A Comprehensive Classification and Evolutionary Analysis of Plant Homebox Genes

Krishanu Mukherjee,*†† Luciano Brocchieri,*1 and Thomas R. Bürglin†††1

*Genetics Institute, Department of Molecular Genetics and Microbiology, College of Medicine, University of Florida; †Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden; ‡School of Life Sciences, Södertörn University, Sweden

The full complement of homeobox transcription factor sequences, including genes and pseudogenes, was determined from the analysis of 10 complete genomes from flowering plants, moss, Selaginella, unicellular green algae, and red algae. Our exhaustive genome-wide searches resulted in the discovery in each class of a greater number of homeobox genes than previously reported. All homeobox genes can be unambiguously classified by sequence evolutionary analysis into 14 distinct classes also characterized by conserved intron–exon structure and by unique codomain architectures. We identified many new genes belonging to previously defined classes (HD-ZIP I to IV, BEL, KNOX, PLINC, WOX). Other newly identified genes allowed us to characterize PHD, DDT, NDX, and LD genes as members of four new evolutionary classes and to define two additional classes, which we named SAWADEE and PINTOX. Our comprehensive analysis allowed us to identify several newly characterized conserved motifs, including novel zinc finger motifs in SAWADEE and DDT. Members of the BEL and KNOX classes were found in Chlororibionta (green plants) and in Rhodophyta. We found representatives of the DDT, WOX, and PINTOX classes only in green plants, including unicellular green algae, moss, and vascular plants. All 14 homeobox gene classes were represented in flowering plants, Selaginella, and moss, suggesting that they had already differentiated in the last common ancestor of moss and vascular plants.

Introduction

Homeobox genes encode a typical DNA-binding domain of 60 amino acids, known as homeodomain (HD), that characterizes a large family of transcription factors. The homeodomain folds into a characteristic 3D structure containing three alpha-helices, of which the second and third form a helix-turn-helix motif. The first homeobox genes were isolated from the fruit fly Drosophila melanogaster and were subsequently found to be involved in many aspects of development (for review, see Gehring et al. 1994; Bürglin 2005). Many more homeobox genes have been subsequently identified from all major eukaryotic lineages (Derelle et al. 2007). Based on sequence differentiation and fusion with characteristic codomain sequences, animal homeodomain proteins have been classified into several distinct classes, including TALE, Antp, PRD, SIX, LIM, POU, ZF, CUT, HNF, and PROS (Bürglin 1994, 2005; Holland et al. 2007; Takatori et al. 2008). A distinction has been made between “typical” homeodomains, characterized by a length of 60 amino acids, versus “atypical” ones of different lengths (Bürglin 1994). The latter include a group characterized by homeodomains of 63 aa, with three extra residues inserted between helix 1 and 2 (Bürglin 1995), that have been named TALE (Three Amino acid Loop Extension) homeobox genes (Bertolino et al. 1995; Chen et al. 2003). Both TALE and typical homeobox genes were found to be present in all major eukaryotic lineages including plant, fungi, and animals, suggesting that these two types of homeobox were present in the eukaryote ancestor (Bürglin 1995; Bharathan et al. 1997; Bürglin 1997, 1998a; Derelle et al. 2007). Besides TALE, other homeobox genes of noncanonical length have emerged from the analysis of animal sequences (Bürglin 1997; Bürglin and Cassata 2002). These can be clustered into separate classes of atypical homeobox genes characterized by unique homeodomain insertions and by class-specific codomain architectures, making it apparent that insertions in homeodomain loops have independently occurred multiple times in evolution.

Plant homeodomain proteins have been classified in the literature into various groups based on sequence similarity of their homeodomains and on the presence of characteristic codomains. Bharathan et al. (1997) classified them into seven classes: KNOX and BEL, belonging to the TALE superclass (Bürglin 1997), ZM-HOX, HAT1, HAT2, ATHB8, and GL2. The HAT1, HAT2, ATHB8, and GL2 genes are all characterized by a leucine-zipper motif downstream of the homeodomain (Ruberti et al. 1991) and have been successively renamed HD-ZIP I, HD-ZIP II, HD-ZIP III, and HD-ZIP IV, respectively (Bharathan et al. 1997; Meijer et al. 1997; Aso et al. 1999; Sakakibara et al. 2001). Chan et al. (1998) proposed an alternative classification into five groups (HD-ZIP, GLABRA, KNOTTED, PHD, and BEL).

Although many homeobox genes have been reported from plants, a complete survey and classification of all homeobox genes in plant species from disparate evolutionary groups is lacking. The completion of several high-quality plant genome sequencing projects provided us with the unique opportunity to make a complete assessment and thorough comparative analysis of the homeodomain proteins encoded in plants. The analysis of the full set of homeobox genes in genomes from diverse species allows for a definitive classification of plant homeodomain proteins and an assessment of their origins, evolutionary relations, patterns of differentiation, and proliferation in the various phylogenetic groups. We are interested in finding answers to the following questions: 1) What are the evolutionary relations among plant homeodomain proteins? 2) How many classes of plant homeodomain proteins can we distinguish in the complete collections from multiple genomes? 3) When did each class appear in evolutionary time? 4) What classes are present in which plant groups? 5) How did each class proliferate in the different plant groups? 6) Can we classify plant homeodomain sequences...
based on a characteristic pattern of insertions in homeodomain loops?; 7) Can we characterize homeodomain evolution based on intron–exon structure?; 8) What is the relation of plant homeodomain sequence evolution with the evolution of their codomain architecture?; 9) Can we relate homeobox gene and class presence and differentiation within each genome with the developmental properties of the species bearing them?

Our thorough analyses provide a comprehensive classification scheme and suggest a reconstruction of the evolutionary history of plant homeodomain protein codomain architecture applicable to all plant homeodomain proteins. The results of our study provide a clear picture of the patterns of origin and differentiation of homeodomain classes in different plant groups and suggest interpretations linking their appearance to organism differentiation and increasing developmental complexity. We recognize appearance of all present-day classes of plant homeodomain proteins early in land plant evolution and their differential proliferation into variable numbers of paralogs within each class in different plant groups. We identified and classified in this study a higher number of homeobox genes than previously reported, allowing us to identify several subclasses based on phylogenetic clustering and presence of subclass-specific motifs. We found that, compared with animal homeobox genes, insertions in the homeodomain loops are quite common among plant homeobox genes. We newly characterized several conserved codomain motifs, specific to each of the classes, and predicted their functionality based on sequence similarity to proteins available in databases. The resulting rich data set of plant homeodomain proteins, comprising over 1,000 sequences, is available for download.

Methods

Database Searches and Retrieval of Protein Sequences

Thorough TBlastN searches with several divergent homeodomain proteins of plants and animals were performed to retrieve homeobox genes of Arabidopsis through the TAIR database server (http://www.arabidopsis.org) and of rice through the Gramene database server (http://www.gramene.org). Likewise, analogous searches were performed for the maize genome, through the TAIR maize genome server (http://tigrblast.tigr.org/cgi/blast/PlantGDBblast). The plant genome database server was also used to retrieve homeobox genes of Arabidopsis through the TAIR database server (http://www.arabidopsis.org) and reformatted for the local Blast. In order to remove the redundancy in the database, each sequence in the database was blasted against the database, and redundant entries were consolidated. New rounds of BlastP and TBlastN searches of the nr protein and GenBank databases at NCBI restricted to Arabidopsis thaliana using default values were carried out using representative homeodomains of different classes from plants and animals as a query (e.g., POU, LIM, CUT, Antp, HD-ZIP, BEL [Bürglin 1994]). Hits from these searches were checked against data in the local Filemaker database using the “Search” feature (e.g., accession numbers, or N-terminal protein sequences), or a local Blast search of a retrieved hit was performed against the local database. New sequences identified in this fashion were added to the database. A preliminary Neighbor-Joining tree was generated, and PSI-Blast searches with a member of each evolutionary group used as a query were carried out against the nr protein database at NCBI, and iterations were stopped when no new sequences were detected. Every hit for the BlastP, TBlastN, and PSI-Blast searches was manually examined for its potential as a homeodomain, according to the criteria outlined in the text. No particular e value was taken as an automatic cut-off point, as the goal was to detect as many homeobox genes as possible. In a few instances, Blast matrix and expected value were relaxed to include additional sequences, which were then checked manually for their potential as homeodomain sequences. Arabidopsis thaliana ESTs at NCBI were searched using each entry of our local database as a query using TBlastN. Searches of newly discovered conserved domains/motifs linked to homeobox genes were carried out using BlastP with default values and without species restriction, unless the results would lead to too many hits (e.g., PHD, DDT). All A. thaliana homeodomain protein entries (mislabeled as Hox) in the SMART database were also checked against the local database.

For rice, maize, and poplar, the genomic sequences of the few thousand nucleotides upstream and downstream of the homeodomain were analyzed to predict the complete homeodomain protein sequences using either the MIT Genescan server or the softberry FGENESH+ server, using the most similar plant gene from the blast searches as a guide. In the multiple sequence alignment, if the homeodomain showed obvious misalignment, the most similar plant sequence was used as a guide to correct the homeodomain following a three-frame translation of the sequence and manual determination of potential intron and exon boundaries. A similar procedure was used in some sequences that showed obvious gaps and misalignments within conserved domains upon sequence alignment. In those cases, the genomic sequence was examined, and it was compared either as three-frame translation with a closely related protein sequence or as DNA against a closely related DNA sequence, using the dot matrix program PPCMatrix (Bürglin 1998b). The genomic sequence was translated in all three frames in PPCMatrix and examined for splice sites in the regions where the sequence similarity terminated. In the case of maize, part of the homeodomain could be found in one contig and the other part in a different contig. In this case, manual contig assemblies were carried out and finally confirmed by Blast searches. In several cases, Arabidopsis homeobox
genes were repredicted with FGENSEH+ using the most similar gene ortholog as a guide to obtain the full-length protein.

The European Molecular Biology Open Software suite was used for pairwise alignment and for locating the homeo-domain sequence within a protein sequence. Sequences were submitted also to the SMART and NCBI CDD (Schultz et al. 1998; Marchler-Bauer et al. 2003) to identify conserved domains. These databases did not contain or identify many of the conserved motifs and unusual homeodomain sequences.

The homeodomain DNA from representative members of each class were blasted against the JGI Chlamydomonas project Blast server (http://genome.jgi-psf.org/Chlre3/Chlre3.home.html) to retrieve the Chlamydomonas reinhardtii homeodomain sequences. Similarly, to recover the homeodomain protein sequences of Physcomitrella patens, TBLastN searches were performed at JGI (http://genome.jgi-psf.org/).

Multiple Sequence Alignment and Phylogenetic Analysis

The multiple sequence alignment tool MUSCLE (Edgar 2004) was used for multiple protein sequence alignment. Sequences were further edited and aligned manually, when necessary, using the “Seaview” multiple sequence editor (Galtier et al. 1996). For phylogenetic analyses of conserved domains, sequences were trimmed so that only the relevant protein domains remained in the alignment. Phylogenetic relationships were inferred using the maximum likelihood (ML) method as implemented in PHYML (Guindon and Gascuel 2003). For the ML trees, the JTT substitution model was used and results were evaluated with 100 or 1,000 bootstrap replicates. Use of alternative substitution models (WAG, LG) did not affect the results in any of the cases tested (results not shown). The generated trees were displayed using TREEVIEW 1.6.6 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) and NJPLOT by M. Gouy (http://pbil.univ-lyon1.fr/software/njpplot.html).

Secondary Structure Prediction and Protein Homology Modeling

We predicted the secondary structures of several divergent homeodomain sequences using the PHD, JPRED, and Predict Protein servers (Rost and Sander 1993; Cuff et al. 1998). For homology modeling, the crystal structure of the engrailed homeodomain (PDB id: 1enl) obtained from Protein Data Bank (PDB) was used as a template. The aligned sequences were submitted to SWISS-MODEL (http://www.expasy.org/swissmod/) to obtain the 3D structure of some of the atypical homeodomains. The model was viewed in Swiss-PDB Viewer (Kaplan and Littlejohn 2001), and the quality of the model was judged by the phi–psi angle represented in Ramachandran Plots.

Results

We searched homeodomain sequences in several plant genomes (table 1) including the genomes of the red alga Cyanidioschyzon merolae (Rhodophyta), the unicellular green algae Ostreococcus lucimarinus, Ostreococcus tauri and Chl. reinhardtii (Chlorophyta), the moss P. patens (Bryophyta), the spikemoss Selaginella moellendorffii (Lycopodiophyta), and the flowering plants A. thaliana, poplar (eudicots), maize and rice (monocots), using the TBLastN (Gertz et al. 2006) search engine with a variety of homeodomain sequences as queries. In most cases, whenever significant similarity to a homeodomain sequence was identified, the genomic sequence was excised and homology-based gene predictions were performed using the most similar query as a guide. In many instances, regions of conservation were manually reconstructed by multiple sequence alignment of three-frame translations of the genomic sequence including the Blast hits. The results of our extensive database searches are summarized in table 2. In the genome of Arabidopsis (eudicot), we found 110 unique homeobox genes, and in the genome of rice (monocot), we found 117 sequences (110 genes and 7 pseudogenes). Searches in the genomes of poplar (eudicot) and maize (monocot) each yielded 149 homeobox sequences (148 genes and 1 pseudogene) (table 2). In the genome of Selaginella (Lycopodiophyta), we found 45 homeobox genes and 1 pseudogene. In the genome of moss P. patens (Bryophyta), we identified 66 homeobox genes and 1 pseudogene. Among unicellular green algae (Chlorophyta), we found five homeobox genes in the genome of Chl. reinhardtii and seven homeobox genes in the genomes of O. lucimarinus and O. tauri. We retrieved a similar number of homeobox genes (six genes) from the genome of the red alga (Rhodophyta) Cyanidioschyzon merolae. An alignment of HD sequence representatives from one monocot (rice), one dicot (Arabidopsis), moss, and Selaginella, showing sequence features unique to each class, is shown in fig. 1. The alignment of all sequences found in Arabidopsis, moss, Selaginella, unicellular algae, and red algae is shown in supplementary fig. 1 (Supplementary Material online). See supplementary table 1 (Supplementary Material online) for a complete catalog of all sequences identified in the plant genomes analyzed in this study. Several groups of homeodomains do not fit into the typical 60-residue pattern, as they have insertions between helix 1 and helix 2, and/or

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Plant Genomes Analyzed in This Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Genome Size (Mbp)</td>
</tr>
<tr>
<td>Eudicots</td>
<td>Arabidopsis thaliana (thale cress)</td>
</tr>
<tr>
<td></td>
<td>Populus trichocarpa (poplar)</td>
</tr>
<tr>
<td>Monocots</td>
<td>Oryza sativa (rice)</td>
</tr>
<tr>
<td></td>
<td>Zea mays (maize)</td>
</tr>
<tr>
<td>Lycopodiophyta</td>
<td>Selaginella moellendorffii</td>
</tr>
<tr>
<td>Bryophyta</td>
<td>Physcomitrella patens</td>
</tr>
<tr>
<td>Chlorophyta</td>
<td>Chlamydomonas reinhardtii</td>
</tr>
<tr>
<td>Uncellular green algae</td>
<td>Ostreococcus lucimarinus</td>
</tr>
<tr>
<td></td>
<td>Ostreococcus tauri</td>
</tr>
<tr>
<td>Rhodophyta (red algae)</td>
<td>Cyanidioschyzon merolae</td>
</tr>
</tbody>
</table>
helix 2 and helix 3 (fig. 1). However, several criteria identify all these sequences as homeodomains: 1) The amino acid sequences fit the profile of amino acids established from 346 animal homeodomains (Bürglin 1994) when the loop regions are removed; 2) The pattern of conservation, that is, which positions are conserved and which ones are not, conforms to the profile established for the animal homeodomains (Bürglin 1994), with positions 16, 20, 48, 50, 51, and 53 highly conserved among plant and animal homeodomains; 3) Secondary structure prediction and homology modeling conform to the homeodomain structure; 4) Blast searches using as a query each newly identified homeodomain confirmed their highest similarity to homeodomain proteins from other plants. All plant homeodomain proteins that we retrieved conserve at least 36% identity with animal homeodomain proteins.

By phylogenetic analyses based on the ML procedure implemented in PHYML (Guindon and Gascuel 2003) and other approaches (see Methods), we found that all plant homeodomain proteins reliably group into 14 distinct classes with robust (generally 70% or more) bootstrap support (fig. 2 and supplementary figs. 2 and 3). Supplementary Material online). Furthermore, from the multiple sequence alignments of full-length homeodomain proteins belonging to the 14 classes, distinctive motifs were found conserved and uniquely associated with each class across monocots, eudicots, Selaginella, and moss and even among unicellular green algae or red algae (fig. 3). Finally, the intron positions within all the 110 homeomain genes of Arabidopsis were strikingly conserved within each class (fig. 4). Based on previous classifications (see Introduction) and on the results of our analysis, we propose the following classification scheme for the plant homeobox genes.

The HD-ZIP Superclass (HD-ZIP I, HD-ZIP II, HD-ZIP III, HD-ZIP IV)

The HD-ZIP superclass is composed of four individual classes, HD-ZIP I, HD-ZIP II, HD-ZIP III, and HD-ZIP IV (Sessa et al. 1994), all of which are characterized by presence of a leucine-zipper adjacent to the C-terminus of the homeodomain (fig. 3). The HD-ZIP II class is distinguished by a “CPSCE” motif conserved downstream of the leucine-zipper (Chan et al. 1998). The HD-ZIP III and HD-ZIP IV classes were characterized by the START (Steroidogenic Acute Regulatory protein–related lipid Transfer) (Ponring and Aravind 1999) and HD-SAD (START associated conserved domain) (Schrick et al. 2004; Mukherjee and Bürglin 2006) domains. The HD-ZIP III class is distinguished from HD-ZIP IV by an additional conserved C-terminal domain, previously named MEKHLA domain (Mukherjee and Bürglin 2006). Our exhaustive database searches yielded the 48 leucine-zipper homeodomain proteins recently identified in the genome of Arabidopsis (Baima et al. 2001; Henriksson et al. 2005; Nakamura et al. 2006; Ciarbelli et al. 2008). We identified a similar number of leucine-zipper homeodomain proteins (47 genes) in the genome of rice. In the genomes of poplar, maize, Selaginella, and moss, we identified 61, 70, 13, and 33 leucine-zipper homeodomain proteins, respectively (table 2). Thus, the HD-ZIP superclass comprises from 40% to 50% of all homeobox genes of flowering plants and moss (table 2). The classification of HD-ZIP proteins was based on the evolutionary tree relations among HD sequences and was also supported by the association of these sequences with codomains characteristic of the different HD-ZIP classes (see above). From the alignment of all known and newly identified proteins we also discovered that HD-ZIP III proteins were distinguished from other HD-ZIP class genes by four extra residues inserted between helices 2 and 3 of the homeodomain (fig. 1). We propose this insertion as a diagnostic new feature for this class. We could not find any HD-ZIP gene in unicellular green algae or in red algae.

The PLINC Zinc Finger Class

A PLINC zinc finger HD class protein was first identified in Flaveria trinervia (Asteraceae) (Windhoven et al.

<table>
<thead>
<tr>
<th>Super Class</th>
<th>Class</th>
<th>Eudicots</th>
<th>Monocots</th>
<th>Bryophyta</th>
<th>Lycoperidiophyta</th>
<th>Unicellular Green Algae</th>
<th>Red Algae</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>At</td>
<td>Poplar</td>
<td>Rice</td>
<td>Maize</td>
<td>Sm</td>
<td>Cr</td>
</tr>
<tr>
<td>HD-ZIP</td>
<td>HD-ZIP I</td>
<td>17</td>
<td>22</td>
<td>14</td>
<td>24</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>HD-ZIP</td>
<td>HD-ZIP II</td>
<td>10</td>
<td>16</td>
<td>14 (2)</td>
<td>17</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>HD-ZIP</td>
<td>HD-ZIP III</td>
<td>5</td>
<td>8</td>
<td>7 (3)</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>HD-ZIP</td>
<td>HD-ZIP IV</td>
<td>16</td>
<td>15</td>
<td>12</td>
<td>21</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>PLINC</td>
<td>14</td>
<td>17</td>
<td>11</td>
<td>17</td>
<td>—</td>
<td>11 (1)</td>
<td>5</td>
</tr>
<tr>
<td>WOX</td>
<td>16</td>
<td>19</td>
<td>15</td>
<td>16</td>
<td>—</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>TALE</td>
<td>KNOX</td>
<td>8</td>
<td>15</td>
<td>12 (1)</td>
<td>13</td>
<td>5</td>
<td>5 (1)</td>
</tr>
<tr>
<td></td>
<td>BEL</td>
<td>13</td>
<td>19</td>
<td>14</td>
<td>17 (1)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Uncharacterized</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>DDT</td>
<td>4</td>
<td>7</td>
<td>3 (1)</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>PHD</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>NDX</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>LD</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>PINTOX</td>
<td>1</td>
<td>1 (1)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>SAWADEE</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>Uncharacterized</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>148 (1)</td>
<td>110 (7)</td>
<td>148 (1)</td>
<td>66 (1)</td>
<td>45 (1)</td>
<td>5</td>
</tr>
</tbody>
</table>
2001), where it was named ZF-HD for two highly conserved zinc finger–like motifs upstream of the homeodomain (fig. 3). Subsequently 14 members of the same class have been characterized in Arabidopsis (Tan and Irish 2006). In each of the four flowering plant genomes that we analyzed, we identified between 11 and 17 gene sequences belonging to this class, along with 5 genes from Selaginella and 11 genes and 1 pseudogene in moss (table 2). The corresponding proteins were characterized by clustering in the ML tree reconstruction and by the presence of the two diagnostic putative ‘‘zinc finger motifs’’ upstream of the homeodomain (fig. 3). The first putative zinc finger has a consensus sequence C-X_3-H-X_9-D-X_1-C, where H is not always conserved, and the second putative zinc finger has the conserved pattern C-X_2-C-X_1-C-H-X_3-H. We propose to name this class and the corresponding domain ‘‘PLINC’’ (Plant Zinc Finger), to distinguish it from the unrelated zinc finger class of homeodomain proteins, called ZF, found in animals (Bu¨rglin 1994). Most of these proteins were also characterized by the substitution of the usually conserved homeodomain residue F49 with a methionine residue and by insertion of one amino acid between helix 1 and helix 2, which we propose to be two strong diagnostic features of this class. Exceptions were two genes, one from A. thaliana (At1g14687) and one from rice (03g50920), which at the same position showed, respectively, 3 or 6...
amino acid insertions (supplementary fig. 1, Supplementary Material online). We could not identify any gene belonging to this class in unicellular green algae or in red algae, suggesting that this class originated within the Streptophyta clade and that it was already present in the last common ancestor of moss and vascular plants.

The WOX Class

The homeodomain proteins of this class have one or two extra residues between helices 1 and 2, and four to five extra residues between helices 2 and 3 (fig. 1). Recently, Haecker et al. (2004), based on their genome-wide search, reported and annotated 14 WOX genes from Arabidopsis. We found in Arabidopsis 16 WOX genes (table 2), characterized by clustering in the evolutionary tree and by distinctive conserved motifs both upstream and downstream of the homeodomain, including the WUS Box (Haecker et al. 2004). The WOX class can be subdivided into several families (supplementary fig. 4, Supplementary Material online). Among these, we found that families such as WOX 11/12 and WOX 8/9 had a C-terminal motif of about 60 aa in length conserved between monocots and eudicots. We found six WOX genes in Selaginella, three in moss, and one WOX gene in the unicellular green algae Ostreococcus lucimarinus and Ostreococcus tauri (fig. 2 and supplementary fig. 1, Supplementary Material online) but none in Chl. reinhardtii, indicating that the WOX class of proteins was present in the last common ancestors of green plants (Chlorobionta) and was successively lost in the Chlamydomonas lineage.

The TALE Superclass (KNOX and BEL)

TALE homeodomain proteins are characterized by three extra residues between helix 1 and helix 2 (Bertolino et al. 1995; Bürglin 1997). In plants, TALE genes have been

Fig. 2.—Circular representation of the evolutionary tree of plant homeodomain sequences (branch lengths not drawn to scale), based on all homeodomain sequences identified in Arabidopsis, Selaginella, moss, unicellular green algae, and red algae supplemented by selected sequences from other flowering plant species wherever only one protein from Arabidopsis was found (see supplementary table 1 [Supplementary Material online] for a complete list of species). The tree, which should be considered unrooted, was obtained with the ML procedure implemented in PHYML with the JTT substitution model using an alignment of homeodomain sequences as shown in supplementary fig. 1 (Supplementary Material online). Bootstrap support was based on 1,000 replicates and is indicated for relevant branches. Bootstrap support obtained after excluding all sequences from unicellular green algae and red algae is shown in parentheses. Clades supported by robust bootstrap values (70% or more) are shown with thicker lines. Bootstrap values for intraclass branches are omitted. Bootstrap values less than 5% for branches connecting different classes are not shown. All homeodomain classes are identified as separate clades in this analysis and are consistently supported by conserved, class-specific domain architecture (see fig. 3) and unique splice junctions (see fig. 4). Red-colored branches indicate sequences from red algae, blue-colored branches refer to the unicellular green alga Chlamydomonas reinhardtii, the light-blue color identifies the unicellular green algae Ostreococcus lucimarinus and Ostreococcus tauri, gold-colored branches indicate moss proteins, and a light-green color is used to represent sequences from Selaginella.
extensively studied and classified into the two classes KNOX and BEL. Among the homeobox gene classes, KNOX and BEL appeared to be the oldest classes, with members present in single-cell green algae and in red algae (fig. 2).

Members of the KNOX class have been previously classified into two families, KNOX I and KNOX II (Kerstetter et al. 1994). We found eight members of the KNOX class in *A. thaliana*, 15 in poplar, 12 (and one pseudogene) in rice, 13 in maize, 5 (and 1 pseudogene) in *Selaginella*, 5 in moss, and a single unambiguously classified gene in unicellular green and red algae (fig. 2). Five additional genes from *Selaginella*, one from *Chlamydomonas*,

---

**Fig. 3.**—Schematic overview of the domain architecture of all 14 classes of plant homeodomain proteins. The following domains and motifs are indicated: HD, leucine-zipper (LZ), CPSCE motif, CESV motif, START domain, homeodomain-START associated domain (HD-SAD), MEKHLA domain, PLINC zinc finger, BEL domain (A & B), KNOX domain (A & B), ELK motif, DDT domain, WSD motif, D-TOX ZF, PEX-PHD, PHD, LUMI, conserved motifs in LD homeodomain proteins (LD1, LD2, LD4, AND LD5). Among DDT proteins, only D-TOX A is indicated with its full symbol; D-TOX B, D-TOX C, D-TOX D, D-TOX E, D-TOX F, and D-TOX G are indicated as B, C, D, E, F or G, respectively.

**Fig. 4.**—Intron positions in the homeobox sequences of *Arabidopsis thaliana*. A consensus (majority) 60-residue homeodomain sequence is shown above a ruler where individual codon base positions are separated by tick marks. Intron positions are indicated by vertical lines labeled by the class names where each intron was found. The intron positions are generally conserved within each class. Exceptions are ATHB1, which has one intron, whereas other HD-ZIP I genes are single-exon genes; and At4g12750, a DDT class member, which has an extra intron at the beginning of the homeodomain. Gene classes not shown (HD-ZIP I, WOX, PLINC) do not have introns in the homeodomain. One of the splice sites in the NDX genes lies in the loop between helix 2 and helix 3 where it cannot be placed accurately on the consensus sequence. Its approximate position is marked by a crossbar at the end of the line.
and two from the red alga clustered with the TALE class. However, they did not confidently cluster within the KNOX or BEL classes (fig. 2), and we could not recognize in their sequence a typical KNOX or BEL codomain architecture. From the analysis of the codomain structure of KNOX genes from flowering plants, we could further classify KNOX I genes into two subclasses, KNAT 2/6 and KNAT1, while the KNOX II class could be separated into the two subclasses KNAT7 and KNAT3/4/5 (fig. 5). Except for the STM subclass, which appeared to be dicot specific, all other subclasses had representatives from monocots and eudicots, suggesting that they were present in flowering plants prior to the separation of the two groups. KNOX gene representatives from both Selaginella and moss clustered within the KNOX I or the KNOX II subclasses (fig. 5), confirming the results of a recently published evolutionary analysis (Sakakibara et al. 2008) and suggesting that both KNOX I and KNOX II genes were present in the common ancestor of moss and vascular plants. Both KNOX families conserve a diagnostic KNOX domain upstream of the homeodomain, composed of two blocks (KNOX A and KNOX B) separated by a variable region (Bürglin 1997), as well as a shorter motif adjacent to the homeodomain, named ELK (Vollbrecht et al. 1991). The multiple sequence alignment of KNOX I and KNOX II proteins (supplementary fig. 5, Supplementary Material online) showed that in KNOX II proteins the KNOX B subdomain is characterized by a 30-aa insertion. In addition, we could newly identify in all KNOX II genes a highly conserved motif consisting of 20 amino acids that we named “KNOX II C” (supplementary fig. 5, Supplementary Material online).
The KNOX domain shares similarities with the MEIS and PBC domains of animal TALE homeobox genes, suggesting that these domains derived from a common ancestral domain that has been named MEINOX (Bürglin 1997, 1998a). The ELK motif, situated between the KNOX and HD domains of KNOX class proteins, is in a position analogous to the PBC-B motif of animal PBC class homeobox genes. From sequence analysis we found that the ELK domain showed similarities with the PBC-B motif, including highly conserved hydrophobic residues, and that both motifs are predicted to fold into helical conformations (fig. 6). This reinforces the notion that genes of the PBC class are also derived from an ancestral MEINOX TALE homeobox gene (Bürglin 1998a). We propose that the ELK domain, as well as the C-terminal end of the PBC-B domain, should be considered part of the MEINOX domain.

The BEL class of proteins is uniquely characterized by two conserved domains upstream of the homeodomain, called SKY and BEL (Bellaoui et al. 2001; Becker et al. 2002). We identified 13–19 BEL genes in flowering plants, 2 genes in Selaginella, 4 genes in moss, 1 or 2 genes in unicellular green algae, and 1 gene in red alga (table 2). By ML evolutionary tree reconstructions based on the homeodomain and codomain amino acid sequences and by similarities in codomain architecture, we could distinguish within the BEL class five newly classified subclasses, common to monocot and eudicot flowering plants (supplementary fig. 6, Supplementary Material online). In addition to SKY and BEL, we have detected in this class a third highly conserved 10-aa motif, which we named “ZIBEL”, repeated at both the C-terminal and N-terminal ends of the BEL proteins (fig. 7, panel a). This motif was conserved in all BEL subclasses except at the C-terminal end of the ATH1 subclass, where the SKY domain was also missing (supplementary fig. 6, Supplementary Material online). Furthermore, we also found the ZIBEL motif at the N-terminal end of HD-ZIP II proteins (fig. 7, panel a).

Considering that the KNOX, PBC, and MEIS domains are all bipartite, with a variable-length linker region connecting two conserved regions, we propose that a corresponding BEL domain should also be redefined as a bipartite domain composed of a conserved N-terminal domain (BEL-A), also including the SKY motif, and a conserved C-terminal domain (BEL-B). This definition is also consistent with the observation that both parts of the newly defined BEL domain are necessary for interaction with KNOX homeodomain proteins (Bellaoui et al. 2001; Smith et al. 2002). We also identified in all BEL subclasses except BLR a third conserved motif, that we named BEL-C, situated between the BEL-B and the homeodomain sequences (supplementary fig. 6, Supplementary Material online). No obvious conservation had been so far described between the BEL and MEINOX domains (Becker et al. 2002). However, when the redefined BEL domain was aligned with the MEIS domain, the BEL-A region aligned with the MEIS-A domain, with two adjacent leucine residues highly conserved in BEL-A and in the MEIS-A domain within a segment predicted in both classes to be helical (fig. 7, panel b).

New Classes of Plant Homeodomain Proteins

The DDT Class

One group of homeodomain proteins is characterized by the presence of the DDT domain (Doerks et al. 2001), located downstream of the homeodomain. Our extended searches and evolutionary analyses allowed us to newly define this group of proteins as a separate class of HD proteins. This group encompasses the largest plant homeodomain proteins, with sequences of lengths up to about 1,900 amino acids. We identified three to seven genes encoding a DDT domain in flowering plants, two genes in Selaginella, three genes in moss, and one gene in each genome of unicellular green algae (table 2). No DDT class genes were identified in red algae. In addition to the DDT domain, from the alignment of these proteins we identified seven other conserved motifs, distributed throughout the entire length of the protein, which we named D-TOX A to G (DDT Homeobox Class Domain, fig. 3). The newly identified D-TOX F motif is a zinc finger motif highly conserved between monocots and eudicots, characterized by the pattern C-X$_5$-C-X$_{10}$-C-X$_5$-C. To distinguish it from other zinc finger domains, we propose to name the newly identified zinc finger motif “D-TOXZF” (fig. 8, panel a). Using D-TOX E as a query in database searches, we identified other proteins from rice, Arabidopsis, and slime mold that did not contain the homeodomain but conserved the D-TOX E and the DDT domains, with 36–43% amino acid identity. We also identified similarity with the animal Williams-Beuren syndrome transcription factor proteins, which also contain the DDT domain, as well as PHD and bromo domains, with 25–28% sequence identity (fig. 8, panel b). We propose to uniquely identify the D-TOX E motif as the Williams-Beuren syndrome DDT (WSD) motif. We classified genes belonging to the DDT class by phylogenetic analyses and codomain architecture into three subclasses (D-TOX1, D-TOX2, D-TOX3), of which D-TOX1 and D-TOX2 are represented in both eudicots and monocots, whereas D-TOX3 is eudicot specific (fig. 9). The D-TOX3 subclass has lost all motifs characteristic of the DDT class of genes except for the D-TOX A motif. The DDT class of genes appears to have evolved early in green plant (Chlorobionta) evolution because members of this class were found in unicellular green algae and in land plants (fig. 2).

The PHD Class

The pathogenesis-related homeobox gene A (PRHA) (Plesch et al. 1997) and HAT3.1 gene (Schindler et al. 1993) from Arabidopsis were characterized by the presence of the PHD codomain, located several hundred amino acids upstream of the homeodomain. We found from two to five PHD genes encoded in flowering plant genomes, one gene in Selaginella, and two genes in moss. We did not detect any PHD HD gene in unicellular green algae or in red algae. These genes formed a distinct clade in the ML tree reconstruction (fig. 2), which, together with the diagnostic presence of the PHD domain, allowed us to newly define the class of PHD homeodomain proteins. Adjacent to the PHD domain proteins of this class lies an additional
conserved motif of about 90 aa, which we called PEX-PHD (fig. 3). This region, which we found only in plant PHD homeodomain proteins, is rich in charged residues. Other smaller conserved motifs are found mainly between the PHD domain and the homeodomain (fig. 3).

The ML evolutionary tree obtained from the alignment of the homeodomain and PHD sequence regions suggests that PHD proteins divide into two distinct subclasses (supplementary fig. 7, Supplementary Material online), named after the corresponding Arabidopsis genes: PRHA, with a high bootstrap support (98%), and HAT3.1, with a lower support (58%). Both subclasses are represented in all flowering plant genomes examined, but only proteins classified in the PRHA group are found in moss (two genes) and in Selaginella (one gene). The most parsimonious scenario is that the HAT3.1 subclass evolved in the flowering plant lineage after separation of moss and Selaginella. However, the topology of the rooted trees (fig. 2 and supplementary fig. 7, Supplementary Material online) suggests that both subclasses were already present before separation of the moss and Selaginella lineages and that HAT3.1 genes were successively lost from these genomes.

The NDX Class

Homeobox genes expressed in the nodule (Nodulin Homeobox genes, NDX) were first reported by Jørgensen et al. (1999) from soybean and from Lotus japonicus. We identified two NDX genes in the genomes of poplar and moss; one NDX gene in A. thaliana, rice, maize, and Selaginella; and none in unicellular green or red algae, suggesting that this class appeared early in the evolution of the land plant clade before separation of moss and vascular plants. We also searched NDX genes in several other unfinished genome sequences and EST collections of other plant species. A multiple sequence alignment of the corresponding proteins is presented in supplementary fig. 8 (Supplementary Material online). The homeodomains of the NDX class are atypical and highly divergent. We found that NDX proteins have a diagnostic insertion of six amino acids between helices 2 and 3 (fig. 1 and supplementary fig. 1, Supplementary Material online). The NDX-B was a 540-aa-long domain located upstream of the homeodomain. NDX-B was an 80-aa-long domain located downstream of the homeodomain. Both domains are highly conserved among flowering plants and in moss (supplementary fig. 8, Supplementary Material online).

The LD Class

A homebox gene called LD (Luminidependens) has been previously identified in Arabidopsis and in maize (van Nocce et al. 2000). In addition to the genes from these species, we found a single copy of LD genes in rice, maize, Selaginella, and moss and two copies of the gene in poplar. We did not find any LD gene in unicellular green algae or in red algae (table 2). The distribution of LD genes suggests that they were present in the last common ancestor of moss and vascular plants and that they evolved after the divergence of Chlorophyta and Streptophyta. The homeodomain of this class is of typical length, but it is characterized by several unusual substitutions of otherwise conserved residues, such as W48F and N51X (fig. 1). The sequence alignment of LD proteins revealed five conserved codomains, which we named LD1–LD5 (supplementary fig. 9, Supplementary Material online). PSI-Blast database searches revealed that the 80-aa-long LD3 region was also conserved within an unrelated group of plant transcription factors (fig. 10, panel a). We refer to this newly identified conserved domain as the “LUMI domain.”

The PINTOX Class

Homeodomain proteins of the PINTOX class formed another distinct clade with a strong bootstrap support of 84% (fig. 2). We named this class PINTOX from one of the genes represented in this class, the Plant Interactor Homeobox rice gene GF14c-int., which has been isolated and named for its interaction with GF14-c in a two-hybrid screen (Cooper et al. 2003). We newly identified at least one full-length gene as belonging to this class in all other examined green plant genomes, including Chl. reinhardtii and other unicellular green algae, indicating that PINTOX genes originated before the divergence of Chlorophyta and Streptophyta. PINTOX homeodomain proteins are characterized by the substitution N51D within the homeodomain region (fig. 1) and by a highly conserved basic domain of about 70 aa (named PINTOX domain) newly identified upstream of the homeodomain (fig. 10, panel b). Further upstream of the PINTOX domain is a conserved acidic domain, which we named “Acid Pint” (fig. 3), whereas the N-terminal region has conserved hydrophobic and basic residues.

The SAWADEE Class

In all examined genomes of flowering plants, in Selaginella and in moss we identified from one to three genes that clustered with high (99%) bootstrap support in the ML evolutionary tree (fig. 2) into a newly identified class that we named SAWADEE. These genes are characterized by diagnostic insertions of 10 aa between the second and third homeodomain helices (fig. 1), extending the second loop more so than in any other plant homeodomain protein. C-terminal to the homeodomain we identified a 130- to 140-aa-long conserved region, the SAWADEE domain (fig. 10, panel c). We found homologs of the SAWADEE domain also in non-HD genes conserved among monocots and eudicots.
The SAWADEE domain is characterized by several conserved cysteine and histidine residues, suggesting that it may be involved in metal coordination.

Discussion

By exhaustive searches of complete genomes and other sequence data we have identified the full repertoire of homeodomain proteins found in diverse plant species, improving previous reports of homeobox genes in Arabidopsis thaliana and in rice, and newly identifying the complete set of homeobox genes from the genomes of poplar and maize (eudicot and monocot flowering plants); spikemoss S. moellendorfii (Lycopodiophyta); moss P. patens (Bryophyta); the unicellular green algae O. lucimarinus, O. tauri, and Chl. reinhardtii (Chlorophyta); and the red alga Cy. meroideae (Rhodophyta). From these analyses we obtained a comprehensive perspective of the evolutionary history of plant homeobox genes and gene classes and a reliable, exhaustive, nonredundant source catalog of plant homeodomain protein sequences. Our searches expanded the number of plant homeobox genes previously reported in the literature. In Arabidopsis we identified 110 genes versus the 87 homeobox genes previously reported (Shiu et al. 2005). In rice, the most recent report identified 107 homeobox genes (Jain et al. 2008). From this collection we excluded eight zinc finger domain proteins that did not contain a conserved homeodomain, and we added 11 other homeobox genes not previously reported. Among rice homeobox genes, 31 sequences have been recently identified as HD-ZIP genes (Agalou et al. 2008). In our analysis of the rice genome, we identified 47 HD-ZIP genes and 5 HD-ZIP pseudogenes. Although the genome of rice (400–430 Mb) is more than three times larger than the genome of Arabidopsis (125 Mb), the total number of its predicted genes (37,544 predicted genes) is only 34% more than the number estimated in Arabidopsis (28,000 predicted genes). We identified in both genomes 110 homeobox genes. To our knowledge, no data were previously available on the number of homeobox genes in the genomes of poplar and maize. Although poplar and maize have genomes of substantially different sizes (550 Mb and 2,900 Mb, respectively), they are thought to encode a similar total number of genes (58,000 and 54,000 predicted genes, respectively). Consistent with their greater total number of predicted genes compared with Arabidopsis or rice, we identified in both genomes 148 homeobox genes (vs. 110 genes in Arabidopsis and rice). However, in the genome of the moss P. patens (Bryophyta), with a genome size of 500 Mb and 30,000–35,000 estimated genes, we found only 66 homeobox genes, whereas we identified 45 homeobox genes in the 110-Mb genome of
Selaginella. In unicellular algae we identified a much smaller number of homeobox genes. Although the unicellular green alga *Chl. reinhardtii* has a genome size similar to *Arabidopsis* at 121 Mb and about half as many predicted genes (15,000 genes), we identified in its genome only five homeobox genes. Similarly, the other unicellular green algae *O. lucimarinus* and *O. tauri*, with a genome size of about 13 Mb and 7,700–7,900 predicted genes, contained only seven homeobox genes each. Finally, six homeobox genes were found in the genome (of 16.5 Mb and 5,331 predicted genes) of the red alga *Cy. merolae* (Rhodophyta). Expression of about half of the homeobox genes of *Arabidopsis* has been experimentally verified. Searching EST databases we found expression data for 92 *Arabidopsis* homeobox genes (84%), and sequence conservation suggests that all 110 genes play a biological role.

FIG 9.—ML tree obtained using the homeodomain and associated signature codomains of DDT class proteins, outgrouped by PINTOX class proteins (At_At5g11270 and St_Pint1). Proteins from monocots are shown within light-green–colored boxes; from eudicots in light-blue–colored boxes, and from *Selaginella* yellowish green–colored boxes. The DDT class can be subdivided into three families: D-TOX1, D-TOX2, and D-TOX3. D-TOX3 is eudicot specific and has secondarily lost all signature codomains found in D-TOX1 and D-TOX2 with the exception of the D-TOX A codomain. A DDT protein from *Chlamydomonas reinhardtii* (Cr_C460085), three proteins from moss (Pp_sca_419, Pp_sca_15b and Pp_sca_89), and two proteins from *Selaginella* (Sm_422900 and Sm_447019) cannot be classified with certainty within any of the three families.

New Classes and Conserved Motifs

Based on our analysis, we propose that plant homeobox genes can be classified into 14 classes (HD-ZIP I to IV, PLINC, WOX, NDX, DDT, PHD, LD, PINTOX, SAWADEE, KNOX, and BEL), of which four are grouped into the HD-ZIP superclass (HD-ZIP I to IV) and two (KNOX and BEL) belong to the TALE superclass. All homeobox genes found in the genomes of flowering plants, *Selaginella*, and moss were unambiguously clustered by evolutionary tree reconstructions into 14 distinct groups (classes) with generally high bootstrap support (70% or more). Each class was additionally supported by distinctive signature motifs or codomains and by typical intron–exon structures. Among (see Adams and Wendel [2005] for a review), and one duplication in moss (Rensing et al. 2008). These duplications are likely to be related to the increased number of homeobox genes found in flowering plants compared with moss. However, conservation of a greater number of genes in flowering plants is also likely to be related to the increased developmental and organizational complexity of these species.
these, the classes PHD, DDT, NDX, LD, SAWADEE, and PINTOX are here for the first time defined. We discovered and characterized numerous new motifs, which appeared to be highly conserved in moss and flowering plants, and, whenever present, also in unicellular green and red algae, suggestive of a common, evolutionarily conserved functional role. Notable examples are the LUMI domain, the ZIBEL motif, the WSD motif, and the PINTOX domain. The ZIBEL motif is conserved at both the N-terminal and C-terminal ends of the BEL class of proteins and at the N-terminal end of the HD-ZIP II class of proteins. We speculate that through the ZIBEL motif the HD-ZIP

![Multiple alignments of (a) the LUMI domain of the LD class; (b) the PINTOX domain of PINTOX class proteins; and (c) the SAWADEE domain of SAWADEE class proteins. The position of conserved cysteine and histidine residues is highlighted below the alignment. At each position conserved amino acid types with similar physicochemical properties are highlighted in different shades of gray.](https://academic.oup.com/mbe/article-abstract/26/12/2775/1538706/fig-10)
II and BEL homeodomain proteins may interact with each other or may interact with the same target proteins. We found that the WSD motif of plant DDT proteins shows high sequence similarities with a protein that causes Williams-Beuren syndrome (WSTF) in humans (Cus et al. 2006), suggesting that the WSD motif was already present in the common ancestor of plants and animals. The WSTF gene was found to play an important role in chromatin remodeling and was associated with other motifs such as the PHD, bromo, MBD, and ring finger domains. Little experimental information on the function of the SAWADEE homeodomain proteins is available, and their roles can only be speculated from sequence features and expression data. Conservation of basic, cysteine, and histidine residues suggest that SAWADEE may be a DNA-binding domain. EST data indicate that SAWADEE is expressed in roots, leaves, developing seeds, embryos, flower buds, and endosperm.

Some of the newly identified domains, as well as previously known domains, are also found fused with other nonhomeodomain proteins (DDT, LUMI, PHD, and WSD) or as independent proteins (KNOX, SAWADEE, PLINc, START, and HD-SAD). It is possible that these proteins act as negative regulators of the corresponding HD proteins by competing for protein–protein interaction or DNA-binding sites. This has been recently shown in the case of the non-HD, KNOX domain–containing protein PTS of tomato, which competes for protein–protein interaction sites with the HD proteins KNOX I and BIP (Kimura et al. 2008).

Evolutionary History of Plant Homeobox Genes

Identification of the different classes and codomains of homeodomain proteins in various plants and other evolutionary groups allows us to coherently reconstruct the overall evolutionary history of this protein family in plants, summarized in fig. 11. The overall picture that emerges from our analysis is that the family of homeobox genes has expanded and differentiated together with the differentiation of plants into organisms of increasing complexity. The functional differentiation of homeodomain proteins appears to have first been achieved through the acquisition of different codomain architectures. This stage had already achieved its current state during the evolution of land plants before separation of moss and vascular plants, when all 14 different codomain architectures represented in modern homeodomain classes must have been already present. Further specialization was successively achieved by the proliferation of gene paralogs within each class, as observed in modern flowering plants. It must be noted that our interpretation of a progressive increase in the number of homeobox genes in the lineage leading to flowering plants could be equivalently interpreted as corresponding loss of genes in the respective lineages. Although this interpretation is in principle coherent, we find the alternative view of a progressive lineage-specific expansion of the homeobox gene
family more plausible because 1) it is more parsimonious, requiring for each new gene one lineage-specific duplication event rather than one duplication in a common ancestor plus one or more lineage-specific gene loss(es), and 2) it correlates with a corresponding increase in organism developmental and functional complexity. Nevertheless, some of the results suggest gene loss events. For example, a representative of the WOX class is found in the genome of the Ostreococcus unicellular green algae, but not in the other green alga Chl. reinhardtii.

**HD Classes and Domains Common to All Examined Plant Groups (Rhodophyta, Chlorophyta, Moss, and Vascular Plants)**

The two classes of TALE HD proteins found in plants, KNOX and BEL, are characterized by a conserved codomain structure that is also recognized in two of the several classes of animal TALE HD proteins, MEIS and PBC. This suggests that these proteins derived from a common ancestor, and their codomain was named MEINOX (Bürglin 1997). The duplication of MEINOX genes observed in animals and plants has been previously described as two independent events in the plant and animal lineages (Bürglin 1998a). However, the presence of a common intron insertion position in the plant BEL class and in the animal PBC class but not in the KNOX (plant) and MEIS (animal) classes suggests that their common ancestor duplicated and that one intron was acquired by one of the two gene copies before separation of animals and plants. In either case, KNOX and BEL HD proteins must have been already present in the last common ancestor of the plant clade. Typical (non-TALE) HD proteins are also found in animals, plants, including unicellular green and red algae, and many other anciently diverged eukaryotic groups (Derelle et al. 2007), suggesting that TALE and non-TALE HD proteins were already present in the common ancestor of plants and animals. However, none of the animal non-TALE HD proteins can be reliably associated with a specific non-TALE plant class. The START, PHD, and DDT-WSD protein domains are found associated with HD in modern land plants and associated with other proteins (PHD and DDT-WSD) or absent (START) in unicellular green and red algae. Because these domains are also found in animals not associated with HD, we most parsimoniously predict that they were present but separated from HD in the common ancestor of plants and animals (fig. 11). Among these domains, START is not found in unicellular green algae and red algae, suggesting that it has been secondarily lost in these groups.

**Appearance of Three New HD Classes in the Common Ancestor Lineage of Green Plants**

In moss, vascular plants, and unicellular green algae, we found the DDT-WSD domain associated with HD, identifying the presence of the homeodomain protein DDT class in the last common ancestor of green plants (Chlorobionta). We also infer that at this time HD was also associated with the green plant–specific PINTOX and WUS motifs, giving rise to the corresponding PINTOX and WOX classes. The evolution of green plants can then be associated with the addition of three more HD classes to the KNOX and BEL classes present in all plants. Presumably each of these classes was originally represented by one gene, as observed in modern-day unicellular green algae. In the genome of Chl. reinhardtii, we found the bacterial PAS-like MEKHLA domain, so far found associated with HD in land plants, including the liverwort Marchantia polymorpha, and in the charophycean green alga Chara corallina (Floyd et al. 2006). We speculated that this domain was acquired from the cyanobacterial genome of the chloroplast at the onset of the green plant clade (Mukherjee and Bürglin 2006).

**Multiple HD Genes and HD Classes in Embryophyta (Land Plants)**

Embryophyta are characterized by the appearance, at least in the lineage leading to moss and vascular plants, of nine classes of HD proteins not present in Rhodophyta or Chlorophyta, namely HD-ZIP I to IV, LD, SAWADEE, NDX, PHD, and PLINC. These classes are identified by fusion of HD with codomains (PHD, START, MEKHLA, CPSC; HD-SAD, LD, SAWADEE, NDX, and PLINC). All HD classes not found in Rhodophyta and Chlorophyta were found both in moss and in flowering plants, suggesting that they originated early in land plant evolution (at least before separation of moss and vascular plants and possibly before transition to land) in relation to newly acquired developmental features. In fact, a full-length HD-ZIP III gene with all functional domains conserved has been recently detected also in Cha. corallina (Floyd et al. 2006), a freshwater green alga belonging to the charophycean group Charales, believed to be a sister group to land plants (Karol et al. 2001; McCourt et al. 2004). This suggests that at least HD-ZIP III homeobox genes (and presumably all HD-ZIP homeobox gene classes) must have been present in the last common ancestor of Charales algae and land plants. Interestingly, no new HD classes (defined by their codomain associations) developed in the evolution of vascular plants or moss after their separation. Instead, further complexity and functional differentiation have been achieved, at least in flowering plants and in moss, through a great proliferation of gene paralogs within each of the original 14 classes.

**Rooting the Tree of Plant Homeodomain Proteins**

The evolutionary trees based on homeodomain sequences (fig. 2 and supplementary figs. 2 and 3, Supplementary Material online) robustly identify the different classes of plant homeodomain proteins, which are also fully supported by their characteristic codomain architectures. However, the trees do not provide reliable information on the evolutionary relations among most of the classes, which are connected by poorly supported topologies. These are also contradicted by the partial information provided by intron insertion analysis, including two parsimony-informative intron insertion sites (fig. 4). One intron insertion site is conserved between the non-TALE classes LD, SAWADEE, and PHD. The same intron insertion site is also found in the TALE class BEL and in some animal fungus TALE HD classes—PBC, TGIF, IRO (Bürglin 1997),
and in MKX (Mukherjee K, unpublished data). However, this intron is not present in KNOX, the other plant TALE class. KNOX instead has an intron insertion position in common with the non-TALE class PINTOX, neither present in BEL nor in any animal or fungus HD gene. This arrangement would question the monophyletic origins of plant TALE as well as of plant non-TALE proteins. The evolutionary significance of these few common intron insertions is, however, dubious because we do not know the likelihood of parallel intron insertion or deletion events in different lineages. For example, we also find an intron insertion position conserved between the SAWEDEE class and the Arabidopsis gene ATHB1, which clearly belongs to the HD-ZIP I class, suggesting homoplasry or multiple intron losses in different classes. In fig. 12 we show a summary representation of the results from the trees of fig. 2 and supplementary figs. 2 and 3 (Supplementary Material online), where each class is represented by a single branch, and the topology of branches connecting TALE or non-TALE classes with a bootstrap support of less than 50% is replaced by circles representing their uncertainty. Sequence representatives of animal TALE and non-TALE proteins cluster with plant TALE and non-TALE proteins, respectively (supplementary fig. 3 [Supplementary Material online] and Derelle et al. [2007]), consistent with previous observations that both types of homeodomain proteins are found in anciently diverged eukaryotic groups (Derelle et al. 2007) and must have been present early in eukaryote evolution. However, TALE proteins from plants, animals, or any other phylogenetic group cannot be used as an outgroup to root the subtree of plant non-TALE proteins unless we assume that all modern classes of plant non-TALE homeodomain proteins originated monophyletically after separation of TALE and non-TALE genes. With this assumption, the root of plant non-TALE proteins would lie in the branch connecting the two types (bold faced in fig. 12). If this is not the case (i.e., some non-TALE classes separated before TALE genes separated from other non-TALE classes), the root of all non-TALE proteins should be positioned within the circle of uncertain topology shown in fig. 12 connecting the different non-TALE plant classes.

Similar arguments apply to the rooting of TALE proteins, although conservation of their signature insertion suggests that all TALE genes have a monophyletic origin, a notion that is challenged, however, by the pattern of intron conservation (see above). The two TALE gene classes found in plants resulted either from a duplication in the common ancestor of plants and animals or after separation of the two phyla (Bürglin 1997).

Throughout plant evolution, the homeobox gene family has proliferated and diversified in accordance with the growth in structural and developmental complexity of the organisms in which they were expressed. Some homeodomain proteins have been intensely studied, but little or nothing is known about the functionality of many other homeobox genes and classes. In this study of multiple plant genomes, we newly uncovered HD protein classes and a greater abundance of homeobox genes than previously known. We also identified many previously unnoticed conserved motifs whose specific role in protein–protein or protein–DNA interaction remains to be experimentally verified. Our findings provide a rich data set for future experimental analyses and characterizations.

**Supplementary Material**

Supplementary figures 1–9 and table 1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We thank Maria Thalén and Tobias Klingenduss for the initial compilation of sequences, Shaina Wallach and Samir Saha for critical reading of the manuscript, and two anonymous reviewers for their constructive suggestions. T.R.B. was supported by the Swedish Foundation for Strategic Research and the Karolinska Institutet, Söder-töms Högskola. L.B. was supported by the University of Florida Genetics Institute.

**Literature Cited**


Schrick K, Nguyen D, Karrowski WM, Mayer KF. 2004. START lipid/sterol-binding domains are amplified in plants and are predominantly associated with homeodomain transcription factors. Genome Biol. 5:R41.


Neelima Sinha, Associate Editor

Accepted August 20, 2009