Computational and molecular analysis of Myb gene family

by

Cizhong Jiang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Thomas A. Peterson, Major Professor
Xun Gu
Volker Brendel
Xiaoqiu Huang
Daniel F. Voytas

Iowa State University
Ames, Iowa
2004

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of
Cizhong Jiang
has met dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

ABSTRACT

Myb proteins are defined by a highly conserved DNA-specific binding domain termed Myb, which is composed of approximately 50 amino acids with constantly spaced tryptophan residues. Multiple copies of Myb domains often exist as tandem repeats within a single protein. There are up to four tandem Myb repeats present in Myb proteins identified to date (termed R0R1R2R3 hereafter). Each Myb repeat can form three α-helices, the third of which plays a recognition role in specifically binding to DNA. In contrast to the conservation of Myb domains, C-terminal coding regions and non-coding regions (promoter, 5'UTR, introns, and 3'UTR) are dramatically divergent.

In our study, we collected more inclusive Myb genes than previous research by screening and sequencing Myb clones from sorghum and maize with a targeted approach. Then, we conducted a series of phylogenetic analyses to explore the evolutionary origin of Myb genes. The results suggest that the whole Myb gene family originated from an ancient Myb-box DNA-binding motif. One and two intragenic duplications produced R2R3 and R1R2R3 Myb genes, respectively, which then co-existed in the primitive eukaryotes and gave rise to the currently extant Myb genes. Based on our results, we proposed that plant R1R2R3 Myb genes were derived from R2R3 Myb genes by gain of R1 repeat through an ancient intragenic duplication; this gain model is more parsimonious than the previous proposal that R2R3 Myb genes were derived from R1R2R3 Myb genes by loss of R1 repeat. The phylogenetic analysis of isolated individual Myb repeats indicates that the R2 repeat has evolved more slowly than the R1 and R3 repeats. However, it is not clear which repeat is the most ancient one.

Myb proteins are one of the largest transcription factor families critical for the regulation and control of gene expression. However, little is known about their functions. Here, we clustered the closely-related Myb genes into subgroups from *Arabidopsis* and rice on a basis of sequence similarity and phylogeny. The gene structure analysis revealed that both the positions and phases of introns are conserved in the same subgroup, but are different between subgroups. Moreover, there is a significant excess of phase 1&2 introns as well as an excess of non-symmetric exons in Myb domains. Conserved motifs were detected in C-terminal coding regions of Myb genes within subgroups. EST blast search shows that the identified C-

terminal motifs exist specifically in Myb genes, and could serve as an additional identifying characteristic of Myb genes. We also found that Mybs with similar functions are clustered together as subgroups. Taken together, the conserved motifs and splicing sites may reflect functional constraints upon Myb domains. The functional classification table obtained from this result, could provide a reference to assign putative function to newly identified Mybs. In contrast, the non-coding regions are too divergent to enable identification of common regulatory motifs. Finally, the distribution pattern of introns in the phylogenetic tree indicates that Myb domains originally had a compact size without introns. Non-coding sequences were inserted and the splicing sites were conserved during evolution. Overall, these results provide significant new insights into the origin, evolution, and functional roles of Myb-encoding genes.

CHAPTER 1. GENERAL INTRODUCTION

Introduction

Like all eukaryotes, plants employ complex and highly controlled mechanisms of gene regulation which are critical for growth, development, and response to environmental changes. To date, many results from EST and genomic sequencing projects indicate that differential gene expression plays a major role in determining plant traits (Doebley and Lukerns 1998), and plants commonly contain multiple copies of genes encoding highly-similar proteins. Regulation of gene expression at the level of transcription influences many important biological processes in a cell or organism. Our present knowledge of plant regulatory sequences is based largely on the analysis of individual genes. However, single gene analysis is time-consuming, expensive, and provides only fragmentary information. Our project focused on a gene family named Myb encoding transcription factors with multiple copies in plant genomes. With the assistance of computational techniques, we attempt to determine the number of Myb genes in *Arabidopsis*, and the major cereals: rice, sorghum, and maize; infer the evolutionary origin of Myb genes; identify the conserved motifs; classify and predict their functions; and correlate the regulatory elements with the differential gene expression in transgenic assays. This study should help to elucidate the evolutionary history patterns and constraints of Myb genes; understand the relationship of Myb gene sequences and their functional diversity; and improve some agronomic traits involved Myb genes. This system-wide approach to gene expression represents the next logical step in genome studies.

Dissertation Organization

This dissertation covers the major projects during my Ph.D. study, and comprises seven chapters. Each chapter is composed of several components such as abstract, introduction, results and discussion, and so forth. Chapter one addresses the background and significance of the projects, and states the organization of this dissertation as well. Chapter two describes screening and sequencing of Myb genes in sorghum and maize. In chapter three, we infer the evolutionary origin of Myb genes by analyzing the inter- and intra-organismal divergence of Myb genes. In chapter four, we identify the complement of Myb genes in *Arabidopsis* and rice, classify and predict their functions. Chapter five describes the detailed transgenic assays

aimed to identify regulatory elements and correlate them with gene expression patterns. The intern project in Pioneer Hi-Bred Inc., a DuPont company, in the summer of 2002 is presented in chapter six. All the member genes of the 37 transcription factor families in maize, rice and *Arabidopsis* are identified and classified by type in this transcription factor project. The general conclusions of the research topics are summarized in chapter seven. Finally, the dissertation ends with the acknowledgements to all the people who have given their kind help and support.

Literature Review

1. The discovery and characteristics of Myb

The Myb gene was first identified in the form of the v-Myb oncogene of the avian myeloblastosis virus (Klempnauer et al. 1982). Subsequently, members of the Myb gene family were found existing widely in both plants and animals (Rosinski and Atchley 1998). A small number of Myb genes have also been found in fungi (Lipsick 1996), mycetozoa (Braun and Grotewold 1999, Kranz et al. 2000), and microsporidia (Jiang et al. 2003). There are no Myb genes recognized in prokaryotes so far.

A conserved Myb domain of approximately 50 amino acids with constantly spaced tryptophan residues characterizes the family of Myb proteins. Multiple copies (up to four) of the Myb domains are frequently present as tandem repeats within a single protein. Most such Myb proteins contain two or three tandem repeats near their amino terminus. To date, no two-repeat Myb genes were detected in animals. Only a single four-repeat Myb gene (termed R0, R1, R2, and R3 hereafter) was reported in *Arabidopsis* (Stracke et al. 2001). Interestingly, each of the Myb repeats is more closely related to other members of the same family than to other repeats within the same protein (Lipsick 1996). Within each repeat form three α-helices (Ogata et al. 1955). The third α-helix is thought to play a recognition role in binding to a short DNA sequence (Robinowicz et al. 1999). The previous result suggested that each repeat folds into a helix-turn-helix (HTH) variant related to that of prokaryotic repressors (Ogata et al. 1955), and similar to that of the homeodomain-containing proteins (Gabrielsen et al. 1991, Ogata et al. 1995). The HTH motif has since been reported to bind directly to the major groove of DNA (Ogata et al. 1992). Furthermore, some homology

between the C-terminal region of R2 repeat and the basic DNA-binding motif of leucine zipper proteins has been reported (Carr and Mott 1991).

The sixth residue C-terminal to the third tryptophan in R2 repeat is highly conserved in both animal and plant Myb proteins (Hegvold and Gabrielsen 1996, Williams and Grotewold 1997). Only a conserved proline (P) residue was found in this site in all animal Myb proteins, and more than 95% in plant, which is sometimes replaced by alanine (A), serine (S), or arginine (R). The substitution of P to A was reported by Robinowicz (1999). Accordingly, the Myb proteins are designated as P- and A-type, respectively (Dias et al. 2003, Jiang et al. 2003).

Interestingly, the first tryptophan (with polar side chain) of R3 repeat in three-repeat Myb proteins is replaced in two-repeat Myb proteins by phenylalanine, isoleucine, and leucine (all with nonpolar side chains). These R1R2R3 Myb proteins in *Arabidopsis* exhibit structural features of the vertebrate c-Myb preoncoprotein, and were first designated as pc-Myb (plant-c-Myb-like genes) to distinguish them from the larger R2R3 Myb genes in plants (Braun and Grotewold 1999).

It is worth pointing out that some proteins have been identified containing one or partial Myb repeat in *Drosophila* (England et al. 1992), plants (da Costa et al. 1993; Baranowskij et al. 1994; Lugert and Werr 1994; Kirik and Bäumlein 1996; Feldbrügge et al. 1997), and yeast (Morrow et al. 1993). Additionally, some R2R3 Myb proteins in fungi and plants lack the typical constantly-spaced tryptophan residues. The completion of more and more genome sequencing projects, and advances in molecular biology over the past twenty years, have enabled a more comprehensive understanding of the evolution and function of the whole Myb gene family in different species.

2. The evolutionary history of Myb

Recent results from EST and genomic sequencing projects indicate that plants commonly contain multiple copies of gene encoding highly-similar proteins. These multiple gene copies may have arisen by whole-genome polyploidizations, or regional duplications affecting sub-genomic segments. Myb proteins are such genes with multiple copies in eukaryotic genomes, and widely exist in different organisms. An important suggestion is, how was the Myb gene

family expanded through gene duplication and divergence? These latter processes have been proposed to be major forces in genomic and organismal evolution (Ohno 1970).

The first attractive model for evolution of Myb gene family was presented by Lipsick (1996). It proposed that all current Myb genes arose from a common ancient one Myb repeat designated R1/2 (mixed R1 and R2). The one or partial repeat Myb proteins were derived from this ancestor through substitutions. With two successive intragenic duplications, R1/2 evolved into R1/2R3, then R1R2R3 which became the recent ancestor of current Myb genes. The plant and animal R1R2R3 Myb genes resulted from the subsequent duplications of the entire gene. In contrast, the ancestral R1R2R3 lost R1 to become the ancestral R2R3, which gave rise to the currently extant plant R2R3 Myb genes with the duplications.

This loss-of-R1 model received support from other groups' work. Rosinski and Atchley (1998) conducted a series of phylogenetic analyses of 42 Myb proteins from both plants and animals using complete protein sequences, the conserved DNA-binding domain, and the flanking regions, respectively. Their result suggested that Myb proteins are a polyphyletic group originated from a "Myb-box" DNA-binding motif, and that the plant Myb ancestors had three repeats, but first repeat was lost to produce two-repeat Myb genes. Subsequently, Braun and Grotewold (1999) discovered pc-Myb genes (plant R1R2R3 Myb), and re-investigated the evolution of the plant Myb gene family. Their analyses indicated that pc-Myb gene was duplicated into multiple copies in plants after the divergence of plants from other eukaryotic groups, and also supported the loss-of-R1 model. Similar results follow analyses of new identified Myb genes, remained consistent with the Lipsick model (Jin and Martin 1999). With the identification of a number of R1R2R3 Myb genes in all major plant lineages, plant R2R3 Myb genes were again thought to have evolved from plant R1R2R3 Myb genes by loss of R1 (Kranz et al. 2000).

All the previous researchers agreed on the following conclusions: 1.The Myb domain is likely to have arisen after the divergence of eubacteria and eukaryotes (Lipsick 1996). 2. The duplications which generated these tandem repeats occurred prior to the divergence of these distantly related species based on the fact that each R2 repeat is more homologous to other R2 repeats than to any R3 repeat; this conclusion holds true for the R3 repeat as well. 3. The origins of the Myb gene preceded the origins of the major plant phyla. 4. The ancestral R1R2R3 Myb formed prior to the divergence of plants and animals. 5. The duplications that

resulted in the amplification of Myb genes in plants occurred prior to the divergence of monocots and eudicots. Then, R2R3 Myb genes further evolved during plant speciation.

Recently, Dias et al. (2003) showed that the origin of A-type Myb genes from P-type Myb genes may precede the divergence of monocots and eudicots. This group of P-to-A Myb genes recently underwent an expansion in the grasses, involving genome duplication, tandem gene duplication and a more ancient duplication. Additionally, the evolutionary pattern of Myb-homologous genes in cotton seems to occur independently in the allopolyploid nucleus, rather than via concerted evolution (Cedroni et al. 2003). However, the general evolutionary history, patterns, and constraints of the Myb gene family still remain unclear. Especially, the relationship (mycetozoa, (animal, plant)) on which the loss model was based, is no longer compelling (Baldauf et al. 2000). Previous researchers (Braun and Grotewold 1999, Kranz et al. 2000) supported the loss model in part based on the observation that R1R2R3 Myb genes were more widely present in different organisms than R2R3 Myb genes. However, with more and more sequences available, we found this conclusion is no longer tenable. Specifically, the ancient origin of the plant R2R3 Myb gene is subject to question. With a more inclusive data set which has become available from recent genome projects, we can discern a more reliable scenario of the evolution of the Myb gene family.

3. The functional diversity of Myb transcription factors in plants

To date, hundreds of Myb genes have been broadly identified in all kinds of eukaryotes: microsporidia, insects, slime moulds, fungi, plants and animals. The Myb genes comprise one of the largest transcription factor families in eukaryotes. In some well-studied examples, the Myb proteins have been shown to make base-specific contacts target genes by tandem R2R3 repeats' binding to the major groove of the DNA double helix (Ogata et al. 1994).

A review by Lipsick (1996) listed the functions of some Myb-related genes. Two Myb genes have been found in the budding yeast *Saccharomyces cerevisiae*: *BAS1* cooperates with *BAS2/PHO2* (homeodomain) to regulate *HIS4*; *REB1* (partial Myb) regulates rRNA initiation and termination, and affects pol II transcription by altering chromatin structure. Only *CDC5* (an ancient paralog of Myb) was identified in fission yeast *Schizosaccharomyces pombe*; *CDC5* is involved in control of cell cycle. Another partial Myb gene *flbD* in the

fungus *Emericella nidulans* regulates condiophore initiation. In *Drosophila*, the Myb gene *stonewall* related to *Adf-1* affects oocyte differentiation.

Sequence comparisons have shown that v-Myb may have been transduced from the host vertebrate gene, then evolved into part of the virus. Myb genes closely related to v-Myb have been detected in many vertebrate genomes (Weston 1998). Although the number of Myb genes in vertebrates is relatively few (Thompson and Ramsay 1995, Rosinski and Atchley 1998), they are categorized into three forms based on their function: A-Myb, B-Myb and C-Myb (Lipsick 1996). For example, it has been reported that C-Myb regulates proliferation and differentiation of hemopoietic cells (Duprey and Boettiher 1985). B-Myb is poorly understood; it may down-regulate C-Myb (Foos et al. 1972). The function of A-Myb is less clear, although it is correlated with the proliferation of spermatogonial cells in testis (Takahashi et al. 1995).

In contrast to animals, a huge number of Myb genes have been recognized in plants. For example, 125 R2R3 Myb genes have been identified in *Arabidopsis* (Stracke et al. 2001). Additional set of plant Myb genes include more than 80 Myb ESTs obtained by RT-PCR from maize (Rabinowicz et al. 1999), approximately 30 in *Petunia hybrida* (Avila et al. 1993), 14 different Myb related cDNAs in tomato, *Lycopersicon esculentum* (Lin et al. 1996), and approximately 200 Myb genes predicted in cotton (Cedroni et al. 2003). When compared with their counterparts in animals, plant Myb homologues are structural and functional more variable (Martin and Paz-Ares 1997). Most of their functions are still unknown, except for some well-studied examples. Research to date indicates that there are four major roles for plant Myb transcription factors (Martin and Paz-Ares 1997, Meissner et al. 1999): 1) Controlling secondary metabolism, particularly in phenylpropanoid pathway (Paz-Ares et al 1987 [maize gene *C1*], Grotewold et al. 1994 [maize gene *p1*], Sablowski et al. 1994, Moyano et al. 1996, Tamagnone et al. 1998, Mol et al. 1998, Quattrocchio et al. 1999 [*Petunia hybrida* gene *AN2*], Borevitz et al. 2000 [*Arabidopsis* genes *PAP1*, *PAP2*], Zhang et al. 2000 [maize gene *p2*], Aharoni et al. 2001 [strawberry gene *FaMyb1*], *Nesi et al. 2001* [*Arabidopsis* gene *TT2*]) and tryptophan biosynthesis (Walker 1993, Bender and Fink 1998 [*Arabidopsis* gene *ATR1*]); 2) Regulating cellular morphogenesis, for example, *Arabidopsis* genes *GL1* (Oppenheimer et al 1991) and *Wer* (Lee and Schiefelbein 1999) are critical for hair cells of the leaf and in the stem; *Antirrhinum majus* gene *MIXTA* controls cell

shape (Noda et al 1994) and *PHAN* regulates cell differentiation (Waites et al. 1998); 3) Participating in signal transduction, such as responses to plant growth (Iturriaga et al. 1996 [*Craterostigma plantagineum* gene *cpm10*], Gubler et al. 1995 [barley gene *Gam1*]); 4) Communicating with the environment, such as responses to drought, abiotic stress (Urao et al. 1993, Magaraggia 1997, Hoeren et al. 1998) and resistance to pathogen attack (Yang and Kleissig 1990). In addition, some plant Myb proteins may play a role as structure factors. A recent paper showed that the Myb-like domain of plant telomeric protein RTBP1 might play an important role in plant telomere function *in vivo* (Yu et al. 2000).

Previous results have shown that there is a strong correlation between functionality of an amino acid site in Myb proteins and its conservation through evolution (Rosinski and Atchley 1998). For example, the constantly spaced tryptophan residues in R2R3 domains are central to the formation of a hydrophobic core of amino acids which in turn is required in the protein's sequence-specific DNA binding (Ogata et al. 1992). Additionally, the linker region that joins R2 and R3 repeats (approximately nine residues) plays a fundamental role in positioning the DNA-recognition helices on the DNA (Hegvold and Gabrielsen 1996) by providing flexibility between R2 and R3 repeats (van Aalten et al. 1998). Further insight was provided by studies of two closely related genes in maize, *C1* and *P*. The protein C1 regulates target genes together with either of two bHLH proteins R or B, whereas P does not interact with R or B. However, when the four predicted solvent-exposed residues in the first α-helix in R3 repeat of *p* gene are substituted with corresponding residues from *C1*, the mutated P protein is able to physically interact with R (Grotewold et al. 2000).

Another interesting feature of plant R2R3 Myb proteins is that the highly conserved R2R3 domains are coupled with dramatically divergent C-terminal regions. Does this divergence difference reflect functional constraints upon the R2R3 domains with the C-terminal regions? Again, this question has been experimentally tested to some extent. When the C-terminal regions of the maize genes *C1* and *p* were exchanged, the activation potentials of both hybridized genes were changed (Grotewold et al. 2000).

To gain insight into Myb genes' functions, the complement of Myb genes in *Arabidopsis* and other plants were clustered into subgroups based on sequence similarity by phylogenetic tree. The conserved motifs were identified in C-terminal regions in each subgroup. Then, their functional characteristics were categorized on the basis of representative sequences with

experimentally proved functions (Stracke et al. 2001). A similar approach was applied in two other gene families, kinesin (Lawrence et al. 2002) and bHLH (Toledo-Ortiz et al. 2003).

Interestingly, the divergent non-coding regions may effect different gene expression patterns. For example, the maize *p1* and *p2* Myb genes have highly homologous coding regions but have divergent promoters, which may be responsible for their distinct expression patterns (Zhang et al. 2000).

Why are there so many plant Myb proteins? Probably, plants used R2R3 Myb transcription factors selectively to control their specialized physiological functions (Martin and Paz-Ares 1997). To assign the biological functions of Myb genes with unknown function is one of the challenging goals in biology. Unfortunately, many techniques of genetic analyses, such as targeted gene disruption, are laborious and inefficient in plants. For example, no knockout alleles for *Arabidopsis* R1R2R3 Myb genes were detected from the several available insertion-mutagenesis populations. It was proposed that the disruption of the 3-repeat Myb genes is lethal because of its essential role in cell cycle (Stracke et al. 2001). In addition, because of the huge number of Myb genes, their redundancy may also reduce the efficiency of the genetic methods by compensating for the loss of a mutated target Myb gene. Therefore, a computational method based on conserved motif sequences as described in this dissertation will facilitate the functional categorization of newly identified Myb genes.

References

Aharoni, A., Ric De Vos, C.H., Wein, M., Sun, Z., Greco, R. Kroon, A., Mol, J.N.M., and O'Connell, A.P. (2001) The strawberry *FaMyb1* transcription factor suppresses anthocyanin and flavonol accumulation in transgenic tabacco. *Plant J.*, **28**: 319-332

Avila, J, Nieto, C, Canas, L, Benito MJ, Paz-Ares J. (1993) *Petunia hybrida* genes related to the maize regulatory *C1* gene and to animal myb-oncogenes. *Plant J.*, **3**:553-562

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., Doolittle, W.F., 2000. A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. Science 290, 972-977.

Baranowskij, N., Frohberg, C., Prat, S., and Willmitzer, L. (1994) A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. *EMBO* J. **13**:5383-9392

Bender, J., and Fink, G.R. (1998) A Myb homologue, ATR1, activates typtophan gene synthesis in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA,* **95**: 5655-5660

Borevitz, J.O., Xia, Y.J., Blount, J., Dixon, R.A., Lamb, C. (2000) Activation tagging identifies a conserved Myb regulator of phenylpropanoid biosynthesis. *Plant Cell,* **12**:2383-2393

Braun, E.L. and Grotewold, E. (1999) Newly discovered plant c-*myb*-like gene rewrite the evolution of the plant *myb* gene family. *Plant Physiology*, **121:**21-24

Carr, M.D., and Mott, R.F. (1991) The transcriptional control proteins c-Myb and v-Myb contain a basic region DNA binding motif. *FEBS Lett*, **282**: 293-294

Cedroni ML, Cronn RC, Adams KL, Wilkins TA, Wendel JF (2003) Evolution and expression of Myb genes in diploid and polyploid cotton. *Plant Molecular Biology* **51:**313-325

da Costa, E, Silva, O, Klein, L, Schmelzer, E, Trezzini, GF and Hablbrock, K. (1993) BPF-1, a pathogen-induced DNA-binding protein involved in the plant defense response. *Plant J.* **4:**125-135

Dias, A.P., Braun, E.L., McMullen, M.D., Grotewold, E., 2003. Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication. Plant Physiology **131**: 610-620.

Doebley J, Lukens L. (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell*, **10**: 1075-1082

Duprey, SP and Boettiher, D. (1985) Developmental regulation of c-myb in normal myeloid progenitor cells [published erratum appears in Proc Natl Acad Sci USA, 1986 Apr; 83(7):2281]. *Proc Natl Acad Sci USA*, **82:**6937-6941

England, BP, Admon, A and Tjian, R. (1992) Cloning of *Drosophila* transcription factor Adf-1 reveals homology to Myb oncoproteins. *Proc. Natl. Acad. Sci.* USA, **89:**683-687

Feldbrügge, M, Sprenger, M, Hahlbrock, K and Weisshaar, B. (1997) PcMYB1, a novel plant protein containing a DNA-binding domain with one MYB repeat, interacts in vivo with a light-regulatory promoter unit. *Plant J.* **11:**1097-1093

Foos, G, Grimm, S, Kempnaauer K.H. (1992) Functional antogonism between members of the myb family: B-myb inhibits v-myb induced gene activation. *EMBO J* **11:**4619-4626

Gabrielsen, OS, Sentenac, A and Fromageot, P. (1991) Specific DNA binding by c-Myb: evidence for a double helix-turn-helix-related motif. *Science*, **253:**1140-1143

Grotewold, E., Drummond, B.J., Bowen, B., and Peterson, T. (1994) The Myb-homologous *P* gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell*, **76:**543-553

Grotewold, E., Sainz, M.B., Tagliani, L., Hernandez, M., Bowen, B., and Chandler, V.L. (2000) Identification of the residues in the Myb domain of maize C1 that specify the interaction with the bHLH cofactor R. *Proc Natl Acad Sci USA*, **97**: 13579-13584

Gubler, F., Kalla, R., Roberts, J.K., and Jacobsen, J.V. (1995) Gibberellin-regulated expression of a Myb gene in barley aleurone cells: Evidence for Myb transactivation of a high-pl alpha-amylase gene promoter. *Plant Cell*, **7:**1879-1891

Hegvold, A.B. and Gabrielsen, O.S. (1996) The importance of the linker connecting the repeats of the c-Myb oncoprotein may be due to a positioning function. *Nucleic Acids Res.* **24:**3990-3995

Hoeren, F.U., Dolferus, R, Wu, U, Peacock, W.J. and Dennis, E.S. (1998) Evidence for a role for AtMyb2 in the induction of the *Arabidopsis* alcohol dehydrogenase gene (*ADH1*) by low oxygen. *Geniteics* **149:** 479-490

Iturriaga, G., Leyns, L., Villeas, A., Gharaibeh, R., Salamini, F., and Bartels, D. (1996) A family of novel myb-related genes from the resurrection plant *Craterostigma plantagineum* are specifically expressed in callus and roots in response to ABA or desiccation. *Plant Mol. Biol.*, **32:**707-716

Jiang C, Gu J, Chopra S, Gu X, and Peterson T (2003) Ordered Origin of the Typical Two- and Three-Repeat Myb Genes. *Gene* **(submitted)**

Jin, H., and Martin, C. (1999) Multifunctionality and diversity within the plant Myb-gene family. *Plant Mol. Biol.* **41**: 577-585

Kirik, V and Bäumlein, H. (1996) A novel leaf-specific myb-related protein with a single binding repeat. *Gene*, **183:**109-113

Klempnauer, K-H, Gonda, T.J. and Bishop, J.M. (1982) Nucleotide sequence of the retroviral leukemia gene *v-myb* and its cellular progenitor *c-myb*: the architecture of a transduced oncogene. *Cell* **31:** 453-463

Kranz, H, Scholz, K and Weisshaar, B. (2000) c-Myb oncogene-like genes encoding three Myb repeats occur in all major plant lineages. *Plant Journal*, **21:** 231-235

Lee, M.M., Schiefelbein, J. (1999) WEREWOLF, a Myb-related protein in *Arabidopsis*, is a position-dependent regulator of epidermal cell patterning. *Cell*, **99**: 473-483

Lin, Q., Hamilton W.D., Merryweather, A. (1996) Cloning and initial characterization of 14 myb-related cDNAs from tomato (*Lycopersicon esculentum* ev Alisa Craig). *Plant Mol Biol* **30:**1009-1020

Lipsick, J.S. (1996) One billion years of Myb. *Oncogene*, **13**:223-235

Lawrence CJ, Malmberg RL, Muszynski MG, and Dawe, RK (2002) Maximum likelihood methods reveal conservation of function among closely related kinesin families. *J Mol. Evol.* **54**: 42-53

Lugert, T. and Werr, W. (1994) A novel DNA-binding domain in the Shrunken initiator-binding protein (IBP1). *Plant Mol Biol* **25:**493-506

Magaraggia, F., Solinas, G., Valle, G., Giovinazzo, G. and Coraggio, I. (1997) Maturation and translation mechanisms involved in the expression of a myb gene of rice. *Plant Mol. Biol.* **35:**1003-1008

Martin, C. and Paz-Ares, J. (1997) Myb transcription factors in plants. *Trends Genet*, **13:** 67-73

Meissner, R.C., Jin, H., Cominelli, E. et al. (1999) Function search in a large transcription factor gene family in *Arabidopsis*: assessing the potential of reverse genetics to identify insetional mutations in R2R3 Myb genes. *Plant Cell*, **11**: 1827-1840

Mol, J., Grotewold, E. and Koes, R. (1998) How genes paint flowers and seeds. *Trends Plant Scie.* **3:**212-217

Morrow, B.E., Ju, Q. and Warner, J.R. (1993) A bipartite DNA-binding domain in yeast Reb1p. *Mol. Cell. Biol.*, **13**: 1173-1182

Moyano, E., Martinez Garcia, J.F., and Martin, C. (1996) Apparent redundancy in myb gene function provides gearing for the control of flavonoid biosynthesis in Antirrhinum flowers. *Plant Cell*, **8:**1519-1532

Nesi, N., Jond, C., Debeaujon, I., Caboche, M., Lepiniec, L. (2001) The *Arabidopsis TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. Plant Cell 13, 2099-2114.

Noda, K., Glover, B.J., Linstead, P., and Martin, C. (1994) Flower color intensity depends on specialized cell shape controlled by a Myb-related transcription factor. *Nature*, **369:**661-664

Ogata, K, Morikamura, H, Sekikawa, A, Inoue, T, Kanai, H, Sarai, A, Ishii, S, and Nishimura, Y. (1992) Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. *Proc Natl Acad Sci USA*, **89:**6428-6432

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. and Nishimura, Y. (1994) Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell*, **79:** 639-648

Ogata, K., Morikamura, H., Nakamura, H., Hojo, H., Yoshimura, S., Zhang, R., Aimoto, S., Ametani, Y., Hirata, Z. and Sarai, A. (1995) *Nat. Struct. Biol.*, **2**: 309-320

Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag, Berlin and New York, pp. 59-60

Oppenheimer, D.G., Herman, P.L., Sivakumaran, S., Esch, J., and Marks, M.D. (1991) A myb gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules. *Cell*, **67:**483-493

Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P.A., and Saedler, H. (1987) The regulatory c1 locus of *Zea mays* encodes a protein with homology to Myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J*, **6**:3553-3558

Quattrocchio, F., Wing, J., van der Woude, K., Souer, E., de Vetten, N., Mol, J., and Koes, R. (1999) Molecular analysis of the anthocyanin2 gene of petunia and its role in evolution of flower color. *Plant Cell*, **11**: 1433-1444

Rabinowicz, PD, Braun EL, Wolfe, AD, Bowen, B, and Grotewold, E. (1999) Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants. *Genetics*, **153:** 427-444

Rosinski, JA, and Atchley, WR. (1998) Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin. *J Mol Evol*, **46**:74-83

Sablowski, R.W.M., Moyano, E., Culianez Macia, F.A., Schuch, W., Martin, C., and Bevan, M. (1994) A flower-specific Myb protein activates transcription of phenylpropanoid biosynthetic genes. *EMBO J.*, **13:**128-137

Stracke, R., Werber, M., and Weisshaar, B. (2001) The R2R3-Myb gene family in *Arabidopsis thaliana*. *Curr Opin Plant Biol*, **4**: 447-456

Takahashi, T, Nakagoshi, H, Sarai, A, Nomura, N, Yamamoto, T, Ishii, S. (1995) Human A-Myb gene encodes a transcriptional activator containing the negative regulatory domains. *FEBS Left* **358:**89-96

Tamagnone, L, Merida, A, Parr, A, Mackay, S, Culianez-Macia, FA, Roberts, K and Martin, C. (1998) The AmMYB308 and AmMYB330 transcription factors from Antirrhinum regulate phenylpropanoid and lignin biosynthesis in transgenic tobacco. *Plant Cell*, **10:**135-154

Thompson, M.A., and Ransay, R.G. (1995) Myb: An old oncoprotein with new roles. *BioEssays* **17:**341-350

Toledo-Ortiz G, Huq E, and Quail PH (2003) The *Arabidopsis* Basic/Helix-Loop-Helix transcription factor family. *Plant Cell* **15**: 1749-1770

Urao, T., Yamaguchi-Shinozaki, K., Urao, S., and Shinozaki, K. (1993) An *Arabidopsis* myb homolog is induced by dehydration stress and its gene product binds to the conserved Myb recognition sequence. *Plant Cell*, **5:**1529-1539

Van Aalten, D.M.F., Grotewold, E., and Joshua-Tor, L. (1998) Essential dynamics from NMR structures: dynamic properties of the Myb DNA-binding domain and a hinge-bending enhancing variant. *Methods*, **14**: 318-328

Walker, J.C. (1993) Receptor-like protein kinase genes of *Arabidopsis thaliana*. *Plant J.* **3**: 451-456

Weston, K. (1998) Myb proteins in life, death and differentiation. *Curr Opin Genet Dev*, **8**: 76-81

Williams, CE and Grotewold, E. (1997) Differences between plant and animal Myb domains are fundamental for DNA-binding and chimeric Myb domains have novel DNA-binding specificities. *J. Biol. Chem.* **272:**563-571

Yang, Y. and Klessig, D.F. (1996) Isolation and characterization of a tobacco mosaic virus-inducible myb oncogene homolog from tobacco. *Proc. Natl. Acad. Sci. USA*, **93:**14972-14977

Yu, EY, Kim SE, Kim, JH, Ko, JH, Cho, MH and Chung, I.K. (2000) Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1. *J Biol Chem*, **275:** 24208-24214

Zhang, P., Chopra, S. and Peterson, T. (2000) A segmental gene duplication generated differentially expressed *myb*-homologous genes in maize. *Plant Cell*, **12:**2311-2322

CHAPTER 2. SCREENING AND SEQUENCING OF MYB GENES IN SORGHUM AND MAIZE

Abstract

With the completion of the *Arabidopsis* genome sequencing, 125 R2R3 Myb genes have been identified in this simple dicotyledenous plant. In contrast, relatively a few Myb genes are known in sorghum and maize; these latter monocotyledenous plants have much larger genomes than that of *Arabidopsis*, but they have yet to be sequenced. In order to collect more inclusive Myb gene data sets, we took advantage of available resources to isolate sorghum and maize Myb genes in a targeted strategy. A cDNA fragment of maize *P1-wr* gene spanning the R2R3 Myb domain was radiolabelled and used as probe to screen Myb-homologous genes from sorghum and maize BAC libraries. These Myb-hybridizing BACs were sequenced by a primer-walking strategy. Finally, the sequences were assembled, annotated and deposited into GenBank for public use. Our results indicate that R2R3 Myb gene family has been greatly expanded in sorghum and maize during evolution. The number of Myb genes in these two cereals is estimated at >64 and >200, respectively.

Introduction

To this date, a number of crop plant genome projects have been initiated. However, due to their size, complexity and high proportion of repetitive elements, the genomes of many crop plants will be difficult to finish than that of *Arabidopsis*. Therefore, unlike *Arabidopsis*, the precise number of Myb genes is still unknown in sorghum and maize. Relatively few Myb genes of these two species have been identified so far, including sorghum *y1* (Chopra et al. 2002), maize *C1* (Paz-Ares et al 1987), *p1* (Grotewold et al. 1994), *p2* (Zhang et al. 2000), *rs2* (Timmermans et al. 1999, Tsiantis et al. 1999), and so forth. In the present era of high-throughput data acquisition and information processing, more shared resources and capabilities are available. It should be possible to infer a rough estimate of the number of Myb genes in sorghum and maize, although the final answer awaits completion of the ongoing genome projects.

Libraries that contain large insert DNA carried in a suitable vector are critical for the analysis of complex genomes. The dominant vectors include cosmids (Collins and Hohn

1978), yeast artificial chromosomes (YAC) (Burke et al. 1987), bacterial artificial chromosomes (BAC) (Shizuya et al. 1992), and P1-derived artificial chromosomes (PAC) (Ioannou et al. 1994). The BAC system has many potential advantages over YAC system, and has been favored for construction of DNA libraries of plant genomes.

The first plant BAC library published was the sorghum SB_BBa-Library (Woo et al. 1994) and was donated to CUGI by R. Wing. This library was constructed in the *Hin*dIII site of the vector pBeloBAC11. It contains 14,208 clones with an average insert size of 157 kb covering 3 genome equivalents. A BAC library of the inbred maize line B73 was constructed by Tomkins et al. (2000). Similarly, this library was constructed in *Hin*dIII site of pBACindigo536 (vector pCUGI). It contains 247,680 clones covering 13.5 genome equivalents. The average insert size is 137 kb with a range of 42 to 379 kb.

Most plant Myb genes share highly conserved R2R3 domain into which two small introns inserted. This conserved Myb domain allows ready access to both promoter sequences and C-terminal regions using common degenerate sequencing primers through a primer-walking strategy. Taken together, the BAC libraries can serve as a starting point for the isolation and characterization of Myb genes.

In this study, we screened public sorghum and maize BAC libraries for clones hybridizing with a Myb probe. First, BAC library filters were probed with fragments of cDNAs covering Myb R2R3 domains. Then, the hybridizing BAC clones were identified and the corresponding BAC templates were prepared for sequencing. Finally, the sequencing results were assembled, annotated, and deposited into GenBank.

Materials and Methods

Screening Myb-positive BAC clones from sorghum and maize libraries

BAC libraries, sorghum SB_BBa-Library and maize ZMMBBb-Library were obtained from Clemson University Genomics Institute (CUGI, https://www.genome.clemson.edu/orders/). The BAC membrane filters were hybridized with a Myb-homologous probe, maize gene *P1-wr* cDNA containing R2R3 domain. The probes were radiolabeled with an oligo-labeling kit (Amersham Pharmacia Biotech, Inc.), and membrane hybridizations were performed based on the protocol from CUGI

(http://www.genome.clemson.edu/protocols/hyb_filter.html). Those clones with intense hybridization signals were designated as Myb-positive BACs (Fig. 1). Their clone IDs were identified according to the protocol provided by CUGI (http://www.genome.clemson.edu/groups/bac/protocols/addressnew.html), and cultures containing these clones were ordered for the following sequencing.

Sequencing Myb-positive BAC clones from sorghum and maize libraries

Sequencing was performed on BAC DNA prepared using a QiaGen Plasmid Midi Kit. A degenerate primer close to the beginning of exon II in Myb genes was designed according to the 78 maize Myb ESTs from Dr. Grotewold (Rabinowicz et al. 1999) and used to directly sequence from BAC templates. The degenerate primer sequence is 5'GAKGYCSGGSCGVAGGTAGTT3'. This sequence reads toward the 5' direction of Myb genes. After the first sequencing run, we obtained specific sequences for each clone. Then, two specific primers (upstream and downstream) were designed to be complementary to the previous sequences obtained from each BAC clone. This procedure was repeated until approximately 1.2 kb upstream to the start codon and complete Myb domains were obtained, respectively. The sequencing was carried out with the Applied Biosystems fluorescent sequencing system at the Iowa State University Nucleic Acid Facility.

Assembling and Annotating Myb genes

Individual sequences from the same BAC were assembled into contigs based on the overlap of sequences from the two successive primer-walking sequencing. The features of each assembled sequence were annotated, such as the number of exons and introns, their locations, and other comments. Finally, they were submitted to GenBank using the program sequin (ftp://ftp.ncbi.nih.gov/sequin/).

Results and Discussion

We obtained 85 Myb-positive clones from one sorghum BAC membrane filter containing 12,288 clones (equivalent to 2.6 genomes) (Fig. 1). The sequencing yielded 64 unique sorghum Myb genes (their GenBank accession numbers: AF474125-AF474133, AF470058-AF470071, and AY363121-AY363161). In contrast, 143 Myb-hybridizing clones were

identified in two maize BAC filters containing 36,864 clones (equivalent to 2 genomes), 31 of which were confirmed to represent unique Myb genes by sequencing (their GenBank accession numbers: AF474115-AF474124, AF470072-AF470092). The redundancy is about 30%. Based on this result, there may be approximately 96 unique Myb genes existing in the two hybridized filters. The whole maize BAC ZMMBBb-Library has 247,680 clones in total (equivalent to 13.5 genomes). Taken together, we estimated >200 Myb genes in the maize genome. This number is consistent with a previous estimate based on RT-PCR detection result of expressed Myb genes, our results confirmed that the Myb gene family has been highly amplified in plants (Rabinowicz et al. 1999).

The sequencing strategy employed here produced some sequences for a significant number of clones. However, in relatively few cases were the complete R2R3 Myb domains obtained (9 out of 64 sorghum clones, and 10 out of 31 maize clones). Most sequences extended into the 5'UTR region, but stopped in intron 2 before reaching exon 3, which comprises part of the R3 repeat. The sequencing difficulties could be attributed to several causes. In some cases, degenerate-sequencing primers did not work on all clones. In other cases, the high GC content in monocot introns may have caused difficulties in sequencing. Some secondary structures might also hinder sequencing by failing to completely denature and thus preventing annealing of the primer to the template.

Conclusion

The Myb gene family has been expanded in sorghum and maize: at least 64 Myb genes are present in sorghum genome, and more than 200 exist in maize genome.

Acknowledgements

We thank Dr. Chopra for initiation of the project, and Terry Olson for technical assistance in the BAC membrane filter hybridization.

References

Burke, D.T., Carle, G.F., and Olson, M.V. (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science*, **236**: 806-811

Chopra, S., Gevens, A., Svabek, C., Peterson, T., and Nicholson, R. (2002). Excision of the Candystripe1 transposon from a hyper-mutable *Y1-cs* allele shows that the sorghum *Y1* gene controls the biosynthesis of both 3-deoxyanthocyanidin phytoalexins and phlobaphene pigments. *Physiological and Molecular Plant Pathology,* **60**: 321-330

Collins, J, Hohn, B. (1978) Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Biotechnology*. 1992; **24**: 193-197

Grotewold, E., Drummond, B.J., Bowen, B., and Peterson, T. (1994) The Myb-homologous *P* gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell*, **76:**543-553

Ioannou, Pl.A., Amemiya, C.T., Garnes, J., Kroise, P.M., Shizuya, H., Chen, C., Batzer, M.A., de Jong, P.J. (1994) A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat. Genet.*, **6**: 84-90

Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P.A., and Saedler, H. (1987) The regulatory c1 locus of *Zea mays* encodes a protein with homology to Myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J*, **6**:3553-3558

Rabinowicz, PD, Braun EL, Wolfe, AD, Bowen, B, and Grotewold, E. (1999) Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants. *Genetics*, **153:** 427-444

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA*, **89**: 8794-8797

Timmermans MCP, Hudson A, Becraft PW, Nelson T (1999) Rough sheath2: a Myb protein that represses knox homebox genes in maize lateral organ primordia. *Science* **284**: 151-153

Tsiantis M, Schneeberger R, Golz JF, Freeling M, Langdale JA (1999) The maize rough sheath2 gene and leaf development programs in monocot and dicot plants. *Science* **284**: 154-156

Tomkins, J.P., Frisch, D., Jenkins, M., Barnett, L., Luo, M., Wing, R.A. (2000) A BAC library for the maize inbred line B73. *Maize Genetics Cooperation Newsletter*, **74**: 18-19

Woo, S.-S., Jiang, J., Gill, B.S., Paterson, A.H. and Wing, R.A. (1994) Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. *Nucleic Acids Res*. **22**: 4922-2931

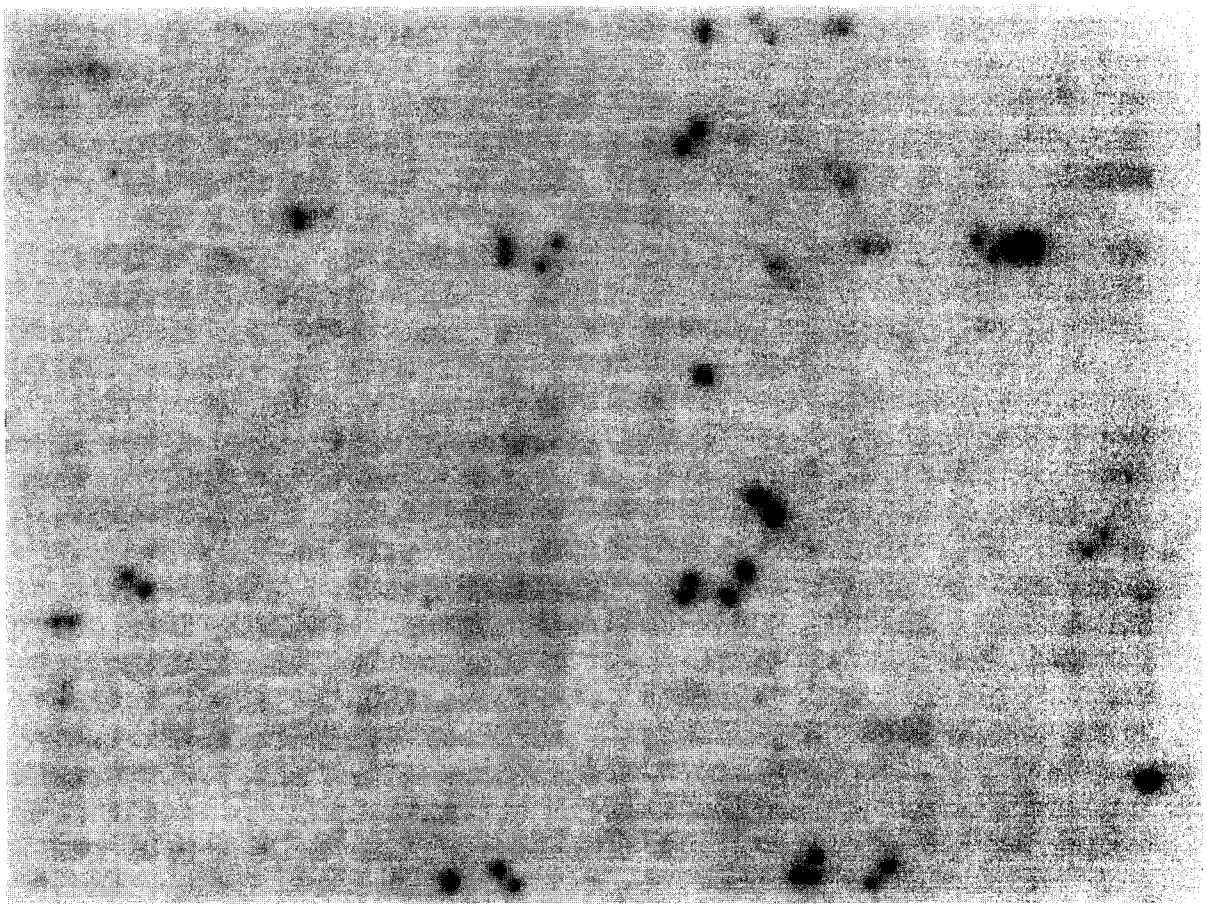


Figure 1. Autoradogram of a maize BAC ZMMBBb-Library membrane following hybridization with a Myb domain probe. The portion shown contains two of the six fields from one library membrane. Each field contains 384 squares. Within each square there are 16 positions where 8 clones are spotted in duplicate. The spot pattern indicates the plate number. Once the plate number is determined, the square locations are identified either by using a grid or counting the rows and columns. Then the specific clone ID can be obtained from the corresponding filter column in the library plate decoder table.

CHAPTER 3. ORDERED ORIGIN OF THE TYPICAL TWO- AND THREE-REPEAT MYB GENES

A paper accepted by *Gene*

Cizhong Jiang, Jianying Gu, Surinder Chopra, Xun Gu, and Thomas Peterson

Abstract

Myb domain proteins contain a conserved DNA-binding domain composed of one to four conserved repeat motifs. In animals, Myb proteins are encoded by a small gene family and commonly contain 3 repeat motifs (R1R2R3); whereas, plant Myb proteins are encoded by a very large and diverse gene family in which a motif containing 2 repeats (R2R3) is the most common. In contrast to the conservation in the Myb domain, other regions of Myb proteins are highly variable. To explore the evolutionary origin of Myb genes, we cloned and sequenced Myb domains from maize and sorghum, and conducted a comprehensive phylogenetic analysis of Myb genes. The results indicate that the origins of individual Myb repeats are strikingly distinct, and that the R2 repeat has evolved more slowly than the R1 and R3 repeats. However, it is not clear which repeat is the most ancient one. The evidence also suggests that R2R3 and R1R2R3 Myb genes co-existed in eukaryotes before the divergence of plants and animals. Based on our results, we propose that R1R2R3 Myb genes were derived from R2R3 Myb genes by gain of the R1 repeat through an ancient intragenic duplication; this gain model is more parsimonious than the previous proposal that R2R3 Myb genes were derived from R1R2R3 Mybs by loss of the R1 repeat. A separate group of diverse non-typical Myb proteins exhibits a polyphyletic origin and a complex evolutionary pattern. Finally, a small group of ancient Myb paralogs prior to the amplification of current Myb genes are identified. Together, these results support a new model for the ordered evolution of Myb gene family.

Key word: evolution; phylogeny; gene duplication; gene family; domain

1. Introduction

The Myb gene super-family comprises a group of related genes found in plant, animal, and fungal genomes. The archetype is the *v-myb* oncogene from Avian Myeloblastosis Virus (Klempnauer et al.1982). Subsequently, members of the Myb gene family were identified in diverse plants and animals (Rosinski and Atchley 1998). Myb genes encode proteins with DNA-binding domains composed of one, two or three semi-conserved motifs of approximately 50 amino acids each. Each motif is capable of forming 3 α–helices; the 3rd α–helix is thought to play a recognition role in binding to a short DNA sequence (Rabinowicz et al. 1999). Many Myb proteins contain two or three tandem repeats of the motif (referred to as R1, R2, and R3 repeat, respectively) within a single protein (Lipsick 1996). A 4-repeat Myb gene was reported in *Arabidopsis* (Stracke et al. 2001).

Vertebrate genomes are reported to contain relatively few Myb genes (Rosinski and Atchley 1998); these are classified on a functional basis into three general forms, A-Myb, B-Myb and C-Myb (Lipsick 1996). In contrast to animals, flowering plants contain large numbers of Myb genes. For example, 125 R2R3 Myb genes have been identified in Arabidopsis (Stracke et al. 2001). Surveys based on expressed sequences have detected more than 80 Myb genes in maize (Rabinowicz et al. 1999), approximately 30 in *Petunia hybrida* (Avila et al. 1993), and approximately 200 Myb genes in cotton (Cedroni et al. 2003). Moreover, plant Myb genes are structurally and functionally very diverse. The functions of most plant Myb genes are unknown, although analysis of some well-studied examples indicates three well-defined roles for plant Myb transcription factors to date: 1) Controlling secondary metabolism, particularly in flavonoid biosynthesis (Grotewold et al. 1994); 2) Regulating cellular morphogenesis (Oppenheimer et al. 1991); 3) Mediating signal transduction pathways, such as in response to abiotic stress (Urao et al. 1993) and pathogen attack (Yang and Klessig 1990). In addition, some plant Myb proteins may serve structural roles as in the case of RTBP1, a Myb-domain protein implicated in plant telomere function (Yu et al. 2000).

It should be noted that a group of Myb proteins containing only one or a partial Myb repeat was found in *Drosophila* (England et al. 1992) and in plants (Baranowskij et al. 1994). Additionally, some Myb proteins with 2 or 3 repeats in fungi and plants lack the typical

constantly-spaced Trp residues. These proteins with atypical Trp spacing or partial Myb domains are non-typical Myb proteins.

In this article we focus on the typical R2R3 and R1R2R3 Myb proteins, which have sufficient sequence divergence to investigate molecular evolution of the Myb gene family. Recently, Dias et al. (2003) provided evidence for expansion of the maize R2R3 Myb gene family through genome duplication, tandem gene duplication, and a more ancient duplication. Additionally, Myb homologous-genes in cotton appear to evolve independently in the allopolyploid nucleus, rather than via concerted evolution (Cedroni et al. 2003). However, the general evolutionary scenario of the Myb gene family still remains unclear; specifically, the ancient origin of the plant R2R3 Myb gene family is subject to controversy. Here, we analyzed new sorghum and maize Myb gene sequences obtained in our laboratory and available Myb sequence data to explore Myb gene evolutionary history, and to study the relationship of Myb gene sequences and their functional diversity.

2. Materials and Methods

2.1 Isolation and Sequencing of Myb Genes from Sorghum and Maize

Maize and sorghum Myb genes were isolated from BAC (Bacterial Artificial Chromosome) clones obtained from CUGI (Clemson University Genomics Institute, Clemson, South Carolina; www.genome.clemson.edu). Membrane filters containing arrayed clones of sorghum (SB_BBa) and maize (ZMMBBb) genomic libraries were hybridized with a radiolabeled fragment of a maize *P1-wr* gene cDNA containing a typical R2R3 Myb domain. Radiolabeled probes were prepared using an oligo-labeling kit (Amersham Pharmacia Biotech, Inc.), and filter hybridizations were done according to a protocol provided by CUGI. BAC DNA was prepared using a QiaGen Plasmid Midi Kit.

Myb genes were directly sequenced from BAC clone templates using a degenerate primer (5'-GAKGYCSGGSCGVAGGTAGTT-3') complementary to sequences within exon 2 of 78 maize Myb EST sequences (Rabinowicz et al. 1999) provided by Dr. Erich Grotewold (Ohio State University). The degenerate sequencing primer was used to prime sequencing reactions proceeding in the 5' direction of Myb genes; sequences obtained were used to design new primers specific for each clone for further sequencing in the upstream and downstream

directions. Further rounds of sequencing followed by specific primer design continued until complete Myb domain sequences were obtained (Figure 1). Sequencing was carried out with the Applied Biosystems fluorescent sequencing system at the Iowa State University DNA Sequencing and Synthesis Facility. BAC clone sequences that did not contain two Myb repeats (e.g. R2R3) were not considered here.

2.2 Myb Proteins Collection

Myb gene nucleotide sequences from various species were obtained from GenBank by homologous tblastn search using known Myb domains as query sequences. A perl script program (available from C. Jiang) was used to extract all CDSs (peptide) from the GenBank files. The unique GI numbers were used as sequence identifiers for each CDS. For *O. sativa*, we performed tblastn search on the Monsanto draft sequence of *O. sativa* cv. Nipponbare (www.rice-research.org) using as query the sequences of 29 maize Myb genes determined in this study, plus the maize *p1* and *p2* gene sequences reported previously. Because our main aim is to address the origin of plant R2R3 Myb genes, for clarity only Myb genes with complete R2R3 domains were included in the analysis; virtually the same results are obtained when partial R2R3 domains are included (not shown). Due to space limitations, the phylogenetic trees shown here comprise selected representative Myb genes from *Arabidopsis thaliana* and *Oryza sativa*, plus all available Myb gene sequences from other organisms to represent the major clades. Nevertheless, during the entire analysis we retained all the informative of Myb genes to avoid any potential bias. Our final dataset includes 139 Myb proteins from 38 species, including lower plants (moss and fern), angiosperms (monocots and dicots), invertebrate (*Drosophila*) and vertebrate, fungi (*Neurospora crassa*), *Mycetozoa* (*Dictyostelium discoideum*, i.e., slime mold), *Microsporidia* (*Encephalitozoon cuniculi*), and *Alveolata* (*Plasmodium falciparum*). To our knowledge, our analysis is based on a much more inclusive Myb sequence data set than those of previous studies.

2.3 Sequence Alignment and Phylogenetic Tree Inference

Myb proteins were aligned using Clustal X (version 1.81) with the default settings. No manual adjustment was found to be necessary. Phylogenetic analyses were conducted using MEGA version 2.0 (Kumar et al. 2001, http://www.megasoftware.net/). Primarily we use the

neighbor-joining method with *p*-distances, which has been suggested for the analysis of large numbers of genes with relatively short sequences (Nei 1996, Rabinowicz et al. 1999). Bootstrapping (1000 replicates) was performed to evaluate the statistical reliability of the inferred topology. The ancient paralog of Myb gene of *Plasmodium falciparum* was used as an outgroup for the phylogenetic trees.

3. Results

3.1 Identification of Myb Domains, Conserved Coding Sequences and Variant Non-coding Sequences in Myb Genes

Currently there are relatively few complete Myb gene sequences available from cereal grain species. To increase the representation of Myb genes from these important crop plants, we isolated and sequenced maize and sorghum genes encoding Myb-homologous proteins. First, we isolated BAC clones that hybridized with a typical R2R3 Myb probe; second, we used a degenerate oligonucleotide primer complementary to a highly conserved region of Myb genes to sequence directly from each BAC clone. The sequences obtained from the degenerate primer were then used to design specific oligonucleotide primers for additional rounds of sequencing reactions (Figure 1). A total of 95 sorghum and maize BAC clones were analyzed in this way; partial sequences were obtained from 76 clones, and the complete R2R3 Myb domain sequences were determined for 9 sorghum and 10 maize Myb genes. Additional Myb gene sequences from diverse plant and animal species were obtained through database searches. Multiple sequence alignment and comparisons of the R2R3 domains were performed on a total of 139 Myb proteins from 38 species. A multiple alignment of 77 sequences representing each clade is presented in Figure 2. (The multiple alignment of the complete sequence set is available at http://pc21955.zool.iastate.edu/origin/MixR2R3Algn.pdf.).

Our results confirm the highly conserved nature of the Myb R2R3 repeat sequences. For example, the Myb domains of the human and mouse A, B and C types (sequences 11-12, 14-15, and 8-9) are identical across their respective Myb types. Likewise, several of the plant Myb proteins also have identical R2R3 domains, including the *p1* and *p2* genes of maize and teosinte (sequences 48-51); sorghum Sb19073330 and Sb19073336 (sequences 37 and 38); and three pairs of *Arabidopsis* Myb genes (two pairs are shown in Figure 2: sequences 17

and 18; and sequences 19 and 20). In general, the Myb domains are not interrupted by deletions or insertions; an interesting exception to this is the maize Myb gene Zm19072746 (sequence 47) which has a Leucine (L) insertion in the 1st α–helix in the R3 repeat. The B73 inbred line used for BAC library has no *C1* function, and the lesion in *C1* in B73 has not been identified (Rabinowicz et al. 1999). The sequences of Zm19072746, *C1*, ZmMYB-IP50 (Rabinowicz et al. 1999) are identical except for this insertion. Therefore, this could be a non-functional allele of *C1*. The effect of this insertion on the function of this protein is not known.

Within each Myb repeat, the third α-helix is more conserved than the other two; this implies that the sequence of this region is critical for Myb proteins function. Consistent with this, the third α-helix has been assigned a recognition role in DNA binding (Ogata et al. 1994), and residue changes in this α-helix are known to impair DNA-binding activity. Specifically, five residue substitutions were introduced in this α-helix in the maize *p1* gene; the mutation of Leu^{55} to Glu was shown to interfere with the binding of the p1 protein to DNA (Williams and Grotewold 1997).

The Myb genes present an interesting disparity in which the Myb domain is highly conserved, whereas the coding sequences downstream of the R2R3 domains are often very divergent. Because we isolated genomic clones, we could also assess sequence conservation in non-coding regions. The positions of introns 1 and 2 are most often conserved, whereas the intronic sequences are highly diverged. Additionally, we found that some genes lacked intron 1, intron 2, or both. The 5' flanking regions are also highly divergent. Such divergence in 3' or 5' flanking regions could occur through gradual sequence changes, or alternatively as a direct outcome of a gene duplication event as described for the maize *p1* and *p2* Myb genes (Zhang et al., 2000). In this example, sequences that were 3' of the single ancestral *p* gene were duplicated and inserted 5' of one of the new gene copies, resulting in the formation of two genes with highly conserved coding regions, but completely different 5' regulatory sequences. The distinct promoters of these genes are thought to be responsible for their distinct expression patterns (Zhang et al. 2000).

3.2 Phylogenetic Analysis of Myb R2R3 Domains

To assess the phylogenetic relationships of Myb domain proteins, we used the R2R3 sequences to infer a phylogenetic tree using neighbor-joining (NJ) methods. Other phylogenetic methods (e.g., the parsimony and BIONJ methods in PAUP*) yielded similar results (data not shown). Apparently, the phylogeny of Myb gene family can be described best by four major clades: R2R3 clade, R1R2R3 clade, rs2-like clade, and U (Unusual) clade. The bootstrap value supporting each clade as monophyletic is very high, from 91% to 100% (Figure 3, arrows).

Within the R2R3 clade, Mybs from monocot and dicot species are not found in distinct groups, but instead are interspersed. This suggests that significant expansion of plant R2R3 Myb genes occurred before the divergence of monocots and dicots. This conclusion is in agreement with previous studies (Lipsick 1996; Rabinowicz et al. 1999). In contrast, we note that a subgroup of R2R3 Myb genes has a replacement at a conserved proline residue (Figure 2, bottom arrowhead). In this subgroup, the proline residue is replaced by alanine, serine, or arginine, resulting in a series of Mybs termed the A-, S-, and R-type, respectively. These distinctive types comprise a subclade within the R2R3 clade; the bootstrap value supporting this topology is high (96%; Figure 3, upper arrowhead). These results indicate that the A-, R- and S-type Mybs diverged from the progenitor P-type Mybs after the amplification of R2R3 Myb genes in plants. Otherwise, they would be likely interspersed within the R2R3 clade. However, their origin may precede the divergence of monocots and eudicots because of the presence of A-type Mybs in both monocots and eudicots. This result is consistent with a recent study (Dias et al. 2003).

Within the R1R2R3 clade, there are two distinct subclades that contain Myb genes from plants and animals (Figure 3, diverged at the arrow). This supports a common ancestor for R1R2R3 Mybs of both plants and animals. The topology for the animal subclade is highly ordered and indicates a paralogous pattern of gene divergence to generate the A-, B- and C-Myb types (paralogous genes result from gene duplication, not speciation.) These three Myb forms have been found in human, mouse and chicken to date. Evolution of the animal Myb genes was shaped by two independent duplications: the first duplication gave rise to the B-Myb lineage and a second lineage; the second lineage underwent a second duplication to produce the A-Myb and C-Myb genes. We have shown that these two gene duplication

events occurred in the early stage of vertebrates (Wang and Gu 2000). Interestingly, the observation that the two v-Myb (viral-Myb) genes are within the C-Myb clade is consistent with the idea that the retroviral v-Myb gene sequences were acquired from their hosts.

Some R2R3 Mybs from fungi (Nc2253310) or microsporidia (Ec19069694) seem to be more closely related to the R1R2R3 clade rather than to the R2R3 clade (Figure 3); however, the low bootstrap value for Nc2253310 renders its phylogenetic position uncertain. Moreover, the rs2-like clade (purple) is outside of the clade including R2R3 and R1R2R3 Mybs; this result is supported by a fairly high bootstrap value (71%). Finally, Figure 3 suggests that the U (unusual) clade that contains one 4-repeat *Arabidopsis* Myb and three *CDC5* orthologs may have been derived from an ancient paralog of current widely-distributed R2R3 Myb genes.

3.3 Phylogenetic Tree of Isolated Myb Repeats

To address the evolutionary relationship of Myb repeats, we reconstructed a phylogenetic tree containing 317 individual Myb repeat sequences (Figure 4), based on comparisons of 47 sites. Due to the space limitation, the tree presented here is partially compressed. (The complete tree is available at http://pc21955.zool.iastate.edu/origin/isolatedMybRepeats_uc.emf). The inferred phylogeny shows that R1, R2 and R3 repeats are monophyletic (Figure 4, solid circles), implying that the duplications that generated these tandem repeats occurred prior to the separation of the plant and animal kingdoms. Within the R2 (or R3) clade, repeats from the R1R2R3 genes are completely separate from repeats from the R2R3 genes, denoted by R2_R1R2R3 and R2_R2R3, respectively (or R3_R1R2R3 and R3_R2R3). Moreover, within the R1 clade, the R1 repeats (from R1R2R3 Myb genes) are divided into genes from both plant and animal species; one may observe the same pattern within the R3_R1R2R3 subclade (Figure 4). In contrast, the subclade R2_R1R2R3 cannot be divided into exclusively plant- and animal-subclades. This result may imply that R2 is evolving more slowly, and did not accumulate any changes that unite the animals to the exclusion of the plants. Besides, the one-repeat Myb genes (Figure 4, taxa without the prefix R1, R2 or R3) and Myb genes from lower level organisms (such as Microsporidia, Mycetozoa and Fungi) may have been highly diverged. Because the bootstrap values are low, partially due to the low number of sites, the detail of

their evolution cannot be determined with certainty. Additionally, the identity of the most ancient Myb repeat could not be inferred based on analysis of the individual repeats.

4. Discussion

4.1 Expansion of Myb gene family in monocot

Previous reports have described the expansion of Myb-homologous genes in various plants including maize and *Arabidopsis* (e.g., Rabinowicz et al. 1999). For example, 125 R2R3 Myb genes have been identified in *Arabidopsis* (Stracke et al. 2001). Our results confirm and extend those studies, with new sequence data from 64 sorghum and 31 maize Myb genomic clones (accession#: AF474125-AF474133, AF470058-AF470071, AY363121-AY363161, and AF474115-AF474124, AF470072-AF470092, respectively) which we isolated in a targeted strategy. By screening 12,288 sorghum genomic BAC clones (equivalent to 2.6 genomes), we obtained 85 Myb-hybridizing clones, yielding 64 unique sorghum Myb genes. In addition, we identified 51 unique Myb genes in the Monsanto rice database (DraftRiceData). One interesting question is, how many Myb genes exist in the maize genome? Though the final answer awaits the ongoing maize genome projects, we may infer a rough estimate from our data. We screened 36,864 maize BAC clones (2 genome equivalents) and isolated 143 Myb-hybridizing clones. We obtained sequences from 44 BAC clones and detected 31 unique Myb genes among them (redundancy = 30%), suggesting the presence of approximately 100 unique Myb genes from two screened filters. Based on these results, we estimate that the maize genome contains >200 Myb genes; this number is consistent with a previous estimate based on cDNA data (Rabinowicz et al. 1999).

The expansion of the Myb gene family in monocots that is evident from these data supports the idea that gene duplication and divergence are major forces in genomic and organismal evolution (Ohno 1970). The newly created genes arising from duplication may undergo one of several possible fates (Prince and Pickett 2002): 1) Non-functionality. This fate is most common, and may result from mutation or loss from the genome due to chromosomal restructuring. 2) Neo-functionality. This may occur when a population acquires a new advantageous allele under distinct selective constraints. 3) Sub-functionalization. This is an alternative outcome in which a pair of duplicated genes exhibit

complementary, independent sub-functions such that both genes together provide the complete function of the single ancestral gene. For example, the Myb-homologous genes *c1* and *pl1* in maize originated from an ancient gene duplication (Cone et al. 1993). Both genes regulate the expression of anthocyanin structural genes, but in distinct parts of the plant. The *c1* gene controls pigmentation in the aleurone layer of the kernel, whereas the *pl1* gene regulates pigmentation in the vegetative and floral tissues. It is noteworthy that their Myb domains differ only at five residues in the R2R3 domains (Figure 2 sequence 45, 46).

4.2 Important Residue Changes In Myb Protein

It has previously been shown that residue substitutions at the A- and P-type determining site can reduce the DNA binding affinity of Myb proteins (Hegvold and Gabrielsen 1996). However, the maize p1 protein is an A-type Myb, and it binds DNA with high affinity (Williams and Grotewold 1997). Another distinctive site is located in the linker between the third α-helix of the R2 repeat and the first α-helix of the R3 repeat (Figure 2, asterisk). In animal R1R2R3 Myb domains this site always contains a Trp residue, whereas it contains a non-Trp hydrophobic residue in plant R2R3 Myb proteins. In plants, presence of a Trp residue at this site is associated with the occurrence of an additional upstream exon that encodes an R1 repeat. These plant R1R2R3 Myb proteins have been termed pc-Myb (plant c-Myb-like; Braun and Grotewold 1999); dozens of pc-Myb proteins were identified in major land plant lineages: bryophytes (mosses, liverwort), pteridophytes (lycopsid, ferns, horsetail), monocots (barley, rye) (Kranz et al. 2000), and dicots (*Arabidopsis*, tobacco, *Papaver rhoeas*). Thus, this particular Trp residue appears to be a characteristic feature of R1R2R3 Mybs.

In addition, other important residue changes were identified between R2R3 and R1R2R3 Mybs. The alignment of their protein sequences (Figure 2) shows that both types of Myb proteins have highly conserved residues at ten sites within their DNA-binding domains. Six of these sites are in the two DNA recognition α-helices (Figure 2, arrows). At a particular site within the R2 repeat, plant R1R2R3 Myb proteins have a different residue (proline) from that of the animal counterparts (Figure 2, upper arrowhead). At the position located two residues further C-terminal, there is a one amino acid residue insertion/deletion polymorphism that distinguishes R1R2R3 proteins from nearly all of the R2R3 proteins

(Figure 2, second arrow). In addition to providing identifying characteristics of the R2R3 and R1R2R3 Myb proteins, these residue differences are expected to affect the Myb domain structure, and may therefore distinguish the functional properties of the R2R3 and R1R2R3 Myb proteins. For example, these differences would probably impact DNA recognition and binding of the R2R3- and R1R2R3-type Mybs (Williams and Grotewold 1997). The 3D structure of mouse C-Myb protein from PDB (protein database bank; http://www.rcsb.org/pdb) shows that the DNA recognition α-helix interacts with the DNA major groove. Moreover, five amino acid residues in the helix-turn-helix motif bind directly to the major groove of DNA (Ogata et al. 1994). These observations suggest a strong correlation between the functionality of specific amino acid residues and their conservation through evolution. The functional properties of these sites may be tested by introducing mutations and analyzing the properties of the resulting proteins.

Moreover, site-specific rate shifts between some Myb clusters have been detected by the method developed by Gu (1999); using the software DIVERGE (Gu and Vander Velden 2002). For example, the coefficient (θ) of functional divergence between R2R3 and R1R2R3 type Myb genes is $\theta = 0.24 \pm 0.06$, which is significantly larger than zero. It implies that the change of conservation through evolution may be correlated with the change in the functionality of specific amino acid residues (Wang and Gu 2001).

4.3 Conserved amino acid motifs in carboxy-terminal regions

Although the Myb gene superfamily comprises one of the largest families of transcription factors known to date, the functional roles of most plant Myb genes are largely unknown. It would be extremely helpful to be able to use phylogenetic and computational methods to predict the functions of plant Myb genes. Interestingly, plant R2R3 Myb genes possess highly conserved N-terminal Myb domains, but have dramatically divergent C-terminal regions. Recently, conserved motifs within the C-terminal regions of paralogous and orthologous Myb genes have been identified (Stracke et al. 2001). These conserved motifs may reflect constraints upon the functions of related Myb proteins. To test this idea, we are currently analyzing the sequences, expression patterns, and functions, where known, of over 250 R2R3 Myb genes from various plants. The preliminary results support the idea that genes with similar functions are clustered together as subgroups based on their phylogeny

and similarity. Additionally, many of the individual subgroups contain C-terminal motif sequences which can be used as distinct markers for Myb gene subgroups. Details and further results of this analysis will be published elsewhere (Jiang, C. and Peterson, T., in preparation).

4.4 Origin of Typical Myb Proteins: Gain-of-R1 Model vs. Loss-of-R1 Model

It is thought that the Myb domain first originated over 1 billion years ago, shortly after the divergence of eubacteria and eukaryotes (Lipsick 1996). Rosinski and Atchley (1998) argued that typical Mybs might have a polyphyletic origin, with only the Myb DNA-binding motif deriving from a common origin. With substantial sequence data available, we are able to investigate the origin of two major types of Myb genes, R1R2R3 and R2R3. Although the evolutionary relationship of the non-typical Myb genes (one-repeat and partial Myb genes) remains unclear, the topology (Figure 3) of our study suggests that the R2R3 and R1R2R3 Myb genes were generated by successive gain of repeat units. This "gain" model is illustrated schematically in figure 5(B): first, the ancestral R2R3 Myb (2R) was produced by an intragenic domain duplication; subsequently, the ancestral R1R2R3 Myb (3R) was formed by a further intragenic domain duplication. Both the R2R3 and R1R2R3 Mybs co-existed in primitive eukaryotes, and gave rise to the currently extant Myb genes.

In contrast, a previous model proposed that R2R3 Myb genes originated from R1R2R3 genes by loss of the R1 repeat (Lipsick 1996; Rosinski and Atchley 1998); this model has received further support (Braun and Grotewold 1999; Kranz et al. 2000). According to the loss model, R1R2R3 Mybs were generated by successive intragenic duplication events in the primitive eukaryotes, and these evolved into today's R1R2R3 Myb genes in plants and animals. Whereas, the R2R3 Myb genes commonly found in plants were produced by the loss of the first repeat (R1) from R1R2R3 Mybs in the plant evolutionary lineage. However, when these two models were applied to the topology indicated by our results, the gain model can be accommodated with two changes, whereas the loss model requires five changes (Figure 5). Thus, the gain model provides a more parsimonious explanation for Myb gene evolution and hence is favored over the loss model.

The presence of a R1R2R3 Myb gene in the mycetozoan (slime mold) *Dictyostelium* was previously cited as evidence in support of the loss model (Lipsick 1996); this conclusion was

based on the assumption that the mycetozoa are an outgroup to an animal + plant clade. However, recent analyses more strongly support an animal + mycetozoa clade (Baldauf et al. 2000), hence the evidence for the domain loss model is no longer compelling. In subsequent studies, support for the domain loss model has been inferred based on the observation that R1R2R3 Mybs are more broadly distributed (in animals, mycetozoa and plants) than typical R2R3 Mybs (in plants). (Braun and Grotewold 1999; Kranz et al. 2000). However, this reasoning neglects the phylogeny and is dependent on sample size. For instance, our blast search found that R2R3 Mybs are also present in microsporidia, mycetozoa and fungi (Figure 3). In summary, the gain model can provide a reasonable explanation for the phylogenetic distribution of different types of Mybs: the second ancient duplication of 2-repeat Myb genes produced the 3-repeat Myb genes via gain of R1; subsequently, the 2-repeat Myb genes were lost in an early ancestor of the animal kingdom. Although our analysis can not rule out other possibilities completely, taken together, the gain model is favored based on the present evidence.

Acknowledgements

We thank Terry Olson for technical assistance in the isolation and sequencing of Myb genes. We acknowledge Monsanto/Pharmacia rice-research.org (www.rice-research.org) for providing the *O. sativa* genomic sequences. This work was supported by a grant from the USDA-NRICGP to T.P. and S.C, and an NIH grant to X. G. X.G is the DuPont young faculty. This is Journal Paper No. J-19751 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3297, and supported by Hatch Act and State of Iowa funds.

References

Avila, J., Nieto, C., Canas, L., Benito, M.J., Paz-Ares, J., 1993. *Petunia hybrida* genes related to the maize regulatory *C1* gene and to animal myb proto-oncogenes. Plant J. 3, 553-562.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., Doolittle, W.F., 2000. A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. Science 290, 972-977.

Baranowskij, N., Frohberg, C., Prat. S,, Willmitzer. L., 1994. A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. EMBO J 13, 5383-9392.

Braun, E.L., Grotewold, E., 1999. Newly discovered plant c-*myb*-like gene rewrite the evolution of the plant *myb* gene family. Plant Physiology 121, 21-24.

Cedroni, M.L., Cronn, R.C., Adams, K.L., Wilkins, T.A., Wendel, J.F., 2003. Evolution and expression of Myb genes in diploid and polyploid cotton. Plant Mol Bio 51, 313-325

Cone, K.C., Cocciolone, S.M., Burr. F.A., Burr, B., 1993. Maize anthocyanin regulatory gene *pl* is a duplication of *c1* that functions in the plant. Plant Cell 5, 1795-1805.

Dias, A.P., Braun, E.L., McMullen, M.D., Grotewold, E., 2003. Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication. Plant Physiology 131, 610-620.

England, B.P., Admon, A., Tjian, R., 1992. Cloning of *Drosophila* transcription factor Adf-1 reveals homology to Myb oncoproteins. Proc. Natl. Acad. Sci. USA 89, 683-687.

Grotewold, E., Drummond, B.J., Bowen, B., Peterson, T., 1994. The Myb-homologous *P* gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell 76, 543-553.

Gu, X., 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. 16, 1664-1674.

Gu, X., Vander Velden, K., 2002. DIVERGE: Phylogeny-based Analysis for Functional-Structural Divergence of a Protein Family. Bioinformatics 18, 500-501.

Hegvold, A.B., Gabrielsen, O.S., 1996. The importance of the linker connecting the repeats of the c-Myb oncoprotein may be due to a positioning function. Nucleic Acids Res. 24, 3990-3995.

Klempnauer, K.-H., Gonda, T.J., Bishop, J.M., 1982. Nucleotide sequence of the retroviral leukemia gene *v-myb* and its cellular progenitor *c-myb*: the architecture of a transduced oncogene. Cell 31, 453-463.

Kranz, H., Scholz, K., Weisshaar, B., 2000. c-Myb oncogene-like genes encoding three Myb repeats occur in all major plant lineages. Plant Journal 21, 231-235.

Kumar, S., Tamura, K., Jakobsen, I.B., Neim, M., 2001. MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 17, 12:1244-1245.

Lipsick, J.S., 1996. One billion years of Myb. Oncogene 13, 223-235.

Nei, M., 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. 30, 371-403.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S., Nishimura, Y., 1994. Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. Cell 79, 639-648.

Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag, Berlin and New York, pp. 59-60

Oppenheimer, D.G., Herman, P.L., Sivakumaran, S., Esch, J., Marks, M.D., 1991. A myb gene required for leaf trichome differentiation in *Arabidopsis* is expressed in stipules. Cell 67, 483-493.

Prince, V.E., Pickett, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. 3, 827-837.

Rabinowicz, P.D., Braun, E.L., Wolfe, A.D., Bowen, B., Grotewold, E., 1999. Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants. Genetics 153, 427-444.

Rosinski, J.A., Atchley, W.R., 1998. Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin. J. Mol. Evol. 46, 74-83.

Stracke, R., Werber, M., Weisshaar, B., 2001. The R2R3-Myb gene family in *Arabidopsis thaliana*. Current Opinion in Plant Biology 4, 447-456.

Urao, T., Yamaguchi-Shinozaki, K., Urao, S., Shinozaki, K., 1993. An *Arabidopsis* myb homolog is induced by dehydration stress and its gene product binds to the conserved Myb recognition sequence. Plant Cell 5, 1529-1539.

Wang, Y., Gu, X., 2000. Evolutionary patterns of gene families generated in the early stage of vertebrates. J Mol Evol, 51, 88-96.

Wang, Y., Gu, X., 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158, 1311–1320.

Williams, C.E., Grotewold, E., 1997. Differences between plant and animal Myb domains are fundamental for DNA-binding activity, and chimeric Myb domains have novel DNA-binding specificities. J. Biol. Chem. 272, 563-571.

Yang, Y., Klessig, D.F., 1996. Isolation and characterization of a tobacco mosaic virus-inducible myb oncogene homolog from tobacco. Proc Natl Acad Sci USA 93, 14972-14977.

Yu, E.Y., Kim, S.E., Kim, J.H., Ko, J.H., Cho, M.H., Chung, I.K., 2000. Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1. J. Biol. Chem. 275, 24208-24214.

Zhang, P., Chopra, S., Peterson, T., 2000. A segmental gene duplication generated differentially expressed *myb*-homologous genes in maize. Plant Cell 12, 2311-2322.

Figure 1. Structure of typical plant R2R3 Myb genes and sequencing strategy.
Start codon (ATG), 3 exons (large rectangle), 2 introns (if present) and stop codon (*) are shown. Six gray/dark boxes are the 6 α–helices in R2 and R3 repeats. The arrows show the bi-directional primer walking sequencing strategy. The bottom line indicates the assembled sequences.

Figure 2. Multiple alignment of 77 representative Myb domains.
The first 25 sequences are R1R2R3 Myb proteins from plants and animals. Sequences 28~74 are plant R2R3 Myb proteins. Sequences 69-74 are rs2-like Myb proteins. Sequence 77 from *Plasmodium falciparum* is used as an outgroup for phylogenetic analysis.
Top: The brackets indicate the extent of the R2 and R3 repeats, while the bars indicate the positions of the 3 α–helices in each repeat (Ogata et al. 1994). The arrowhead and arrows show the residue changes described in Section 4.2. **Bottom**: The vertical bars show the positions of introns I and II if present. The six asterisks indicate the constantly spaced Trp residues. The arrowhead indicates the A-, P-, R-, and S-type determining site. The sequence identifiers are composed of initials for genus and species, followed by GI number in GenBank. The common names of 7 well-known maize genes are also indicated after the GI number, as well as sorghum *y1* gene, and *Arabidopsis CDC5* gene. v-Myb indicates viral Myb proteins.

Figure 3. A NJ tree of the complete R2R3 repeats from 139 Myb proteins of 38 species. The open arrowhead indicates the divergence of R2R3 and R1R2R3 Myb genes. The solid arrowhead shows the A-, R- and S-type clade. The four arrows show the branch values of the corresponding clades. Bootstraps (1000 replicates) of >50% are shown. The U clade has *CDC5* orthologs and one 4-repeat *Arabidopsis* Myb gene. Taxa not from plants and animals are indicated in parentheses. Pf11595856 from *Plasmodium falciparum* is used as an outgroup.

Figure 4. A partially compressed NJ tree of 317 isolated R1, R2 and R3 Myb repeats.
Solid circles show the three monophyletic groups of R1, R2 and R3. Two large triangles indicate R2 and R3 subclades from plant R2R3 Mybs. Within the R1 and R3 subclasses from

R1R2R3 Mybs, plant and animal taxa are indicated separately. In contrast, R2 subclass cannot be divided into exclusively plant- and animal- subclades. (Animal taxa are in gray shading.) Two triangles with vertical lines are from rs2-like Mybs. Outside of the three monophyletic groups, taxa with the prefix R1, R2 or R3 correspond to the U (Unusual) group in figure 3. The rest of taxa indicate the previously-reported one-repeat Myb genes. Taxa from species other than plants and animals are indicated in parentheses.

Figure 5. Loss model (A) vs. gain model (B) for Myb gene evolution.
The arrow at left indicates the origin of R2R3 Mybs from an ancient Myb domain. Thin lines and thick lines indicate 2-repeat and 3-repeat Mybs, respectively. Each black dot indicates a subsequent change (gain or loss of Myb repeat, or extinction event). The dashed lines indicate the proposed extinction of the 2-repeat lineage in animals. The outgroup includes representatives from plants, Alveolata and Mycetozoa.
A. Loss model: There is one change (gain) from 2R to 3R at the 2nd duplication. The resulting 3R Myb became the common ancestor of current Myb genes (Lipsick 1996), and underwent 3 subsequent losses of R1: one to generate the rs2-like clade, one to generate the plant/animal clade, and one additional loss leading to the Microsporidia. (We ignore the Fungi Myb Nc2253310 because of its low bootstrap value. See figure 3). The 2R Myb died out in the animal lineage. In total there are 5 changes: one gain, three losses, and one extinction.
B. Gain model: There is one change (gain) to form the 3R Mybs in plants, mycetozoa and animals. The ancient 2R and 3R Mybs coexisted in primitive eukaryotes and descended into present-day organisms. The 2R Myb died out in the animal lineage. In total there are 2 changes: one gain, and one extinction. Clearly, the gain model is more parsimonious than the loss model.

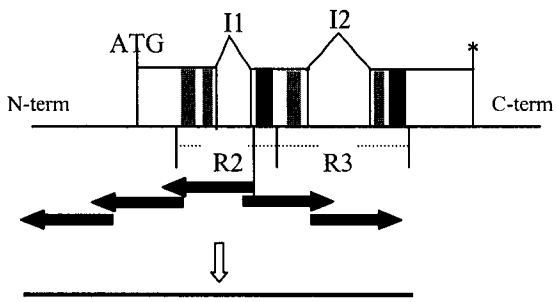


Figure 1

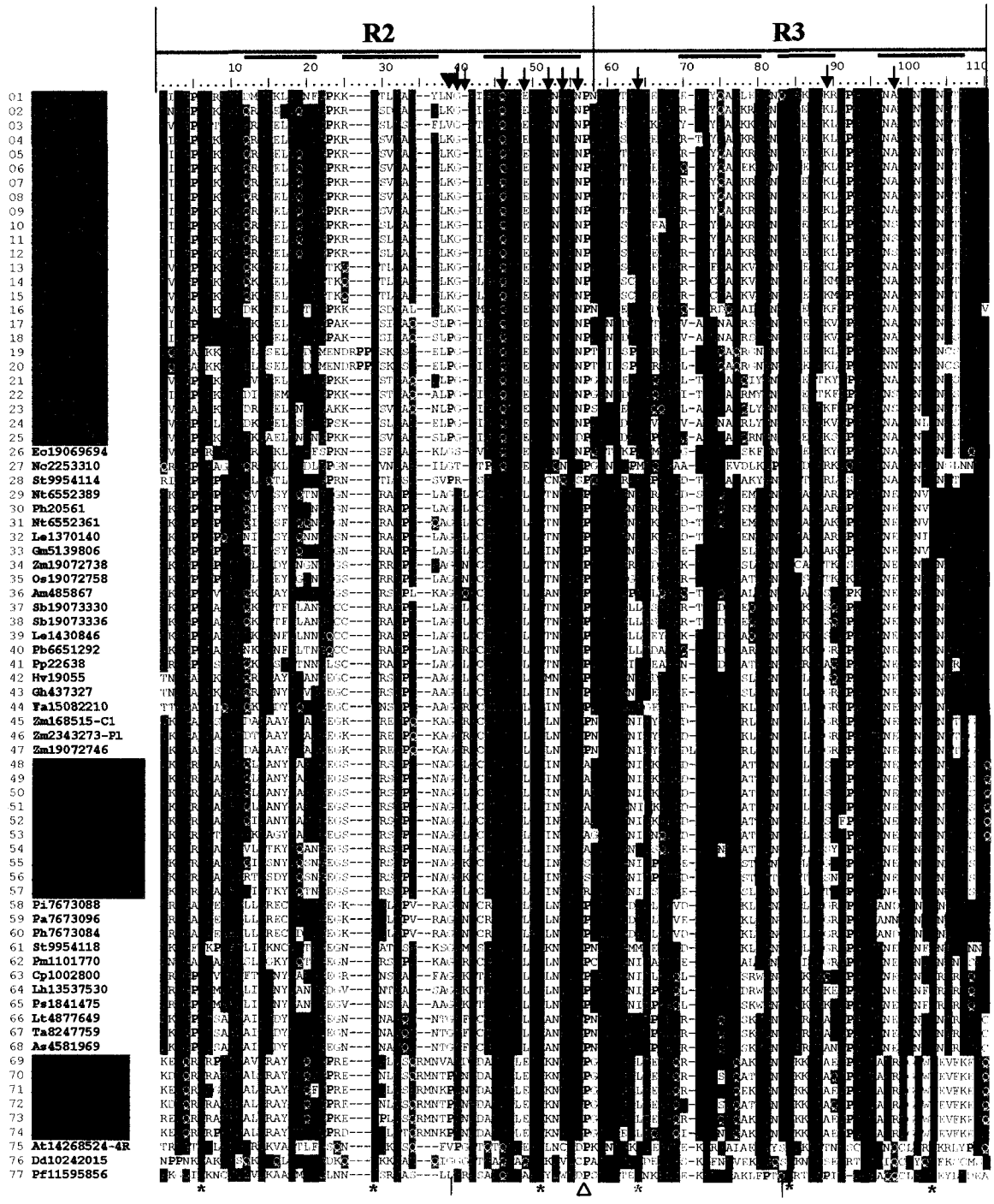

R2
R3
10
20
30
40
50
60
70
80
90
100
110
Eo19069694
Nc2253310
St9954114
Nt6552389
Ph20561
Nt6552361
Le1370140
Gm5139806
Zm19072738
Os19072758
Am485867
Sb19073330
Sb19073336
Le1430846
Pb6651292
Pp22638
Hv19055
Gh437327
Fa15082210
Zm168515-C1
Zm2343273-P1
Zm19072746
Pi7673088
Pa7673096
Ph7673084
St9954118
Pm1101770
Cp1002800
Lh13537530
Ps1841475
Lt4877649
Ta8247759
As4581969
At14268524-4R
Dd10242015
Pf11595856

Figure 2

Figure 3

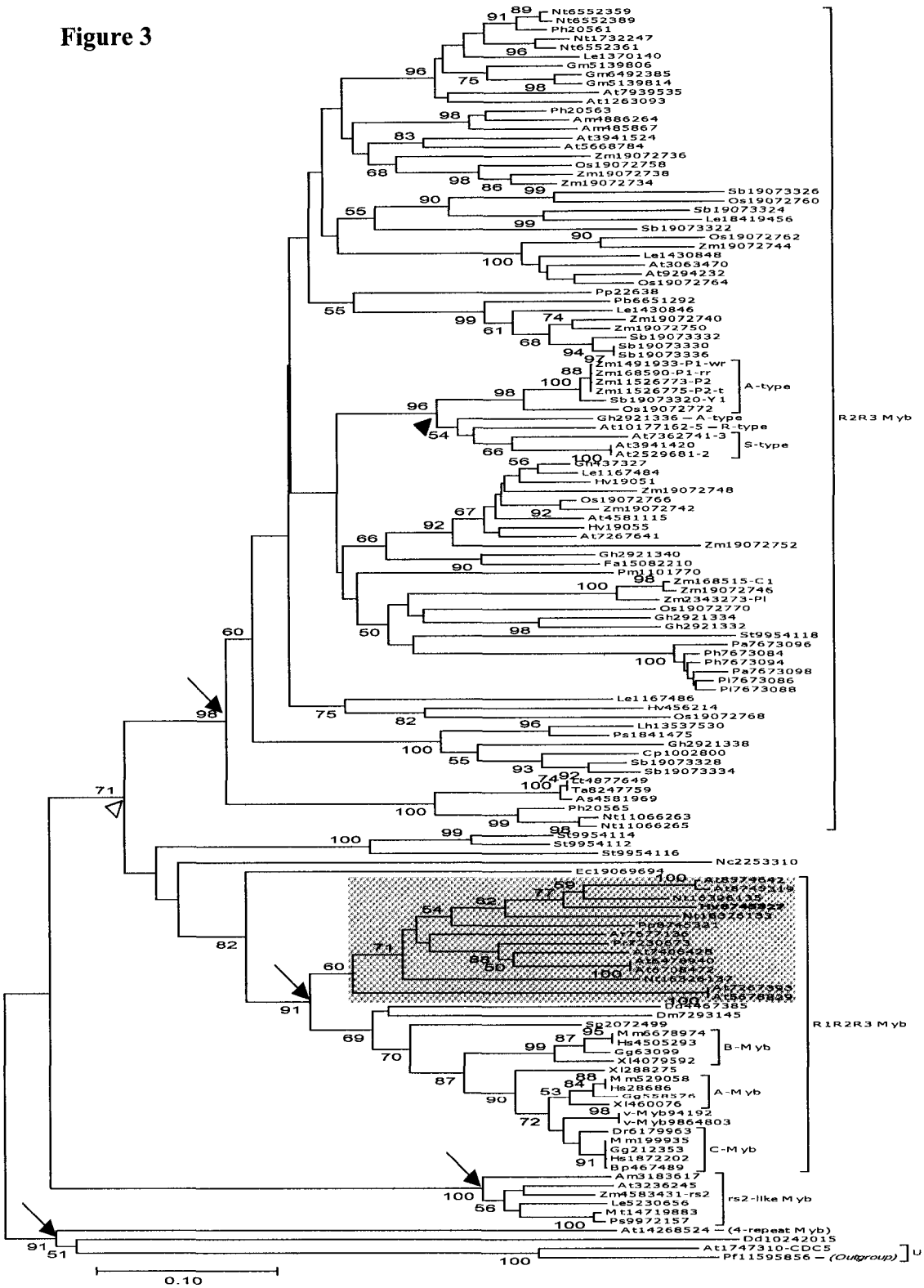

Nt6552359
Nt6552389
Ph20561
Nt1732247
Nt6552361
Le1370140
Gm5139806
Gm6492385
Gm5139814
At7939535
At1263093
Ph20563
Am486264
Am485867
At3941524
At5668784
Zm19072736
Os19072758
Zm19072738
Zm19072734
Sb19073326
Os19072760
Sb19073324
Le18419456
Sb19073322
Os19072762
Zm19072744
Le1430848
At3063470
At9294232
Os19072764
Pp22638
Pb6651292
Le1430846
Zm19072740
Zm19072750
Sb19073332
Sb19073330
Sb19073336
Zm1491933-P1-wr
Zm168590-P1-rr
Zm11526773-P2
Zm1152675-P2-t
Sb19073320-Y1
Os19072772
A-type
Gh2921336 — A-type
At10177162-5 — R-type
At7362741-3
At3941420
At2529681-2
S-type
R2R3 Myb
Gm137327
Le1167484
Hv19051
Zm19072748
Os19072766
Zm19072742
At4581115
Hv19055
At7267641
Zm19072752
Gh2921340
Fa15082210
Pm1101770
Zm168515-C1
Zm19072746
Zm2343273-P1
Os19072770
Gh2921334
Gh2921332
St9954118
Pa7673096
Ph7673084
Ph7673094
Pa7673098
Pl7673086
Pl7673088
Le1167486
Hv456214
Os19072768
Lh13537530
Ps1841475
Gh2921338
Cp1002800
Sb19073328
Sb19073334
Et4877649
Ta8247759
As4581969
Ph20565
Nt11066263
Nt11066265
St9954114
St9954112
St9954116
Nc2253310
Ec19069694
Dm67385
Dm7293145
Sp2072499
Mm6678974
Hs4505293
Gg63099
Xl4079592
Xl288275
Mm529058
Hs28686
Xl460076
v-Myb94192
v-Myb9864803
Dr6179963
Mm199935
Gg212353
Hs1872202
Bp467489
B-Myb
A-Myb
C-Myb
R1R2R3 Myb
Am3183617
At3236245
Zm4583431-rs2
Le5230656
Mt14719883
Ps9972157
rs2-like Myb
At14268524 — (4-repeat Myb)
Dd10242015
At1747310-CDC5
Pf11595856 — (Outgroup)
U
0.10

Figure 4


At2129650
R3 From R2R3
R3
R3 From Plant R1R2R3
87
81
R3 From Animals R1R2R3
R3-Dd4467385 – (Mycetozoa)
R3-St9954116
R3-St9954114
(Divergent R3 From Potato)
96
R3-St9954112
96
R3-Nc2253310 – (Fungi)
99
R3 From rs2-like
R3-Dd10242015 – (Mycetozoa)
99
St1076660
St7705206
R3-Pf11595856 – (Alveolata)
R3-At1747310-CDC5
99
R0-At14268524-4R – (4R-Myb)
R3-At14268524-4R – (4R-Myb)
R2-Dd10242015 – (Mycetozoa)
68
R1 From Plant R1R2R3
R1
54
R1 From Animals R1R2R3
Dr6179963
Mm199935
Hs1872202
Gg212953
Rn467485
Xl450076
Hs28686
Gg238576
Mm549059
51
53
80
60
50
86
78
97
Dd4467385 – (Mycetozoa)
R2 From R1R2R3
Pp8745321
Nt16326133
At7677136
Pr7230673
At7406428
At6708472
At6478940
99
Nt16326137
70
Hv8745327
Nt16326135
86
51
At8745319
At8574642
99
99
At7267393
At5678829
Ec19069694 – (Microsporidia)
R2
R2 From R2R3
99
R2 From rs2-like
R2-Nc2253310 – (Fungi)
R1-At14268524
R2-At14268524
(4R-Myb)
R2-Pf11595856 – (Alveolata)
R2-At1747310-CDC5
99
Pc2224899
Dm283636-Adf-1
57
Zm483444
99
Pc396197
0.10

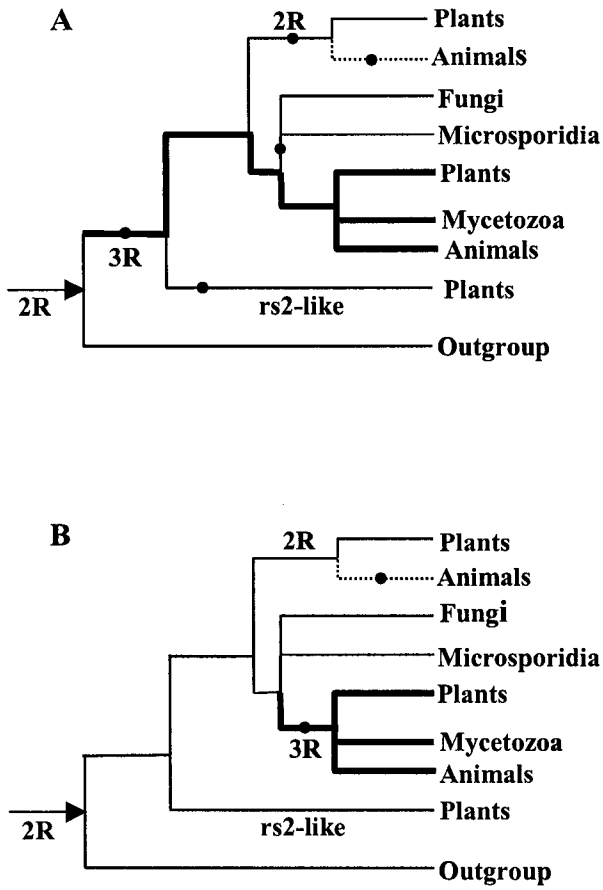

A
2R
Plants
Animals
Fungi
Microsporidia
Plants
Mycetozoa
Animals
3R
2R
Plants
rs2-like
Outgroup
B
2R
Plants
Animals
Fungi
Microsporidia
Plants
3R
Mycetozoa
Animals
2R
Plants
rs2-like
Outgroup

Figure 5

CHAPTER 4. FUNCTIONAL CLASSIFICATION AND PREDICTION OF MYB GENE FAMILY IN *ARABIDOPSIS* AND RICE

A paper to be submitted to *Genome Biology*

Cizhong Jiang and Thomas Peterson

Abstract

Myb proteins contain a conserved DNA-binding domain composed of one to four repeat motifs (referred to as R0R1R2R3); each repeat is approximately 50 amino acids, with constantly spaced Tryptophan residues. It is one of the largest families of diverse transcription factors in plants. However, with the exception of a few well-studied cases, little is known about the functions of most Myb genes. Here, we used computational techniques to classify and predict the functions of Myb genes from *Arabidopsis* and rice. In our study, 130 Myb genes were identified in *Arabidopsis* and 85 in rice. The collected Myb proteins were clustered into subgroups based on sequence similarity and phylogeny. Conserved motifs were detected in C-terminal coding regions of Myb genes within subgroups. Moreover, EST blast search confirmed that those motifs only specifically existed in Myb genes. In contrast, no common regulatory motifs were identified in the non-coding regions due to the high divergence. Interestingly, Gene structure analysis revealed that there was a significant excess of phase 1&2 introns as well as an excess of non-symmetric exons in Myb domains. The exon-intron structure differed between subgroups, but was conserved in the same subgroup. Additionally, the distribution pattern of introns in the phylogenetic tree suggested that Myb domains originally had a compact size without introns. Non-coding sequences were inserted and the splicing sites were conserved during evolution. Taken together, the conserved motifs and splicing sites may reflect functional constraints upon Myb domains. Based on our analyses, we have predicted the functions of a large number of previously uncharacterized Myb genes. The resulting functional classification table may provide a useful starting point for determination of Myb gene function.

Introduction

Regulation of gene expression at the level of transcription controls many important biological processes in a cell or organism. The process of transcription recruits a number of different transcription factors, which can be activators, repressors, or both (Stracke et al. 2001). Genome-wide comparisons have revealed the diversity in the regulation of transcription during evolution. With the completion of *Arabidopsis* genome sequencing, 5% of its genome was found to encode more than 1500 transcription factors (Riechmann et al. 2000). On the basis of similarities, transcription factors have been classified into families (Pabo and Sauer 1992). In plants, Myb factors comprise one of the largest of these families (Romero et al. 1998; Riechmann et al. 2000).

The first Myb gene was found as the v-Myb oncogene from Avian Myeloblastosis Virus (Klempnauer et al. 1982). Subsequently, members of the Myb gene family were identified in diverse plants and animals (Thompson and Ramsay 1995; Martin and Paz-Ares 1997; Rosinski and Atchley 1998). Previous research showed that animal genomes encode relatively few Myb genes (Thompson and Ramsay 1995; Rosinski and Atchley 1998). In contrast, flowering plants contain large numbers of Myb genes with very diverse structure and function (Martin and Paz-Ares 1997). Moreover, most plant Myb genes have unclear functions, although some well-studied examples imply important roles for plant transcription factors to date in regulation of secondary metabolism, cellular morphogenesis, and pathogen resistance, and in responses to growth regulators and stresses (Martin and Paz-Ares 1997; Meissner et al. 1999). With the completion of rice genome sequencing (Yu et al. 2002; Goff et al. 2002), the entire complement of Myb genes can be identified and described. However, a great deal of experimental work is required to determine the specific biological function of each Myb gene. We have employed phylogenetic and computational methods to classify Myb genes in subgroups, and we propose that genes within a subgroup will likely share similar functions. The resulting subgroup classification and functional prediction may be very useful for research on agronomic traits in rice, which is the most important crop for human consumption, and an important model for other cereal grains.

Materials and Methods

Myb proteins used in the analysis

Rice genome sequences (Scaffold data set) were obtained from http://btn.genomics.org.cn:8080/rice/. FGeneSH (http://www.softberry.com/berry.phtml) has been used successfully to predict genes in rice (Yu et al. 2002), and GenScan (http://genes.mit.edu/GENSCAN.html) were used together to predict genes by taking rice genomic sequences as input. The two prediction results were combined as the complement of rice proteins. Two programs, blastp (Altschul et al. 1997) and HMMER (Eddy 1996), were used to identify Myb genes from this rice protein data set. For blastp, we used a set of Myb R2R3 domains as query sequences. For HMMER, we used Myb profile from Pfam. We parsed and combined the results of both searches, and obtained the final complement of rice Myb proteins with manual inspection of each sequences to confirm the identification of bona fide Myb genes. In the end, 85 typical Mybs with complete R2R3 domains (one R0R1R2R3 and 84 R2R3) and 28 partial Mybs were detected in the rice genome. The partial Mybs are those genes that only contain a segment similar to one or a partial Myb repeat. The sequences of rice Mybs are available at (http://pc21955.zool.iastate.edu/Myb/BGI-Mybs-113.fas).

The complement of *Arabidopsis* proteins from GenBank were used to identify Myb proteins with complete R2R3 domains. The same methods as above were applied. We obtained 130 typical Myb proteins containing complete R2R3 domains (1 R0R1R2R3 Myb, 5 R1R2R3 Mybs, 124 R2R3 Mybs) and 11 partial Mybs. The results are consistent with the previous findings (Stracke et al. 2001).

To collect reference information on Myb gene functions, we used blastp search against non-redundant data set in GenBank. The search yielded 43 plant Myb proteins with complete R2R3 domains; for most of these, some experimental information regarding functions or expression patterns was deposited by individual researchers.

Construction of phylogeny and motif identification

For phylogenetic analysis, the above 258 Myb proteins (130 *Arabidopsis*, 85 rice, and 43 from various other plants) with complete R2R3 domains were included. The sequences were

aligned by ClustalX (version 1.81). The phylogenetic tree was constructed by the neighbor-joining method (Saitou and Nei 1987) using MEGA version 2.0 (Kumar et al. 2001) with the setting of pairwise gap deletion and Poisson distance. Bootstrapping (1000 replicates) was performed to evaluate the degree of support for a particular grouping in the neighbor-joining analysis. To enable the identification of motifs in the C-terminal regions within each subgroup, we did not employ complete gap deletion as this may tend to exclude the contribution of C-terminal residues because of their high divergence. P-distance represents the simplest sort of distance calculation and can be highly biased, so it was not use it, either.

The sequences were classified into subgroups based on their sequence similarity and the topology of the phylogeny. Within each subgroup, motifs were detected using MEME (Bailey and Elkan 1994). With the following parameter settings: the distribution of motifs: zero or one per sequence; maximum number of motifs to find: 16; minimum width of motif: 6; maximum width of motif: 117, in order to identify non-breaking R2R3 domains; minimum number of sites for each motif: the number of sequences, i.e., the motif must be present in all members within the same subgroup. Other options used the default values.

Motif analysis: similarity scores and ratio of nonsynonymous (dN) to synonymous (dS) substitution

To confirm the reliability of the 47 motif candidates identified by MEME, we used PlotSimilarity from GCG package from Genetics Computer Group, Inc. (http://www.accelrys.com/products/gcg_wisconsin_package/) to calculate the similarity score of each motif plus its 10-residue flanking fragments (protein sequences). There were 42 motifs with values above the average score in the motif region and below the average score in the flanking regions, and these were tested further using the dN/dS ratio. The program YN00 from PAML package (Yang and Nielsen 2000) was applied to analyze the conservation of each motif plus its flanking regions (coding DNA sequences). There were 35 motifs with dN/dS <1 in the motif region and >1 in flanking regions.

EST search of motifs

To test whether the C-terminal motifs are found exclusively in Myb genes and not in other non-Myb genes, motif sequences were used to perform tblastn search of the complete

EST data set from GenBank. Low complexity was turned off for optimal short sequence search. For motifs less than 15 residues 10 downstream residues were appended and this elongated sequence was used as query sequence to perform an additional EST search. The corresponding Myb R2 repeats were used in a tblastn EST search as an internal positive control.

Results and Discussion

Expansion of Myb genes in Arabidopsis and rice

The Myb gene family has broadly expanded in plants during evolution. The amplification of Myb gene family occurred prior to the divergence of monocots and dicots (Jiang et al. 2003). In our study, 130 Mybs were found in *Arabidopsis* genome, 85 in *Oryza sativa.* The large size of this gene family was also confirmed in *Zea mays* (Rabinowicz et al. 1999; Jiang et al. 2003), sorghum (Jiang et al. 2003), and cotton (Cedroni et al. 2003). Approximately 30 Myb genes have been described to date in *Petunia hybrida* (Avila et al. 1993). Although most plant Mybs contain only two repeats, there have been 3-repeat Mybs reported in *Arabidopsis* (Stracke et al. 2001), maize (Braun and Grotewold 1999), and other plants (Jiang et al. 2003). However, no 3-repeat Mybs were detected in rice in our study.

Topology of Myb gene phylogeny

Based on the sequence similarity and the topology of the phylogeny, we clustered the Myb genes into 42 subgroups ranging in size from 2 to 14 Myb genes (Fig.1). The phylogenetic topology and subgroup structures are consistent with previous reports (Kranz et al. 1998; Stracke et al. 2001). The comparison result can be accessible here (http://pc21955.zool.iastate.edu/Myb/Subgroups.pdf).

Interestingly, AtMyb33, 65, 101, 104 and At3g60460 were complementary, with few mismatches, to *Arabidopsis* Myb microRNA (noncoding RNA) miR159 (Rhoades et al. 2002). The sequence is 21 nucleotides in length and located in 3'UTR of all 5 genes. MicroRNAs are proposed to act as regulators of gene expression through interactions with complementary mRNA sequences. Importantly, these five *Arabidopsis* Mybs are located in

subgroup G17 (Fig. 1). This clustering further provides a line of evidence for the reliability of the subgroup designation in our analysis.

Conserved gene structure within each subgroup supporting the designation of subgroups

The phylogenetic topology and subgroup structures are based upon sequence comparisons of the complete predicted Myb genes. To test the reliability of the designation of subgroups using independent criteria, we investigated the exon-intron structure of Myb genes subgroup by subgroup. A majority of *Arabidopsis* (59%) and rice (53%) Myb genes have a conserved splicing pattern of 3 exons and 2 introns in R2R3 domains (represented by subgroup G18; Fig. 2A). Either or both of the two introns are absent in 19% of *Arabidopsis* Myb genes and 12% of rice Myb genes. Variable splicing patterns different from G18 were detected in 22% of *Arabidopsis*, and 35% of rice Myb genes, respectively (Data not shown). Strikingly, the exon-intron structure is conserved within each subgroup, but varies between subgroups (Fig. 2B, C). This supports the designation of subgroup from the independent criterion of splicing pattern.

Interestingly, the Myb gene splicing patterns constitute four major blocks in the Myb gene phylogeny (Fig. 1). Block A lacks both intron 1 and 2. There are three splicing patterns in block B: Subgroup G21 lacks both introns, subgroup G24 lacks only intron 2, and the remaining genes have altered splicing sites when compared to subgroup G18. Myb genes in block C have the major splicing pattern (81.2%) typified by G18, with some individual genes lacking intron 1 (9.4%), intron 2 (4.7%), both (1.9%), or having non-major splicing patterns (2.8%). In contrast 58.2% of Myb genes in block D retain the typical splicing sites, and the rest lack only intron 1 (G02, G05, half of the genes in G25). In addition to splice site locations, we also examined the position of splicing with respect to the open reading frame (phase). Splicing can occur between two codons or within a codon: phase 0 intron, occur after the third nucleotide of the first codon; phase 1 intron, occur after the first nucleotide of the single codon; phase 2 intron, occur after the second nucleotide. Figure 2A shows not only the conserved locations in the Myb domain protein sequences but also the conserved phases of introns within the same subgroup. Moreover, there is a significant excess of phase 1&2 introns as well as an excess of non-symmetric exons in Myb genes. (Symmetric exons are exons that are flanked by introns of the same phase.) According to the intron-early

theory (Gilbert 1987, Souza 2003), with exons multiple of three bases, an excess of phase 0 introns and symmetric exons facilitates exon shuffling by avoiding interruptions of the open reading frame, and thus could accelerate the rate of recombinational fusion and exchange of protein domains. In contrast, our results suggest that ancient Myb genes had a compact size without introns. During evolution, under some unknown mechanisms, introns were inserted into Myb domains and resulted in the observed splicing patterns. At the same time, transposition of introns occurs very infrequently during evolution (Fedorova and Fedorov 2003). One splicing pattern remained unchanged in the subsequent gene amplification, resulting in the major splicing pattern as G18. This intron-gain model consistent with previous results shows that numerous introns have been inserted into plants and retained in the genome (Rogozin et al. 2003). Besides, the conserved gene structure within the same subgroup may be related to their conserved functions in an unknown mechanism. A similar approach to gene classification using intron/exon structure has been applied in the kinesin family (Lawrence et al. 2002) and the bHLH family (Toledo-Ortiz et al. 2003), and the results support a similar evelutionary pattern.

Although the splicing sites are conserved, the sizes of both introns vary greatly for different Myb genes. Approximately 85% of intron 1 & 2 of Myb genes is shorter than 300 bp in *Arabidopsis* and rice. Detailed information about the distribution of intron sizes of Myb genes is available at (http://pc21955.zool.iastate.edu/Myb/Introns.pdf). It is worth noting that the size of intron 2 of maize *p1* and *p2* orthologs is very large, ~ 5 kb. This intron size information may be helpful for aligning ESTs with genomic sequences. Strikingly, a 743-base fragment was found in intron 2 of maize *P1-rr* and *P1-wr* alleles, but not in *P1-rw* and *p2* alleles. A 10-base direct repeat (5'TGATTT TGAC3') flanks this fragment. Interestingly, no Ac elements were found inserted in its adjacent 3.2-kb intronic region, but frequent Ac insertion occurred in other regions. This could be due to a particular chromatin structure refractory to Ac insertion in this region (Athma et al. 1992). Blast search detected this fragment (94% identity over 723 base pairs) at one other locus in maize genome, but with a new flanking direct repeat (5'GGATATCCA3'). Its GenBank accession number is AF466202 (located 84795..85689, 12-MAR-2002 version). These results are consistent with a previous proposal that some transposable elements could insert into genome as intronic

sequences, a mechanism which has been proposed for the insertion of nuclear introns (Menssen et al. 1990, Fridell et al. 1990).

C-terminal motifs

The extent of the Myb R1, R2 and R3 repeats is based on similarity to the previously-published consensus Myb repeat sequences (Frampton et al. 1991; Ogata et al. 1992). We used computational methods to identify conserved sequences downstream of the Myb repeats. A total of 30 motifs were identified in the C-terminal regions, with each motif ranging in size from 7 to 32 amino acids (Fig. 1, blue and pink boxes). An exceptionable large (91 residues) domain was found in subgroup G31, gene *CDC5*, which is a conserved Myb paralog that originated prior to Myb-family amplification (Jiang et al. 2003). In addition, Myb genes maize *C1*, *Pl*, and AtMyb123 (*TT2*) in subgroup G12 have a nine-amino acid motif previously reported (Dias et al. 2003, Stracke et al. 2001). This motif is highly conserved in the above three genes, although it is absent in the rice Myb genes in this subgroup (Fig. 1 green box 31).

Interestingly, five short fragments directly following the Myb R3 repeat are highly conserved in some subgroups (Fig. 1, yellow boxes). We designated these extension motifs, E1, E2, etc. Subgroups with an extension motif contain few or zero motifs in their C-terminal coding regions when compared to those subgroups without extension motifs (Fig. 1). In G04 and G26, two short conserved segments following E1 are termed E2 and E3. The five extension motifs are relatively small, ranging from 8-13 residues, but they are much more conserved than other motifs' (Table 1). In the group of five extension motifs, 47 (of 54) sites are occupied by a single residue in more than 50% of the Myb proteins, and this value is greater than twice the relative frequency of the second most frequent residue.

To test the reliability of the motif predictions, the similarity scores were calculated over the motif plus its flanking regions. The similarity plots produced much higher scores in the motif region than the flanking regions (Fig. 3), thus supporting the identified motifs. Similar results were observed using the ratio of nonsynonymous (dN) to synonymous (dS) substitution, which is a typical way to examine the degree of functional constraint upon proteins using evolutionary comparisons (Dias et al. 2003). The results showed that motif regions were subject to purifying selection (ratio < 1), whereas the flanking regions were

subject to positive selection (ration > 1) (Fig. 3). This suggests that the identified motifs contribute to the conserved function within each subgroup.

Specificity of motifs to Myb genes

We wished to determine whether the detected motifs are specifically present in Myb genes, but not in non-Myb genes. Therefore, we used motif sequences (cDNA) as query sequences and performed tblastn search of the EST data set from GenBank. As a positive control we used the Myb R2 domain in tblastn search, and we always obtained ESTs identical to the query. No EST hits were returned when we used the intact five extension motifs in tblastn search. This could be due to the short size of motifs, 8-13 residues, which diminished the performance of tblastn. Therefore we appended 10 downstream amino acids to each extension motif, repeated the search, and obtained 21 ESTs with high identity (>85%). When translated into proteins, all the 21 ESTs from extension motif search also contain a Myb domain. Interestingly, we detected more ESTs from E1 than from the other extension motifs (Table 2). Most likely this is due to the presence of E1 in more Myb genes than other motifs (Fig. 1).

Similarly, 12 (of 30) C-terminal motifs detected homologous ESTs in tblastn search. Most of these ESTs did not contain Myb domains, however because the C-terminal motifs are located downstream some distance from the Myb domains, the returned ESTs are very likely not long enough to reach Myb domains (Table 2). However, the alignment of ESTs with the known Myb showed high identity not only in motif region but also in flanking and farther regions. This suggests that such ESTs are in fact from Myb genes. Moreover, both the extension and C-terminal motifs detected Myb genes specifically from within the same subgroup, rather than from other subgroups. These motifs seem to be specific characteristics of each Myb subgroup. In addition, this result suggests that the conserved motifs may contribute to the same function within each subgroup.

Functional classification and prediction

With the completion of *Arabidopsis* and rice genome sequencing, more than two hundred plant Myb genes have been predicted. However, little is known about the function of most plant R2R3-Myb genes. Therefore, it would be very useful to be able to predict their putative

functions. In principle, the conservation of amino acid motifs should reflect some functional constraints in the evolutionary history of these proteins. Interestingly, the Myb domains are highly conserved, whereas the C-terminal regions are usually quite divergent. As shown here, certain subgroups contain conserved motifs present in carboxy-terminal region (Fig. 1). This characteristic could be used to classify and predict Myb gene function. The topology of Myb phylogeny (Fig. 1) confirms that the Myb genes with similar function are located in the same subgroup. For example, two Myb orthologs, snapdragon *PHAN* gene and maize *rs2 gene*, are located in subgroup G30, and both are involved in organ development: *PHAN* has been shown to regulate the development of the proximodistal axis and dorsoventral asymmetry of lateral organs such as leaves, bracts, and petal lobes (Waites et al, 1998), while the *rs2* gene controls the development of maize lateral organ primordia by repressing expression of *knox* (*knotted1*-like homeobox) genes that are required for the normal initiation and development of lateral organs (Tsiantis et al. 1999; Timmermans et al. 1999). In another example, the *Arabidopsis* genes *GL1* and *WER* located in subgroup G08 are both involved in epidermal cell development: *GL1* activates the *GLABRA2* homeobox gene for trichome (hair cell) development in some parts of the leaf and in the stem (Oppenheimer et al. 1991, Noda et al. 1994), while *WER* controls the formation of the root epidermis by regulating expression of the *GLABRA2* gene (Lee and Schiefelbein 1999). In contrast, another snapdragon Myb gene *MIXTA* located in subgroup G03; controls the formation of conical cells on the petal increasing flower color intensity (Noda et al. 1994). Although *GL1, WER,* and *MIXTA* are all involved in development, differentiation, and/or fate determination of epidermal cells, their immediate functions may differ; for example, regulating different target genes, or at different developmental stages. Therefore, they are located in two distinct subgroups.

Similar results are observed for Myb genes involved in the phenylpropanoid biosynthetic pathway (Fig. 1, Subgroups 12, 13, 14): *C1* (Paz-Ares et al. 1986, Paz-Ares et al. 1987), *Pl, TT2 (Nesi et al. 2001), AN2* (Quattrocchio et al. 1999), *p1* (Grotewold et al. 1991), *p2* (Zhang et al. 2000), *FaMyb1* (Aharoni et al. 2001), *PAP1* and *PAP2* (Borevitz et al. 2000). They all encode a transcription factor to activate enzymes for the phenylpropanoid synthesis except that *FaMyb1* transcription factor suppresses anthocyanin and flavonol accumulation (Aharoni et al. 2001). However, they may be expressed in different tissues. For instance, Zm-*p1* is expressed in pericarp and silk, whereas Zm-*p2* is expressed in anther and silk (Zhang et al.

2000). Strikingly, the cluster (Fig. 1, red arrow) of Myb genes involved in flavonoid-derived pigment biosynthesis also includes *Craterostigma plantagineum* Myb gene *cpm10* in response to ABA. Previous research has revealed that *C1* gene is also regulated by ABA during seed development (Kao et al. 1996). Maize genes *C1*, *p1*, and *Arabidopsis* gene *TT2* each control flavonoid biosynthesis in certain tissues of the developing seed (Paz-Ares et al. 1987; Grotewold et al. 1991; *Nesi et al. 2001*). Thus, members in this cluster may play a role in pigmentation and /or response to ABA.

Additionally, the functional conservation among some Myb genes during evolution could be observed in the cell cycle protein *CDC5* (Fig. 1, G31), found in yeast (Burns et al. 1995; Ohi et al. 1994), human (Lei et al. 2000), *Arabidopsis* (Hirayama and Shinozaki 1996), and *Plasmodium falciparum* (GI#: 11595856). The CDC5 protein performs an essential function in cell cycle control at G2/M, and also participates in pre-mRNA splicing (Burns et al. 1999, McDonald et al. 1999).

Most plant Myb genes are thought to encode transcription factors which activate or repress target gene expression either independently or together with cofactors. The conserved motifs described here may identify the recognition and binding sites through which Myb protein interact with cofactors. Therefore, through the conserved motifs, Myb genes share similar functions or participate in the same pathway. Therefore, we could assign a putative function to subgroups based on the 'marker' Myb genes with know function. For example, members in subgroup G08 are predicted to be involved in epidermal cell patterning and differentiation. Mybs from subgroup G14 are predicted to regulate the accumulation of anthocyanins or flavonoid-derived pigments. The detailed functional classification and prediction is in Table 3, which may serve as a reference to predict the functions of newly identified Myb genes.

Identification of regulatory elements in non-coding regions

In addition to the C-terminal motifs detected in the predicted Myb proteins, we wanted to test whether any conserved DNA sequence motifs could be identified among the Myb gene subgroups. We applied motif-searching tools to detect conserved regulatory elements in the promoter plus 5'UTR, and intron regions of the Myb genes. In contrast to the C-terminal coding regions, no conserved DNA sequence motifs were identified in the Myb gene non-

coding regions except in the subgroup containing the maize *p1* and *p2* genes, and orthologs from sorghum and rice (Fig. 1 G14). Within this subgroup, a highly conserved scheme of TATA-box, Transcription Start Site sequences, and 5' UTR CA-box were found (data not shown here). Otherwise, no significant regulatory elements were detected in non-coding regions of other Myb genes. Our results suggest that it will be difficult to directly identify regulatory motifs in non-coding regions using only existing computational techniques. Possibly, the identification of co-regulated genes using microarray analysis will assist in the identification of common regulatory elements.

Conclusion

The expansion of Myb genes in plants makes it one of the largest families of transcription factors known to date. However, the specific roles of Myb genes in regulating plant traits are still unclear. Here, we used overall sequence similarity to cluster Myb genes from *Arabidopsis* and rice into 42 subgroups. The subgroup designations were well supported by sequence similarity, exon-intron structure, and microRNA complementarity. Furthermore, we found that the splicing sites and the phase of introns are conserved in Myb domains within the same subgroup but different between subgroups. The phylogenetic topology of splicing patterns suggested that Myb domains may originally have a compact size without introns, which were inserted and remained during evolution. Computational searches were used to identify conserved C-terminal motifs present in the different subgroups. These motifs appear to be specific characteristics of Myb subgroup. In contrast to the C-terminal motifs specifically present in Myb genes, no conserved regulatory elements were identified in the non-coding regions. Myb genes with the similar function are clustered in the same subgroup. The obtained functional classification table could serve as a reference to predict the functions of newly identified Myb genes.

References

Aharoni A, Ric De Vos CH, Wein M, Sun Z, Greco R, Kroon A, Mol JNM, and O'Connell AP (2001) The strawberry *FaMyb1* transcription factor suppresses anthocyanin and flavonol accumulation in transgenic tabacco. *Plant J.*, **28**: 319-332

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**:3389-3402

Avila J, Nieto C, Canas L, Benito MJ, Paz-Ares J (1993) *Petunia hybrida* genes related to the maize regulatory *C1* gene and to animal myb proto-oncogenes. *Plant J* **3**: 553-562

Athma P, Grotewold E and Peterson T (1992) Insertional mutagenesis of the maize *P* gene by intragenic transposition of Ac. *Genetics* **131**: 199-209

Bailey TL and Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* AAAI Press, Menlo Park, California, pp 28-36

Ballesteros ML, Bolle C, Lois LM, Moore JM, Vielle-Calzada J-P, Grossniklaus U, and Chua N-H (2001) LAF1, a MYB transcription activator for phytochrome A signaling. *Genes & Development*, **15**: 2613-1625

Borevitz JO, Xia YJ, Blount J, Dixon RA, Lamb C (2000) Activation tagging identifies a conserved Myb regulator of phenylpropanoid biosynthesis. *Plant Cell* **12**: 2383-2393

Braun, E.L. and Grotewold, E. (1999) Newly discovered plant c-*myb*-like gene rewrite the evolution of the plant *myb* gene family. *Plant Physiology*, **121:**21-24

Burns CG,Ohi R, Krainer AR, Gould KL (1999) Evidence that Myb-related CDC5 proteins are required for pre-mRNA splicing. *Proc Natl Acad Sci USA* **96**: 13789-13794

Cedroni ML, Cronn RC, Adams KL, Wilkins TA, Wendel JF (2003) Evolution and expression of Myb genes in diploid and polyploid cotton. *Plant Molecular Biology* **51:**313-325

Chopra, S., Gevens, A., Svabek, C., Peterson, T., and Nicholson, R. (2002). Excision of the Candystripe1 transposon from a hyper-mutable *Y1-cs* allele shows that the sorghum *Y1* gene controls the biosynthesis of both 3-deoxyanthocyanidin phytoalexins and phlobaphene pigments. *Physiological and Molecular Plant Pathology,* **60**: 321-330

de Souza SJ (2003) The emergence of a synthetic theory of intron evolution. *Genetica* **118**: 117-121

Eddy SR (1996) Hidden Markov models. *Curr Opin Struct Biol* **6**: 361-365

Fedorova L and Fedorov A (2003) Introns in gene evolution. *Genetica* **118**: 123-131

Gilbert W (1987) The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.* **52**: 901-905

Goff SA, Riche D, Lan T-H et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L ssp *japonica*). *Science* **296**: 92-100

Frampton J, Gibson TJ, Ness SA, Döderlein G and Graf T (1991) Proposed structure for the DNA-binding domain of the Myb oncoprotein based on model building and mutational analysis. *Protein Engineering* **4**: 891-901

Fridell RA, Pret AM, and Searles LL (1990) A retrotansposon 412 insertion within an exon of the *Drosophila melanogaster Vermilion* gene is spliced from the precursor RNA. Genes Devel **4**: 559-566

Grotewold E, Athma P, and Peterson T (1991) Alternatively spliced products of the maize *P* gene encode proteins with homology to the DNA-binding domain of myb-like transcription factors. *Proc Natl Acad Sci USA* **88**: 4587-4591

Gubler F, Kalla R, Roberts JK, Jacobsen JV (1995) Gibberellin-regulated expression of a myb gene in barley aleurone cells: evidence for Myb transactivation of a high-pI alpha-amylase gene promoter. *Plant Cell*, **7**:1879-1891

Hirayama T and Shinozaki K (1996) A cdc5+ homolog of a higher plant, *Arabidopsis thaliana. Proc Natl Acad Sci USA* **93**: 13371-13376

Iturriaga G, Leyns L, Villegas A, Gharaibeh R, Salamini F, Bartels D (1996) A family of novel myb-related genes from the resurrection plant Craterostigma plantagineum are specifically expressed in callus and roots in response to ABA or desiccation. Plant Mol. Biol. **32**: 707-716

Jiang C, Gu J, Chopra S, Gu X, and Peterson T (2003) Ordered Origin of the Typical Two- and Three-Repeat Myb Genes. *Gene* **(accepted)**

Jin H, Cominelli E, Bailey P, Parr A, Mehrtens F, Jones J, Tonelli C, Weisshaar B and Martin C (2000) Transcriptional repression by AtMyb4 controls production of UV-protecting sunscreens in *Arabidopsis*. *EMBO J.* **19**: 6150-6161

Joshi CP, Zhou H, Huang X, and Chiang VL (1997) Context sequences of translation initiation codon in plants. *Plant Molecular Biology*, **35**: 993-1001

Kao C-Y, Cocciolone SM, Vasil IK, and McCarty DR (1996) Localization and interaction of the *cis*-acting elements for abscisic acid, VIVIPAROUS1, and light activation of the C1 gene of maize. *Plant Cell*, **8**: 117-1179

Kirik V, Kölle K, Miséra S and Bäumlein H (1998) Two novel MYB homologues with changed expression in late embryogenesis-defective *Arabidopsis* mutants. *Plant Mol. Biol.* **37**: 819-827

Klempnauer K-H, Gonda TJ and Bishop JM (1982) Nucleotide sequence of the retroviral leukemia gene *v-myb* and its cellular progenitor *c-myb*: the architecture of a transduced oncogene. *Cell* **31**: 453-463

Kranz HD, Denekamp M, Greco R, Lin HL, Leyva A, Meissner RC, Petroni K, Urzainqui A, Bevan M, Martin C, Smeekens S, Tonelli C, Paz-Ares J, Weisshaar B (1998) Towards the functional characterization of the members of the R2R3-MYB gene family from *Arabidopsis thaliana. Plant J* **16**: 262-276

Kumar S, Tamura K, Jakobsen IB and Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. *Bioinformatics* **17**: 1244-1245

Lawrence CJ, Malmberg RL, Muszynski MG, and Dawe, RK (2002) Maximum likelihood methods reveal conservation of function among closely related kinesin families. *J Mol. Evol.* **54**: 42-53

Lee MM, Schiefelbein J (1999) WEREWOLF, a Myb-related protein in *Arabidopsis*, is a position-dependent regulator of epidermal cell patterning. *Cell* **99**: 473-483

Lei XH, Shen X, and Xu XQ (2000) Human Cdc5, a regulator of mitotic entry, can act as a site-specific DNA binding protein. *J Cell Sci* **113**: 4523-4531

Martin C and Paz-Ares J (1997) Myb transcription factors in plants. *Trends Genet* **13**: 67-73

McDonald WH, Ohi T, Smelkova N, Frendewey D, and Gould DL (1999) Myb-related fission yeast cdc5p is a component of a 40S snRNP-comtaining complex and is essential for pre-mRNA splicing. *Mol. And Cell Biol.* **19**: 5352-5362

Meissner RC, Jin H, Cominelli E et al. (1999) Function search in a large transcription factor gene family in *Arabidopsis*: assessing the potential of reverse genetics to identify insetional mutations in R2R3 Myb genes. *Plant Cell* **11**: 1827-1840

Menssen A, Hohmann S, Martin W, Schnable PS, Peterson PA, Saedler H, and Gierl A (1990) The En/Spm transposable elements of *Zea mays* contains splice sites at the termini generating a novel intron from a dSpm element in A2 gene. *EMBO J*, **9**: 3051-3057

Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L (2001) The *Arabidopsis TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* **13**: 2099-2114

Noda K, Glover BJ, Linstead P, and Martin C (1994) Flower color intensity depends on specialized cell shape controlled by a Myb-related transcription factor. *Nature* **369:** 661-664

Ogata K, Morikamura H, Sekikawa A, Inoue T, Kanai H, Sarai A, Ishii S, and Nishimura Y (1992) Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. *Proc Natl Acad Sci USA*, **89:** 6428-6432

Ogata K, Morikawa S, Nakamura H, Sekikawa A, Inoue T, Kanai H, Sarai A, Ishii S and Nishimura Y (1994) Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79**: 639-648

Ohi R, McCollum D, Hirani B, Den Haese GJ, Zhang X, Burke JD, Turner K, and Gould KL (1994) The Schizosaccharomyces pombe cdc5+ gene encodes an essential protein with homology to c-Myb. *EMBO J* **13**: 471-483

Oppenheimer DG, Herman PL, Sivakumaran S, Esch J, Marks MD (1991) A myb gene required for leaf trichome differentiation in *Arabidopsis* is expressed in stipules. *Cell* **67**: 483-493

Pabo CO and Sauer RT (1992) Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* **61**: 1053-1095

Paz-Ares J, Ghosal D, Wienand U, Peterson PA, and Saedler H (1987) The regulatory c1 locus of *Zea mays* encodes a protein with homology to Myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J* **6**: 3553-3558

Paz-Ares J, Wienand U, Peterson PA, and Saedler H (1986) Molecular cloning of the c1 locus of *Zea mays*: a locus regulating the antocyanin pathway. *EMBO J* **5**: 829-834

Penfield S, Meissner RC, Shoue D, Carpita NC, and Bevan MW (2001) MYB61 is required for mucilage deposition and extrusion in the *Arabidopsis* seed coat. *Plant Cell*, **13**:2777-2791

Quaedvlieg N, Dockx J, Keultjes G, Kock P, Wilmering J, Weisbeek P and Smeekens S (1996) Identification of a light-regulated *Myb* gene from an *Arabidopsis* transcription factor gene collection. *Plant Molecular Biology* **32**: 987-993

Quattrocchio F, Wing J, van der Woude K, Souer E, de Vetten N, Mol J, and Koes R (1999) Molecular analysis of the anthocyanin2 gene of petunia and its role in evolution of flower color. *Plant Cell* **11**: 1433-1444

Rabinowicz PD, Braun EL, Wolfe AD, Bowen B, and Grotewold E (1999) Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants. *Genetics* **153**: 427-444

Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, and Bartel DP (2002) Prediction of plant microRNA targets. *Cell* **110**: 513-520

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C-Z, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Milgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G-L (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* **290**: 2105-2110

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, and Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology* **13**: 1512-1517

Romero I, Fuertes A, Benito MJ, Malpical JM, Leyva S, Paz-Ares J (1998) More than 80 R2R3-Myb regulatory genes in the genome of *Arabidopsis thaliana. Plant J* **14**: 273-284

Rosinski JA, and Atchley WR (1998) Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin. *J Mol Evol* **46**: 74-83

Saitou N, and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406-425

Schmitz G, Tillmann E, Carriero F, Fiore C, Cellini F, and There K (2002) The tomato *Blind* gene encodes a Myb transcription factor that controls the formation of lateral meristems. *Proc Natl Acad Sci USA* **99**: 1064-1069

Stracke R, Werber M, and Weisshaar B (2001) The R2R3-Myb gene family in *Arabidopsis thaliana. Curr Opin Plant Biol* **4**: 447-456

Thompson MA, and Ransay RG (1995) Myb: An old oncoprotein with new roles. *BioEssays* **17**: 341-350

Timmermans MCP, Hudson A, Becraft PW, Nelson T (1999) Rough sheath2: a Myb protein that represses knox homebox genes in maize lateral organ primordia. *Science* **284**: 151-153

Toledo-Ortiz G, Huq E, and Quail PH (2003) The *Arabidopsis* Basic/Helix-Loop-Helix transcription factor family. *Plant Cell* **15**: 1749-1770

Tsiantis M, Schneeberger R, Golz JF, Freeling M, Langdale JA (1999) The maize rough sheath2 gene and leaf development programs in monocot and dicot plants. *Science* **284**: 154-156

Vailleau F, Daniel X, Tronchet M, Montillet J-L, Triantaphylidès C, and Roby D (2002) A R2R3-MYB gene, AtMYB30, acts as a positive regulator of the hypersensitive cell death program in plants in response to pathogen attack. *Proc Natl Acad Sci USA* **99**: 10179-10184

Waites R, Selvadurai HRN, Oliver IR and Hudson A (1998) The *PHANTASTICA* gene encodes a MYB transcription factor involved in growth and dorsoventrality of lateral organs in *Antrirrhinum. Cell* **93:** 779-789

Walker JC (1993) Receptor-like protein kinase genes of *Arabidopsis thaliana. Plant J.* **3**: 451-456

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

Yu J, Hu S, Wang J et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L ssp *Indica*). *Science* **296**: 79-91

Zhang P, Chopra S and Peterson T (2000) A segmental gene duplication generated differentially expressed *myb*-homologous genes in maize. *Plant Cell* **12**: 2311-2322

Figure 1. Phylogeny, subgroup designations, and C-terminal motifs in Myb proteins from *Arabidopsis* and rice. **Left**: The phylogenetic tree represents 130 Myb genes from *Arabidopsis*, 85 from rice, and 43 from other plants, which are clustered into 42 subgroups (triangles). The checked-triangle subgroups contain conserved C-terminal motifs. The arrow indicates a large cluster of genes involved in the phenylpropanoid biosynthetic pathway or ABA response. Some 'landmark' Myb proteins are listed in parentheses for functional references. The uncompressed tree with full taxa names are available at (http://pc21955.zool.iastate.edu/Myb/UnCom.doc). Comparison of the subgroup designations used in this study with that in Stracke et al. 2001 can be accessed at (http://pc21955.zool.iastate.edu/Myb/Subgroups.pdf). **Middle**: The four blocks (A-D) indicate the distribution of the splicing patterns in the Myb R2R3 domains; see text for details. **Right**: Motifs were detected using MEME and drawn to scale. The gray boxes indicate the Myb R2 and R3 repeats. The yellow boxes indicate the extension motifs following the R3 repeat. The pink boxes represent the motifs identified in the previous report (Stracke et al. 2001), and the blue boxes are the motifs newly discovered here. The thin lines indicate coding regions lacking a detectable motif, with a polypeptide length indicated by the number above the diagonal slash marks. The scale bar is equivalent to 50 residues.

Figure 2. **A**, Locations of intron 1 and 2 splicing sites in R2R3 domains. **Left**: Six representative Myb R2R3 domain sequences are shown. The extent of the R2 and R3 repeats is indicated by the brackets at bottom. The triangles indicate the positions of the splicing sites, and the numbers above the triangles indicate the phases of introns. Subgroup G18 represents the major splicing pattern, i.e., 77 (of 130) Mybs in *Arabidopsis*, and 45 (of 85) Mybs in rice have this splicing pattern. The shaded W residues indicate the constantly spaced Tryptophan residues. The representative sequences of the six subgroups are, **G18**: At15223612; **G12**: Scaffold479_5; **G17**: At15225687; **G20**: At15225582; **G22**: At15233911; **G23**: At15224693. **Right**: The table lists the number of Myb genes of each splicing pattern and the total number of Myb genes in *Arabidopsis* and rice. *Note*: 22 *Arabidopsis* and 18 rice genes have the typical G18 splicing pattern, except that they lack either intron 1 or intron 2. Additionally, 6 *Arabidopsis* and 12 rice genes have no introns within the R2R3 domain.

Finally, 2 *Arabidopsis* and 4 rice Myb genes have other non-typical splicing patterns. These splicing patterns are not shown here.

B, C, The conserved exon-intron structure of all member genes in subgroups G18 and G23. Boxes and lines indicate exons and introns, respectively. Additional examples are available at (http://pc21955.zool.iastate.edu/Myb/geneStru.pdf).

Figure 3. The similarity scores and dN/dS ratios of motifs plus 10-residue flanking fragments. The curve indicates the similarity scores along the upstream-flanking, motif (peak), and downstream-flanking regions. The dashed line shows the average similarity value for the entire alignment. The vertical axis is the score obtained from scoring matrix BLOSUM62, with scores ranging from –4 to 11. The horizontal axis indicates the position of the alignment. The three values indicate the ratios of nonsynonymous to synonymous substitutions in the upstream-flanking, motif (peak), and downstream-flanking regions, respectively. Diagrams for other motifs are available at (http://pc21955.zool.iastate.edu/Myb/SimiScores-dNdS.pdf).

Table 1. Consensus sequences of C-terminal motifs

Motif	Alias	Consensus Sequences
E1*	24	LxxMGIDPVTH[KR]P
E2*	11	RLDLLD[IL]SS[IL]L
E3		FSHLMAEI
E4		IHTYRRKYTA
E5*	18	QRAGLPLYPpE[IV]
1*	9	Gq[SA]KnAAxLSH[MT]AQWESARLEAEARLARESKL
2		exe[DE]NKNYWNSI[LF]NlV[ND]SSpSdSs
3*	4	cPDLNL[ED]LrISPP
4*	12	SSS[ST][AS]RLLN[KR]VA
5*	15	WV[HL][ED]D[DE]FELS[ST]L[TV][MN]M
6*	1.1	YASSTENI[SA][RK]LLQ[GN]W
7*	1.2	QGsLSL[IF]EKWLFd[DE]Q[SG]
8*	3	WFKHLESELGLEEdDNQQQ
9		IDdSFWsETL
10*	2	DISNsNKDsatSsEDvlAiIDeSFWSeVv
11		KNEKKIEnWEG
12*	2	drNdKgYNhDMEFWFD
13*	6	[QL]K[IN][NG]VxKPRPRSF[ST]VNNxC[NS]H
14		AIW[DG][SG]LWNLD
15*	7	KRRGGRTx[GR]xsxK
16*	19	DQ[ST]gENYWg[MV]DD[IL]W[PS]
17*	20	LWMPRLVERI
18*	16	PxLfFSEWl
19		[LV]fRPVARvGAFsx[CY]Npg
20		S[LV]LGPEFVDYE[DE]h
21*	22.1	Vx[TA]GLYM
22		PGSP[ST]GSD[VR]SD[SL]S[HT][GI]
23		[ED]DPPTSLSLSLPGaD
24*	22.2	GEFM[AT][VA][VM]QEMI[KR][AT]EVRSYMAe[MV][QG]xx[NA]G[GC]G
25*	25	L[EQ][ND]YI[KR]S[IL]xIN
26*	21	[PV]pF[FI]DFLGVG
27		Pixx[GS][KR]Y[DE][HW][IL]LExFAEKLVKERP
28		SPSVTLSL[SA][PS][SA][TA]VA[PA]aP[PA]aP
29		YDa[AN]DdPRkLRPGEIDPNPEaKPARPDPVDMDEDEKEMLSEARARLANTrGKKA KRKAREKQLEeARRLAsLQKRRELKAAGIdgrhrKRK
30		IDYNAEIPFEK[KR][AP]paGFYDTaDEDRp[AN]D
31*	5	DE[DE]WLRChT

Note: Motifs with * were identified as the alias by Stracke et al. (2001). Consensus sequences follow the criteria of Joshi et al. 1997: a single capital letter is given if the relative frequency of a single residue at a certain position is greater than 50% and greater than twice that of the second most frequent residue. When no single residue satisfied these criteria, a pair of residues were assigned as capital letters in brackets if the sum of their relative frequencies exceeded 75%. If neither of these two criteria was fulfilled, a lower case letter was given if the relative frequency of a residue is greater than 40%. Otherwise, x is given.

Table 2. Homologous EST hits with >85% identity

Motif	GenBank GI No. of ESTs
E1	5843677[1], 21377108[2], 18257159[2], 906292[1], 9609685[2], 9363004[2], 8725610[2], 13119940[2], 13117218[1], 9609609[1], 22261008[2], 19878767[2]
E2	4383290[1], 18257119[1], 19271091[1], 14126137[1], 19865055
E3	16310275
E4	1053668, 19873274
E5	19877001
2	24047454[#2], 22934063[#3], 9278726[#4]
4	5840751[4], 8723410[#4]
6	24067636[2], 13351242[2], 13243677[#3], 22011333[#3], 12632086[2], 18099225[3]
9	5845671[#]
16	19798907[#]
21, 22	8723927[#], 8716394[#], 5851413[#], 8724987, 26019636[1], 8726119[#1], 8723277[1], 28617714[1], 27446064[1], 18734995[1], 17021688[1], 16347988[1], 22525500[1], 13480528[1], 22525278[1], 15200035[1], 10232029[#1], 21285929[1], 17518756[1], 10236319[1], 9899127[1], 10709579[1], 8685347[1], 5820271[1]
23	8723927[#], 8715463[#], 5843769[#], 13382621[#], 8716394[#], 19805828[#], 8702130[#], 24009285[#3], 22411885[1], 8666370[#3]
24	19805828[#], 8723927[#], 8715463[#], 5843769[#], 8716394[#], 19824493[#1], 19836808[#1], 8702130[#1], 8703249[#1], 8703516[#4]
27	23999937[2], 24012501[2], 24009684[2], 19876981[2]

Notes: The superscript digits indicate the number of mismatches over C-terminal motifs at nucleotide level. Those incomplete ESTs without reaching Myb domains are indicated by #. Motif 29 and 30 detected >100 ESTs not listed here. Underlined ESTs are identical over length to a Myb gene in the training data. (We assume a few mismatches as sequencing error if any.)

Table 3. Characterization of Myb gene family from *Arabidopsis* and rice

Clade	Characterization	Myb Genes
G03	controls the formation of conical cells on the petal increasing flower color intensity	Am-*Mixta* (Noda et al 1994)
G04	induced by light, may have a role in processes like chloroplast maturation or leaf expansion	At-*M4* (Quaedvlieg et al. 1996)
G05	downregulates the target gene to control production of UV-protecting sunscreens (sinapate esters) in leaves in *Arabidopsis*	At-*Myb4* (Jin et al. 2000)
G06	activates tryptophan gene expression	At-*ATR1* (Walker 1993)
G07	controls the formation of lateral meristems	Le-*Blind* (Schmitz et al. 2002)
G08	controls epidermal cell patterning and differentiation by regulating expression of the *GLABRA2* homeobox gene	At-*WER* (Lee and Schiefelbein 1999), At-*GL1*(Oppenheimer et al 1991)
G09	a positive regulator of hypersensitive cell death in response to pathogen attack	At-*Myb30* (Vaillean et al. 2002)
G11	induced by ABA and IAA	At-*Myb14* (Kranz et al. 1998)
G12	activator of anthocyanin biosynthesis	Zm-*C1* (Paz-Ares et al 1987), Zm-*Pl*
	partially determines proanthocyanidin accumulation in developing seed by inducing ectopic expression of *BAN*	At-*TT2 (Nesi et al. 2001)*
G13	regulator of phenylpropanoid, anthocyanins biosynthesis	At-*PAP1*, At-*PAP2* (Borevitz et al 2000); Ph-*AN2* (Quattrocchio et al 1999)
G14	activates enzymes for the phenylpropanoid synthesis except that *FaMyb1* suppresses anthocyanin and flavonol accumulation	Sb-*y1*(Chopra et al. 2002); Zm-*p1* (Grotewold et al 1991), Zm-*p2* (Zhang et al 2000);Fa-*Myb1* (Aharoni et al. 2001)
G16	expressed in undifferentiated callus tissue up-modulated by ABA	Cp-*cpm10* (Iturriaga et al. 1996)
G17	gibberellin-regulated transcriptional activator of alpha-amylase gene promoter	Hv-*Gam1* (Gubler et al. 1995)
G18	required for mucilage deposition and extrusion in the Arabidopsis seed coat	At-*Myb61* (Penfield et al. 2001)
G19	transcription activator for phytochrome A signaling	At-*LAF1* (Ballesteros et al. 2001)
G21	may serve roles in embryo longevity or during seed germination	At-*R1*, At-*R2* (Kirik et al. 1998)
G30	regulates the development of the proximodistal axis and dorsoventral asymmetry of lateral organs	Am-*PHAN* (Waites et al. 1998)
	controls the development of maize lateral organ primordia by repressing expression of knox genes	Zm-*rs2* (Tsiantis et al 1999; Timmermans et al 1999)
G31	function in cell cycle control at G2/M, and participate in pre-mRNA splicing	At-*CDC5* (Hirayama and Shinozaki 1996); Sp-*CDC5* (Burns et al. 1999, Ohi et al. 1994)

Notes: Only those subgroups are listed which contain Myb genes with experimental proof and published literature. The specific myb genes in each subgroup can be obtained from the uncompressed tree (http://pc21955.zool.iastate.edu/Myb/UnCom.doc). If Myb genes in the same subgroup have the same function or expression profile, only one description is given.

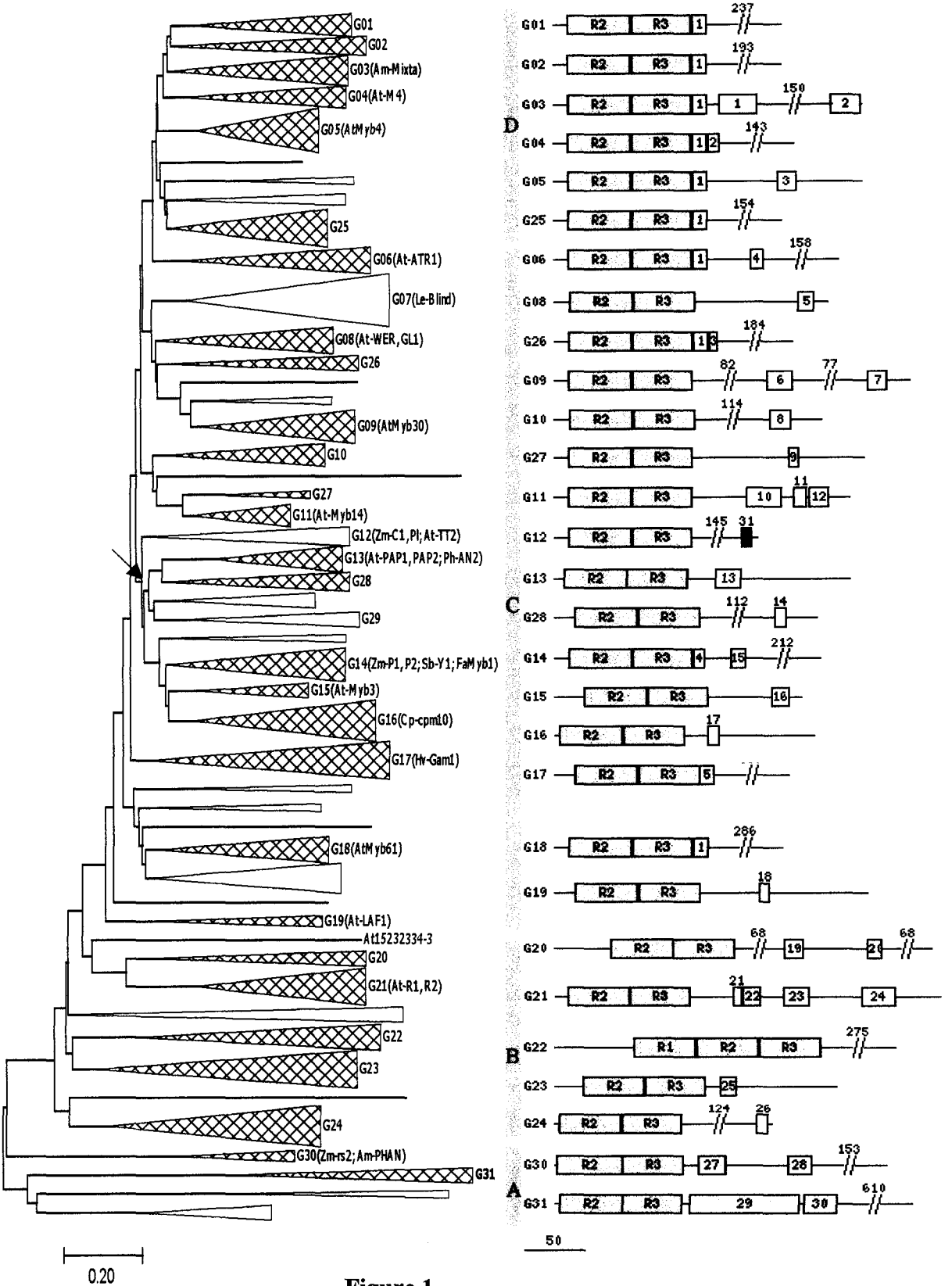
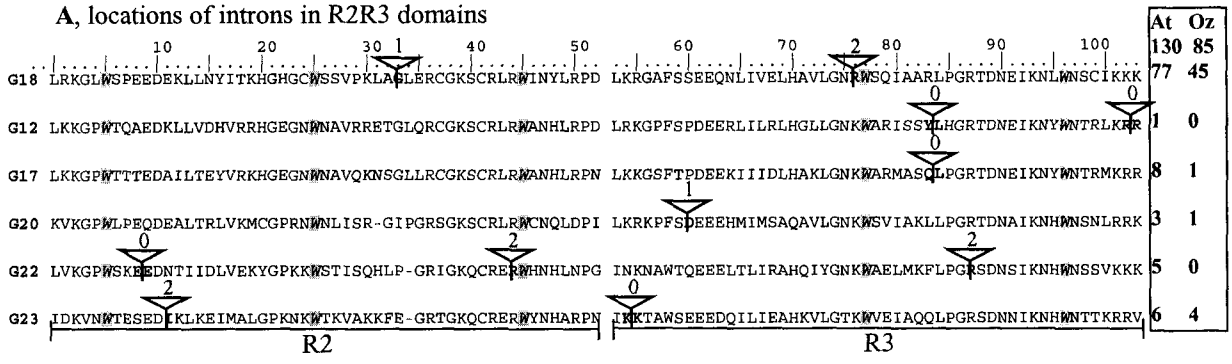
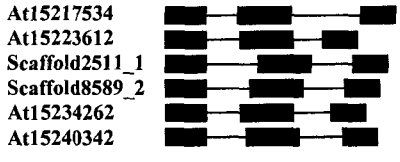
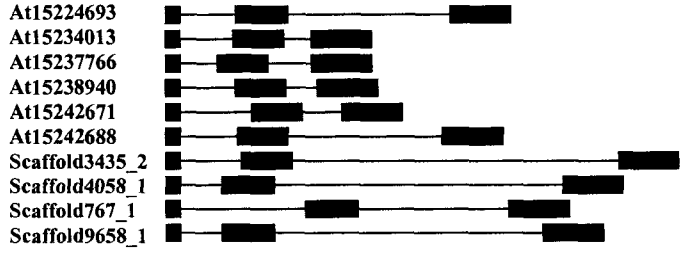

G01
G02
G03(Am-Mixta)
G04(At-M4)
G05(AtMyb4)
G25
G06(At-ATR1)
G07(Le-Blind)
G08(At-WER,GL1)
G26
G09(AtMyb30)
G10
G27
G11(At-Myb14)
G12(Zm-C1,Pl;At-TT2)
G13(At-PAP1,PAP2;Ph-AN2)
G28
G29
G14(Zm-P1,P2;Sb-Y1;FaMyb1)
G15(At-Myb3)
G16(Cp-cpm10)
G17(Hv-Gam1)
G18(AtMyb61)
G19(At-LAF1)
At15232334-3
G20
G21(At-R1,R2)
G22
G23
G24
G30(Zm-rs2;Am-PHAN)
G31
0.20
D
C
B
A
R1
R2
R3
50

Figure 1



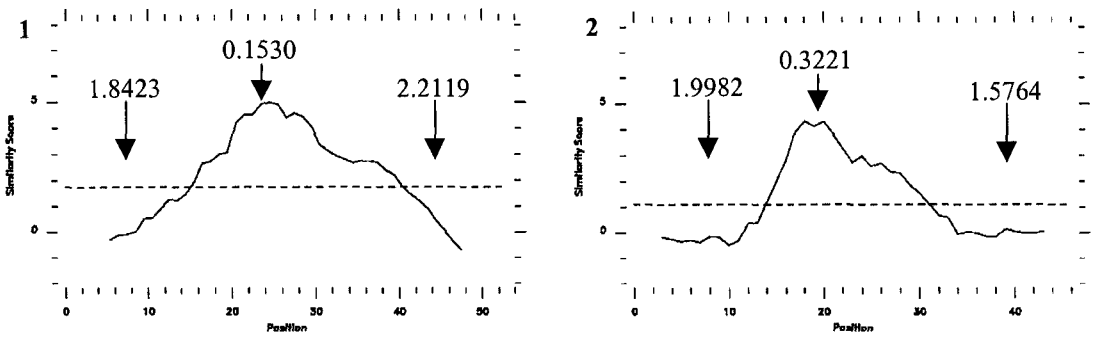


B, exon-intron structure of R2R3 domain in clade G18



C, exon-intron structure of R2R3 domain in clade G23

Figure 2



A, Similarity Score of motif 1

B, Similarity Score of motif 2

Figure 3

CHAPTER 5. IDENTIFICATION AND ISOLATION OF REGULATORY ELEMENTS OF CLOSELY RELATED MYB GENES

Abstract

Non-coding sequences of genes (promoter, 5'UTR, introns and 3'UTR) are thought to contain elements which regulate and control expression. A collection of Myb genes available identified in *Arabidopsis*, rice, and other plants were clustered based on sequence similarity and phylogeny for identifying orthologs or closely-related candidates. We used computational methods to search for conserved regulatory motifs in the genes from each clustered group. No regulatory elements were detected due to the divergence of the non-coding regions. However, there was one exception, *p1/p2* alleles from maize, rice, and sorghum, in which a few conserved regulatory elements were identified including TATA-box, the transcription start site sequences, CA-box, and another unknown motif. Our analysis suggested that motif-searching tools alone have difficulties in identifying regulatory elements in the highly divergent regions. The opportunity could be largely increased in the co-regulated genes identified from microarray analysis.

Introduction

Temporal and spatial regulations of transcription factor expression are critical to plant growth, development, and responses to environmental stress (Meisel and Lam 1997). This sophisticated and accurate biological process is controlled by several kinds of regulatory elements. TATA-like sequence has been identified in the basal promoter regions of plant nuclear genes transcribed by RNA pol II. The TATA-binding proteins (TBPs) bind to TATA-like sequence and interact with other TBP-associated factors (TAFs) which in turn interact with transcriptional activator proteins that bind to short-sequence motifs in the promoter (Weaver 2001).

Previous research has shown that changes in regulatory elements could result in variant expression patterns, and these changes ear important in determining plant traits (Doebley and Lukens 1998). However, the modes of evolution of eukaryotic regulatory sequences are not well understood. It is already clear that promoters evolve under different constraints than

coding sequences. It should be noted that intragenic or intronic sequences may also contain regulatory sequences, as was demonstrated for expression of the *Arabidopsis Agamous* gene (Sieburth and Meyerowitz 1997). However, the non-coding sequences are much more divergent when compared with the coding regions. Thus, it is difficult to detect any significant match of long regulatory elements, and hard to distinguish critical short regulatory elements from fortuitous sequence matches. For example, the motifs that mediate TAF binding are frequently found in promoter regions. One way to increase the chances of success chance is to search for sequence motifs within sets of the closely-related genes, such as orthologs/paralogs, or in coordinately expressed genes. Spellman et al. (1998) identified 800 genes that exhibited cell cycle-specific expression patterns in budding yeast; cell cycle-specific transcription factor binding sites were then found in the majority of the genes' promoters. Similarly, 22 PHO-regulated genes were identified in a whole-genome DNA microarray analysis, and all of their promoter sequences have at least one copy of the Pho4 recognition site (Ogawa et al. 2000).

In addition to studying coordinately-expressed genes, comparison of orthologous Myb genes from different species should aid the identification of conserved elements important for regulation, as well as sequence divergences and replacements that may be responsible for different expression patterns.

Most plant promoters are thought to be relatively compact and generally contained within approximately 2 kb of the 5' transcription start site (Brendel et al. 1998). One exception is the maize *p* gene, which encodes a Myb large and complex regulatory region (Sidorenko et al. 1999). Sidorenko et al. (1999) isolated three regulatory fragments from maize Myb gene *P-rr*, the basal *P-rr* promoter fragment (-235 to + 326, termed Pb), the proximal segment (-1252 to -236, termed P1.0), and the distal segment (-6110 to -4842, termed P1.2). They designed four expression constructs: P0::GUS contains *Adh1-S*, GUS coding region, and *PinII* terminator, but without promoter sequences of *P-rr*; Pb::GUS, P1.0b::GUS, and P1.2b::GUS contain the corresponding *P-rr* regulatory fragments fused to P0::GUS. The transgenic assays indicated that the fragment Pb is required to drive expression of GUS, and both the proximal and distal segments enhance expression levels. Further analyses (Sidorenko et al. 2000) revealed that the P1.0 and Pb fragments have a complex structure including the 3' half of a Tourist element, two large inverted repeats containing two pseudo-

tRNA coding sequences in which A- and B-boxes are embedded, and large palindromes. Several palindromes were also found in the Pb element. It was thought that the interactions between these elements may modulate *P1-rr* expression. Interestingly, the transgenic assays indicated that the 1.2-kb *Sal*I fragment in P1.2b::GUS construct can induce silencing and a paramutant state of *P1-rr*. But, it is still not clear if the same *Sal*I sequences in the endogenous *P1-rr* allele are crucial for subsequent paramutation effects (Sidorenko et al. 2001). Another well-studied case is two closely-related maize Myb genes *p1* and *p2* (Zhang et al. 2000). Sequence comparison showed that *p1* and *p2* have highly similar coding regions, and approximately 400 bp of 5'UTR regions including 90 bp short fragment upstream adjacent to transcription start site. The divergence of promoter regions was thought to cause the distinct tissue-specific expression of the two genes.

Interspecific comparisons are one of the most powerful ways to identify conserved non-coding sequences. The functional importance of conserved regions can be tested experimentally. Sorghum and maize are ideal choices for this purpose. Sorghum is sufficient distant from maize such that non-functional regions of DNA will have accumulated many mutations and be recognizably different, whereas functional sequences will still be identifiable. In our study, we attempted to identify conserved regulatory motifs in non-coding regions with the assistance of computational tools. First, we constructed a phylogenetic tree and identified clades of closely-related Myb genes. Then, motif-finding tools were applied to search for regulatory elements within each clade. In addition, transgenic expression assay in cultured maize BMS (Black Mexican Sweet) cells were conducted to test the functionality of the promoter regions of some sorghum and maize Myb genes.

Materials and Methods

Expression constructs

Based on the Myb gene sequences obtained in a related study (Jiang et al. 2003), we designed two PCR primers to amplify approximately 1-kb segments assumed to contain the promoter sequences. One primer is complementary in the region of the start codon and contains *Nco*I site to enable fusion with a GUS reporter gene. The second primer is complementary to a region approximately 1 kb upstream of the start codon; a *Sal*I or *Xho*I

site was introduced into the upstream primer. The PCR reaction products were purified and ligated with pGEM T-easy vector, and transformed into DH5α *Ecoli*. The plasmid was harvested after incubation in LB medium at 37°C overnight. The amplified Myb gene promoter was excised from the plasmid with the enzyme pair *SalI*/*Nco*I or *Xho*I/*Nco*I, and ligated with the expression vector dp1492 (Pioneer Hi-Bred Inc., a DuPont company), cut with the same enzyme pair, and transformed into DH5α *Ecoli*, and incubated as above. The amplified expression constructs were collected and purified by using a Wizard Mini-Prep kit (Promega). We designed 9 and 6 expression constructs from sorghum and maize, respectively (Table 1).

Microprojectile bombardment

Bombardment conditions were as described previously (Klein et al. 1988, Bowen 1992, Sidorenko et al. 1999). For each bombardment, 10 μl of gold particles (60 mg/ml stock) were combined with 0.3 μg of construct of interest and the same amount of selection plasmid 35S::CRC containing the maize *C1* and *R1* coding regions driven by double cauliflower mosaic virus (CaMV) 35S promoter and eliciting anthocyanin production (Bowen 1992) for normalization and internal positive control. The mixture was held 5 min on ice followed by addition of 50 μl of 2.5 M $CaCl_2$ and 40 μl of 0.1 M spermidine. The mixture was vortexed for 20 sec and held on ice for 10 min for precipitation, and centrifuged for 5-10 sec at 10,000 rpm. The supernatant was discarded, and the pellet washed twice with 200 μl of 100% ethanol. Finally, the pellet was resuspended in 100% ethanol ready for use. Bombardments were performed with a PDS 1000 Biolistic particle gun (DuPont) using a rupture disk with pressure of 650 PSI. Bombardments for each construct were performed in triplicate.

Transient assay

Suspension maize BMS cells were obtained from Pioneer Hi-Bred Inc., a DuPont company, Johnston, Iowa. Cells were incubated in MS medium in darkness at 28°C for 3-4 days, then harvested and treated with 0.25 M mannitol in MS medium (Murashige et al. 1962) for 12 h before bombardment. After bombardment, the cells were kept in MS medium in darkness at 28°C for 2-3 days. For each plate, the numbers of red spots (anthocyanin-producing foci) were counted. Then, the cells were stained with 5-bromo-4-chloro-3-indolyl

glucuronide (x-Gluc) (Jefferson 1987, Schmidt and Willmitzer 1988) in darkness at 28°C for 2-3 days. Finally, the number of blue spots (GUS-positive foci) were counted and the levels of expression of each plasmid were set as the ratio of #blue spots/#red spots.

Regulatory element identification

Myb genes were clustered into subgroups on a basis of sequence similarity and phylogeny (Jiang et al. 2003). Within each subgroup, the motif searching tools MEME and MACAW were applied to detect conserved sequences in the non-coding regions.

Results and Discussion

Components of the expression constructs

Previous research has suggested that most plant promoters are relatively compact and generally contained within approximately 2 kb of the 5' transcription start site (Brendel et al. 1998). Sidorenko et al. (1999) has shown that the basal promoter fragment (-235 to + 326) of the maize Myb gene *p1* is capable to drive the expression of GUS in a transient assay. Therefore, we cloned Myb promoter segments containing at least 900 bp upstream of the start codon into the expression plasmid dp1492. The size range of the Myb promoter regions varies from 0.9 to 1.6 kb. Each of the expression constructs contains a Myb basal promoter, GUS gene and *Pin*II terminator (Fig. 1).

Weak activity of Myb basal promoters in transient assays

Transient assay techniques have proven useful in functional analysis of promoter segments of plant regulatory genes and in the identification of *cis*-regulatory elements responsible for tissue-specific expression of maize genes (Radicella et al. 1992, Sidorenko et al. 1999). Previous results have indicated that the 561 bp basal fragment of maize Myb *p1* containing the transcription start site (-235 to +326) can drive weak expression of fused GUS gene in floral organs such as husk, silk, kernel pericarp, cob and male inflorescence (Sidorenko et al. 2000). Here, we compared the activities of 15 Myb gene promoters in a transient assay system following microprojectile bombardment of maize BMS cells. No background CRC or GUS foci were found in BMS cells in the negative control (Fig. 2, A).

Red spots from CRC plasmid as the internal positive control confirmed the efficiency of the bombardment (Fig. 2, B-D in panel I). The negative control of intact plasmid dp1492 showed no expression of GUS (Fig. 2, B in panel II). In contrast, the CaMV 35S promoter used as the internal positive control gave high expression of GUS as expected (Fig. 2, C in panel II). However, only weak activities of the Myb basal promoter fragments were detected in this transient assay system (Fig. 2). For example, the sorghum Myb9 basal promoter gave only weak GUS expression (Fig. 2, D in panel II). Similar results were obtained from the remaining of constructs of sorghum and maize Myb basal promoter fragments (data not shown). The weak activities of these Myb basal promoter fragments could be due to the absence of the more upstream enhancers as described for maize *p1* basal promoter regions (Sidorenko et al. 1999, 2000). It is also possible that other factors regulating the expression of these Myb genes in a tissue-specific or inducible manner are absent in BMS cells.

Putative regulatory elements identified in maize p1/p2 alleles

A characteristic feature of the Myb gene family is the highly conserved R2R3 Myb domains followed by divergent C-terminal regions. However, a number of conserved motifs have been identified in the C-terminal regions of subgroups of Myb genes in *Arabidopsis* (Stracke et al. 2001). Members of these gene subgroups may often share a similar function or expression pattern, this idea is supported by the available information on Myb genes' functions (Stracke et al. 2001, Lawrence et al. 2002). Sequence comparisons also show a high divergence in the non-coding regions such as promoter, intronic sequences, and 3'UTR sequences. It will be interesting to determine whether if there are any conserved sequences present in these regions may regulate the Myb genes' expression patterns.

To identify conserved regulatory elements, we clustered Myb genes from plants as described in chapter 2. Myb genes within the same subgroup are closely-related in function or expression pattern. However, no conserved sequences were detected among Myb genes in any subgroup, except the *p1/p2* containing subgroup. Within this subgroup, TATA-box, the transcription start site sequences, CA-box, and a downstream motif were found in all *p1/p2* alleles from rice, sorghum, and maize. Although insertions (39-54, and 86-94) and deletion (107-209) occurred in rice, the scheme of these identified elements is conserved as well. The

sequence similarities also imply that sorghum is more closely related to maize than rice (Fig. 3) as expected in the Kellogg lab (Mason-Gamer and Kellogg 1996).

Our results suggest that, it will be difficult to identify regulatory elements of large gene family clustered by sequence similarity and phylogeny. The divergence and complexity of promoter and intronic sequences could diminish the efficiency of motif-searching tools. One way to increase the opportunity of finding regulatory motifs is to search among candidate genes with co-ordinate expression patterns. This approach has been used successfully in yeast (Spellman et al. 1998, Ogawa et al. 2000) and *Arabidopsis* (Schaffer et al. 2001). However, orthologs or paralogs identified by phylogeny may have a greater chance to contain conserved regulatory elements in non-coding regions, as we have shown for the *p1/p2* alleles.

Conclusion

Regulatory elements (TATA-box, transcription start site, CA-box) were identified only in orthologs and paralogs of the maize *p1/p2* genes. Thus, the efficiency of detecting regulatory motifs in non-coding regions of related genes is low when using only computational tools. The chances may be increased by searching within groups of orthologs or paralogs; or by searching within groups of co-regulated genes identified by global expression studies.

References

Bowen, B. (1992) Anthocyanin genes as visual markers in transformed maize tissues. In: Gallagher SR (ed) GUS protocols: Using the GUS gene as a reporter of gene expression, pp. 163-177. Academic Press, San Diego, CA

Brendel, V., Carle-Urioste, J.C., and Walbot, V. (1998) Intron recognition in plants. In: J. Bailey-Serres & D.R. Gallie, Eds. A Look Beyond Transcription: Mechanisms determining mRNA stability and translation in plants, pp. 20-28. Amer. Soc. Plantt Physiol., Rockville, MD

Doebley, J., and Lukens, L. (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell*, **10**: 1075-1082

Jefferson, R.A. (1987) *Plant Mol. Biol. Rep.*, **5**: 387-405

Jiang C, Gu J, Chopra S, Gu X, and Peterson T (2003) Ordered Origin of the Typical Two- and Three-Repeat Myb Genes. *Gene* **(accepted)**

Klein, T.M., Fromm, M., Weissinger, A., Tomes, D., Schaaf, S., Sletten, K., Sanford, J.C.: Transfer of foreing genes into intact maize cells with high-velocity microprojectiles. *Proc. Natl. Acad. Sci. USA*, **85**: 4305-4309

Lawrence CJ, Malmberg RL, Muszynski MG, and Dawe, RK (2002) Maximum likelihood methods reveal conservation of function among closely related kinesin families. *J Mol. Evol.* **54**: 42-53

Mason-Gamer, R.J., and Kellogg, E.A. (1996) Testing for phylogenetic conflict among molecular data sets in the *Triticeae*. *Syst. Biol.* **45**: 522-543

Meisel, L., and Lam, E. (1997) Switching of gene expression: analysis of the factors that spatially and temporally regulate plant gene expression. *Genet. Eng. (NY)* **19**: 183-199

Murashige, T., Scoog, F. (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant*, **15**: 473-497

Ogawa, N., DeRisi, J., and Brown, P.O. (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol. Biol. Cell*, **11**: 4309-4321

Radicella, J.P., Brown, D., Tolar, L.A., and Chandler, V.L. (1992) Allelic diversity of the maize *B* regulatory gene: different leader and promoter sequences of two *B* alleles determine distinct tissue specificities of anthocyanin production. *Genes and Development*, **6**: 2152-2164

Schaffer, R., Landgraf, J., Accerbi, M., Simon, V., Larson, M., and Wisman, E. (2001) Microarray analysis of diurnal and circadian-regulated genes in *Arabidopsis*. *Plant Cell*, **13**: 113-123

Schmidt, R., and Willmitzer, L. (1988) *Plant Cell Rep.*, **7**: 583-586

Sidorenko, L., Li, X., Tagliani, L., Bowen, B., and Peterson, T. (1999) Characterization of the regulatory elements of the maize P-rr gene by transient expression assays. *Plant Mol. Biol.* **39**: 11-19

Sidorenko, L., Li, X., Cocciolone, S.M., Chopra, S. Tagliani, L., Bowen, B., Daniels, M., and Peterson, T. (2000) Complex structure of a maize Myb gene promoter: functional analysis in transgenic plants. *Plant Journal.* **22**: 471-482

Sidorenko, L., and Peterson, T. (2001) Transgene-induced silencing identifies sequences involved in the establishment of paramutation of the maize *p1* gene. *Plant Cell.* **13**: 319-335

Sieburth, L.E., and Meyerowitz, E.M. (1997) Molecular dissection of the AGAMOUS conrol region shows that *cis* element for spatial regulation are located intragenically. *Plant Cell*, **9**: 355-365

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**: 3273-3297

Stracke, R., Werber, M., and Weisshaar, B. (2001) The R2R3-Myb gene family in *Arabidopsis thaliana. Curr Opin Plant Biol*, **4**: 447-456

Weaver, R.F. (2001) Molecular Biology. 2nd edition, McGraw-Hill Science/Engineering/Math, ISBN: 0072345179, pp. 304-312

Zhang, P., Chopra, S. and Peterson, T. (2000) A segmental gene duplication generated differentially expressed *myb*-homologous genes in maize. *Plant Cell*, **12:**2311-2322

Table 1. Expression constructs of Myb gene promoters from sorghum and maize

Name	Accession	Enzymes	Insert (kb)
pdSbMyb4	AF470058	*Sal*I/*Nco*I	1.264
pdSbMyb9	AF470059	*Sal*I/*Nco*I	1.31
pdSbMyb20	AF474127	*Sal*I/*Nco*I	0.875
pdSbMyb23	AF470060	*Sal*I/*Nco*I	1.267
pdSbMyb34	AF474128	*Sal*I/*Nco*I	1.284
pdSbMyb44	AF470061	*Xho*I/*Nco*I	1.184
pdSbMyb53	AF474130	*Sal*I/*Nco*I	1.071
pdSbMyb54	AF474131	*Sal*I/*Nco*I	1.17
pdSbMyb67	AF474133	*Xho*I/*Nco*I	1.2
pdZmMyb1	AF474115	*Xho*I/*Nco*I	1.572
pdZmMyb18	AF474117	*Xho*I/*Nco*I	1.382
pdZmMyb27	AF470081	*Xho*I/*Nco*I	1.307
pdZmMyb30	AF474120	*Xho*I/*Nco*I	1.119
pdZmMyb41	AF470088	*Xho*I/*Nco*I	1.190
pdZmMyb57	AF470092	*Sal*I/*Nco*I	0.996

Notes: All constructs were inserted into vector dp1492, and transformed into *E. coli* DH5α using Ampicillin resistance for selection.

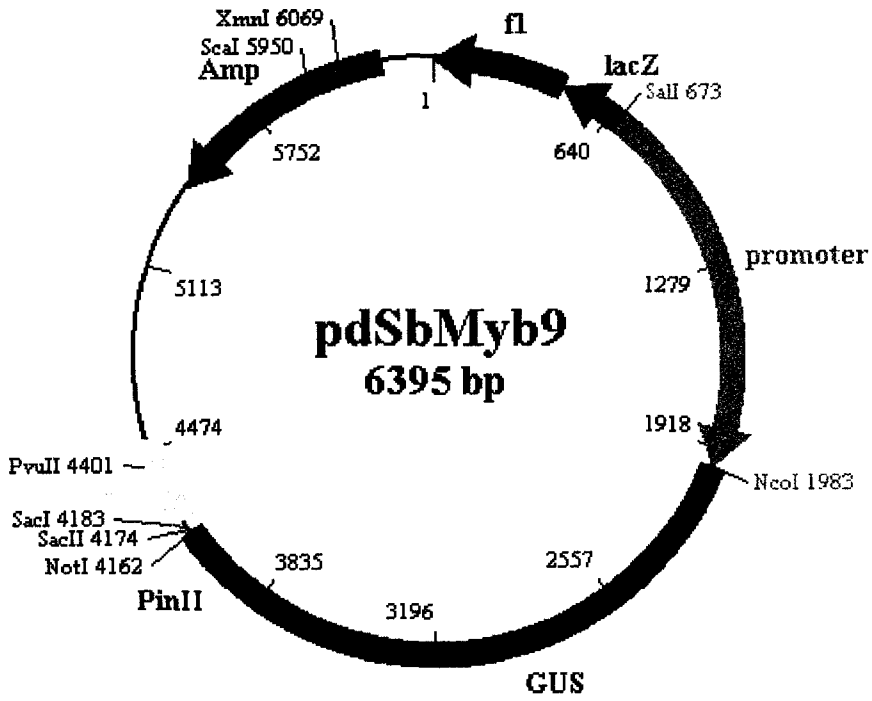


Figure 1. The expression construct of sorghum Myb9 promoter fragment indicates the promoter, GUS gene, and *Pin*II terminator from the plasmid dp1492 backbone. The promoter is 1310 bp long starting upstream of the start codon. The full size of the construct is 6395 bp. The marker gene is Ampicillin.

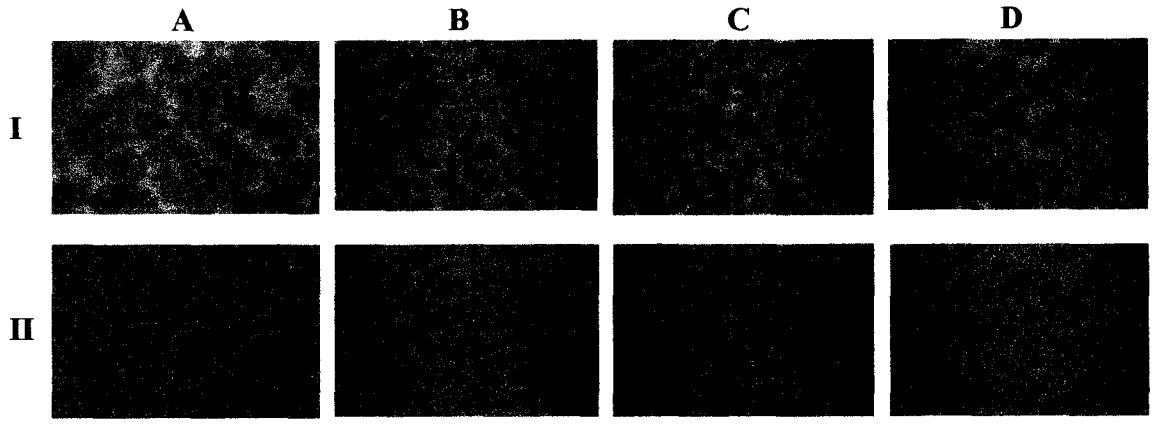


Figure 2. Results of transient assays in maize BMS cells. Panel **I** shows CRC red spots (as normalization control) before staining GUS. Panel **II** shows GUS blue spots. **A**, bombardments without any plasmids (negative control); **B-D**, CRC co-bombardments with plasmid dp1492 (negative control), Ubi::GUS (positive control), and pdSbMyb9 (sample), respectively.

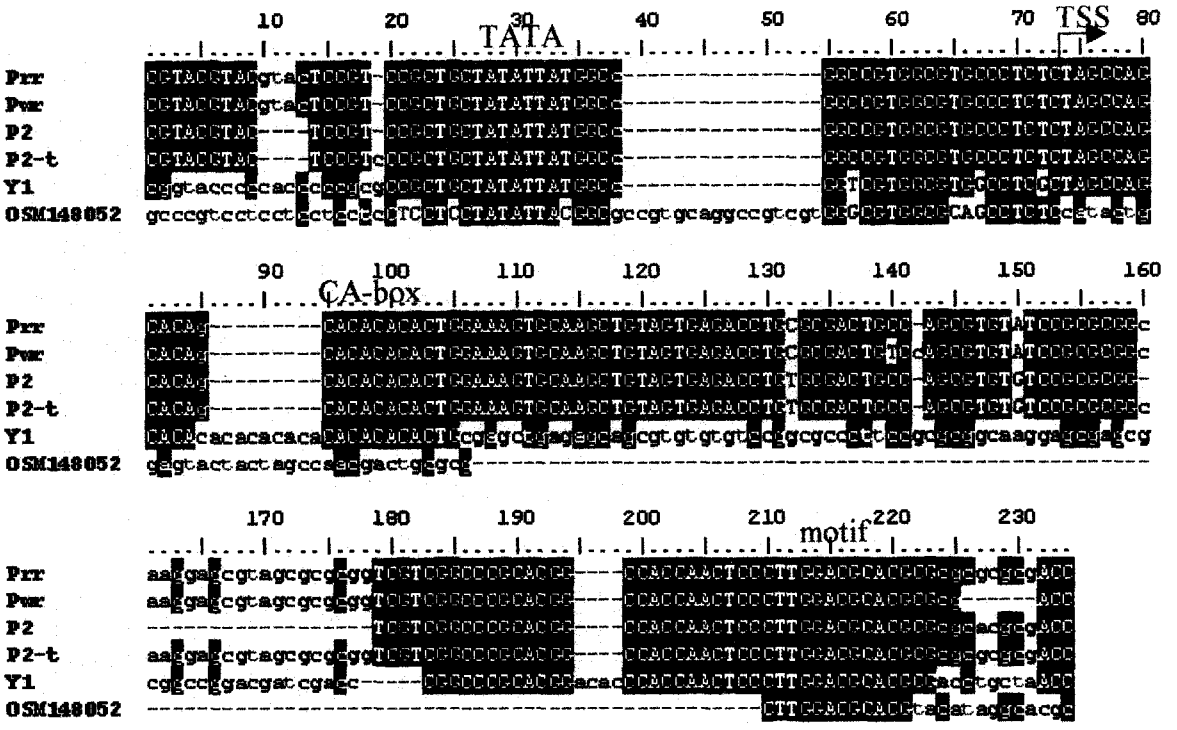


Figure 3. Conserved regulatory elements in the orthologs and paralogs of maize *p1/p2* genes. Four possible regulatory elements were identified in *p1* and *p2* alleles from rice (*OSM148052*), sorghum (*y1*), maize (*Prr, Pwr, p2*), and teosinte (*p2-t*): presumptive TATA-box (27-34), transcription start site (74), CA-box (95-102), and an unknown motif (210-221).

CHAPTER 6. IDENTIFICATION AND CATEGORIZATION OF TRANSCRIPTION FACTORS IN MAJOR PLANTS

Abstract

Using hmmsearch and BLAST search, we identified and classified all genes of known 37 transcription factor families in maize, rice and *Arabidopsis*. We compared and combined the results from the two methods, and resolved the sequences into a non-redundant list. All the biologically useful information about these transcription factors (TFs) was constructed as a shared source (Oracle Database). Three search methods were provided for identifying TFs of interest: 1). Detect all TFs in a given sequence,. 2). Find all sequences containing the given TF in a specified species. 3). List all sequences and all TFs contained in those sequences from a given organism(s). An automated method was devised for the identification and classification of all members of any gene family.

Introduction

Expression of a coding sequence as a functional protein is a long and complex process involving several levels of regulation and control. Among them, transcriptional regulation is a key step which may involve the participating of a variety of transcription factors controlling growth, development, disease-resistance, stress-response, tissue-specificity, and so on. The term 'transcription factors' usually describes a class of proteins that recognize and specifically bind to DNA sites in the promoter regions, and then regulate the frequency of transcription initiation of the target gene. These transcription regulators can be either activators or repressors of transcription (Stracke et al. 2001). Riechmann et al. (2000) reported that different TF families in the major three eukaryotic lineages (plants, fungi, and animals) can control similar biological functions. Conversely, similar DNA binding domains present in these three kingdoms often control different functions in each one. Therefore, the difference of content, sequence, and structure in TF families may contribute to functional diversity. In addition, changes in transcriptional regulators were thought to be a major contributor to functional diversity and evolutionary changes (Carroll 2000).

With the completion of several major eukaryotic genome sequencing projects, the entire complement of genes coding for transcription factors in these genomes can be identified and

described. The results indicate that each eukaryotic lineage employs a sizeable fraction of transcription factors to regulate and control gene expression at the transcription level. For example, the genes encoding TFs number 1533 (5.9% of the total number of genes in the genome) in *A. thaliana*, 209 (3.5%) in *S. cerevisiae*, 669 (3.5%) in *C. elegans*, 635 (4.5%) in *D. melanogaster* (Riechmann et al. 2000), and 1306 in rice (Goff et al. 2002). These numbers could still be underestimates because some uncharacterized proteins could be transcriptional regulators (Schauser et al. 1999, Boggon et al. 1999). Most of these TFs haven't been characterized genetically, especially in plants. To date, only approximately 5% of TFs in *Arabidopsis* have been analyzed by mutation (Riechmann et al. 2000, web supplemental). In contrast, ~25% of known TFs in *Drasophila* and *C. elegans* (Ruvkun et al. 1998) have been characterized genetically.

Transcription factors have been classified into families based on their sequence-specific DNA-binding domains (Pabo et al. 1992). Based on the list of transcriptional regulators in a previous report (Riechmann et al. 2000), we used their DNA-binding domains as queries and performed searches over *Arabidopsis*, maize and rice database to identify the complement of these transcription factors encoded in each genome. The resulting gene dataset were resolved into a non-redundant list, and classified by type. The biologically useful information about these transcription factors was collected into a shared source (Oracle Database). Three query methods were provided so that biologists can search for particular transcription factor(s) of interest.

Materials and Methods

Data sets

Transcription factors families comprised two groups: for group 1 [AP2/EREBP, ARID, bHLH, bZIP, C2C2 (Zn)-GATA, C2H2 (Zn), C3H-type 1 (Zn), CAAT C (CBFB_NFYA), GRAS, HB, HMG-box, HSF, LFY (FLO_LFY), MADS, Myb, Myb-related (TEA/ATTS), SBP, TCP, TUB, WRKY (Zn)], we retrieved the profiles of their DNA-binding domain from pFam database (http://pfam.wustl.edu/index.html). The group 2 [ABI3/VP1, Alfin-like, ARF, Aux/IAA, C2C2-CO-like, C2C2-Dol, C2C2-YABBY, C3H-type 2 (Zn), CPP (Zn), E2F/DP, EIL, GARP-ARR-B class, GARP-G2-like, JUMONJI, NAC, Nin-like, TriHelix] did not have

corresponding pFam profiles; in these cases, one representative member of each family was used as query sequence.

The target sequence databases include arabidopsisNR (*Arabidopsis* proteins annotated by TIGR), BGI_rice_transl.fa (proteins translated from rice genome deposited by Beijing Genomics Institute), ncbi_rice_cds.fa (public rice nucleotides in GenBank), and unicorn4.1_6frame.fa (maize proteins owned by Pioneer Hi-Bred Inc.).

Identification of transcription factors

The pFam profiles from group 1 were used to identify all members of each TF family in the above four databases using program hmmsearch. To identify TFs which may have been missed by hmmsearch, the retrieved sequences were in turn used as query sequences for blast searches of the four databases. Both results were compared and combined, then resolved into a non-redundant list. In contrast, only blast search was performed for the group 2 TF families, followed by removal of redundant. Finally, the categorized non-redundant TFs were collected into an Oracle relational database for biological-based queries. Perl scripts were implemented to automate and fulfill the tasks.

Results and Discussion

Categorization of transcription factors in Arabidopsis, maize, and rice

We identified 1217 TFs in *Arabidopsis*, 1173 in rice from Beijing Genomics Institute, 536 in rice from GenBank, and 797 in maize from Pioneer Hi-Bred Inc., a DuPont company. The complement of TFs in *Arabidopsis*, maize, and rice are similar to that reported previously for *Arabidopsis* (Riechmann et al. 2000), and are similar in number and representation of specific family types (Table 1). The results show that Myb and AP2/EREBP are the largest and second largest families in *Arabidopsis* (181 and 144) and rice (161, 141), respectively. The smallest family LFY, has only one copy in *Arabidopsis* and rice; this is consistent with previous results (Riechmann et al. 2000, Goff et al. 2002).

It has been proposed that each of the major eukaryotic lineages (plants, fungi, and animals) has a characteristic set of particular transcription factor families (Meyerowitz 1999). Excepting the lineage-specific families and the families present in all lineages, some

transcription factors may exist in two of the three kingdoms. For example, the XOX/TCF (SRY-related HMG box/T cell factor) group exists in fungi and animals, but is absent in plants (Riechmann et al. 2000, web supplemental). In contrast, there were no cases of transcriptional regulators present in both plants and fungi, but missing in animals; this is consistent with the topology (plants, (fungi, animals)) (Riechmann et al. 2000). However, at least three families (TUBBY-like, CPP-like, and E2F/DP) were found in plants and animals, but are missing in yeast (Boggon et al. 1999, Cvitanich et al. 2000, Dyson 1998). It is unclear whether these families were specifically lost from the yeast genome, or were never included in the fungal lineage. The lineage-specific TF families contain 45% of the total TFs in *Arabidopsis*, 47% in *C. elegans*, and 32% in yeast (Riechmann et al. 2000). There are 20 plant-specific TF families identified (Table 1, asterisks). Some of them such as Myb, and AP2/EREBP have been greatly amplified. In addition, the function of most of the transcription factors is unclear. However, sequence comparison with *Arabidopsis* TFs can identify the corresponding rice orthologs with considerable similarity.

Relational database of transcription factors in Arabidopsis, maize, and rice

The inclusion of redundant information in databases can lead to a number of problems; including redundant storage, update anomalies, insertion anomalies, and deletion anomalies (Ramakrishnan and Gehrke 2000). Therefore, the TFs compiled here were resolved into non-redundant sequences, and classified by type. Four entities were created: DOMAIN_TF (each transcription factor stored as a record), DOMAIN_SEQ (each sequence stored as a record), DOMAIN_HMM (each TF identified by hmmsearch in a sequence and stored as a record), and DOMAIN_BLAST (similar to DOMAIN_HMM but identified by blast). Obviously, each domain identified by hmmsearch (DOMAIN_HMM) or blast (DOMAIN_BLAST) uniquely belongs to one of the TF families (DOMAIN_TF), but each TF may have multiple copies in a source sequence. This indicates one-to-many relationship between DOMAIN_TF and DOMAIN_HMM / DOMAIN_BLAST. Similarly, each domain detected in either method is from one source sequence, and each sequence may contain multiple copies of the same TFs, or more than one TFs from different families. Therefore, there is also one-to-many relationship between DOMAIN_SEQ and DOMAIN_HMM / DOMAIN_BLAST. However, entities DOMAIN_TF and DOMAIN_SEQ do not have direct mapping relationship, but are

associated through DOMAIN_HMM or DOMAIN_BLAST. More detailed structure of the database is demonstrated in the schema (Fig. 1).

A user-friendly graphical interface was created for biologists to search for any transcription factors (TFs) of interest (data not shown). Three query methods were provided: 1, given a sequence, find all TFs contained inside; 2 given a TF, find all sequences containing this TF from the specified organism(s); 3, list all sequences and their corresponding TFs contained inside from the specified organism(s). There are other options for each method such as E-Value (the E-value of the domain alignment), Score (the identical score of the domain alignment), and so forth.

In order to provide the most complete detection of TFs, we used both hmmsearch and blast. Comparison of results from group 1 indicates that both methods can identify some transcription factors which are missed by the other method. Hmmsearch is more powerful than blast in some families, whereas blast perfoms better for other families. In addition, the perl scripts written for this projects can provide an automated method to identify and classify all members of a given family.

Conclusion

All the member genes of 37 transcription factor families were identified and classified by type in *Arabidopsis*, maize, and rice. The retrieved sequences were resolved into a nun-redundant list, and collected into an Oracle relational database for biological-based inquery. An automated method was devised for the identification of all the members of a given gene family.

Acknowledgements

I thank Dr. Dave Selinger for kind help and constructive guidance. I also appreciate Dr. Carl R. Simmons for setting up the facilities for this project, Kent Vander Velden for setting up the nodes for computation, Dr. Ningning Han for Oracle database help, and Drs. Hai Zhu, Andy Coats, and David A. Curiel for their technical help.

References

Boggon, T.J., Shan, W.S., Santagata, S., Myers, S.C., Shapiro, L. (1999) Implication of tubby proteins as transcription factors by structure-based functional analysis. *Science.* **286**: 2119-2125

Carroll, S.B. (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, **101**: 577-580

Cvitanich, C., Pallisgaard, N, Nielsen, K.A., Hansen, A.C., Larsen, K., Pihakaski-Maunsbach, K., Marcker, K.A., Jensen, E.O. (2000) CPP1, a DNA-binding protein involved in the expression of a soybean leghemoglobin c3 gene. *Proc Natl Acad Sci USA.* **97**: 8163-8168

Dyson, N. (1998) The regulation of E2F by pRB-family proteins. *Genes Dev.* **12**: 2245-2262

Goff, S.A., Riche, D., Lan, T.-H. et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296:** 92-100

Meyerowitz, E.M. (1999) Plants, animals and the logic of development. *Trends in Genetics*, **15**: M65-M68

Pabo, C.O., Sauer., R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem*, **61**: 1053-1095

Ramakrishnan, R. and Gehrke, J. (2000) Database management systems. 2nd Edition, R.R. Donnelley & Sons Company, pp. 418

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C.-Z., Keddie, J., Adam, L., Pineda, Ol, Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., and Yu, G.-L. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**: 2105-2110

Riechmann, J.L., Heard, J., Martin, G., et al. (2000) Web supplemental: *http://www.sciencemag.org/cgi/content/full/290/5499/2105/DC1* (10-10-2003 version)

Ruvkun, G., Hobert, O. (1998) The taxonomy of developmental control *in Caenorhabditis elegans. Science,* **282**: 2033-2041

Schauser, L, Roussis, A, Stiller, J, Stougaard, J. (1999) A plant regulator controlling development of symbiotic root nodules. *Nature.* **11**: 191-195

Stracke, R., Werber, M., and Weisshaar, B. (2001) The R2R3-Myb gene family in *Arabidopsis thaliana. Curr Opin Plant Biol*, **4**: 447-456

Table 1. Transcription factor classes in Arabidopsis, rice, and maize

Group 1	*Arabidopsis*	Rice (BGI)	Rice (GenBank)	Maize
AP2/EREBP*	144	141	33	87
ARID	4	2	0	3
bHLH	93	74	23	9
bZIP	35	50	14	15
C2C2 (Zn)-GATA	29	28	6	19
C2H2 (Zn)	26	32	54	5
C3H-type 1 (Zn)	21	16	4	6
CAAT C (CBFB_NFYA)	10	10	1	20
GRAS*	32	54	10	34
HB	50	44	17	20
HMG-box	15	12	2	15
HSF	23	23	7	22
LFY (FLO_LFY)*	1	1	2	1
MADS	89	67	56	55
Myb	181	161	51	77
Myb-related (TEA/ATTS)	3	2	1	1
SBP*	16	18	2	17
TCP*	23	19	7	12
TUB	11	13	2	24
WRKY (Zn)*	73	95	168	50
Group 2	***Arabidopsis***	**Rice (BGI)**	**Rice (GenBank)**	**Maize**
ABI3/VP1*	6	5	1	7
Alfin-like*	7	12	6	31
ARF*	24	26	3	43
Aux/IAA*	26	28	5	34
C2C2-CO-like*	24	19	16	20
C2C2-Dof*	36	22	10	13
C2C2-YABBY*	4	6	2	13
C3H-type 2 (Zn)*	6	6	0	8
CPP (Zn)	8	5	1	9
E2F/DP	5	6	2	4
EIL*	6	8	2	4
GARP-ARR-B class*	31	27	4	10
GARP-G2-like*	34	32	2	19
JUMONJI	7	6	0	2
NAC*	87	84	18	57
Nin-like*	17	14	2	25
TriHelix*	9	5	2	6
Total	**1216**	**1173**	**536**	**797**

Notes: The asterisk indicates the plant-specific transcription factor families (Riechmann et al. 2000).

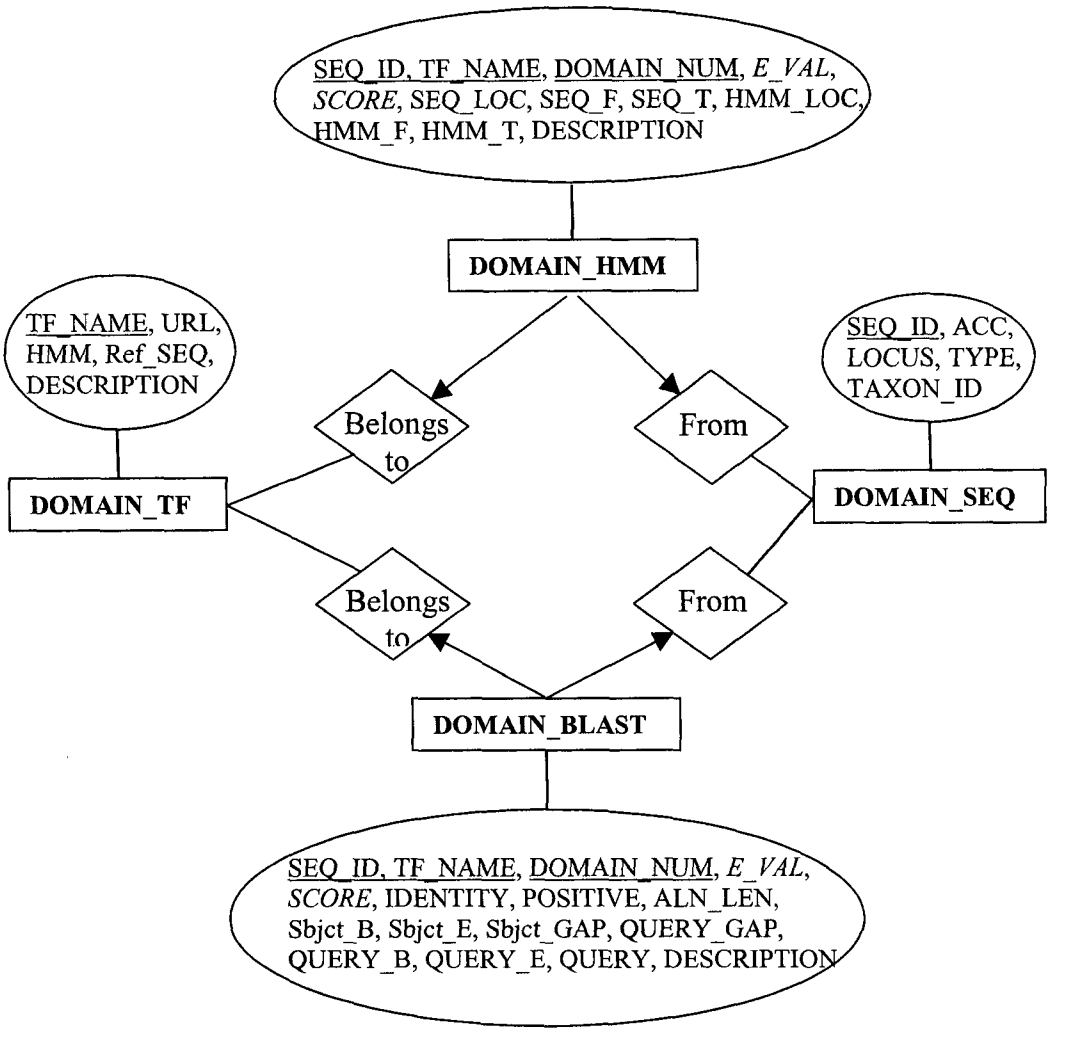


Figure 1. Schema of Oracle database. Rectangle stands for entity, diamond for relationship, and oval for attributes of each entity. The italic attributes (E_VAL and SCORE) are indexed.The primary keys are underlined. It should be noted that two foreign keys (SEQ_ID and TF_NAME) and another attribute (DOMAIN_NUM) together comprise a primary key. There are two types of key constraints (arrows). One states that given a domain identified by hmmsearch 'DOMAIN_HMM', we can uniquely determine the transcription factor to which it belongs. The other one indicates the unique source sequence where the given domain was identified by hmmsearch or BLAST.

CHAPTER 7. GENERAL CONCLUSIONS

Summary

Myb genes encode one of the largest transcription factor families in plants. Myb proteins are defined by a highly conserved DNA-specific binding domain termed Myb, which is composed of approximately 50 amino acids with constantly spaced tryptophan residues. Multiple copies of Myb domains often exist as tandem repeats within a single protein. There are up to four tandem Myb repeats present in Myb proteins identified to date (termed R0R1R2R3 hereafter). Each Myb repeat can form three α-helices, the third of which plays a recognition role in specifically binding to DNA. In contrast to the conservation of Myb domains, C-terminal coding regions and non-coding regions (promoter, 5'UTR, introns, and 3'UTR) are dramatically divergent.

Since the first Myb gene was identified in the avian Myeloblastosis virus (Klempnauer et al. 1982), other members of the Myb gene family have been found existing widely in major lineages. However, in contrast to 125 *Arabidopsis* R2R3 Myb identified after the completion of its genome sequencing (Stracke et al. 2001), there are only a few known Myb genes in sorghum and maize. In order to expand the available Myb gene dataset in these two cereals, we screened Myb-homologous BAC clones from sorghum and maize BAC libraries with a maize *P1-wr* cDNA fragment spanning the Myb R2R3 domains. BAC clones hybridizing to the probe were initially sequenced using a degenerate primer complementary to R2 repeat. Additional sequencing was carried out using a primer walking strategy. Finally, we obtained 64 and 31 unique Myb genes from sorghum and maize BAC libraries (their GenBank accession numbers: AF474125-AF474133, AF470058-AF470071, AY363121-AY363161, and AF474115-AF474124, AF470072-AF470092, respectively). In addition, we identified 51 unique Myb genes in the Monsanto rice database (DraftRiceData) through blast homologous search. Similarly, 85 unique Myb genes were found in the rice genome from the Beijing Genomics Institute. Taken together, these numbers suggest that Myb gene family has undergone significant amplification in the grasses during evolution. It is noteworthy that no R1R2R3 Myb genes were detected in the above cereals, whereas 5 three-repeat Myb genes were identified in *Arabidopsis* (Stracke et al. 2001).

When and how did the expansion of Myb genes occur during evolution? What processes generated the multiple copies of Myb domains arranged as tandem repeats within a single protein? What is the origin of the whole Myb gene family? All these questions are of great interest from an evolutionary point of view.

In this study, we collected the most inclusive set of Myb genes to date and attempted to infer the evolutionary origin of the Myb gene super family. Sequence comparison shows that each of the Myb repeats is more closely related to other members of the same repeat family from other proteins than to other repeats within the same protein. This implies that the duplications which generated these tandem repeats occurred before the divergence of these distantly related species. All plant R2R3 Myb genes were clustered together as a monophyletic group across species. This topology indicates that the R2R3 Myb gene family was amplified before the divergence of monocots and eudicots. Within this monophyly, a small clade includes all P-to-A Myb genes which implies that the substitutions of P-to-A preceded the bulk of the expansion. For R1R2R3 Myb genes, both plant and animal genes form a monophyly. This indicates that the R1R2R3 Myb genes from these two lineages shared a common ancestor. Interestingly, rs-like orthologs form another monophyletic group outside of the above two; this could be due to long-branch attraction. The high bootstrap values for each clade support the reliability of the topology. (chapter 3, figure 3).

The phylogenetic analysis of isolated Myb repeats produced three monophylies for R1, R2 and R3, respectively. R2 repeats from plant R2R3 Myb genes clustered separately from all R1R2R3 Myb genes. The R1 and R3 repeats from R1R2R3 Myb genes can be divided into exclusive plant- and animal-subclades, but not R2 (chapter 3, figure 4). This result may imply that R2 is evolving more slowly, and did not accumulate any changes that unite the animals to the exclusion of the plants. In summary, we here proposed that the Myb gene family evolved via a gain model as follows: beginning with a single Myb repeat, the first intragenic domain duplication produced the ancestral R2R3 Myb. A further intragenic domain duplication formed the ancestral R1R2R3 Myb. The R2R3 and R1R2R3 Myb genes co-existed in primitive eukaryotes, and gave rise to the currently extant Myb genes. Interestingly, R2R3 Myb genes died out in animal lineage, whereas they underwent massive amplification in the plant lineage. In contrast, a previous loss model (Lipsick 1996) proposed that the ancestral R1R2R3 Myb genes arose from the original one Myb repeat by two

intragenic domain duplications. Then R1R2R3 Myb genes existed in primitive eukaryotes, and gave rise to today's R1R2R3 Myb genes in plants and animals. Within the plant lineage, however, R2R3 Myb genes resulted from the loss of R1 from a plant R1R2R3 Myb gene. Since the loss model was proposed, further investigations have shown that some assumptions and observations on which the loss model was based are no longer compelling or true. Moreover, the gain-of-R1 model is more parsimonious and hence favored in our analysis.

Unlike animal Myb genes, the functions of most plant Myb genes are not clear except for a few well-studied cases. Therefore, determination of Myb genes' functions has become a challenging topic in this field. Our analysis has shown that ten conserved important residue sites in R2R3 domains may be correlated with changes in functionality through evolution (chapter 3, figure 2). However, do the dramatically divergent C-terminal regions reflect functional constraints upon the R2R3 domains? Can we classify and predict the functions of Myb genes on the basis of conserved motifs in C-terminal regions? In order to address these questions, we collected and analyzed 258 Myb genes, 130 from *Arabidopsis*, 85 from rice, and 43 from other plants. These genes were clustered into 42 subgroups based on sequence similarity and phylogeny. The size of the subgroups varies from 2 to 14 genes. Interestingly, not only the locations but also the phases of introns in the R2R3 domains are conserved within the same subgroup, but may be different between subgroups. Using computational searches we detected conserved motifs in the C-terminal coding regions. Also by consulting the published literatures we found that Myb genes sharing similar functions are clustered within a subgroup. Both the conserved motifs and the exon-intron structures may reflect functional constraints upon the Myb structure. Further analysis indicates that the identified C-terminal motifs specifically exist in Myb genes, and could serve as an additional identifying characteristic of Myb genes. Taken together, these results from the basis for a functional classification table, which could provide a starting place for characterization of the newly identified Myb genes. Additionally, our analysis suggests a plausible scenario for the evolution of introns in Myb genes (chapter 4, figure 1). We suggest that Myb domains originally contained no introns, and that introns were inserted during evolution. One intron splicing pattern remained unchanged upon the subsequent gene duplications, and gave rise to the currently extant major splicing pattern.

In addition to the conserved motifs found in the coding regions, the non-coding regions (promoter, 5'UTR, introns, and 3'UTR) are expected to contain some elements that regulate Myb gene expressions crucial for plant growth, development, and responses to environmental stress. In our project, we used both transient assays and computational tools to identify the conserved regulatory elements in the closely-related genes. In transient assays, we tested a series of expression constructs containing basal promoter fragments (~ 1 kb sequence upstream to the start codon) from different Myb gene from sorghum and maize. Each of the basal promoter fragments was fused to GUS report gene and bombarded into maize BMS cells. Unfortunately, the expression of GUS driven by the Myb basal promoter fragment was low, and showed little distinguishable difference in quantity between different Myb promoters. This may reflect the weak activity of the basal promoter fragments in the absence of more upstream enhancer sequences. In the computational approach, first, the closely-related Myb genes were identified and clustered on the basis of sequence similarity and phylogeny. Second, motif-searching tools were applied to identify the conserved regulatory elements in each group of closely-related genes. No regulatory motifs were detected among groups of closely-related genes except for those related to the maize *p1/p2* genes. Four presumptive regulatory elements were identified in this group of orthologs/paralogs: TATA-box, the transcription start site, CA-box, and an unknown motif (chapter 5, figure 3). These results suggest that motif-searching tools alone are inadequate to identify regulatory elements in non-coding regions due to the high sequence divergence. The chances for successful identification of regulatory sequences may be increased in groups of orthologs. Finally, performance may be improved by using global expression analysis to identify co-regulated genes.

Future Research

How to efficiently characterize the functions of newly identified genes is always a topic of interest to molecular biologists. Our results suggest that comparison of orthologous genes with considerable similarity may be one of the most powerful ways to classify and predict the functions of genes. The most important step is to identify the authentic candidate orthologs. For example, in subgroup G12, the topology of gene tree is perfectly consistent with that of species tree (chapter 4, figure 1). This subgroup has four Myb genes (AtMyb123, (riceMyb,

(maize *C1*, *Pl*))). Obviously, the species tree is (*Arabidopsis*, (rice, maize)). In such a case, we predict that AtMyb123 likely shares a similar function with maize *C1* and *Pl*. Further homology search found that AtMyb123 actually is *Arabidopsis* gene *TT2* which has been experimentally shown to induce ectopic expression of *BAN*, and thus result in proanthocyanidin accumulation (Nesi et al. 2001). Therefore, it is reasonable to design experiments to test whether the rice Myb gene has a similar function. We also observed that the gene tree was consistent with the species tree in subgroup G30 (chapter 4, figure 1, rs2-like gene), which includes AtMyb91. The maize *rs2* gene controls the development of cells by repressing expression of *knox* (*knotted1*-like homeobox) genes which are required for the normal initiation and development of lateral organs (Timmermans et al. 1999). Therefore, AtMyb91 likely shares a similar function. If a mutant line for AtMyb91 were available from the public insertion-mutagenesis populations, one could observe its phenotype and further detect its effect on expression of *knotted1*-like homeobox genes.

It has been demonstrated in maize *C1*, *Pl* genes that the divergent C-terminal regions contain some conserved residues which may reflect constraints on the functions of these proteins (Grotewold et al. 2000). Based on the identified conserved C-terminal motifs and their consensus sequences, one could introduce mutations at some sites, or make chimeras by exchanging C-terminal segments of related genes, and then test the activities of the modified genes with transient assays. This approach could aid the identification of functionally conserved residues, and understand how they impact interactions with cofactors.

In addition to the conserved motifs in the C-terminal coding regions, there are conserved regulatory elements present in non-coding regions important for accurate regulation and control of gene expression. Global expression approaches may increase the chances to identify regulatory motifs by clustering genes with co-ordinate expression (Spellman et al. 1998, Ogawa et al. 2000). Correlation can be conducted in several ways, such as tissue specificity, growth conditions (drought stress, pathogen resistance), and so forth. For example, a Myb gene that is expressed highly only in root tissue would be a likely candidate regulator of root-specific genes. Once such co-regulated genes are identified, then motif-searching tools can be applied on them to find conserved regulatory elements. Finally, the functions of the candidate regulatory elements can be tested in transient assays or transgenic plants. It should be noted that traditional microarray examines only steady mRNA level in

different tissues or under environmental stresses. It does not distinguish between transcription and stability as the reason for the presence of the mRNA, not does it determine whether the mRNA is being translated. To distinguish among these levels of control, one can probe the microarray with RNA transcribed from isolated nuclei to measure transcription, independent of stability, and polysome-associated mRNA to measure translatable mRNA levels. Due to the high divergence of non-coding regions, it is often difficult to identify conserved motifs at sequence level. Therefore, it may be important to compare the secondary structures, as these may share common structural features.

Intronic sequences may contain regulatory elements, as was demonstrated for expression of the *Arabidopsis Agamous* gene (Sieburth and Meyerowitz 1997). Additionally, in plants, an accurately spliced intron enhances protein expression compared to intronless transcription units (Callis et al. 1987, Schuler 1998). Our results show that although introns are largely divergent over their whole length, conservation can be observed in the regions proximal to exons in some subgroups. Site substitutions may help to identify the nucleotides important for accurate splicing.

Interestingly, a possible transposon-like 743-bp fragment is inserted in the second intron of maize *P1*-rr and *P1*-wr alleles, but is absent in other *p1/p2* orthologs. This fragment is flanked by a direct 10-bp repeat in *p1* alleles, but by a distinct 9-bp repeat in another maize locus. Its GenBank accession number is AF466202 (located 84795..85689, 12-MAR-2002 version). Standard molecular biology approaches may be applied to test whether this represents a novel transposable element.

Taken together, with the assistance of computational tools, we can acquire data and process information in a high-throughput approach. The findings may facilitate the discovery of new problems, and guide us to solve those problems. However, the combination of bioinformatics findings and molecular biological experiments will provide the most in-depth understanding of complex biological problems. Through these complementary approaches, one may interpret bioinformatics data in a biological sense, instead of merely generating lists of observations or facts from bioinformatics.

References

Callis, J., Fromm, M., and Walbot, V. (1987) Introns increase gene expression in cultured maize cells. *Genes Dev.* **1**: 1183-1200

Grotewold, E., Sainz, M.B., Tagliani, L., Hernandez, M., Bowen, B., and Chandler, V.L. (2000) Identification of the residues in the Myb domain of maize C1 that specify the interaction with the bHLH cofactor R. *Proc Natl Acad Sci USA*, **97**: 13579-13584

Klempnauer, K-H, Gonda, T.J. and Bishop, J.M. (1982) Nucleotide sequence of the retroviral leukemia gene *v-myb* and its cellular progenitor *c-myb*: the architecture of a transduced oncogene. *Cell* **31:** 453-463

Lipsick, J.S. (1996) One billion years of Myb. *Oncogene*, **13**:223-235

Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L (2001) The *Arabidopsis TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* **13**: 2099-2114

Ogawa, N., DeRisi, J., and Brown, P.O. (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol. Biol. Cell*, **11**: 4309-4321

Schuler, M.A. (1998) Intron recognition in plants. In J. Bailey-Serres and D.R. Gallie (eds.), A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants, pp. 1-19

Sieburth, L.E., and Meyerowitz, E.M. (1997) Molecular dissection of the AGAMOUS conrol region shows that *cis* element for spatial regulation are located intragenically. *Plant Cell*, **9**: 355-365

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**: 3273-3297

Stracke, R., Werber, M., and Weisshaar, B. (2001) The R2R3-Myb gene family in *Arabidopsis thaliana*. *Curr Opin Plant Biol*, **4**: 447-456

Timmermans MCP, Hudson A, Becraft PW, Nelson T (1999) Rough sheath2: a Myb protein that represses knox homebox genes in maize lateral organ primordia. *Science* **284**: 151-153

ACKNOWLEDGEMENTS

I thank my major advisor Dr. Thomas Peterson for his effective guidance, generous supports, inspiring discussion, and encouragement during my Ph.D. study. I also thank my co-major advisor Dr. Xun Gu for his suggestive and inspiring discussion, and kind help for my bioinformatics research. It is a great pleasure to have the opportunity to conduct my Ph.D. research under their supervision from which I have learned both scientific knowledge and the rigorous scientific approach from which my future research will certainly benefit.

I am also thankful to Dr. Volker Brendel, Dr. Xiaoqiu Huang, Dr. Daniel Voytas, and Dr. Vasant Honavar for serving as my POS committee members and their great suggestions.

I appreciate for the excellent suggestions from Dr. Jianying Gu in understanding the evolution patterns of Myb gene family.

I would like to express my thanks to the following P-Team members: Dr. Surinda Chopra, Dr. Suzy Cocciolone, Dr. Lyuda Sidorenko, Dr. Erica Unger-Wallace, Dr. Yongli Xiao, Dr. Jianbo Zhang, Lakshminarasimhan Krishnaswamy, Zhuying Li, Terry Olson, Diane Sickau, Yibin Wang, Feng Zhang for their advice and help on my research and friendship during my stay in the lab. I have really had a wonderful time in working with them, and am going to miss them all.

Finally, I have the most truly thanks from the bottom of my heart to my wife Wei Li who has stayed alone in China for almost one year when I prepared for my graduation. It is really a hard time for her. I could not finish my research without her understanding, unselfish supports, concerns, encouragement, and endless love. I am also thankful to my parents, sisters and brothers for their endless love, supports and encouragement.