Evolutionary Dynamics of the DNA-Binding Domains in Putative R2R3-MYB Genes Identified from Rice Subspecies indica and japonica Genomes

Li Jia*, Michael T. Clegg, and Tao Jiang

Department of Biological Sciences, Wichita State University, Wichita, Kansas 67260 (L.J.); Departments of Botany and Plant Science (M.T.C.) and Computer Science (T.J.), University of California, Riverside, California 92521

The molecular evolution of the R2R3-MYB gene family is of great interest because it is one of the most important transcription factor gene families in the plant kingdom. Comparative analyses of a gene family may reveal important adaptive changes at the protein level and thereby provide insights that relate structure to function. We have performed a range of comparative and bioinformatics analyses on R2R3-MYB genes identified from the rice (Oryza sativa subsp. indica and indica) and Arabidopsis genome sequences. The study provides an initial framework to investigate how different evolutionary lineages in a gene family evolve new functions. Our results reveal a remarkable excess of non-synonymous substitutions, an indication of adaptive selection on protein structure that occurred during the evolution of both helix1 and helix2 of rice R2R3-MYB DNA-binding domains. These flexible α-helix regions associated with high frequencies of excess non-synonymous substitutions may play critical roles in the characteristic packing of R2R3-MYB DNA-binding domains and thereby modify the protein-DNA interaction process resulting in the recognition of novel DNA-binding sites. Furthermore, a co-evolutionary pattern is found between the second α-helix of the R2 domain and the second α-helix of the R3 domain by examining all the possible α-helix pairings in both the R2 and R3 domains. This points to the functional importance of pairing interactions between related secondary structures.

Recent studies have shown that gene duplication, as one of the most important mechanisms for the emergence of novel gene functions (Ohno, 1967, 1970; Nei, 1969; Zhang et al., 1998), has resulted in substantial genetic redundancy in all sequenced genomes (Lynch and Conery, 2000; Prince and Pickett, 2002). After gene duplication, a relaxed selective constraint on newly duplicated gene copies allows for an exploration of the space of adaptive protein structures and thereby promotes functional diversification of the duplicates through the origin of biochemical innovation (Sankoff, 2001; Wagner, 2001). Genetic redundancy is common in the rice (Oryza sativa) genome and in most other plant genomes such as Arabidopsis (Feng et al., 2002; Goff et al., 2002; Sasaki et al., 2002; Yu et al., 2002). Rice offers important advantages for basic research in monocotyledonous plants because of its diploid nature, transformability, and the existence of genetic and molecular resources (Sasaki and Burr, 2000; Yuan et al., 2001). Rice also represents the only case where two subsp. (japonica and indica) have been sequenced and are publicly available. The genomic information of rice provides a rich resource not only for dissecting the rice genome but also for performing comparative studies between the two subsp. of rice and between rice and other cereal species (Bennetzen, 2002; Yuan et al., 2003). Systematic analyses of the Arabidopsis genome have facilitated detailed studies of genome structure and evolution in the plant kingdom (Blanc et al., 2000; Ball and Cherry, 2001; Castresana, 2001; Sreekumar et al., 2001). It is now possible to study the evolutionary dynamics of rice gene families by using corresponding genes from Arabidopsis genome to identify rice orthologs, thus facilitating performing comparative analyses.

Arabidopsis R2R3-MYB (R2R3-AtMYB) has provided an excellent model for studying molecular mechanisms underlying the acquisition of divergent functions by paralogous genes during evolution (Jia et al., 2003). Selection is expected to have a differential impact on synonymous and non-synonymous changes. Therefore, the comparison of synonymous and non-synonymous substitution rates among genes from different evolutionary lineages provides a powerful means of analyzing protein evolution (Sutton and Wilkinson, 1997; Yang and Nielsen, 1998; Lukens and Doebley, 2001). It is well known that R2R3-MYB genes play important roles in the regulation of secondary metabolism, the control of cell shape, disease resistance, and hormone responses in the plant king-
dom (Urao et al., 1993; Gubler et al., 1995; Martin and Paz-Ares, 1997; Kranz et al., 1998; Romero et al., 1998; Braun and Grotewold, 1999; Jin and Martin, 1999; Stracke et al., 2001). It also is speculated that transcription factors play crucial roles in generating morphological and physiological variations through their control on developmental processes (Doebly and Lukens, 1998; Stern, 1998; Cubas et al., 1999; Arthur, 2002). The above features make the R2R3-MYB gene family an ideal subject for comparative analyses. Because more than 100 R2R3-MYB genes have been identified from the Arabidopsis genome (Kranz et al., 1998; Stracke et al., 2001) and our initial results suggest that a minimum of 80 R2R3-MYB genes exist in rice (either japonica or indica) genomes (data not shown), it is of great interest to ask how patterns of protein change differ in rice as compared with that in Arabidopsis. To address this question, we have identified rice orthologs of R2R3-AtMYB and analyzed non-synonymous and synonymous substitution rate differences in the duplication process to search for evidence of positive selection on different regions of R2R3-MYB DNA-binding domains. A remarkable excess of non-synonymous substitutions was identified in both helix1 and helix2 of rice R2R3-MYB DNA-binding domains when compared with that of helix3. This is similar to the findings in the study of the R2R3-AtMYB family (Jia et al., 2003) and indicates that greater adaptive flexibility is a property of the first two helices. These flexible α-helix regions associated with excess non-synonymous substitution frequencies are thought to play critical roles in the characteristic packing of DNA-binding domains during the protein-DNA interaction process. Moreover, a striking co-evolutionary pattern is found between the second α-helix of the R2 domain and that of the R3 domain from both Arabidopsis and rice R2R3-MYB genes, evidently pointing to the functional importance of the pairing correlations between related secondary structures in preserving the conformation of the specific protein folding pocket. Our results suggest that, like their Arabidopsis counterparts, adaptive protein evolution has occurred in the R2R3-MYB DNA-binding domains during the evolutionary history of rice.

RESULTS

Conserved DNA-Binding Domains of Rice and Arabidopsis R2R3-MYB Genes

R2R3-MYB is one of the most important subfamilies of HLH transcription factor superfamilies. The DNA-binding domains of R2R3-MYB genes consist of two structurally identical repeats: the R2 domain and the R3 domain (Ogata et al., 1994; Lipsick, 1996; Kranz et al., 1998). The protein sequence alignment of the rice R2R3-MYB DNA-binding domains revealed that each repeat consists of three α-helices and the last two α-helices on each repeat with a loop between them form a characteristic structure of the HLH DNA-binding motif (Fig. 1A). It is known that the first and second helices (i.e. helix1 and helix2) do not directly bind to DNA. However, they contribute to the characteristics of the HLH motif folding that are important in the recognition of a specific gene target. On the other hand, the third helix (i.e. helix3) of each repeat makes direct contact with the major groove of its DNA target (Ogata et al., 1994; Williams and Grotewold, 1997). A full-length R2R3-MYB gene contains two DNA-binding domains, an activation domain, and several flexible domains (Fig. 1B). A sequence comparison of DNA-binding domains among the putative R2R3-MYB genes from rice and Arabidopsis suggests that the R2R3-MYB genes have evolved in a very conserved manner, especially in their DNA-binding domains, and this conservation continues after the divergence of dicot and monocot about 200 million years ago (Wolfe et al., 1989).

Heterogeneous Positive Selection Pressure in the Phylogeny of Rice R2R3-MYB DNA-Binding Domains

Rooted phylogenetic trees of R2R3-MYB DNA-binding domains were reconstructed using the maximum likelihood method from fastDNAml for both japonica and indica rice R2R3-MYB genes (Figs. 2 and 3). The maximum likelihood-based free ratio model that assumes selective pressures are different among all the lineages was found to fit significantly better (at 0.01 level) than the maximum likelihood-based one ratio that assumes selective pressures are the same for all lineage models for both japonica and indica. For japonica, the maximum likelihood of the free ratio model is −10,494.99, compared with −11,353.40 for the one ratio model, and twice the log-likelihood difference, 2Δl, is 858.41. For indica, the maximum likelihood of the free ratio model is −4,959.52, compared with −5,160.64 for the one ratio model, and twice the log-likelihood difference is 201.12. The statistical test of the difference was calculated from a χ² distribution with degrees of freedom = 106 for japonica and degrees of freedom = 42 for indica. Because the maximum likelihood-based free ratio model was found to fit significantly better than the maximum likelihood-based one ratio model for both japonica and indica phylogenies at the 0.01 level, the dN:dS (non-synonymous substitution rate: synonymous substitution rate) ratio for each branch of the given tree topologies was estimated using the free ratio model. The positively selected branches A to G were inferred with dN:dS ratios of 4.2, 2.7, 5.6, 40.9, 11.3, 4.1, and 13.5, respectively, for japonica (Fig. 2), and the branches A to D were inferred with dN:dS ratios of 6.3, 2.1, 17.2, and 20.8, respectively, for indica (Fig. 3). The high dN:dS ratios suggest that these branches have experienced excess non-synonymous substitutions, which indicate episodes of accelerated protein evolution. We also confirmed these putative positive selection branches by using a pair-wise com-
parison method based on estimated ancestral sequences and Fisher’s exact test of statistical significance (Zhang et al., 1997, 1998; Jia et al., 2003).

Unequal Positive Selection Pressure along the DNA-Binding Domains

The identification of sites of accelerated amino acid change should give us a clue about the relationship between sites experiencing positive selection episodes and their consequent effects on protein secondary structure. This information can assist experimental biologists by pointing to potentially fruitful avenues for structural research. Because predominant purifying and neutral selection signals can mask positive selection signals, a noise reduction technique (Jia et al., 2003) was used to identify sites that may have experienced excess non-synonymous substitutions. In our analysis (Table I), the subtrees of each selected clade with eight or less taxa were completely sampled. However, because of computational limitations, the larger clades were only partially sampled by ran-

Figure 1. The amino acid sequence alignment of DNA-binding domains of the R2R3-MYB genes from Arabidopsis and rice. A, Amino acid sequence alignment of the DNA-binding domains (the R2 and R3 domains) of 78 R2R3-MYB genes from Arabidopsis and rice using the ClustalW program. Dark shading, Identical residues; light shading, similar residues. Three helices in the R2 or R3 domains are indicated with white boxes. The R2 and R3 domains are marked with black bars under the corresponding residues. B, Structure of R2R3-MYB genes. R2R3-MYB genes contain two DNA-binding domains (hatch box), an activation domain (shadow box), and multiple flexible domains (white box).
domly choosing a certain number of subtrees. A total of 5,000 subtrees were sampled from a set of approximately $2.73 \times 10^{14}$ subtrees defined in a total of eight positively selected amino acid sites by using the subtree sampling technique. The branch lengths are scaled according to the number of substitutions per nucleotide. AF189786 was chosen as the outgroup. The short bar at the bottom of the figure represents the branch length corresponding to 0.1 (10%) substitutions per site.

A total of 5,000 subtrees were sampled from a set of approximately $2.73 \times 10^{14}$ subtrees defined in a total of eight positively selected clades from the phylogeny of japonica R2R3-MYB DNA-binding domains, and 413 of the sampled subtrees were identified as informative subtrees for the inference of positive selection sites. A total of 1,000 subtrees were sampled from a set of 1,049,065 subtrees defined in a total of four positively selected clades from the phylogeny of indica R2R3-MYB DNA-binding domains, and 107 of the sampled subtrees were identified as informative subtrees. The inferred sites with 99% or higher confidence level were recorded. The total number of excess non-synonymous substitutions on recorded sites was counted and plotted out for both japonica (Fig. 4A) and indica (Fig. 4B). A total of 789 counts of excess non-synonymous substitutions for japonica and 252 for indica, respectively, was found along the 104 amino acid sites in the R2 and R3 domains (Table II). If these substitutions occurred randomly, an average of 7.6 counts per amino acid site would be expected for indica. However, our permutation test illustrated that helix1 and helix2 in the R2 domain have significantly higher counts per site than helix3 in the R2 domain in both japonica and indica. In the R3 domain, helix1 and helix2 have significantly higher counts per site than helix3 as well. These results strongly suggest that excess non-synonymous substitutions are not randomly distributed in the protein secondary structure domains. The incidence of positive selection events is concentrated on certain regions of the R2 and R3 domains. In particular, both helix1 and helix2 in the R2 and R3 domains have experienced more intense positive selection pressure.

Co-Evolution of Certain Protein Secondary Structures in R2R3-MYB DNA-Binding Domains

Co-evolution of certain protein secondary structures (e.g. $\alpha$-helices) may exist because of interactions

Table I. The nos. of relevant subtree topologies (combinations), subtrees sampled, and informative subtrees in the application of the subtree sampling technique on the DNA-binding domains of the R2R3-MYB genes from japonica and indica rice

<table>
<thead>
<tr>
<th>Subspecies</th>
<th>Combinations</th>
<th>Sampled</th>
<th>Informative</th>
</tr>
</thead>
<tbody>
<tr>
<td>japonica</td>
<td>$2.73 \times 10^{14}$</td>
<td>5,000</td>
<td>413</td>
</tr>
<tr>
<td>indica</td>
<td>1,049,065</td>
<td>1,000</td>
<td>107</td>
</tr>
</tbody>
</table>
(e.g., covalent linkage and hydrogen bonding) among these structures within a single protein. The relationship between any two subdomains (or structures) that interact with each other in such a single molecule is a one-to-one relationship. Recall that R2R3-MYB genes function as monomers when binding to their DNA targets, their DNA-binding domains consist of two structurally identical repeats, and each repeat consists of three $\alpha$-helices. Helix1 and helix2 contribute to the characteristics of HLH motif folding, and helix3 of each repeat makes direct contact with the major groove of the target DNA. The function of R2R3-MYB genes depends on an active DNA-binding pocket formed between the R2 and R3 domains. Therefore, any change in one of the two domains that perturbs the activity of the protein must be selected against, or subsequently compensated for, by a correlated change in the other domain. For these reasons, it is very important to identify secondary structures that determine the co-evolution between R2 and R3 domains. The study of these subtle patterns of co-evolution should shed light on the question of how these co-evolutionary pairings were preserved. To study this problem, a correlation analysis was performed to measure quantitatively the strength of co-evolution between $\alpha$-helices in R2R3-MYB DNA-binding domains. The correlation analysis revealed a significant (at the 0.01 level) positive linear regression relationship between the R2 and R3 domains from japonica (and indica) R2R3-MYB genes with the correlation coefficient ($r$) being 0.84 (and 0.78, respectively) and the coefficient of determination ($r^2$) being 0.71 (and 0.61, respectively). As a comparison, the two coefficients for Arabidopsis were found to be 0.79 and 0.62. A further correlation analysis at the secondary structure level of the DNA-binding domains revealed another positive linear regression relationship between helix2 of the R2 domain and helix2 of the R3 domain in both rice and Arabidopsis with a significance at the 0.01 level (Table III).

**DISCUSSION**

Gene duplication and subsequent functional divergence of duplicated genes are major mechanisms for the evolution of novel gene function (Ohno, 1967, 1970; Nei, 1969; Li and Gojobori, 1983; Long and Langley, 1993; Ohta, 1994; Walsh, 1995; Zhang et al., 1998; Lynch and Conery, 2000; Sankoff, 2001; Prince and Pickett, 2002). The functional novelty acquired by duplicate genes may reside in the acquisition of novel catalytic or structural properties or in shifts in developmental expression, thereby facilitating a finer modulation of the developmental program. Furthermore, it has been suggested that evolutionary changes in regulatory genes should be the predominant molecular mechanism governing both physiological and morphological evolution. The R2R3-MYB gene family presents an excellent model for exploring the utility of bioinformatics analyses to reveal the potential structural information latent in genome sequence data. Our investigation of R2R3-MYB DNA-binding domains shows that: (a) accelerated rates of protein evolution exist in both helix1 and helix2 of the R2 and R3 domains, whereas helix3 is more constrained in protein evolution, which is consistent with its direct DNA-binding function; and (b) a strong correlation between separate but functionally interacting protein secondary structures exists in the R2 and R3 domains.

Because R2R3-MYB genes function as monomers when binding to their DNA targets, and alterations in protein sequences play important roles in generating distinct R2R3-MYB functions (Ogata et al., 1994; Williams and Grotewold, 1997; Ganter et al., 1999; Lee and Schiefelbein, 2001), it makes sense to speculate that divergent functions of R2R3-MYB genes most likely are derived from the evolution of their DNA-binding domains. Experimental evidence shows that helix3 of both the R2 and R3 domains is involved in direct interaction with DNA, and mutations occurring in helix3 often abolish MYB DNA-binding ability and result in a functional loss of the MYB protein (Saikumar et al., 1990; Frampton et al., 1991; Gabrielsen et al., 1991; Guehman et al., 1992; Ogata et al., 1994; Oda et al., 1997; Sasaki et al., 2000). These findings suggest that the more flexible helix1 and helix2 structures can accommodate a higher incidence of amino acid change without deleterious effects. Consistent with these results and the concept of “evolutionary filtering” (Zurawski and Clegg, 1987), we have observed that both helix1 and helix2 of the R2 and R3 domains are subjected to significantly higher positive selection pressures than helix3.

It appears likely that subtle changes in DNA-binding characteristics might arise through minor conformational changes associated with the structural interactions between helix1 and helix2 of the R2 and R3 domains. These in turn might provide structural variations in R2R3-MYB proteins and thereby generate novel characteristics for the recognition of new target gene sites. It still remains unclear how positive selection really operated on R2R3-MYB genes. Our speculation on how DNA-binding domains of R2R3-MYB genes evolved to recognize different gene targets is that when an R2R3-MYB gene was duplicated, purifying selection on the duplicate genes was relaxed, and at this stage, a few neutral or nearly neutral mutations resulted in a conformational modification of one of the two redundant proteins to allow the recognition of new gene targets. The acceleration of non-synonymous substitutions followed this event and further enhanced the specialized gene function. Because the analysis on Arabidopsis R2R3-AtMYB genes showed similar evolutionary patterns (Jia et al., 2003), we believe that...
A. R2  R3

<table>
<thead>
<tr>
<th></th>
<th>Helix1</th>
<th>Helix2</th>
<th>Helix3</th>
<th>Helix1</th>
<th>Helix2</th>
<th>Helix3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid position</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive selection counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. R2  R3

<table>
<thead>
<tr>
<th></th>
<th>Helix1</th>
<th>Helix2</th>
<th>Helix3</th>
<th>Helix1</th>
<th>Helix2</th>
<th>Helix3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid position</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive selection counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Legend appears on facing page.)
positive selection was involved in the evolution of novel specific functions of R2R3-MYB genes in both monocot and dicot lineages. However, further experiments are required to examine the functional changes of DNA-binding domains in the evolutionary process.

The functions of proteins in biological systems are determined by the physical interactions that they have with other molecules. Molecular interactions are of primary importance in metabolic and signaling pathways. It has been shown that proteins and their interaction partners must co-evolve so that any divergent change in one molecule can be complemented by the molecule’s interaction partner. Otherwise, the interaction between the molecules will be lost along with its function (Moyle et al., 1994; Atwell et al., 1997; Pazos et al., 1997; Jespers et al., 1999). However, the co-evolutionary pattern of DNA-binding domains in proteins that acts as transcription factors is not well understood. Recent advances in computational methods allow a preliminary identification of interaction partners based on the comparative analysis of genome sequence data, providing a different approach to this fundamental problem. In this paper, we performed a co-evolution analysis to ask whether sequence comparisons can be used to recognize novel co-evolutionary relationships. Operationally, this analysis involves statistical comparisons between pairs of secondary structures that are suspected of interacting with each other in a single molecule. The analysis reveals a strong correlation in amino acid substitution rates between separate but functionally interacting R2 and R3 domains and between helix2 of the R2 domain and helix2 of the R3 domain. These specific co-evolutionary patterns highlight the functional importance of the protein tertiary structures composed by these domains (or subdomains). In addition to the finding that helix2 of both the R2 and R3 domains experienced excess non-synonymous substitutions, the analysis also showed that the strongest co-evolutionary relationship between the second helices of the R2 and R3 domains when compared with other helices in the two domains. The unique properties of helix2 clearly demonstrate its crucial roles during the evolution of the R2R3-MYB gene family.

Molecular genetic analyses have demonstrated that novel morphological and physiological variations may arise from changes in transcription factors. However, recent studies failed to identify evidence of positive selection in the evolution of several transcription factor families (Doebley and Lukens, 1998; Lukens and Doebley, 2001; Zhang et al., 2001). It has been suggested that positive selection might have occurred during a relatively short period of time compared with the much longer time since gene duplication events. Therefore, more frequent purifying and neutral selection events could mask positive selection episodes because most gene families appear to have evolved in an episodic fashion (Gillespie, 1991; Endo et al., 1996; Messier and Stewart, 1997; Zhang et al., 1998, 2001; Lukens and Doebley, 2001). The noise reduction technique employed here is an effective approach to infer amino acid sites that have experienced excess non-synonymous substitutions on specific lineages with good sensitivity and specificity, and it has been used successfully to identify individual amino acid sites that were positively selected during the evolution of the R2R3-MYB gene family (Jia et al., 2003). This approach, combined

Table II. The statistical analysis of positive selection sites based on protein secondary structures of the R2 and R3 domains

<table>
<thead>
<tr>
<th>Category</th>
<th>No. of Sites</th>
<th>Count indica</th>
<th>Percentage indica</th>
<th>Count japonica</th>
<th>Percentage japonica</th>
<th>Count/Site indica</th>
<th>Count/Site japonica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full R2R3 region</td>
<td>104</td>
<td>252</td>
<td>100%</td>
<td>789</td>
<td>100%</td>
<td>2.4</td>
<td>7.6</td>
</tr>
<tr>
<td>R2 domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helix1</td>
<td>15</td>
<td>62</td>
<td>25%</td>
<td>137</td>
<td>17%</td>
<td>4.1**</td>
<td>9.1**</td>
</tr>
<tr>
<td>Helix2</td>
<td>7</td>
<td>49</td>
<td>19%</td>
<td>128</td>
<td>16%</td>
<td>7.0**</td>
<td>18.3**</td>
</tr>
<tr>
<td>Helix3</td>
<td>10</td>
<td>9</td>
<td>4%</td>
<td>13</td>
<td>2%</td>
<td>0.9</td>
<td>1.3</td>
</tr>
<tr>
<td>R3 domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helix1</td>
<td>14</td>
<td>76</td>
<td>30%</td>
<td>289</td>
<td>37%</td>
<td>5.4**</td>
<td>20.6**</td>
</tr>
<tr>
<td>Helix2</td>
<td>6</td>
<td>23</td>
<td>9%</td>
<td>87</td>
<td>11%</td>
<td>3.8**</td>
<td>14.5**</td>
</tr>
<tr>
<td>Helix3</td>
<td>10</td>
<td>7</td>
<td>3%</td>
<td>11</td>
<td>1%</td>
<td>0.7</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Figure 4. Unequal positive selection pressure along the DNA-binding domain. The three helices of each repeat are indicated using gray hatch boxes. The boundary between the R2 and R3 domains is indicated using a vertical dashed black line. A consensus sequence is given to correlate the amino acid positions with the bar chart. The most conserved residues are underlined as landmarks. A, Positively selected amino acid sites in the R2 and R3 domains of the R2R3-MYB genes from japonica rice. B, Positively selected amino acid sites in the R2 and R3 domains of the R2R3-MYB genes from indica rice.
with other protein secondary structure prediction programs, presents an effective method for studying protein-DNA interaction regions of transcription factors based on their primary amino acid sequences.

It has been proposed that manipulations of the gene regulatory system could pave the way for the improvement of rice and other cereals (Tyagi and Mohanty, 2000). Experimental data showed that R2R3-MYB genes carry out important regulatory functions in rice. For example, it has been indicated that the expression of GAMYB gene in rice subsp. indica is stimulated by GA and plays important roles in rice flowering (Gubler et al., 1997; Gocal et al., 1999). Some rice R2R3-MYB genes are known to be involved in proanthocyanidin accumulation in developing seed, the control of the fungal infection and host cell death, and the regulation of the seed maturation (Suzuki et al., 1997; Lee et al., 2001; Nesi et al., 2000). Unfortunately, the functions of most rice R2R3-MYB genes studied in this paper are unknown. Further functional genomics studies need to be performed to identify the important roles of these genes in rice. The identification of positive selection sites (or regions) in the DNA-binding domains of rice R2R3-MYB genes indicates those regions of protein structure most susceptible to adaptive change; therefore, these analyses should assist the protein engineer in narrowing the huge space of potential protein modifications to a manageable subset.

Materials and Methods

Rice (Oryza sativa) R2R3-MYB Gene Sequences

The predicted rice R2R3-MYB protein-coding sequences were retrieved from japonica and indica rice contig databases maintained at the Beijing Genome Institute (http://bti.genomics.org.cn/rice) and the Torrey Mesa Research Institute (http://www.tmri.org), respectively. Both indica and japonica genome sequences are draft sequences, and their quality is unknown. At least one cDNA sequence (i.e. DNA-binding domains of R2R3-AtMYB) was arbitrarily chosen from each distinguishable lineage (i.e. the lineage with bootstrapping support greater than 90%) in the Arabidopsis R2R3-MYB phylogeny (see Fig. 4 in Jia et al., 2003) to serve as queries (with a total of 11 cDNA sequences) to BLAST against the above two databases separately. They are: AtMYB0 (M79448), AtMYB4 (AF062860), AtMYB31 (X90387), AtMYB49 (AF175991), AtMYB68 (AF062901), AtMYB71 (U62743), AtMYB97 (AF176002), AtMYB102 (X90382), AtMYB109 (AF262734), AtMYB113 (A100837), and AtMYB117 (AF334816). The common contig sequences retrieved from the BLAST searches were collected together to form our initial rice R2R3-MYB gene-containing contig libraries, which were separately prepared for japonica and indica rice genomes using our automated program (Perl scripts are available from our Web site at http://www.cs.ucr.edu/~ji/Downloads/Genomics/PlantPhysiology/Perlscripts). GlimmerR (Salzberg et al., 1999), which was trained to recognize genes in the rice genome, was used to annotate genes along R2R3-MYB gene-containing contigs. The gene models above libraries were finally validated. It has been shown that GlimmerR has a sensitivity of 94% and a specificity of 97% at the nucleotide level to predict gene-coding regions on the test set of 172 complete genes (Yuan et al., 2001).

To further guarantee the reliability of GlimmerR-predicted genes for our molecular evolution analyses, expressed sequence tag (EST) databases maintained by GenBank, EMBL, the DNA Data Bank of Japan, The Institute for Genomic Research, the Beijing Genome Institute, and the Torrey Mesa Research Institute were used to confirm genes that have EST evidence. Only sequences with an expectation value close to 0.0 and identities above 90% were selected in our study. A table of japonica and indica putative R2R3-MYB genes and the EST/cDNA clones used to confirm these genes is presented as supplemental material (Supplemental Table I, available in the online version of this article at http://www.plantphysiol.org). In addition, R2R3-MYB cDNA sequences from KOME (http://cDNA01.dna.affrc.go.jp/cDNA; Kikuchi et al., 2003) and GenBank were also incorporated into our data set (Supplemental Table II). All the annotated codon sequences of their DNA-binding domains are available at http://www.cs.ucr.edu/~ji/Downloads/Genomics/PlantPhysiology/Sequences. For convenience, each putative rice R2R3-MYB gene identified above was named after the name of the contig on which it is contained (except for cDNAs retrieved from KOME, where the gene names all start with AK; GAMYB is the gene’s conventional name; CA767403 is the accession no. of the gene in GenBank). All the contig sequences used in our study are available in GenBank, and their accession numbers are provided in the Supplemental Table I.

Phylogeny Reconstruction

Multiple sequence alignments of DNA-binding domains of putative rice R2R3-MYB genes were obtained using ClustalW (Thompson et al., 1994). Protal2dna (http://bioweb.pasteur.fr/sequanal/interfaces/protal2dna.html) was then used to align R2R3-MYB codon sequences according to their protein sequence alignment. Only the R2 and R3 domains consisting of sequences of 312 nucleotides (i.e. 104 codons) were kept for our phylogenetic analyses, similar to the analysis in Jia et al. (2003). The maximum likelihood method from fastDNAml ( Olsen et al., 1994) and the neighboring-joining method from PHYLIP (Felsenstein, 1995) were used to reconstruct phylogenetic trees based on the alignment of the DNA-binding domains of R2R3-MYB nucleotide sequences. The maximum likelihood method used global rearrangements and a transition/transversion ratio of 1.7. Because both methods gave almost the same tree topologies, the ones reconstructed using maximum likelihood method were kept for the subsequent positive selection studies.

Inference of Excess Non-Synonymous Substitution Sites and Unequal Positive Selection Pressure Test

A maximum likelihood-based free ratio model from PAML ( Yang, 1997), which assumes heterogeneous selective pressures among lineages, was used to estimate the \( dN/dS \) ratio (denoted as \( \omega \), i.e. the ratio between the non-synonymous substitution rate and the synonymous substitution rate, where \( dS \) is the number of non-synonymous substitutions per non-synonymous site, and \( dN \) is the number of synonymous substitutions per synonymous site). We used the same lineage-specific branches that were used to estimate non-synonymous substitutions. The identified branches were also confirmed using estimated ancestral codon sequences based on a pair-wise comparison method (Zhang et al., 1997, 1998; Jia et al., 2003). A discrete model ( PAML; Yang, 1997), where three \( \omega \) classes (\( \omega_0 \), purifying selection; \( \omega_1 \), neutral selection; and \( \omega_2 \), positive selection) with probabilities (\( p_0 \), \( p_1 \), and \( p_2 \)) are estimated from the data using maximum likelihood. These estimates identify excess non-synonymous substitutions at amino acid sites likely to have experienced positive selection pressures for each identified lineage. If the third \( \omega \) value is significantly greater than 1.0, non-synonymous substitution

<table>
<thead>
<tr>
<th>Table III. Correlation coefficients for the given domain pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation Coefficients</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>r (R2, R3)</td>
</tr>
<tr>
<td>r (R2-1, R3-1)</td>
</tr>
<tr>
<td>r (R2-2, R3-2)</td>
</tr>
<tr>
<td>r (R2-3, R3-3)</td>
</tr>
</tbody>
</table>
| r (R2, R3), Correlation coefficient between the R2 and R3 domains; r (R2-1, R3-1), correlation coefficient between helix1 of the R2 domain and helix1 of the R3 domain; r (R2-2, R3-2), correlation coefficient between helix2 of the R2 domain and helix2 of the R3 domain; r (R2-3, R3-3), correlation coefficient between helix3 of the R2 domain and helix3 of the R3 domain. ** Significance at 0.01 level. }


Copyright © 2004 American Society of Plant Biologists. All rights reserved.
is inferred to occur more frequently than synonymous substitution at some codons, and positive selection is implicated. Positively selected amino acid sites were then identified using a Bayesian method implemented in PAML. This method calculates the posterior probability of a codon site that is from any one of three \( \omega \) classes. However, when purifying selection or neutrality are the forces governing the evolution of a phylogeny, the third \( \omega \) ratio should be less than or equal to 1.0 in the discrete model; thus, no positively selected sites can be identified (Jia et al., 2003). To infer potential positive selection sites for a given lineage involving excess non-synonymous substitution, the background noise caused by purifying and neutral selection needs to be minimized. A noise reduction technique (Jia et al., 2003) was performed to infer positively selected sites by sampling all or many random subtrees of the clade corresponding to an identified lineage. In this technique, a sampled subtree of a lineage is informative if its third \( \omega \) ratio is significantly \(( \alpha = 0.05 \) level of significance) greater than 1.0, and putative positive selection sites can be determined under discrete model in the subtree. Once a positively selected branch is identified, a lineage that has this branch as its deepest branch is defined. In our subtree sampling procedure, the original deepest branch of the lineage will be kept the same for all sampled subtrees. Given the taxon in a lineage, the subtree topologies can be studied by using different combinations of taxa particle or the “outgroup” taxon. Each tax combination considered induces a subtree of the original lineage. These induced subtrees were checked one by one to identify informative subtree topologies discrete model. For each informative subtree topology, positively selected amino acid sites can be identified and used to infer unequal positive selection pressure along the R2 and R3 domains. Positive selection pressure on a specific protein region is estimated by counting the total number inferred positive selection sites in the region found in the above sampling process.

**Permutation Test on Unequal Positive Selection Pressure**

An adapted permutation test (Churchill and Doerge, 1994; Good, 1994; Jia et al., 2003) was used to examine whether there is significant difference between positive selection pressures on two given regions of DNA-binding domains. The number of positive selection counts at each codon (amino acid) site is the sum of inferences from all the putative positive selection lineages where the subtrees sampled suggested this site as a positive selection site. This number is considered as a data point at the given site. We use \((X_1, \ldots, X_n)\) to denote a set of data points in a region with \( n \) codon sites and \((Y_1, \ldots, Y_m)\) to denote a set of data points in another region with \( m \) codon sites. These two sets of data points form an observation \( Z = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)\). The permutation test was used to check whether there is a significant difference between these two sets of data points (i.e. positive selection counts). The test was performed by generating 1,000 random samples from the set of all permutations of the observed set \( Z \). In our test, Student's \( t \) values for all the permuted sets of data points were calculated, and the distribution of Student's \( t \) values was plotted out. A two-sided test at 0.05 or 0.01 level of significance was conducted to detect a significant difference.

**Correlation Analysis**

A correlation analysis protocol (Goh et al., 2000), which was used to study the co-evolution of proteins and their interaction partners, was modified and introduced into our correlation analysis on the secondary structures of R2R3-MYB DNA-binding domains. All multiple alignments of protein sequences from the domains (e.g. the R2 domain) or subdomains (e.g. helix1 in the R3 domain) analyzed were generated using ClustalW. To compute a distance measure for protein sequences, the PROTDIST program (Felsenstein, 1995) using maximum likelihood estimates based on the Dayhoff PAM matrix was used. To quantify possible co-evolution among key DNA-binding structures (i.e. the R2 and R3 domains; helix1, helix2, and helix3 in the R2 domain; and helix1, helix2, and helix3 in the R3 domain), a linear regression analysis was employed to measure the correlation between the pair-wise evolutionary distances among all sequences in an alignment as follows.

We define \( X \) as a two-dimensional \( N \times N \) matrix of evolutionary distances, where \( N \) is equal to the number of aligned sequences from the subdomains or domains being analyzed (e.g. all the helix2 subdomains of the R2 domains). For the putative subdomains that have co-evolved with the subdomains given in \( X \) (e.g. the helix2 subdomains of the R3 domains), a similar distance matrix, \( Y \), is constructed. \( X_{ij} \) denotes the pair-wise distance between subdomain \( m_i \) and subdomain \( m_j \) subs, likewise, \( Y_{ij} \) signifies the pair-wise distance between subdomain \( n_i \) and subdomain \( n_j \) (where \( n_i \) and \( m_i \) are from the same DNA-binding domain, and \( n_j \) and \( m_j \) are from another DNA-binding domain). The correlation coefficient \( r \) between the matrix \( X \) and the matrix \( Y \) (Press et al., 1988) is defined as:

\[
r = \frac{\sum_{i=1}^{N-1} \sum_{j=1}^{N} (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})}{\sqrt{\sum_{i=1}^{N-1} \sum_{j=1}^{N} (X_{ij} - \bar{X})^2 \sum_{i=1}^{N-1} \sum_{j=1}^{N} (Y_{ij} - \bar{Y})^2}}
\]

where \( 1 \leq i \leq j \leq N \), and \( \bar{X} \) is the mean of all \( X_i \) values, and where \( \bar{Y} \) is the mean of all \( Y_j \) values. Co-evolution is inferred when \( r \) is significantly different from zero. Otherwise, the co-evolution hypothesis is rejected.

**Estimation of Statistical Significance of Correlation**

The significance of a value \( r \) is assessed by a bootstrap analysis that yields an estimate of the size of \( r \) given the size of our data set and by an estimation of the probability of obtaining the observed value of \( r \) by chance (i.e. the \( p \) value; Goh et al., 2000). In the bootstrap analysis, 1,000 data sets containing \( K \) pair-wise distances are randomly drawn (with replacement) from the \( K \) pair-wise distances in the original set. For every such set, the bootstrap correlation coefficient \( r_{\text{boot}} \) is computed. The bootstrap interval (i.e. the interval of \( r_{\text{boot}} \) accounting for 68% of the obtained values of \( r_{\text{boot}} \)) is obtained from the 16% \((a) \) and 84% \((b) \) percentiles in the histogram of the 1,000 values of \( r_{\text{boot}} \) and the mean value of \( r_{\text{boot}} \) is from the 50% percentile. The bootstrap estimate of the so of the observed correlation is then calculated as:

\[
\sigma_{r} = \frac{b - a}{2}
\]

The \( p \) value quantifies the co-evolution between a domain (or subdomain) and its putative interaction partner, and it is obtained by randomly shuffling the pair-wise distances between domains (or subdomains). Thus, the assignment of correspondences (e.g. helix1 is paired with helix2 in the R2 domain and helix1 is paired with helix2 in the R3 domain) is replaced by random assignments, and the correlation coefficient is computed as described above. From the resulting 1,000 values \( r_{\text{rand}} \), the \( z \) score for the actual observed value \( r \) is calculated as:

\[
z = \frac{r - r_{\text{rand}}}{\sigma_{r_{\text{rand}}}}
\]

where \( \sigma \) is the so of \( r_{\text{rand}} \), and \( r_{\text{rand}} \) is the mean. The \( p \) value is then computed as \( p = \text{erf}(z)/\sqrt{\pi} \), where \( \text{erf} \) is the complement error function.

**ACKNOWLEDGMENTS**

We are grateful to Bin Shuai and Haifeng Li for their helpful discussions and to the anonymous referees for many valuable suggestions. Received May 22, 2003; returned for revision August 1, 2003; accepted November 8, 2003.

**LITERATURE CITED**


Molecular Evolution of the R2R3-MYB DNA-Binding Domains

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86: 6201–6205


