Recently Duplicated Maize R2R3 Myb Genes Provide Evidence for Distinct Mechanisms of Evolutionary Divergence after Duplication

Anusha P. Dias, Edward L. Braun, Michael D. McMullen, and Erich Grotewold*

Department of Plant Biology and Plant Biotechnology Center, Ohio State University, Columbus, Ohio 43210 (A.P.D., E.G.); Department of Zoology, University of Florida, Gainesville, Florida 32611 (E.L.B.); and Plant Genetics Research and Plant Science Units, United States Department of Agriculture-Agricultural Research Service, University of Missouri, Columbia, Missouri 65211 (M.D.M.)

R2R3 Myb genes are widely distributed in the higher plants and comprise one of the largest known families of regulatory proteins. Here, we provide an evolutionary framework that helps explain the origin of the plant-specific R2R3 Myb genes from widely distributed R1R2R3 Myb genes, through a series of well-established steps. To understand the routes of sequence divergence that followed Myb gene duplication, we supplemented the information available on recently duplicated maize (Zea mays) R2R3 Myb genes (C1/P11 and P11/P2) by cloning and characterizing ZmMyb-IF35 and ZmMyb-IF25. These two genes correspond to the recently expanded P-to-A group of maize R2R3 Myb genes. Although the origins of C1/P11 and ZmMyb-IF35/ZmMyb-IF25 are associated with the segmental allotetraploid origin of the maize genome, other gene duplication events also shaped the P-to-A clade. Our analyses indicate that some recently duplicated Myb gene pairs display substantial differences in the numbers of synonymous substitutions that have accumulated in the conserved MYB domain and the divergent C-terminal regions. Thus, differences in the accumulation of substitutions during evolution can explain in part the rapid divergence of C-terminal regions for these proteins in some cases. Contrary to previous studies, we show that the divergent C termini of these R2R3 MYB proteins are subject to purifying selection. Our results provide an in-depth analysis of the sequence divergence for some recently duplicated R2R3 Myb genes, yielding important information on general patterns of evolution for this large family of plant regulatory genes.

Gene duplications have long been viewed as a main source of evolutionary novelty (Ohno, 1970). The dramatic expansion of the R2R3 Myb family of regulatory genes in the higher plants provides a striking example of how gene amplifications followed by divergence may have impacted plant evolution. Members of this gene family encode proteins characterized by two 50- to 52-residue-long imperfect repeats. Each of these MYB repeats contains three α-helices, with the second and third helices forming a helix-turn-helix structure when bound to DNA (Ogata et al., 1994). Around 125 R2R3 Myb genes are present in the Arabidopsis genome (Reichmann and Ratcliffe, 2000; Stracke et al., 2001), and more than 200 in maize (Zea mays) and related monocots (E.L. Braun and E. Grotewold, unpublished data). This represents a sharp contrast to the small number of genes encoding MYB homologs in the animal and fungal kingdoms (Lipsick, 1996).

Plant R2R3 Myb genes originated from an ancestral gene encoding a three-MYB repeat protein represented in animals today by c-myc and related genes (Lipsick, 1996), and in plants by the small pc-Myb gene family (Braun and Grotewold, 1999a). After the loss of the R1 repeat, an explosive amplification of the R2R3 Myb gene family occurred 250 to 400 million years ago (Mya) (Rabinowicz et al., 1999). R2R3 Myb genes are more or less evenly distributed throughout the plant genome without forming obvious clusters, as found for plant resistance (R) genes, for example (Bergelson et al., 2001).

R2R3 MYB proteins are characterized by the presence of a conserved MYB domain and a longer divergent C-terminal region. Short conserved motifs in the C-terminal region of these proteins have been identified and were used to classify Arabidopsis R2R3 Myb genes into subgroups (Kranz et al., 1998; Stracke et al., 2001). The dramatic divergence of the C-terminal regions does not appear to have a large influence in the regulatory function of the corresponding proteins. For example, related R2R3 Myb genes that act as regulators of anthocyanin biosynthesis in maize, petunia (Petunia hybrida), and Arabidopsis (encoded by C1, An2, and Pap1, respectively) show little or no detectable identity outside their
MYB domains. Despite this C-terminal divergence, the maize C1 protein can complement petunia An2 mutants, and vice versa (Quattrocchio et al., 1999, 1993). Similarly, the Arabidopsis GL1 and WER proteins are interchangeable in their ability to regulate root hair or trichome formation, despite sharing only 23% C-terminal identity (Lee and Schiefelbein, 2001). Finally, domain-swapping experiments involving maize P1 and C1 suggest regulatory specificity is largely provided by the MYB domains of these proteins (Grotewold et al., 2000).

The high degree of sequence divergence for the C-terminal regions of many R2R3 Myb proteins, coupled with the apparent absence of functional constraints upon these regions, suggests that they may be diverging at the neutral rate. However, some blocks of C-terminal sequence identity in specific plant R2R3 MYB proteins have persisted over very long evolutionary times. The most compelling example is a block of >50 amino acids with >40% identity found in a moss (Physcomitrella patens) MYB protein and the snapdragon (Antirrhinum majus) MIXTA protein (Kranz et al., 1998). Detailed comparison of sequence divergence in the MYB domains and C-terminal regions of R2R3 MYB proteins might provide insights into these apparent contradictions, and the first attempt to compare the evolution of different regions in MYB proteins found substantial disagreement between trees identified using different regions (Rosinski and Atchley, 1998). This prompted the conclusion that MYB proteins are polyphyletic (Rosinski and Atchley, 1998), a term used for groups based upon shared characters that reflect evolutionary convergence (for review, see Page and Holmes, 1998). The conclusion that Myb genes are polyphylectic must be viewed with caution because incongruence between phylogenetic trees estimated using different regions of the same gene could reflect difficulties with the alignment or analysis rather than genuine differences in tree topology. In fact, it has been demonstrated that random data can show significant incongruence with phylogenetically structured data (Dolphin et al., 2000).

Previous studies have provided a number of specific hypotheses regarding the patterns of evolution for different regions of R2R3 MYB proteins, which differ regarding the rate of evolution and degree of functional constraint upon the C-terminal regions (Braun and Grotewold, 1999a; Stracke et al., 2001). Examining the degree of functional constraint upon proteins using evolutionary comparisons is typically accomplished by calculating the ratio of non-synonymous (K_a) to synonymous (K_s) substitutions, a value designated \( \omega \). If there are no constraints upon the amino acid sequence, estimates of \( \omega \) will equal one. In contrast, the purifying selection typical of most proteins will result in estimates of \( \omega < 1 \), whereas the rapid fixation of non-synonymous substitutions that are selectively advantageous will result in estimates of \( \omega > 1 \) for those sites or regions subject to positive selection. Estimates of \( \omega \) will exhibit a high variance when they are calculated using data from ancient gene duplications because synonymous sites are expected to saturate rapidly in plants (see Rabinowicz et al., 1999). However, the identification of a group of R2R3 Myb genes that has undergone a recent amplification in the grasses (Braun and Grotewold, 1999b; Rabinowicz et al., 1999) provides us with an excellent opportunity to investigate the patterns of evolution of recently duplicated R2R3 Myb genes.

In this study, we establish an evolutionary framework to explain events that followed the origin of R2R3 Myb genes from widely distributed c-Myb-like genes. To investigate patterns of evolution responsible for the striking divergence that characterizes plant R2R3 Myb genes, we focused upon recently duplicated R2R3 Myb genes. Toward this goal, we characterized ZmMyb-1F35 and ZmMyb-1F25, two maize R2R3 Myb genes that belong to a gene clade that expanded during the evolution of the grasses (Rabinowicz et al., 1999). The physical map positions of these genes were determined and found to be consistent with an origin during the duplication of the maize genome. Comparisons of these and related R2R3 Myb genes provided evidence that MYB domains and C termini are both subject to purifying selection. Furthermore, our analyses indicate substantial heterogeneity in the number of non-synonymous and synonymous substitutions in different regions of recently duplicated R2R3 Myb genes. Together, these studies provide the first insights, to our knowledge, into the possible mechanisms of divergence that accompanied the dramatic expansion of the R2R3 Myb gene family in the plants.

RESULTS AND DISCUSSION

An Evolutionary Framework for Plant R2R3 MYB Proteins

Plant MYB proteins are encoded by a very large and diverse gene family (Braun and Grotewold, 1999a; Rabinowicz et al., 1999; Stracke et al., 2001), with most plant Myb homologs characterized by two MYB repeats and the insertion of a single amino acid (Leu) in the first (R2) repeat, relative to animal Myb homologs (Fig. 1A). To provide a framework to examine plant Myb gene evolution, we estimated the phylogeny of a large number of Arabidopsis and selected monocot Myb proteins representing all major groups. Two data sets were used, a small alignment (45 sequences) and a large alignment (106 sequences; see supplemental data at www.plantphysiol.org). The topology of the trees obtained using these data sets, or by including additional plant Myb proteins from our own collection or available from public databases, were essentially the same with none of the well-supported clade shown in Figure 1A rearranged (not
shown). Analyses of both alignments revealed two distinct types of plant Myb homologous that lack the Leu insertion in the R2 repeat: (a) the three repeat pc-Myb proteins (Braun and Grotewold, 1999a), also described as 3Rmyb (Kranz et al., 2000); and (b) a subset of MYB proteins designated the "atypical" R2R3 MYB proteins (Braun et al., 2001). Surprisingly, some of these atypical R2R3 MYB proteins (e.g. AtMyb22) have a Trp residue in the first helix of R3, like the pc-Myb proteins but unlike the majority of plant MYB proteins (Braun and Grotewold, 1999a).

The root of the plant Myb phylogeny is likely to be between the pc-Myb proteins and the R2R3 MYB proteins (indicated with an arrow in Fig. 1A). Three repeat MYB homologs are broadly distributed in animals, plants, and slime molds (Lipsick, 1996; Braun and Grotewold, 1999a), whereas typical R2R3 MYB proteins are limited to the green plant lineage (Braun et al., 2001; Stracke et al., 2001). Restricting our consideration to plant MYB proteins, a three-step model for the origin for typical R2R3 MYB proteins is suggested (Fig. 1A). This model would involve loss of the R1 repeat (branch α), mutation of the Trp residue in the first helix of R3 to a Phe (branch β), and insertion of the Leu residue in R2 (branch γ). Loss of the first repeat (branch α) occurred before divergence of land plants and chlorophyte algae because an atypical R2R3 Myb is present in C. reinhardtii (Fig. 1A). The Leu insertion and diversification of the typical R2R3 MYB proteins occurred before the divergence of mosses and angiosperms because a typical R2R3 Myb (PpMyb2) is present in the moss P. patens (Leech et al., 1993).

Previous analyses suggest that many gene duplications in the family encoding typical R2R3 MYB proteins occurred early in the history of land plants (Rabinowicz et al., 1999). This model of rapid diversification early in the history of land plants is consistent with the relatively short and poorly supported (by the bootstrap and posterior probabilities in Bayesian analyses) branches at the base of the typical R2R3 MYB group. Establishing the mechanisms by which the plant R2R3 Myb gene family expanded and diverged is complicated by the ancient nature of...
these duplications. However, several recently duplicated groups of R2R3 Myb genes have been identified in maize (Braun and Grotewold, 1999b), and the analysis of members of one of these groups provides clues on the possible mechanisms of evolutionary divergence after gene duplication of R2R3 Myb genes.

ZmMyb-IF25 and ZmMyb-IF35 Belong to a Group of Recently Duplicated R2R3 Myb Genes

We previously have analyzed more than 80 sequences of Myb<br> and these studies provided evidence that several groups of maize R2R3 Myb genes amplified within the past 50 million years (Braun and Grotewold, 1999b; Rabinowicz et al., 1999). One of these groups corresponds to the P-to-A clade, united by the change of Pro-63 to Ala (Fig. 1A) in the hinge region between the R2 and R3 MYB repeats (Rabinowicz et al., 1999). The analysis of the Myb<br> sequences suggested two types of genes within the P-to-A clade: (a) those diverging at a relatively low rate (e.g. P1, a regulator of 3-deoxy flavonoid biosynthesis; Grotewold et al., 1994); and (b) those diverging more rapidly (e.g. ZmMyb-IF25 and ZmMyb-IF35; Rabinowicz et al., 1999).

To better understand the patterns of evolution of these recently duplicated genes, we cloned ZmMyb-IF35 (accession no. AF521880) and ZmMyb-IF25 (accession no. AF521881; see “Materials and Methods”). A full-length cDNA clone for ZmMyb-IF25 was obtained from a yeast (Saccharomyces cerevisiae) one-hybrid screen, using the high-affinity P-binding sites present in the A1 gene (Grotewold et al., 1994), in an effort to identify additional proteins that recognize this binding site (M.-G. Kim and E. Grotewold, unpublished data). The molecular analysis of the genomic and cDNA clones showed that the intron-exon structure of ZmMyb-IF35 and ZmMyb-IF25 (Fig. 2A) was identical to that of the maize P1, C1, and P11 genes (Paz-Ares et al., 1987; Grotewold et al., 1991; Cone et al., 1993). Like in P1 (Grotewold et al., 1991), the second intron of ZmMyb-IF35 and ZmMyb-IF25 is unusually long, more than 2 kb for ZmMyb-IF35 and over 6 kb for ZmMyb-IF25.

We mapped the ZmMyb-IF35 and ZmMyb-IF25 loci using polymorphisms in the 5’ sequence. Primers designed from the 5’ region of ZmMyb-IF35 detected a direct size polymorphism between B73 and Mo17, the parents of the intermated B73/Mo17 (IBM) mapping population (Lee et al., 2002). Genotypes were determined for the 93 F2 individuals from this population and ZmMyb-IF25 was mapped against a framework map of 96 loci. The IBM and T218 X GT119 maps are available at MaizeDB (http://www.agron.missouri.edu/maps.html). ZmMyb-IF35 mapped to chromosome 3, bin 3.04 and ZmMyb-IF25 mapped to chromosome 8, bin 8.03–8.04. These map positions correspond to duplicated regions of the maize genome, based upon comparative mapping analyses (Helenjartis et al., 1988; Gaut and Doebley, 1997). Thus, similar to the maize regulators of anthocyanin biosynthesis C1 and P11, ZmMyb-IF35 and ZmMyb-IF25 appear to have diverged during the reversion of the segmental allotetraploid ancestor of maize to disomic inheritance, an event that has been estimated to have occurred approximately 11 Mya (Gaut and Doebley, 1997).

Patterns of Evolution of Recently Duplicated R2R3 Myb Genes

The origin of the P-to-A clade precedes the divergence of monocots and eudicots. Evidence for this is provided by the presence of two Arabidopsis sequences (AtMyb11 and AtMyb12) with a similar change of Pro-63 to Ser, AtMyb111 with a change of Pro-63 to Arg, and the cotton (Gossypium hirsutum) GhMyb-J that also has a Pro-63 to Ala substitution (Loguerio et al., 1999). Molecular clock estimates indicated that the expansion of this group to yield a large number of maize paralogs was an event unique to the grasses (Rabinowicz et al., 1999), a clade of plants that diversified within the past 55 million years. Consistent with this hypothesis, AtMyb11, AtMyb12, P1, ZmMyb-IF35, and ZmMyb-IF25 form a well-supported clade, with AtMyb11 and AtMyb12 basal to a monophyletic group of monocot P-to-A R2R3 Myb sequences (Fig. 1, A and B). However, two rice genes encoding P-to-A Mybs (accession nos. BABA64029 and AAL84631) provided evidence that the ZmMyb-IF35 and ZmMyb-IF25 genes diverged from the P lineage, characterized by the recently duplicated P1 and P2 genes (Zhang et al., 2000).

The evidence for orthology of the rice P-to-A sequences and the maize P1, ZmMyb-IF35, and ZmMyb-IF25 genes was strengthened by finding substantial C-terminal identity between these genes (Fig. 2B), prompting us to call the rice genes OsMyb-P and OsMyb-IF. The genomic sequence of OsMyb-P (accession no. AF474141) and OsMyb-IF (accession no. AP002873) shows the presence of relatively long (>4 kb) introns in the same position as the second intron in P1. The OsMyb-IF gene lacks an intron in a position homologous to the first intron in P1 (Fig. 2A), whereas other grass P-to-A Myb genes have short introns in this position.

The sister group to the P-to-A clade includes several anthocyanin regulators (e.g. the products of the maize C1 and the petunia An2 genes). Interestingly,
we found limited support for a clade containing At-Myb123 and C1/P1i from maize, but excluding At-Myb75 (Pap1; Borevitz et al., 2000) and An2 from petunia. Based upon the exchangeability of C1 and An2 (Quattrocchio et al., 1993, 1999) and clustering in previous phylogenetic analyses (Braun and Grotewold, 1999b; Rabinowicz et al., 1999), AtMyb75, petunia An2, and maize C1 had been thought to be orthologous. However, C1, Pl1, and AtMyb123 have two short boxes of identity in the C-terminal region. One of these boxes is a nine-amino acid signature previously reported (Stracke et al., 2001).

Figure 2. Sequence comparisons of recently duplicated R2R3 Myb genes. A. Amino acid sequence alignment in the MYB domain. Comparison of amino acid sequences of the predicted proteins encoded by ZmMyb-IF25, ZmMyb-IF35, OsMyb-IF, OsMyb-P, P1, C1, and Pl1. Dark-shaded boxes indicate identical amino acids and lighter shaded boxes indicate similar amino acids. The positions of the three helices forming the R2 and R3 MYB repeats are shown with the clear boxes and the MYB domain interrupted by two introns are indicated with arrows. The change from Pro to Ala that defines the MybPtoA clade is marked with a star. B. Amino acid sequence alignment for the C termini. Comparison of the predicted C-terminal regions encoded by MybPtoA clade members ZmMyb-IF25, ZmMyb-IF35, OsMyb-IF, OsMyb-P, and P1. Dashes indicate gaps introduced to reflect insertions and deletions during evolution.
second is a 15-amino acid signature starting with amino acid 172 in C1 (E.L. Braun, unpublished data). These conserved motifs in the divergent C-terminal regions of these proteins, together with our phylogenetic analyses (Fig. 1), provide additional evidence that C1 and AtMyb123 are orthologs. Interestingly, AtMyb123 (TT2) was characterized to be a regulator of protoanthercyanidin accumulation in developing seed (Nesi et al., 2001). It is likely that C1 and An2 are paralogs related by a duplication event before the divergence of monocots and eudicots, although they have clearly retained similar functions (Quattrocchio et al., 1999).

Taken together, these findings provide evidence for a model in which: (a) the origin of the P-to-A clade precedes the divergence of monocots and eudicots; (b) this group of R2R3 Myb genes underwent a recent expansion in the grasses; and (c) this expansion involved genome duplication (e.g., ZmMyb-IF35 and ZmMyb-IF25), tandem gene duplication (e.g., P1 and P2; Zhang et al., 2000), and a more ancient duplication (e.g., the P1/P2 and ZmMyb-IF35/ZmMyb-IF25 lineages). The study of the divergence patterns of R2R3 Myb genes that duplicated recently by diverse mechanisms provides a unique opportunity to understand the general evolutionary patterns of plant R2R3 Myb genes.

Divergence of the Conserved MYB Domains and Variable C-Terminal Regions

A general characteristic of plant R2R3 Myb genes is the high conservation of the MYB domains coupled with the dramatic divergence of C-terminal regions. Does this divergence difference reflect functional constraints upon the MYB domain with the C-terminal regions evolving at the neutral rate? Or are MYB domains and C-terminal regions both subject to purifying selection? To investigate the constraints upon the MYB and C-terminal regions of recently duplicated R2R3 Myb genes, the numbers of $K_A$ and $K_S$ substitutions were estimated for pairs of recently duplicated R2R3 Myb genes (Table I). As expected, separate estimates of $\omega$ ($K_A/K_S$) from the MYB and C-terminal regions indicated more constraints upon MYB domains than upon C-terminal regions. However, estimates of $\omega$ for C-terminal regions were lower than expected under the neutral model (Table I), suggesting the existence of moderate purifying selection for these regions.

As a heuristic to examine which regions of these recently duplicated pairs of R2R3 Myb genes had accumulated more non-synonymous substitutions, we conducted a “sliding window” analysis (Fig. 3). These analyses provided evidence for functional constraints upon the MYB domain, consistent with the key role of MYB domains in DNA binding (for review, see Lipsick, 1996) and protein-protein interactions (Grotewold et al., 2000). However, the C-terminal regions of all pairs examined also show regions exhibiting very limited non-synonymous divergence (Fig. 3).

Estimates of $K_S$ were substantially higher for the segments of C1/Pl1 and AtMyb11/AtMyb12 that encode the C-terminal regions of these gene products than for the segments encoding the MYB domains of these proteins (Table I; Fig. 3). The divergent C-terminal regions of AtMyb75/AtMyb90 have accumulated slightly fewer synonymous differences than the highly similar MYB domains. However, the difficulty in estimating numbers of synonymous substitutions suggests that it is reasonable to conclude that any differences in $K_S$ for the MYB domain and the C terminal of both ZmMyb-IF35/ZmMyb-IF25 and AtMyb75/AtMyb90 are extremely modest.

These findings provide strong evidence that the C-terminal regions of all pairs of R2R3 Myb genes studied here are subject to moderate purifying selection and confirm the fact that MYB domains are subject to strong purifying selection. The existence of moderate purifying selection upon C-terminal regions is surprising because the R2R3 Myb genes examined here are members of clades for which experiments have demonstrated limited functional roles for the C-terminal domains (Goff et al., 1991; Sainz et al., 1997; Grotewold et al., 2000). However, based on the evolutionary analyses presented here, we propose that the C termini of these proteins are likely to play a more important role in the function of R2R3 MYB transcription factors than previously anticipated.

These data also provide evidence that $K_A$ and $K_S$ values exhibit a positive intragenic correlation in some genes, extending the results of studies using mammalian sequences (Alvarez-Valin et al., 1998; Smith and Hurst, 1999) to the flowering plants. Clearly, these results challenge the use of $K_S$ values to provide a time scale for evolution, and question the biological basis for the observed infragenic $K_A$-$K_S$ correlation.

| Table 1. Synonymous ($K_S$) and non-synonymous ($K_A$) distances between the pairs of recently duplicated R2R3 Myb genes examined for this study |
|-----------------|-----------------|-----------------|-----------------|
| Pair Comparison | $K_S$ (Mean/Myb/C Term) | $K_A$ (Mean/Myb/C Term) | $\omega$ (Mean/Myb/C Term) |
| C1/Pl1          | 0.2515/0.1239/0.2819 | 0.0707/0.0236/0.1354 | 0.2811/0.1907/0.4803 |
| AtMyb75/AtMyb90 | 0.4368/0.4970/0.3776 | 0.1254/0.0443/0.2558 | 0.2871/0.0891/0.6775 |
| ZmMyb-IF35/ZmMyb-IF25 | 0.4871/0.4708/0.5212 | 0.1136/0.0305/0.1654 | 0.2331/0.0647/0.3174 |
| AtMyb11/AtMyb12 | 1.4303/0.6542/1.3962 | 0.2748/0.0563/0.4383 | 0.1921/0.0861/0.3139 |
Understanding the Divergence of C-Terminal Regions between R2R3 Myb Genes

Although sliding window analyses did show some regions with $\omega = 1$, consistent with specific regions evolving under a neutral model, there were no regions where estimates of $\omega$ greatly exceeded values consistent with the neutral or purifying selection models of evolution. These observations suggest that recently duplicated R2R3 Myb genes related to P1...
and C1 in maize have not diverged because of the action of positive selection, suggested as a possible model for the evolution of specific MADS box transcription factors in species of the genera *Argyroserpium*, *Dubautia*, and *Wilkesia*, a group known as the Hawaiian silverswords (Barrier et al., 2001). Thus, our results suggest that the observed non-synonymous acceleration does not represent a general feature of transcription factor genes in polyploids, because this study did not show evidence for accelerated accumulation of non-synonymous substitutions in maize, an ancient polyploid (Gaut and Doebley, 1997).

The C-terminal regions of R2R3 MYB proteins also accumulate a number of insertion and deletion changes (e.g. Fig. 2), making it difficult to align these regions for proteins that diverged more than 100 Mya. The difficulty in aligning these regions could be largely responsible for the observed incongruence in the evolution rate between different regions of R2R3
MYB proteins. This model cannot be excluded by the results of this study, unlike the neutral model that was excluded by the values of $\omega < 1$ for R2R3 MYB protein C-terminal regions that we observed. The notion that much of the incongruence between the MYB domain and C-terminal regions reflects difficulty in alignment represents an excellent alternative to the polyphyly model of Rosinski and Atchley (1998), and it is consistent with the observation that many R2R3 MYB clades identified by analyses of the MYB domain share C-terminal motifs. It remains possible that domain shuffling have contributed to the divergence of specific Myb genes. In fact, some alleles of the $P1$ gene have a putative Cys-containing metal binding domain at their C terminus that could have arisen by domain shuffling (Chopra et al., 1996). However, there is no evidence that this pattern of evolution is more common than divergence at different rates coupled with substantial numbers of insertion-deletion mutations.

A surprising aspect of the current study is the fact that regions of R2R3 Myb genes that encode the divergent C-terminal regions often exhibit higher $K_s$ values than the regions encoding the MYB domains (see the comparisons of C1/P11 and AtMyb11/AtMyb12; Table I). The observation in this study is difficult to reconcile with neutral models, which postulate that different regions of the genome show different rates of mutation. Likewise, we believe that models that invoke selection for specific mRNA structures (e.g., Eyre-Walker and Bulmer, 1993) represent an unlikely explanation for the observations in this study. The GC-content of genes should have a strong impact upon RNA structure, so it is likely that selection on synonymous sites because of mRNA structure would be stronger and more consistent in the GC-rich genes of grasses. However, the higher apparent rate of synonymous substitution was observed for one pair of genes in maize and one pair of genes in Arabidopsis, despite the very different GC-contents of genes in these organisms.

There are two potential models consistent with all of the available data from paralogous R2R3 Myb genes. It is possible that there are differences in the rate of mutation in different regions of transcription units, a plausible hypothesis considering the involvement of some proteins in both transcription and in DNA repair (Lehmann, 1998). However, we believe a simpler model might be the impact of gene conversions in the conserved regions corresponding to the MYB domains of these regions. This model would explain the more limited divergence of these regions at both synonymous and non-synonymous sites without invoking different rates of mutation. This model would also explain the loss of introns in the conserved region of some R2R3 Myb genes, including OsMYB-IF and ZmMyb-B2 (Rabinowicz and Grotewold, 2000), because this loss could involve gene conversion with the products of reverse transcription after splicing.

Although a general characteristic of the R2R3 Myb gene family is the striking divergence of C-terminal regions, this feature is by no means unique to R2R3 Myb genes. For example, MADS box genes, another large family of genes encoding transcriptional regulators in the green plants (Theissen et al., 1996; Purugganan, 1997), also show extreme sequence divergence in the C termini. Unlike the R2R3 MYB proteins where the C-terminal region can be very large, the divergent C-terminal region of MADS-box proteins usually corresponds to less than one-third of the entire protein. Extending these types of studies to either to a genome-wide scale for Myb genes or other gene families should provide more power to test among distinct evolutionary models. However, detailed analyses of specific genes, such as those conducted here, may provide information that can be difficult to find when conducting large-scale comparative genomic studies.

**CONCLUSIONS**

The expansion of the R2R3 Myb genes early in the history of plants makes it one of the largest families of regulatory proteins known. Evolutionary studies of this gene family have been hindered by the ancient nature of gene duplications. Here, we took advantage of several recently duplicated R2R3 Myb genes to establish some basic patterns of R2R3 Myb gene evolution. We established a pathway that explains the origin of plant specific R2R3 Myb genes from widely distributed three-MYB repeat genes. We established that the divergent C-terminal regions of R2R3 MYB proteins are subject to purifying selection and found that these regions sometimes show higher numbers of synonymous substitutions than the MYB domain. In fact, sliding window analyses indicated the substantial heterogeneity in the accumulation synonymous and non-synonymous substitutions across the coding regions of all pairs of R2R3 MYB proteins investigated. Together, our studies provide the most in-depth analysis of the sequence divergence of recently duplicated R2R3 Myb genes and suggest novel models for the function and evolution of these genes.

**MATERIALS AND METHODS**

**Screening of Bacterial Artificial Chromosome (BAC) Libraries**

The expressed sequence tag (EST) clone corresponding to ZmMyb-IF35 was obtained by searching a proprietary EST database at Pioneer Hi-Bred International (Johnston, IA). After completely sequencing the EST, we established that the 1,577-bp full-length clone encoded a putative protein of 344 amino acids with an amino-terminal R2R3 MYB domain. A probe corresponding to the carboxyl-terminal region of ZmMyb-IF35 cDNA and excluding the amino-terminal MYB domain region was used to screen a maize (Zea mays) B73 inbred BAC library (Incyte Genomics Inc., Palo Alto, CA). Four BAC clones were recovered (91c08, 235g24, 165j16, and 145c17) and were restriction digested to isolate full-length genomic clones. EcoRI
was used to digest 165j16 BAC clone and a 6-kb fragment was subcloned into pBluescript II SK~ (Stratagene, La Jolla, CA), and the presence of the full-length MYB domain was verified by PCR. Sequencing was carried out using primers flanking the two introns of ZmMyb-IF35. A full-length genomic clone of ZmMyb-IF25 in pBluescript II SK~ (Stratagene) was generated by digesting the 145c17 BAC clone with Xhol and inserting a 9-kb fragment into the vector.

**Mapping of ZmMyb-IF35 and ZmMyb-IF25**

The following primers were designed from the 5’ sequences of ZmMyb-IF25 and ZmMyb-IF35: IF25, forward, TTGGCTCTGGTGATCAATAACATGT; IF25, reverse, AGGTGCAACTGAGAAATGC; IF35, forward, GCCATCCCTCTGCCCTT; and IF35, reverse, CTGCCTGGAGAGGAGATC-GAG. These primer pairs were used to amplify the corresponding regions from genomic DNA of 12 inbred lines: B73, Mo17, GT119, T218, IF25, reverse, AGGTGCAACTGAGAAATGC; IF35, forward, GCCATCCCTCTGCCCTT; and IF35, reverse, CTGCCTGGAGAGGAGATC-GAG. The PCR reaction conditions were: 1× PCR buffer, 0.4 mM dNTPs, 50 ng each of SSR primers (forward and reverse), 0.3 units of AmpliTaq Gold (Perkin-Elmer stock), 50 ng genomic DNA, in a total volume of 15 μL. Thermocycling conditions were: 95°C, 15 min; 1 cycle and then a 1°C decrement for the annealing temperature, each repeated once, until the annealing temperature is 55°C followed by 95°C 1’, 55°C 1’, and 72°C 30’ for 20 cycles. The amplification products were resolved on 3.5% (w/v) agarose gels. Genotypes for mapping ZmMyb-IF35 were determined using a size polymorphism of the amplification products between B73 and Mo17; likewise, genotypes for mapping ZmMyb-IF25 were determined using a size polymorphism between T218 and GT119. Linkage maps for placing ZmMyb-IF25 and ZmMyb-IF35 were constructed using MAPMAKER/EXP (version 3.0, Whitehead Institute, Cambridge, MA). Starting with a framework map for each population, the “assign” and “build” commands were used to identify the chromosome and place the locus within the framework order.

**Sequence Analysis**

The analyses of the sequences obtained were conducted using the Oxford Molecular MacVector 6.0 (Accelrys Inc., San Diego) and MacDNAsis (version 2.0, Hitachi Ltd., San Bruno, CA). Sequences used to examine the evolution of MYB domains were aligned using ClustalW (Thompson et al., 1994) and trimmed to exclude sequences outside of the MYB domain. Phylogenetic analyses of these sequences were conducted by weighted neighbor joining (Bruno et al., 2000) of distance estimates obtained using the Wagner model of sequence evolution (Whelan and Goldman, 2001). Previous analyses have suggested significant variance in rates across sites for MYB genes from grasses examined by this study show the extreme codon bias characteristic of other nuclear encoded genes in grasses (Murray et al., 1989), with a third codon position GC content of >80%. In contrast, the Arabidopsis ZR3 Myb genes investigated show a much more limited codon bias, as is generally the case for eudicot genes (Murray et al., 1999). For this reason, Ks and Ka were calculated using the YN00 model of coding sequence evolution (Yang and Nielsen, 2000) because this model accommodates compositional bias and variable transition to transversion ratios. The YN00 program distributed with the PAML package (http://abacus.gene.ucl.ac.uk/software/paml.html) was used for these calculations. When estimates of Ks and Ka were calculated for short segments of the ZR3 Myb coding regions, the transition-transversion parameter and the base composition at each codon position were fixed at the values estimated from the complete sequences. This was accomplished by modifying the YN00 program to allow input of these values from a data file.

**ACKNOWLEDGMENTS**

We are grateful to J. Marcela Hernandez and Katherine Houchnins for excellent technical assistance and Rob Fields for assistance in the assembly of data sets for computer analyses. Anusha P. Dias acknowledges Rasil Warnakulasooriya for helpful discussions. We thank Pioneer Hi-Bred International and Dr. Wes Bruce for the ZmMyb-IF35 cDNA. We also appreciate the comments and suggestions of the two anonymous reviewers.

Received August 1, 2002; returned for revision September 22, 2002; accepted October 19, 2002.

**LITERATURE CITED**


Copyright © 2003 American Society of Plant Biologists. All rights reserved.
Dias et al.


620 Downloaded from on July 15, 2017 - Published by www.plantphysiol.org Copyright © 2003 American Society of Plant Biologists. All rights reserved.