition, some GNA homologs possess an S-locus receptor kinase (SRK) domain (Fig. 2). The domain architecture of one GNA homolog (AT1G11300) contains a tandem repeat of a GNA, S-locus glycoprotein, Pan/Apple, and protein kinase domain (not shown). Almost all GNA homologs are synthesized with a signal peptide, and the majority of them also contain a transmembrane domain (Table 2; Fig. 2).

RNA sequencing data suggest the inducible expression for some GNA homologs (AT1G65790, AT5G60900) after inoculation of Arabidopsis with the fungal pathogen Fusarium oxysporum (Zhu et al., 2013). The expression of two GNA homologs (AT5G60900 and AT5G18470) is upregulated in plants exposed to lipopolysaccharides (Sanabria et al., 2008). Lipopolysaccharides are essential components of the bacterial cell wall, suggesting that these GNA homologs play a role in the defense against bacterial infections. All but one of the GNA homologs (AT5G18470) that show elevated expression levels after Fusarium infection as well as these upregulated by the lipopolysaccharide treatment encode receptor-like kinases (RLKs). Lectin RLKs play a role in plant development, stress, and hormonal responses (Vaid et al., 2013). Another RLK, AT4G12390, is more than 200-fold upregulated after ozone treatment (Xu et al., 2015). Blaum et al. (2014) reported a GNA homolog, AT1G61360, that is coexpressed with BAK1-interacting receptor–like kinase 2, a protein of the leucine-rich repeat receptor–like kinases (RLKs). Leucine-rich repeat receptor–like kinases play a role in development and innate immunity. A GNA homolog called calmodulin-binding receptor–like protein kinase 1 (CBRLK1, AT1G11350) harbors a Ca\(^{2+}\)-dependent calmodulin binding domain in its C-terminus. This protein possesses autophosphorylation sites, as determined with mass spectrometry (Kim et al., 2009). All these results suggest that GNA homologs can play a role in different stress-related responses.

**Hevein Homologs**

Ten hevein homologs were found in the A. thaliana genome (Table 1). All sequences contain a signal peptide but lack a transmembrane domain (Table 2; Fig. 3), suggesting that all hevein homologs are synthesized following the secretory pathway. All hevein homologs represent chimeric lectin sequences (Fig. 3). The hevein domain is linked to a chitinase IV domain of the GH family 19 in nine sequences.

In pathogenesis-related (PR)-4 (AT3G04720), the hevein domain is fused to the Barwin domain of 120 amino acids, named after the Barwin protein of barley, which plays a role in the defense against fungal attacks (Ludvigsen and Poulsen, 1992). Genes that encode the PR proteins are rapidly induced after pathogen attacks and treatment with certain hormones. Recent studies revealed that the expression of PR genes is also regulated by environmental factors like light and abiotic stresses. For example, transcript levels for PR-3 (AT3G12500), a hevein homolog without Barwin domain, and PR-4 are significantly higher after high salt treatment and respond in an ABA-dependent manner (Seo et al., 2008). These genes are also responsive to sulfur dioxide exposure and are involved in the defense mechanism against the fungal pathogen Alternaria brassicicola (Li and Yi, 2012; Mukherjee et al., 2010; Thomma et al., 1999). Price et al. (2015) studied the hevein domain of some class IV chitinases from Arabidopsis by matrix-assisted laser desorption/ionization–time of flight mass spectrometry analysis and revealed the presence of three conserved disulfide bridges in the hevein domain as reported for the model lectin hevein. A whole genome microarray revealed that the expression of another hevein homolog (AT3G54420), called AtEP3, is upregulated after treatment of Arabidopsis with nitric oxide (Parani et al., 2004) and plays a role in programmed cell death (Passarinho et al., 2001).

**Jacalin Homologs**

The jacalin homologs represent a large group of putative lectins in A. thaliana. In contrast to many other lectin families in Arabidopsis, most of the jacalin homologs are composed of one or more jacalin domains only (Fig. 3). Only 6 of the 50 jacalin homologs are chimeric lectins. Four of them contain two to four Kelch motifs, a 44- to 56-amino acid motif that first was discovered in Drosophila and forms a single four-stranded antiparallel β-sheet. Proteins containing Kelch repeats play a role in many aspects of cell function (Adams et al., 2000). The remaining two chimeric jacalin homologs contain an F-box–associated domain (type 1) and/or an F-box domain. The F-box motif consists of approximately 60 amino acids and links the F-box protein to the SCF complex involved in protein degradation (Kipreos and Pagano, 2000). Only three of the JRLs contain a signal peptide, and none of them has a transmembrane domain, suggesting that most of the jacalin homologs are nucleocytoplasmic proteins (Table 2; Fig. 3).

Because ppb1 mutants show yellow phenotypes, de Luna-Valdez et al. (2014) suggested that the jacalin homolog PPB1 (PYK10 binding protein 1, AT3G16420) is involved in the development of chloroplasts. PPB1 may also act as a molecular chaperone that helps the correct polymerization of PYK10 in response to tissue damage and destruction of subcellular structures. PYK10 is a β-glucosidase of the myrosinase family (Nagano et al., 2005). β-glucosidases hydrolyze glucosinolates, a group of secondary plant metabolites that play a key role in the
myrosinase–glucosinolate plant defense system. This hydrolysis only takes place on tissue damage because the myrosinases and glucosinolates have different subcellular localizations. The product of the hydrolysis is an aglucone that can spontaneously rearrange into an isothiocyanate. Isothiocyanates are toxic for microorganisms, nematodes, and insects (Nagano et al., 2005; Wittstock and Burow, 2010). Two jacalin homologs are myrosinase-binding proteins (MBPs), referred to as MBP1 (AT1G52040) and MBP2 (AT1G52030) (Capella et al., 2001). These MBPs interact with myrosinases to form large complexes (Nagano et al., 2005). Four other jacalin homologs are known as A. thaliana nitril specifier proteins (AtNSP1–4, respectively: AT3G16400, AT2G33070, AT3G16390, AT3G16410). They catalyze the formation of nitriles from the aglucone (Kong et al., 2012). The jacalin homolog AtJAC1 (AT3G16470) plays a role in controlling the flowering time of A. thaliana. AtJAC1 regulates the expression of the repressor gene flowering locus C through interaction with glycine-rich RNA-binding protein 7 and as such influences the flowering time (Xiao et al., 2015). The jacalin homolog–restricted tobacco etch potyvirus movement 1 (RTM1, AT1G05760) restricts the long-distance movement of tobacco etch potyvirus in the phloem by acting together with the nonlectin proteins RTM2 and RTM3. As such, the plant can prevent systemic infection (Chisholm et al., 2001). The RTM system can also act toward the plum pox virus and the lettuce mosaic virus (Decroocq et al., 2006; Revers et al., 2003). The expression of the jacalin homolog encoded by AT1G52000 is upregulated after the inoculation of Arabidopsis with the fungal pathogen Fusarium oxysporum (Zhu et al., 2013).

**Legume Lectin Homologs**

The legume lectin homologs represent the largest group of putative lectins in A. thaliana (Table 1). Although 54 homologs are found, their domain architecture is relatively simple (Fig. 4). Only 13 of the 54 homologs contain
the legume lectin domain alone, and 41 legume lectin homologs contain a protein kinase domain C-terminally linked to the legume lectin domain. Except for two legume lectin homologs, all putative legume lectins contain a signal peptide, and 46 of them also contain a transmembrane domain (Table 2). All homologs that contain a protein kinase domain are synthesized with a signal peptide and contain a transmembrane domain.

Mutation of the SGC Lectin RLK (AT3G53810), a legume lectin homolog with a protein kinase domain, resulted in Arabidopsis plants with male sterility as a result of a defect in pollen development (Wan et al., 2008). Several legume lectin homologs (all with a protein kinase domain) are stress-related proteins. The expression of the legume lectin homolog LecRK-b2 (AT1G70130), for example, was upregulated by ABA, salt, and osmotic stress (Deng et al., 2009). He et al. (2004) demonstrated that the expression of AtLecRK2 (AT3G45410) is elevated on salt stress and that this response is regulated by the ethylene signaling pathway. The expression of legume lectin homolog AtLPK1 (AT4G02410) is highly upregulated after ABA, methyl jasmonate, and salicylic acid treatments. Overexpression of this legume lectin homolog in A. thaliana showed a better seed germination under high salt conditions (Huang et al., 2013). The legume lectin homolog named LecRK-VI.2 or LecRKA4.1 (AT5G01540) plays a role in the ABA stress response and the disease resistance of Arabidopsis against Pseudomonas syringae and Pectobacterium carotovorum (Singh et al., 2012; Xin et al., 2009). Also, legume lectin homologs LecRKA4.2 (AT5G01550) and LecRKA4.3 (AT5G01560) have a role in the ABA stress response. These legume lectin homologs are located next to each other on chromosome 5 (Xin et al., 2009). The expression of some of these legume lectin homologs is also influenced by A. brassicicola, F. oxysporum, and ozone (Mukherjee et al., 2010; Xu et al., 2015; Zhu et al., 2013). LecRK-I.9 (AT5G60300) plays a role in the cell wall–plasma membrane adhesions. The destabilization of these adhesions is one of the ways Phytophthora brassicae infects Arabidopsis. Overexpression of LecRK-I.9 in A. thaliana resulted in enhanced resistance to P. brassicae, suggesting that LecRK-I.9 contributes to strengthening of cell wall–plasma membrane adhesions (Bouwmeester et al., 2011).

**Lysin Motif Homologs**

The LysM homologs represent a small group in A. thaliana with diverse domain architectures (Fig. 4). In 5 out of 12 sequences, the LysM domain is linked to a protein kinase domain.
domain. One of these sequences contains two LysM domains. According to InterProscan, one homolog with a protein kinase domain possesses a syndecan/neurexin domain. Although this homolog is well described in the literature, the syndecan/neurexin domain has never been reported. One of the 12 LysM homologs contains an N-terminal F-box domain. Half of the sequences encoding LysM homologs consist of one or two LysM domains. No other protein domains were identified in these coding sequences. Ten LysM homologs are synthesized with a signal peptide, and seven of them contain a transmembrane domain, indicating that most of the LysM homologs follow the secretory pathway (Table 2).

Zhang et al. (2007) described five of the LysM homologs that contain a protein kinase domain. One of these homologs, referred to as AtCERK1, AtLYK1, or LysM RLK1 (AT3G21630), is important in the response of Arabidopsis to fungi. Mutation of RLK1 created more susceptible Arabidopsis plants to fungal pathogens (Wan et al., 2008). More recent research revealed that the expression of AtLYK4 (AT2G23770) and AtLYK5 (AT2G33580) is upregulated by chitin. Pull-down analysis also proved that these LysM homologs can interact with chitin. Only mutants of AtLYK4 showed a reduced expression of chitin-responsive genes (like atlyk1 mutants) and a higher susceptibility to Alternaria brassicicola and Pseudomonas syringae. Because AtLYK4 does not possess an active protein kinase domain, it is hypothesized that AtLYK1 and AtLYK4 may form a chitin–receptor complex with a single active kinase domain (AtLYK1) that starts the downstream chitin signaling (Wan et al., 2012). Furthermore, AtLYK3 (AT1G51940) may be important for the cross talk between the ABA and pathogen stress response (Paparella et al., 2014).

**Nicotiana tabacum Agglutinin Homologs**

Among the 30 Nicotaba homologs retrieved from the Arabidopsis genome, 19 sequences contain an F-box domain (Fig. 5). Four Nicotaba homologs contain an N-terminal Toll/interleukin-1 receptor (TIR) domain, a domain with a role in pathogen detection and defense responses (Burch-Smith and Dinesh-Kumar, 2007). Only one putative Nicotaba lectin contains an avirulence induced gene 1 (AIG1) type G domain, a domain that is also found in GTPases that play a role in the defense against pathogens (Reuber and Ausubel, 1996). The absence of signal peptides or transmembrane domains suggests a nucleocytoplastic localization for the Nicotaba homologs.

Phloem protein 2-A1 (AT4G19840), a Nicotaba homolog that contains only a Nicotaba domain, is part of the phloem protein bodies in the sieve elements. Recombinant protein production and glycan array analysis demonstrated binding of PP2-A1 to N-acetylglucosamine oligomers, high-mannose N-glycans, and 9-acyl-N-acetyleneuraminic sialic acid. Although PP2-A1 did not show insecticidal properties against Acyrthosiphon pisum and Myzus persicae, the weight gain of the nymphs was reduced by adding the recombinant PP2-A1 protein to the aphid diet (Beneteau et al., 2010). According to Lee et al. (2014), PP2-A1 showed molecular chaperone as well as antifungal activity. Because the expression of PP2-A1 is upregulated after pathogen attack and ethylene treatment, the molecular chaperone can play a crucial role in the stress response (Lee et al., 2014). More than half of the F-box–containing Nicotaba homologs were shown to interact with at least one Arabidopsis Skp1-like protein (Dezfuiian et al., 2012; Gagne et al., 2002; Kuroda et al., 2012; Risseeuw et al., 2003; Takahashi et

---

**Fig. 5.** Domain architectures of Nicotiana tabacum agglutinin (Nicotaba) and ricin B homologs. Signal peptides are drawn in yellow. The numbers in brackets indicate the total number of sequences with this domain architecture, followed by the number of sequences that contain a signal peptide and a transmembrane region, respectively. Where only one number is given, none of the sequences with this domain architecture has a signal peptide or a transmembrane region.
Many F-box–containing Nictaba homologs are proven to be stress inducible, mostly by abiotic stresses but to a lesser extent also by biotic stresses (Dezfuijan et al., 2012; Kuroda et al., 2012; Takahashi et al., 2004). Transcript levels for the VIP1 binding F-box protein (AT1G56250) are upregulated by Agrobacterium tumefaciens and help to bring the transfer DNA inside the plant by degrading VirE2 and VIP1 proteins that coat the transfer DNA (Zaltsman et al., 2010). The expression of PP2-B11 (AT1G80110) is elevated by salt stress, and overexpression lines of this gene are more tolerant to high-salinity conditions (Jia et al., 2015). Lee et al. (2014) showed that overexpression lines of PP2-B11 are more sensitive to drought stress, suggesting that PP2-B11 is a negative regulator in the response to drought stress. Carbohydrate-binding activity was reported for the recombinant PP2-B10 or F-box Nictaba (AT2G02360) protein expressed in Pichia pastoris (Stefanowicz et al., 2012). Glycan array analysis demonstrated that F-box Nictaba recognizes N-acetyllactosamine, Lewis A, Lewis Y, and blood type B motifs. It is hypothesized that F-box Nictaba plays a role in nucleocytoplasmic protein degradation (Lannoo et al., 2008). Recently Stefanowicz et al. (2016) showed that overexpression of this F-box Nictaba resulted in a reduction of leaf damage on infection with Pseudomonas syringae. Another Nictaba homolog (AT3G61060) with an F-box domain is downregulated during callus initiation (Xu et al., 2012).

**Ricin B homologs**

Only two ricin B homologs exist in A. thaliana (Table 1). They have the same domain architecture, consisting of an N-terminal GH family 5 domain and a C-terminal ricin B domain (Fig. 5). Only one homolog is synthesized with a signal peptide, and none of the sequences has the characteristics of a transmembrane domain (Table 2). The expression of one of the ricin B homologs (AT3G26140) is upregulated after infection of the plants with plum pox virus (Babu et al., 2008).

**Phylogenetic Analysis and Analysis of the Carbohydrate Binding Site**

Screening of the Arabidopsis genome for lectin sequences and analysis of the protein domains yielded different protein domain combinations composed of a lectin domain linked to the AIG1-type G, Barwin, CID, F-box, F-box–associated domain type 1, GH family 5, GH family 19, Kelch1, Pan/Apple, protein kinase, S-locus glycoprotein, SRK, or TIR domain. Strikingly, the protein kinase and the F-box domains are found in combination with different lectin domains. All the putative lectins with a protein kinase domain are synthesized with a signal peptide, and almost all of them have a predicted transmembrane region, suggesting that these chimeric proteins follow the secretory pathway and probably reside in the plasma membrane or in the apoplast. All putative lectin sequences with an F-box domain do not possess a signal peptide or a transmembrane region and presumably encode nucleocytoplasmic proteins. Although the role of the F-box domain is known, research on genes containing this F-box domain fused to a lectin domain (either a JRL, a LysM, or a Nictaba domain) is far from being fully implemented. No information is available regarding the physiological importance of the F-box JRL and the F-box LysM, but several F-box Nictaba proteins were shown to be stress-related proteins and as such may play a role in plant defense. To check for evolutionary relationships between F-box–lectin domain combinations, phylogenetic trees were created for each lectin family. When a sequence contained multiple lectin domains, they were separated. For instance, the first jacalin domain of AT1G19715.1 is designated AT1G19715.1_1. In addition, the carbohydrate binding site of all the lectin domains was analyzed to check the conservation of the amino acids responsible for interaction with the carbohydrate.

**Evolutionary Relationships of the Jacalin Homologs from Arabidopsis thaliana**

The family of JRLs represents a large group of putative lectins in A. thaliana. Most sequences consist of one or more jacalin domains, but only 12% of the sequences encode chimeric proteins. The dendrogram constructed with the individual jacalin domain sequences from all putative JRLs yielded a very complex tree. The sequences containing the F-box–associated domain type 1 (AT3G59590.1 and AT3G59610.1) are very closely related (Fig. 6). All JRL domains associated with two or more Kelch1 domains are grouped together except for the N-terminal jacalin domain of the only Kelch1 JRL sequence containing two jacalin domains (AT3G16410.1_1). The bootstrap value of the branch is quite high (93%), suggesting that the jacalin domain sequences that occur in combination with Kelch1 domains evolved together, whereas the first jacalin domain of AT3G16410.1 is more similar to the jacalin domains of the nonchimeric JRLs.

Most JRLs composed of a single jacalin domain are separated in two small clusters (Fig. 6). The majority of the lectin sequences contain multiple jacalin domains. Overall, two big clades of jacalin domains belonging to sequences with multiple JRL domains can be distinguished (Fig. 6). Clade A (blue) contains all the first jacalin domains from most nonchimeric JRLs and all second jacalin domains except from one sequence belonging to the JRLs containing four jacalin domains. Clade B (purple) contains all the other jacalin domains of the nonchimeric JRLs and the jacalin domains of the F-box–associated domain type 1 and the Kelch1-containing sequences. Jacalin domains from JRLs containing multiple lectin domains usually do not cluster together in the tree, except for the three jacalin domains of AT1G19715.1. These data suggest that tandemly arrayed jacalin domains within one JRL are not formed by a duplication event. Three JRLs composed of two tandem jacalin domains are synthesized with a signal peptide. The first and second jacalin domains of these three sequences also group
Multiple small tandem duplication clusters are found throughout the tree, grouping closely related jacalin domain sequences (Fig. 1, 6). Together (Fig. 6). Multiple small tandem duplication clusters are found throughout the tree, grouping closely related jacalin domain sequences (Fig. 1, 6).

Conserved Amino Acids in the Carbohydrate Binding Site of Jacalin Homologs from Arabidopsis

The first JRL, referred to as jacalin, was identified from the seeds of jackfruit Artocarpus integrifolia (Bunn-Moreno and Campos-Neto, 1981). Jacalin is composed of four subunits—two α-chains (133 amino acids) and two β-chains (20 amino acids)—and exhibits specificity toward galactose (Houlès Astoul et al., 2002; Sankaranarayanan et al., 1996). Since the discovery of jacalin, many jacalin homologs with specificity toward galactose and mannose have been identified throughout the plant kingdom. In contrast to the galactose-specific jcalins that are confined to the Moraceae, the mannose-specific jcalins are widespread in higher plants (Houlès Astoul et al., 2002). In addition to the difference in carbohydrate binding specificity, galactose- and mannose-specific jcalins differ in the maturation of the lectin polypeptide and their localization in the cell. The post-translational cleavage of the lectin precursor into the α- and β-chain does not take place in the mannose-specific jcalins. Whereas galactose-specific JRLs reside in the vacuole, mannose-specific JRLs are cytoplasmic proteins (Peumans et al., 2000). Houlès Astoul et al. (2002) reinvestigated the carbohydrate binding specificity of jacalin and concluded that the specificity of jacalin is not restricted to galactose and N-acetylgalactosamine but...
extends to mannose and glucose. They reported that the carbohydrate binding site of jacalin consists of an N-terminal glycine and three C-terminal amino acids (Tyr, Trp, and Asp). However, only the N-terminal glycine and the C-terminal aspartic acids were found in the WebLogo resulting from the alignment of 127 jacalin domain sequences from Arabidopsis (Fig. 7A). Judging from the height of the residue, it can be concluded that the N-terminal glycine is conserved for most of the sequences encoding jacalin homologs from Arabidopsis, whereas the C-terminal aspartic acid is not well conserved. The complete WebLogo shows a lot of gaps in the alignment, making it difficult to interpret the data (Supplemental Fig. S1).

Bourne et al. (2004) reported the residues important for carbohydrate binding in the Calystegia sepium agglutinin, a mannose-specific JRL. The carbohydrate binding site consists of an N-terminal glycine (Gly17), Asn96, Tyr141, Tyr142, and Asp144. These residues were not conserved in the JRLs from Arabidopsis. The jacalin domains from Arabidopsis have a well-conserved leucine residue preceding the C-terminal aspartic acid needed for carbohydrate binding (Fig. 7A). Although this residue was not shown to be important for carbohydrate binding, Raval et al. (2004) also noticed that this leucine is largely conserved in 58 sequences of individual jacalin domains from JRLs belonging to 16 different plant species.

**Phylogenetic Analysis of the Lysin Motif Domain Family in Arabidopsis thaliana**

The LysM family represents a small group of putative lectins in Arabidopsis. In addition to the LysM domain, only two protein domains have been identified: an F-box domain and a protein kinase domain. In the phylogenetic tree constructed with the sequences of the LysM domains only, the LysM domain that is associated with the F-box...
domain forms a separate group with the LysM domain from AT5G23130.1 (Fig. 8), representing the only LysM sequence that is synthesized without a signal peptide (similar to F-box–LysM sequence). This LysM domain is linked to a C-terminal sequence with unknown function. The LysM domains that are associated with the protein kinase domains are found in different branches of the tree. The N-terminal LysM domains of the three sequences with a tandem array of LysM motifs cluster together in one branch of the tree (Fig. 8).

Conservation of the Carbohydrate Binding Site in the Lysin Motif Homologs from Arabidopsis

Lysin Motif domains are very short lectin motifs of approximately 40 amino acids. The first LysM domain was discovered in the lysozyme from Bacillus phage φ29 (Garvey et al., 1986). In plants, most LysM proteins for which carbohydrate binding activity is known belong to the group of LysM receptor kinases. These LysM modules can bind peptidoglycans from bacterial pathogens and chitin fragments from fungal pathogens (Lannoo and Van Damme, 2014). Plant LysM modules recognize chitin fragments (GlcNAc)_n, when the degree of polymerization is higher than five (n ≥ 5) (Petutschnig et al., 2010). Kitaoku et al. (2016) investigated the (GlcNAc)_n binding site of a LysM domain containing protein from the green algae Volvox carteri composed of two N-terminal LysM domains (96% sequence homology) and a C-terminal catalytic domain. The binding of chitin fragments to the second LysM domain involved a hydrophobic interaction between Trp96 of the LysM domain and the pyranose ring of (GlcNAc)_n. Nuclear magnetic resonance–based titration experiments revealed that other amino acid residues were important to form the carbohydrate binding site: Gly92, Asp93, Thr94, Phe95, Ala97, Ile98, Ala99, Gln100, Ala119, Arg120, Leu121, Gln122, and Gly124 (Kitaoku et al., 2016). Only residues Asp93, Thr94, Ile98, and Ala99 were conserved in the LysM domains from A. thaliana (Fig. 7B).

Phylogenetic Analysis of the Nicotiana tabacum Agglutinin Family in Arabidopsis thaliana

The Nictaba family groups 13.8% of all putative lectins from A. thaliana. Most Nictaba-related sequences contain an N-terminal F-box domain linked to a Nictaba domain, a domain architecture that is also abundant.

---

Fig. 8. Maximum likelihood phylogenetic tree of Lysin Motif (LysM) domains. The domain architecture of every putative lectin is represented using the same color code as in Fig. 4. The purple frame marks the N-terminal LysM domains of the three sequences with two LysM motifs. The bootstrap values are given at the corresponding branches. The scale bar represents the mean of the number of substitutions per site according to a maximum likelihood estimation.
The phylogenetic tree built from the Nictaba sequences can be divided in two big clades (A and B) (Fig. 9). Clade B mainly contains F-box Nictaba sequences. Only AT2g02280.1 does not possess an F-box domain, but it has an N-terminal region consisting of less than 10 amino acids preceding the Nictaba domain. Maybe At2g02280.1 lost its F-box domain during evolution. Clade A clusters all Nictaba sequences that occur in combination with the TIR domain or the AIG1-type G domain, the Nictaba sequences composed only of a Nictaba domain and a small group of F-box Nictaba homologs. This clustering suggests that the Nictaba domains of the nonchimeric sequences are more closely related to the TIR and AIG1-type G Nictaba sequences. Three out of four TIR Nictaba sequences occur in a tandem duplication on chromosome 5 and group together in the dendrogram (Fig. 9). The subgroup of F-box Nictaba sequences in clade A differs from most F-box Nictaba sequences in clade B in that the latter sequences belong to two tandem duplication clusters on chromosomes 1 and 2 (Fig. 9).

**Conserved Amino Acids in the Carbohydrate Binding Site of Nicotiana tabacum Agglutinin Homologs from Arabidopsis**

Three-dimensional modeling of the Nictaba sequence from *Nicotiana tabacum* suggested that the amino acids Trp15, Trp22, Gliu138, and Gliu145 of Nictaba are important for interaction with the carbohydrate. Mutational analysis indicated that only the tryptophan residues play an important role in the carbohydrate binding site of Nictaba (Schouppe et al., 2010). The WebLogo made for 31 Nictaba domain sequences from *Arabidopsis* revealed that the four residues identified for the lectin from tobacco are quite conserved (Fig. 7C). However, as can be observed in the complete WebLogo (Supplemental Fig. S2), there are gaps in the alignment. It should also be noted that the conserved residues in the Nictaba sequences from *Arabidopsis* are similar to the ones reported for the Nictaba domains from soybean (*Glycine max* (L.) Merr.) (Van Holle et al., 2017). The fact that several amino acids in the binding site are conserved does not allow conclusions to be drawn with respect to the carbohydrate-binding activity of these proteins. Stefanowicz et al. (2012) reported that the specificity of F-box Nictaba (At2G02360) and Nictaba are different despite the fact that the amino acids responsible for carbohydrate binding are conserved. These results suggest that other amino acids in the vicinity of the binding site also play a role in the conformation of the binding site and its interaction with glycans.

**Conclusions**

A total of 217 putative lectin sequences were identified in the genome of *A. thaliana*. Analysis of the domain architectures for each sequence revealed that most sequences contain multiple protein domains. Most of the known protein domains associated with lectin motifs have been...
reported to be involved in stress signaling, development, and defense. Although some domain architectures are unique to a specific lectin domain (AIG1-type G, Barwin, CID, F-box–associated domain type 1, GH family 5, GH family 19, Kelch1, Pan/Apple, S-locus glycoprotein, SRK, and TIR domain), we also retrieved some protein domains that are associated with multiple lectin motifs, in particular the F-box and protein kinase domain.

Judging from the absence or the presence of signal peptides and/or transmembrane regions in the lectin sequences, it is obvious that the putative lectins will end up in different locations in the cell. Taking into account the ambiguity in lectin specificity for many carbohydrate-binding recognition domains, our data do not allow drawing conclusions with respect to the activity of the carbohydrate binding site. However, literature data confirm that at least some lectin motifs exert carbohydrate binding activity and are involved in protein–carbohydrate interactions.

Our findings show that Arabidopsis plants have at their disposal a range of proteins with lectin motifs. These putative lectins are located in different locations in the cell and in different plant tissues and can exert complementary activities. It can be envisaged that these lectins play an important role in plant development and survival under stress conditions.

**Supplemental Information**

Supplementary Table 1. Identification numbers, databases, Pfam ID numbers, and Pfam names of the model sequences for each lectin family.

Supplementary Table 2. Gene names, transcript names, chromosome positions, known protein domains, and protein sequences for all putative lectins in Arabidopsis thaliana.

Supplementary Table 3. Literature related to putative lectins from Arabidopsis thaliana.

Supplementary Fig. 1. Complete WebLogo of the jacalin domains.

Supplementary Fig. 2. Complete WebLogo of the Nictaba domains.

**Acknowledgments**

This research was funded by the Research Council of Ghent University (project BOF15/GOA/005) and the Fund for Scientific Research–Flanders (FWO Grant G006114N). The authors thank Yves Van de Peer and Zheng Li (Bioinformatics and Evolutionary Biology, VIB-Ghent University) for discussions and advice.

**References**


