

# Newly Discovered Plant *c-myb*-Like Genes Rewrite the Evolution of the Plant *myb* Gene Family<sup>1</sup>

Edward L. Braun and Erich Grotewold\*

Department of Plant Biology and Plant Biotechnology Center, The Ohio State University, Columbus, Ohio 43210

Transcription factors containing the conserved Myb DNA-binding domain were first recognized in the form of the *v-myb* oncogene of the avian myeloblastosis virus, but have subsequently been found in diverse eukaryotic groups (Lipsick, 1996). The flowering plants are characterized by the expression of a large number of *myb* genes (Martin and Paz-Ares, 1997), in striking contrast to the animals, fungi, and the cellular slime mold *Dictyostelium discoideum*, all of which express a limited number of regulatory proteins containing conserved Myb motifs (Lipsick, 1996). More than 100 *myb* genes encoding proteins exhibiting 40% to 60% identity with the Myb domain of the vertebrate c-Myb proto-oncoprotein are present in the Arabidopsis genome (Kranz et al., 1998; Romero et al., 1998), and a similar number of *myb* genes are expressed in maize (Rabinowicz et al., 1999). Members of the plant *myb* gene family are involved in the regulation of secondary metabolism, in the control of cell shape and fate, and in responses to hormones, drought, and viral infection (Urao et al., 1993; Gubler et al., 1995; Yang and Klessig, 1997) for review, see Martin and Paz-Ares, 1997. Analyses of synonymous and non-synonymous substitution rates indicate that most *myb* genes present in plants were generated by a series of gene duplications during a period of plant evolution 200 to 550 million years ago (Rabinowicz et al., 1999), prior to the divergence of monocots and dicots. Despite ongoing efforts to elucidate the functions of *myb* genes in plants and other eukaryotes, the origin and evolution of the diverse *myb* gene family in plants, as well as its relationship to *myb* genes in other eukaryotes, remains obscure.

Myb proteins are characterized by the presence of two or three Myb motifs, each of which contains a helix-turn-helix structure with three regularly spaced Trp residues (Lipsick, 1996; Martin and Paz-Ares, 1997). Myb proteins from animals and *D. discoideum*

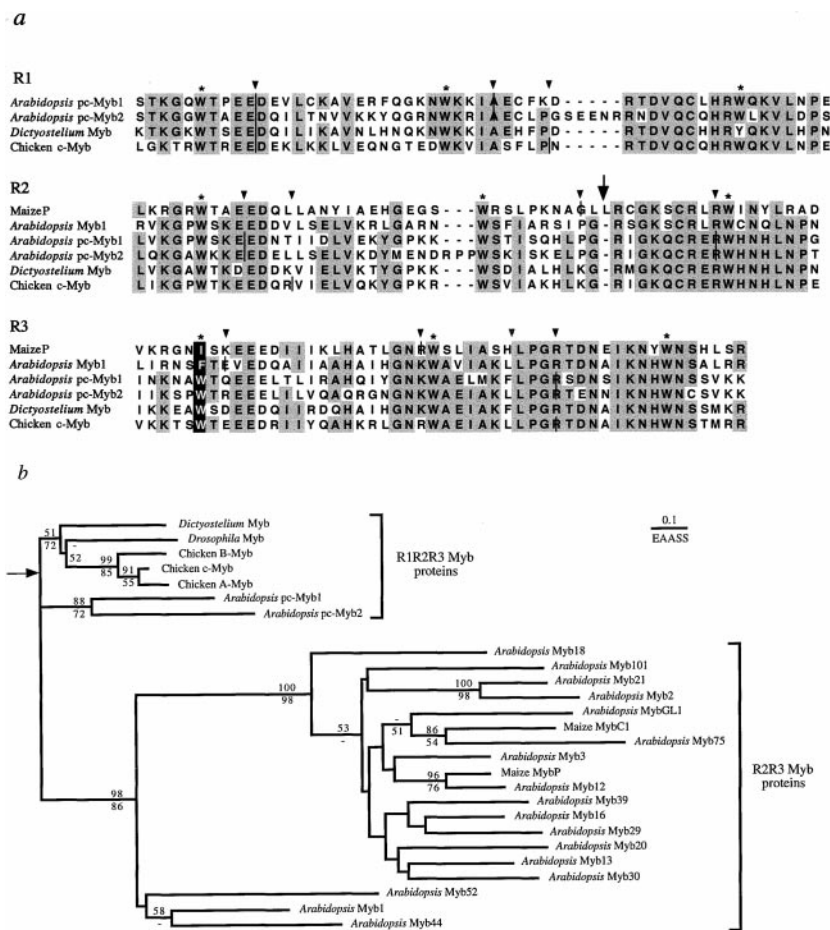
have three Myb motifs designated R1, R2, and R3 (Lipsick, 1996), while plant Myb-domain proteins have two Myb motifs corresponding to R2 and R3 (Martin and Paz-Ares, 1997) and forming the R2R3 *myb* gene family. We report the identification of Arabidopsis genes that encode proteins exhibiting structural features of the vertebrate c-Myb proto-oncoprotein, including the presence of three Myb motifs. We suggest designating these distinctive plant *c-myb*-like genes *pc-myb* to distinguish them from the larger R2R3 *myb* gene family encoding two repeat Myb proteins in higher plants.

The *pc-myb* genes were identified while searching Arabidopsis *myb* genes in the data produced by the genome sequencing project based upon the presence of a Trp residue in the first helix of R3 (Fig. 1a), which is characteristic of vertebrate Myb-domain proteins but occupied by a hydrophobic amino acid in all described plant R2R3 Myb proteins (Martin and Paz-Ares, 1997; Romero et al., 1998; Rabinowicz et al., 1999). Although only two Myb motifs were present in the annotated protein sequences for the *pc-myb* genes that we identified (accession nos. AL022537 for the BAC-containing *pc-myb1* and AF058919 for the BAC-containing *pc-myb2*), careful examination of the genomic sequences revealed the presence of short 5' exon sequences potentially encoding an R1 Myb motif within 1 kb of the first annotated exon. RT-PCR experiments were conducted with primer sequences unique to *pc-myb1* to confirm that these sequences indeed correspond to transcribed exons encoding an R1 motif, a further indication that the Arabidopsis *pc-myb1* gene corresponds to the first identified R1R2R3 *myb* gene in plants that has been shown to be transcribed. Based upon these results, we have deposited the sequences of the ORFs corresponding to *pc-myb1* and *pc-myb2* into the database (accession nos. AF151646 and AF151647).

A direct relationship between the *pc-myb* genes in plants and the vertebrate *c-myb* proto-oncogene is suggested by the presence of conserved introns in the R1 and R3 Myb motifs, one of which is also present in *D. discoideum* (Fig. 1a). The position of these introns is very different from those found in the R2R3 *myb* gene family (Romero et al., 1998; E.L. Braun and E. Grote-

<sup>1</sup> This work was funded in part by a grant from the National Science Foundation (no. MCB-9896111) and by a seed grant from the Ohio State University Office of Research (to E.G.).

\* Corresponding author; e-mail grotewold.1@osu.edu; fax 614-292-5379.



**Figure 1.** Structural and evolutionary relationship of Arabidopsis *pc-myb* genes to other *myb* genes. A, Sequence alignment of *pc-Myb* proteins with selected *Myb* proteins from animals, plants, and *D. discoideum* (Klempnauer et al., 1982; Grotewold et al., 1991; Shinozaki et al., 1992; Stober-Grasser et al., 1992). The positions of the conserved Trp residues are indicated with asterisks, and the Trp residue in R3 that has been replaced by a hydrophobic residue in the plant R2R3 *Myb* protein is indicated with a black background. The position of a single amino acid insertion present in R2 of many plant R2R3 *Myb* proteins is indicated with an arrow. The positions of introns present in the genes encoding these *Myb* proteins are indicated with lines across the relevant sequences and arrowheads above the alignment. Another intron position found in some R2R3 *myb* genes is indicated with a gray arrowhead. Primers used for RT-PCR, which was performed on total RNA isolated from whole Arabidopsis (ecotype Columbia) plants grown for 2 weeks, are positioned outside of the aligned region shown. B, Estimate of phylogeny based upon the amino acid sequences of the R2 and R3 *Myb* motifs. R2R3 sequences from three repeat *Myb*-domain proteins and a diverse set of Arabidopsis and maize R2R3 *Myb* sequences (selected based upon Kranz et al., 1998) were aligned. The phylogeny of *Myb*-domain proteins was estimated by neighbor joining of accepted point mutation (PAM) distances using the PHYLIP program (Phylogeny Inference Package, version 3.57c, Department of Genetics, University of Washington, Seattle). The scale bar corresponds to 0.1 estimated amino acid substitutions per site (EAASS) under the PAM model of sequence evolution (Dayhoff et al., 1978). Bootstrap proportions from 500 replicates using neighbor joining of PAM distances are presented as percentages above the branches and bootstrap proportions from unweighted parsimony analysis of the amino acid sequences using PAUP (Phylogenetic Analysis Using Parsimony, version 4.0b2, Sinauer Associates, Sunderland, MA) are presented below the branches. Bootstrap proportions below 50% are not presented. Although this tree is unrooted, we have indicated the probable position of the root with an arrow. This placement of the root is suggested by our model of *Myb* evolution (see text) and is consistent with the phylogeny of these organisms (Braun et al., 1998).

wold, unpublished observations). The high degree of sequence identity in the *Myb* domain between the *pc-Myb* proteins and the vertebrate c-*Myb* proto-oncoprotein (64% identity for *pc-Myb1* and 62%

identity for *pc-Myb2*) further emphasizes the close relationship between three repeat *Myb* proteins in plants and those in animals. Despite the close relationship within the *Myb* domain between the *pc-Myb*

proteins and the vertebrate Myb proteins, neither pc-Myb1 nor pc-Myb2 have detectable homology to each other or to the vertebrate Myb proteins outside of the conserved Myb domain. However, it is clear that Myb-domain proteins can be extremely divergent outside of the conserved Myb domain (Rosinski and Atchley, 1998), and there is no detectable sequence homology between the *D. discoideum* Myb protein and the animal Myb proteins, despite the closer relationship between these organisms (Braun et al., 1998).

Myb domains formed by multiple Myb repeats probably arose by duplication of an ancestral Myb motif. It has been proposed that the duplication of R2 in an early form of two repeat Myb proteins gave rise to the R1R2R3 Myb domains (Rosinski and Atchley, 1998). Our finding that higher plants express *R1R2R3 myb* genes with intron-exon structures strikingly similar to vertebrate *myb* genes strongly suggests that the R1R2R3 Myb domains formed prior to the divergence of plants and animals. Indeed, the R1 repeats of pc-Myb1, pc-Myb2, and the vertebrate R1R2R3 Myb proteins are more closely related to each other than to the R2 Myb motif in the corresponding protein.

The close relationship between the pc-Myb proteins and the R1R2R3 Myb proteins present in animals and *D. discoideum* suggested by their intron-exon structure was confirmed by phylogenetic analyses of R2R3 Myb sequences (Fig. 1b). These analyses also suggest that the gene duplication resulting in *pc-myb1* and *pc-myb2* occurred after the divergence of plants from other eukaryotic groups. Indeed, the existence of multiple *pc-myb* genes in Arabidopsis suggests that these genes form a small gene family similar to the one encoding the A-Myb, B-Myb, and c-Myb proteins, which resulted from duplications within animals (Rosinski and Atchley 1998; Fig. 1b). Given the close relationship between the pc-Myb proteins and the vertebrate c-Myb proto-oncoprotein, it is possible that the pc-Myb proteins provide functions in plants similar to those provided by the three Myb motif proteins in animals, including the control of cellular proliferation and differentiation (Lipsick, 1996; Weston, 1998). Regulators of such processes in plants remain poorly characterized, but a role for Myb-domain proteins with DNA-binding specificities similar to those of vertebrate Myb proteins has been suggested for the regulation of the plant cell cycle (Chung and Parish, 1995; Ito et al., 1998; for review, see Doonan and Fobert, 1997).

The relationship between the diverse R2R3 *myb* gene family in plants and the R1R2R3 *myb* gene family has been difficult to establish. It has been proposed that either the plant R2R3 *myb* gene family reflects the duplication of an ancestral R1R2R3 *myb* gene after the loss of the R1 motif or the R2R3 *myb*

gene family represents an ancient group of genes that diversified within the higher plants (Lipsick, 1996; Martin and Paz-Ares, 1997; Rosinski and Atchley, 1998). The first hypothesis is more parsimonious when one considers the rates of evolution for both types of Myb proteins, which suggest that the amino acid substitutions have accumulated several times more rapidly in the R2R3 Myb proteins present in higher plants than in the R1R2R3 Myb proteins of animals (E.L. Braun and E. Grotewold, unpublished results). However, additional complexity in the evolution of plant R2R3 Myb proteins is revealed by our phylogenetic analyses (Fig. 1b), which divide these proteins into a larger group of proteins exhibiting a relatively high degree of divergence from the vertebrate c-Myb proto-oncoprotein (40%–52% identity in the Myb domain) and a smaller group of proteins with a two-repeat structure but a higher degree of sequence identity to c-Myb (48%–60% identity in the Myb domain). A major difference between these two groups of R2R3 Myb proteins is a single amino acid insertion in R2 (indicated with an arrow in Fig. 1a).

These analyses suggest a model in which the diverse R2R3 *myb* gene family in plants arose by the loss of the R1 motif from a *pc-myb*-like gene. After the loss of R1, the higher degree of divergence between plant R2R3 Myb proteins and the vertebrate c-Myb proto-oncoprotein suggests that the rate at which amino acid substitutions accumulated in R2R3 Myb proteins increased. Within the R2R3 *myb* gene family, there have been additional substitutions, such as the insertion of a functionally relevant residue in R2 (Williams and Grotewold, 1997), changes in the intron-exon structure (Romero et al., 1998; E.L. Braun and E. Grotewold, unpublished observations), and additional increases in the rate of molecular evolution. Establishing the cellular functions and evolutionary dynamics of these novel *pc-myb* genes is likely to provide additional information about the early evolution of this remarkably diverse family of plant genes.

Received May 24, 1999; accepted May 27, 1999.

#### LITERATURE CITED

- Braun EL, Kang S, Nelson MA, Natvig DO (1998) *J Mol Evol* 47: 531–543
- Chung SK, Parish RW (1995) *FEBS Lett* 362: 215–219
- Dayhoff MO, Schwartz RM, Orcutt BC (1978) In MO Dayhoff, ed, *Atlas of Protein Sequence and Structure*, Vol 5, Suppl 3. National Biomedical Research Foundation, Silver Spring, MD, pp 345–352
- Doonan J, Fobert P (1997) *Curr Opin Cell Biol* 9: 824–830
- Grotewold E, Athma P, Peterson T (1991) *Proc Natl Acad Sci USA* 88: 4587–4591
- Gubler F, Kalla R, Roberts JK, Jacobsen JV (1995) *Plant Cell* 7: 1879–1891

- Ito M, Iwase M, Kodama H, Lavis P, Komamine A, Nishihama R, Machida Y, Watanabe A (1998) *Plant Cell* **10**: 331–341
- Klempnauer K-H, Gonda TJ, Bishop JM (1982) *Cell* **31**: 453–463
- Kranz HD, Denekamp M, Greco R, Jin H, Leyva A, Meissner RC, Petroni K, Urzainqui A, Bevan M, Martin C, Smeekens S, Tonelli C, Paz-Ares J, Weisshaar B (1998) *Plant J* **16**: 263–276
- Lipsick JS (1996) *Oncogene* **13**: 223–235
- Martin C, Paz-Ares J (1997) *Trends Genet* **13**: 67–73
- Rabinowicz PD, Braun EL, Wolfe AD, Bowen B, Grotewold E (1999) *Genetics* (in press)
- Romero I, Fuertes A, Benito MJ, Malpica JM, Leyva A, Paz-Ares J (1998) *Plant J* **14**: 273–284
- Rosinski JA, Atchley WR (1998) *J Mol Evol* **46**: 74–83
- Shinozaki K, Yamaguchi-Shinozaki K, Urao T, Koizumi M (1992) *Plant Mol Biol* **19**: 493–499
- Stober-Grasser U, Brydolf B, Bin X, Grasser F, Firtel RA, Lipsick JS (1992) *Oncogene* **7**: 589–596
- Urao T, Yamaguchi-Shinozaki K, Urao S, Shinozaki K (1993) *Plant Cell* **5**: 1529–1539
- Weston K (1998) *Curr Opin Genet Dev* **8**: 76–81
- Williams CE, Grotewold E (1997) *J Biol Chem* **272**: 563–571
- Yang Y, Klessig DF (1997) *Proc Natl Acad Sci USA* **93**: 14972–14977