Comparative genomics of chalcone synthase and Myb genes in the grass family

Virginia Oberholzer, Mary L. Durbin, and Michael T. Clegg*

Department of Botany & Plant Sciences
University of California, Riverside, CA 92521

(Received 27 December 1999, accepted 27 December 1999)

Most plant genes occur as members of multigene families where new copies arise through duplication. Duplicate genes that do not confer an adaptive advantage to the plant are expected to rapidly erode into pseudogenes owing to the accumulation of transpositions, insertion/deletion mutations and nucleotide changes. Nonfunctional copies will drift to fixation within a few million years and ultimately erode beyond recognition. Duplicate genes that are retained over longer periods of evolutionary time must be positively selected based on some adaptive advantage conferred on the plant species. We explore the dynamics of the recruitment of new duplicate genes for chalcone synthase, the enzyme that catalyzes the first committed step of flavonoid biosynthesis, and for the myb family of transcriptional activators. Our analyses show that new chs genes are recruited into the genome of grasses at a rate of one new copy every 15 to 25 million years. In contrast, the myb gene family is much older and many duplicate copies appear to predate the separation of the angiosperm lineage from other seed plants. The general pattern suggests a rapid adaptive proliferation of new chs genes but a more ancient elaboration of regulatory gene functions. Our analyses also reveal accelerated rates of protein evolution following gene duplication and evidence is presented for interlocus exchange among duplicate gene loci.

INTRODUCTION

Most plant genomes are characterized by a high level of redundancy where the great majority of genes encoding enzymatic or regulatory functions are represented by multiple gene copies. A number of mechanisms are responsible for the generation of genetic redundancy including genome duplication through polyploidy and various recombinational processes (Clegg, 1999). Whatever their origin, new redundant copies of a gene have just two fates. The first and most likely fate is to be converted into a pseudogene through the relentless accumulation of mutations. The second is to differentiate from the parental form through the accumulation of mutations that cause adaptive shifts in enzymatic function or in developmental expression. Adaptive mutations must arise within a relatively restricted window in time that is determined by the rate of mutation to nonfunctional states and by the effective population size of the species (Walsh, 1995). This time window is of the order of several million years. The longer term persistence of duplicate genes is therefore prima facie evidence for an adaptive shift.

Our goal in this article is to explore the tempo of recruitment of duplicate gene copies for two important plant gene families: the chalcone synthase (chs) gene family that encodes the enzyme responsible for the first committed step of the flavonoid biosynthetic pathway, and the genes that encode the myb family of transcriptional activators responsible for the regulation of the flavonoid biosynthetic and other pathways. Our primary focus is on the grass family (Poaceae) where we will present new data on grass chs and myb gene sequences. A major objective is to ask whether there are lessons for the science of plant genomics that can be gleaned from a detailed analysis of gene family evolution.

The chalcone synthase gene family. Chalcone synthase genes have been the objects of considerable study
within dicot species. Many of the species studied are found to contain several copies of chs. For example, twelve (eight complete and four partial) chs genes have been isolated and sequenced in petunia (Koes et al., 1987; 1989), and six genes have been described in the morning glory (Ipomoea purpurea) genome (Durbin et al., 1995; Fukada-Tanaka et al., 1997; Durbin et al., 1999). However, until recently the chs gene family appeared to consist of only two copies in the very limited number of grass species studied (Oryza, Hordeum, Secale and Zea) (Weinand et al., 1986; Franken et al., 1991; Rhode et al., 1991; Christensen et al., direct submission to Genbank 1996). Though only two gene sequences have been reported, Southern blot analysis of the barley genome suggested that there may be as many as seven gene copies comprising the barley chs gene family (Christensen et al., 1998). A complete characterization of gene family redundancy is rarely the goal of most molecular research and it is probable that additional chs genes are present but undiscovered within the grass genome.

Chalcone synthase (CHS) catalyzes the condensation of the acetate groups from malonyl-CoA to form naringenin-chalcone which is the backbone for flavanol and anthocyanin production (Heller and Hahlbrock, 1980). Flavonoid biosynthesis includes synthesis of aurones, flavones, flavonols, and anthocyanins (van Tunen and Mol, 1990; Forkman, 1993). Plants that lack CHS are unable to produce any of these flavonoids including the anthocyanin pigments that determine floral pigmentation. Not only is CHS involved in pigment formation, but it is also part of the pathways that produce coenzymes for functional proteins that aid in plant defense against UV exposure and pathogen attack (Hahlbrock, 1981; Dangl et al., 1989).

The two maize chs genes characterized to date determine the known phenotypes c2 (Weinand et al., 1986) and whp (Coe, 1981). The nucleotide sequences of these genes established that they are duplicate genes at different loci (Franken et al., 1991). The coding region of whp is 94% similar to c2 whereas their introns, and 5' and 3' untranslated regions differ greatly in size and sequence (Franken et al., 1991). The gene structure of c2 in maize is comprised of two exons of 189 and 1210 base pairs, respectively, that are separated by a 1524 base pair intron.

Although, c2 and whp are now known to have duplicated from the same locus within the last 15 - 17 million years (Gaut and Doebley, 1998), they are independently regulated, with different tissue specificity and expression patterns (Franken et al., 1991). C2 is normally expressed in the developing kernel, and whp in the tassel of the maize plant (Franken et al., 1991). Whp is translationally controlled by the anthocyanin intensifying gene In (Franken et al., 1991). As discussed further below, the c1/pl and the r/b regulatory gene families are also involved in chs gene regulation in maize. The C1/Pl and R/B proteins contain myb and myc domains respectively, which are similar to the proto-oncogene binding domains found in mammalian systems (Paz-Ares et al., 1981). Binding sites for these domains (myb: C/TACG/T; myc: NCANNTGN) have been identified in the promoter regions of the c2 and whp genes (Roth et al., 1991; Franken et al. 1991). The myb binding domain has been found in several chs promoter regions and is the region where binding of myb-like regulatory proteins such as C1 in maize (Franken et al., 1991; Marocco et al., 1989) and An 11 in petunia (Quattrochio et al., 1993) occur.

The promoter regions of c2 and whp also contain additional conserved regions known as Box I (CTACCA) and Box II (CACGTG) domains (Franken et al., 1991). These Box domains appear to be important in light inducibility, and studies of the parsley promoter region have shown a remarkable similarity between these sites and those found in other light inducible genes (Shulze-Lefert et al., 1989). Box I is also called the H-box. It has been found in many chs promoter regions (Faktor et al., 1996). This regulatory site has been shown to be involved in UV and elicitor induction in chs for parsley (Schulze-Lefert et al., 1989) and floral pigmentation products in Antirrhinum and Phaseolus (Faktor et al., 1996). The H box is usually linked closely to the G-box domain, which appears to work in combination to elicit floral pigmentation in Antirrhinum and Phaseolus (Faktor et al., 1996). Box II is also known as the G-box, and its core motif (ACGT) is bound by a number of transcription factors that are associated with promoters other than the CHS promoter (Martin, 1993).

Another known binding domain found in the promoter of many light inducible genes, such as chs genes, is the CHS box (TACC[N-]CT) (Dixon et al., 1990). Many of the chs gene promoter regions have more than one copy of this box, for example, Maize c2 and Parsley chs-15. There are also other chs genes such as the soybean chs 1 and chs 2 that only have one copy of this box (Van der Meer et al., 1993).

Another motif, termed the TACPyAT repeat, is believed to be important in organ-specific expression (van der Meer et al., 1992). These repeats have also been found near the TATA box of other genes of the flavonoid pathway (Martin, 1993).

All chs genes characterized to date appear to have a high degree of similarity at the amino acid level particularly at the N-terminal end of exon 2, which contains the known functional domain at cysteine 169 (Lanz et al., 1991; Trolop et al., 1995). Another feature of all chs genes is that an intron always falls between the second and third base of a conserved cysteine codon at amino acid position 63. All CHS proteins also have an absolutely conserved 12 residue motif (WGVLFGFGLT) termed the chalcone synthase/stilbene synthase signature. This motif resides near the carboxy-terminus and is of unknown function.
Finally, duplicate copies of some chs genes appear to have undergone a shift in enzymatic function to stilbene synthase (STS) in some plant lineages (Tropf et al., 1995). Stilbene synthase, like chalcone synthase, catalyzes the condensation of 4-coumaroyl-CoA with malonyl-CoA, however, it creates stilbene rather than chalcone and initiates the pathway leading to the production of stilbene phytoalexins that are used by the plant in defense against pathogens (Schröder et al., 1993). Other shifts in enzymatic function have also been described (Helariutta, et al., 1996; Schröder et al., 1998; Christensen et al., 1998).

**Myb gene family.** As noted above, two plant gene families that regulate the transcription of genes encoding enzymes in the flavonoid biosynthetic pathway both possess binding domains similar to the proto-oncogenes myc and myb. The first myb gene identified in plants was the regulatory gene colorless1 (c1) in maize (Cone et al., 1986; Paz-Ares et al., 1987). The c1 locus was first identified by East and Hayes (1911), but it was not until 1977 that Chen and Coe (1977) showed by genetic analysis that c1 is a functionally active regulatory protein in the aleurone layer of the maize kernel that transactivates both early and late structural genes in the anthocyanin pathway. In 1986, c1 was independently isolated by two research groups via transposon tagging (Cone et al., 1986; Paz-Ares et al., 1986). Paz-Ares et al. (1987) showed that the c1 gene is comprised of three exons of 150, 129, and 720 nucleotide bp respectively, and two introns of 88 and 145 bp. Amino acid sequence analysis revealed a basic region at the amino-terminus that contained a myb domain.

The myb domain of C1 has similarity to the motif found in cMyb. cMyb is a proto-oncogene in mammals that is known to be involved in transactivation and DNA binding by a helix-turn-helix configuration (Shen-Ong, 1990). Like cMyb, C1 also binds DNA and triggers transactivation (Goff et al., 1991). The carboxyl end of the C1 protein has an acidic domain that is homologous to an amphipathic alpha helix (Paz-Ares et al., 1990). This amphipathic alpha helix domain is also involved in transactivation as shown in maize (Goff et al., 1991), as well as in other systems such as yeast (Struhl, 1987).

Maize c1 (Goff et al., 1991) is part of a gene family which includes pl (purple) (Cone et al., 1993), p (pericarp) (Grotewold et al., 1991), zm1 (Zea mays1) (Franken et al., 1994), and zm38 (Zea mays38) (Franken et al., 1994), all of which are involved in regulation of branches of the phenylpropanoid pathway. Cone et al. (1993) demonstrated that pl has a duplicate function to c1 but differs in its tissue specificity. c1 is active in the aleurone layer of the maize kernel and scutellum, whereas pl is active in vegetative tissues. Comparison of the amino acid sequences revealed 80% similarity between c1 and pl overall. In the myb domain (NH2-terminal end) there is 96% similarity, and in the amphipathic region there is 93% similarity. c1 and pl differ in the nucleotide sequence of their promoter regions and also in the placement and presence of cis-acting regions (Cone et al., 1993). These differences in cis-acting regions are most likely responsible for the differences in observed tissue specificity of these genes. Between the two DNA binding regions (the myb domain and the amphipathic alpha helix), c1 and pl are only 66% similar at the amino acid level (Cone et al., 1993). c1 is proposed to be the younger of the two genes and a duplicate of pl that arose as a consequence of the allotetraploid event which is thought to have doubled the maize genome roughly 15 – 17 million years ago (Gaut and Doebley, 1997). The differentiation in tissue specificity observed at these two duplicate regulatory genes has thus arisen over a relatively short period of evolutionary history.

The more divergent pericarp gene (p) is involved in transactivation of structural genes in the flavonoid pathway leading to pholbaphenes. These are flavonoid pigments, like anthocyanins, that are found in the pericarp and the glumes of the cob of maize (Styles and Ceska, 1977). P binds to the promoter regions of c2 (chalcone synthase) and a1 (dihydroflavonol reductase). P and C1 share 70% similarity at the amino acid level in the myb domain and no similarity in the C-terminal end (Grotewold et al., 1991).

zm1 and zm38, two other myb genes found in maize, were isolated from a maize genomic library using the c1 gene as a probe under relaxed hybridization stringency (Marocco et al. 1989). They have only been partially characterized to date but they also appear to be involved in regulation of the anthocyanin pathway (Franken et al., 1994). In vitro, Zm38 can act as a repressor that inhibits production of anthocyanin. The regulatory action of Zm38 is similar to the c1-f allele (an allele that also acts as an inhibitor). Studies in vitro have shown that Zm1 transactivates a1 in the anthocyanin pathway of maize (Franken et al. 1994).

All known Myb proteins, including those found in Drosophila, mice, chickens, frogs, humans, angiosperms, gymnosperms and yeast share certain common structural features in the Myb domain (Shen-Ong, 1990; Martin and Paz-Ares, 1997). These include two or three common subdomains (named R1, R2, R3 as found in mammalian c-myc), conservatively spaced tryptophan residues and a conserved motif (PGRTDXXKNXW) in the R3 subdomain (Shen-Ong, 1990; Martin and Paz-Ares 1997). The Myb domain in plant myb genes all have certain conserved characteristics, the conserved subdomains, R2 and R3, each with three tryptophan residues that are 18–19 amino acids apart, and the R3 consensus sequence upstream of the last tryptophan residue in the subdomain (Martin and Paz-Ares, 1997; Williams and Grotewold, 1997). Also in plant myb genes, but not in animal myb
genes, the site of the first tryptophan residue in the R3 subdomain is replaced by another phenolic amino acid (isoleucine, leucine or phenylalanine) (Martin and Paz-Ares, 1997).

There are not many copies of myb genes in animal genomes but many copies appear to be present in plant genomes (Rosinski and Atchley, 1998). For example, genomic Southern analysis suggests that as many as 100 myb genes exist in the Arabidopsis genome (Martin and Paz-Ares, 1997) and RT-PCR methods suggest that as many as 80 myb genes exist in the maize genome (Rabinowicz, et al., 1999). Gene duplication events give rise to new copies that may take on functionally different roles, this is believed to have occurred in the plant myb gene family, therefore it provides an example of gene duplicates that have evolved functions that are involved in plant specific processes (Martin and Paz-Ares, 1997).

Some plant myb genes are involved in regulation of pathway systems other than the phenylpropanoid pathway. For example, Arabidopsis GL1, a plant myb gene, regulates genes in trichome development (Oppenheimer, et al., 1991). In snapdragon mixta, another member of the myb gene family, regulates the genes involved in floral shape (Noda et al., 1994). Although they all have the Myb domain in common at the NH2-terminus of the protein, there is no similarity at the C-terminal end (Rosinski and Atchley, 1998). This is not surprising given that C-terminal ends of maize C1 and Pl are only 66% similar outside of the amphipathic alpha region even though these two genes are from a recent duplication event in the maize lineage. Many myb genes have also been sequenced and identified in rice and Arabidopsis through the genome projects that are ongoing for these species, but to date few functional characterization studies have been carried out. The myb genes that have been isolated from plants appear to be polyphyletic and ancient, suggesting that some of the myb gene duplication events occurred before the divergence of the seed plants (Rosinski and Atchley, 1998).

The overall picture obtained from a review of the literature is that both the chs and myb gene families are old and both have expanded in numbers and in function over the course of plant evolutionary history. Our goal in this article is to examine the different tempo of evolution for these two gene families that encode important enzymatic functions and fundamental regulatory functions to explore the interface between the science of molecular evolution and the new science of comparative genomics.

MATERIALS AND METHODS

Library Screening. An EMBL-3 lambda library (Clontech Laboratories, Palo Alto, CA) was constructed using Sau3A partially digested genomic DNA from Pennisetum glaucum (pearl millet). Screening for the millet chalcone synthase gene was carried out using a 1.4 kilobase (kb) EcoRI fragment of the cDNA C2 gene from maize (Wienand et al., 1986) following standard methods. Screening for the millet myb gene was carried out using a 1.1 kilobase (kb) EcoRI fragment (base 1 of the transcriptional start site to approximately 60 nucleotides past the translational stop site) of the cDNA C1 gene from maize (Cone et al., 1986).

The lambda clones were subcloned into pUC18 and pUC19 plasmid vectors using several different restriction digests. Sequencing of plasmid inserts was performed using the SequiT therm Excel II DNA Sequencing Kit L-C (Epigentec Technologies, Madison, WI) and the Licor automated sequencer.

Data analysis. Sequence alignments were performed using the CLUSTAL W program (Thompson et al., 1994). The MEGA program (Kumar et al., 1993) was used to calculate genetic distances among sequences. These distances were used to determine the time of duplication for the millet and maize genes.

Distance estimates were also generated by the program PAUP*(Swofford, 1998). These distances were then used to generate neighbor-joining trees using the Kimura 2-parameter estimator (Kimura, 1980). Bootstrap analyses were performed on the data sets to assess the statistical significance of the tree groupings (Felsenstein, 1985). Relative rate tests for gene sequence comparisons were completed using the Naglab program (Gaut et al., 1996). The relative rate tests to compare groups of sequences were performed using the Codrates program (Muse and Gaut, 1994).

RESULTS

Screening and sequencing of candidate chs genes. Multiple screening of the Pennisetum genomic library yielded 10 clones that hybridized with the maize chs cDNA probe. Six of these positive clones were mapped and partially sequenced resulting in the identification of two distinct chs genes. These genes are hereafter referred to as millet chs 1 and millet chs 2.

Both genes contain a single intron located between the second and third base of the cysteine residue at position 63. This is the conserved site of intron placement in all genomic chs genes sequenced to date. The intron in millet chs 1 contains 354 bases, whereas the intron in millet chs 2 contains 596 bases. The difference in length of the introns in these related genes is in part the result of a 58 base pair duplication in the millet chs 2 intron. The coding region consists of two exons, which are identical in size between the two millet chs genes; exon 1 contains 189 nucleotides and exon 2 contains 999 nucleotides giving a combined length of 1188 nucleotides. There are 33 synonymous and 4 non-synonymous nucleotide changes be-
Gene family evolution

Table 1. Chi-squared contingency table testing number of synonymous changes versus number of synonymous sites between exon 1 and exon 2 of the millet CHS 1 and CHS 2 genes.

<table>
<thead>
<tr>
<th></th>
<th>Number of Synonymous Changes</th>
<th>Number of Synonymous Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 1</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>Exon 2</td>
<td>33</td>
<td>295.75</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 5.82 \text{ significant at } 0.025 > p > 0.01, \ \text{d.f.}=1 \]

between the two sequences, although there are only three amino acid replacement changes between the two amino acid sequences (see below).

A contingency table was constructed to determine if there is a significant difference between the number of synonymous sites and the number of synonymous changes present in exon 1 compared with exon 2 (Table 1). The Chi-squared value of 5.82 with one degree of freedom indicates that the difference is significant (0.025 > p > 0.01). The intron region immediately following exon 1 also has no changes, but ten changes were found in the intron region after the 58 bp duplication observed in the chs 2 intron. This heterogeneity in nucleotide differences between exon 1 and exon 2 strongly suggests that an interlocus conversion or recombination event occurred between these two genes subsequent to their original duplication.

A sequence alignment of the 5' non-coding region for both millet chs genes is shown in Figure 1. A CHS box homolog was located 90 nucleotides downstream of a putative TATA box at –46 from the translation start site. Eight nucleotides upstream of the TATA box is a recognizable H box homolog at –157 and also a second H box at –123. Eight base pairs upstream of the H box is a putative myb binding domain at –171 and a putative myc binding domain at –193. Interestingly, no complete G-box is identifiable, although there is a G box core at –84. There is also a TACCAT motif at –122 overlapping the second H box.

Table 1. Chi-squared contingency table testing number of synonymous changes versus number of synonymous sites between exon 1 and exon 2 of the millet CHS 1 and CHS 2 genes.

<table>
<thead>
<tr>
<th></th>
<th>Number of Synonymous Changes</th>
<th>Number of Synonymous Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 1</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>Exon 2</td>
<td>33</td>
<td>295.75</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 5.82 \text{ significant at } 0.025 > p > 0.01, \ \text{d.f.}=1 \]

Amino Acid Sequences. The amino acid sequences of the two millet chs genes were aligned with a consensus sequence for all published chs sequences from grasses and a consensus sequence of chs genes from a subset of published dicot sequences. The millet amino acid sequences contain the conserved cysteine residue (at position 167 on our amino acid alignment) which has been determined to be the site for binding 4-coumaroyl-CoA and is essential for chs activity (Lanz et al., 1991; Tropf et al., 1995). Also present is the absolutely conserved chalcone synthase/stilbene synthase signature found in all CHS/STS proteins in the carboxy-terminus (Martin, 1993). The overall amino acid sequence similarity between the millet CHS sequences and the CHS consensus sequence for grasses is 81% while their similarity to the CHS consensus sequence for dicots is 71%. Interestingly, 43 of the 75 changes found between the millet sequences and the grass consensus sequence are also different between the millet sequences and the dicot consensus sequence but not different between the grass and dicot consensus sequences. This suggests that these two millet genes are distinct from the typical grass chs genes.

![Fig. 1. Alignment of the nucleotide sequences for millet chs 1 and millet chs 2 genes 5' of the translation start site. Putative control elements are denoted by boxes and labeled below.](image-url)
**Chs Gene Family Phylogeny.** The aligned nucleotide sequences of ten grass *chs* coding regions, four orchid ([3 Bromheadia (orchid *chs* 3, *chs* 4 and *chs* 8) and one *Phalaenopsis* (orchid *chs* 1, outgroup)]) were used to generate a neighbor-joining tree using Kimura 2-parameter distance measurements. The resulting tree is shown in Figure 2 with bootstrap values for 100 replicates written above the branches. The two millet *chs* genes cluster on a separate branch outside of the other grass *chs* genes. The fact that the millet *chs* genes do not cluster with the maize genes strongly suggests that the millet *chs* genes sampled are not orthologs of *whp* or *c2*. The two rye *chs* gene copies and the two rice *chs* gene copies also appear to cluster together indicating that relatively recent duplication events have occurred within each of the respective grass lineages represented in the grass data set. Interestingly, the branch leading to the millet lineage and the branch leading to barley *chs* 2 are longer than the branches that lead to the other barley, rye, rice and maize *chs* genes, suggesting an acceleration of nucleotide change along these branches.

A second neighbor-joining Kimura 2-parameter distance tree was generated using 79 dicot and two gymnosperm *chs* coding sequences together with the 14 monocot *chs* coding sequences currently available and is shown in Figure 3. The *chs* genes cluster into plant families with strong bootstrap support (91%–100%) as indicated in the Figure, but with a few noted exceptions. *chs* genes sampled from plants in the order Solanales cluster into two groups. Within Solanales *chs* genes sampled from *Petunia* and *Ipomoea* also form two distinct clusters that we have previously called group I and Group II (Durbin et al., 1999). One gene from *Phalaenopsis* and a *Petunia chs* gene cluster as nearest neighbors, but separate from other orchid and Petunia sequences. The Poaceae (grass family) *chs* gene cluster is almost identical to that shown in Figure 2. The exception is the placement of the barley *chs* 2 that falls outside the main cluster of grass *chs* genes (bootstrap value of 96).

There appear to be groups of genes within some of the plant families that have longer branches suggesting accelerated rates of change, the *Ipomoea* group II (*chs* A, B, C and PS) in order Solanales and the barley *chs* 2 and millet branches in Poaceae (group II) are longer. Within Fabaceae we find also two groupings for pea and soybean (Figure 3).

**Rates of nucleotide change in *chs* genes.** As noted above, the difference in branch lengths between groups of genes in some of the plant families suggests differing rates of nucleotide change among duplicate genes within plant lineages and this is particularly evident in the *Ipomoea*

---

**Fig. 2.** Neighbor-joining distance phylogram for *chs* gene family sequences in the monocots, constructed using PAUP (Swofford, 1998) with Kimura 2-parameter algorithms as described in the Materials and Methods. Branch points with support greater than 70% bootstrap values (100 replicates) are indicated.
Fig. 3. Neighbor-joining distance phylogram for chs gene family sequences in monocots, dicots and gymnosperms constructed using PAUP (Swofford, 1998) with Kimura 2-parameter algorithms as described in the Materials and Methods. Branch points with support greater than 79% bootstrap values are indicated.
group and the Poaceae group. The maximum likelihood relative rates test of Muse and Gaut (1994) was used with various outgroup choices to test the null hypothesis of equal rates among duplicate genes within a genome. Millet chs 1 and chs 2 were found to evolve at equal rates using barley, maize, rice or rye genes as outgroups. Similarly rice chs 1 and chs 2 and rye chs 2 and chs 3 were also found to evolve at equal rates with various other grass outgroups. In contrast, rates of protein evolution were found to be accelerated significantly for barley chs 2 compared with barley chs 1 (p < 0.05) (using millet, maize, rice and rye as outgroups) and for maize whp compared to maize c2 (p < 0.01) (using millet, barley, rice and rye as outgroups). Unequal rates were also detected between plant lineages for non-synonymous change in the millet chs gene lineage compared with the maize chs gene lineage using Li and Bousquet (1992) relative rate tests, based on Nei and Gojobori (1986) distance measures (where the orchid genera Bromheadia and Phaeonopsis provided outgroups). Rates of synonymous nucleotide substitution reached the 5% level in 2 out of 40 comparisons (precisely the expected value for 40 tests). We conclude that synonymous rates are homogeneous among genes within lineages and between lineages.

Based on synonymous site data the time of duplication between the genes within the same grass lineage was estimated (Table 2) using $5 \times 10^{-9}$ synonymous substitutions per site per year as a mutation rate. The maize gene duplication event occurred approximately 17 million years ago with the millet gene duplication occurring approximately 13 million years ago if both exon 1 and exon 2 are combined. If only exon 2 is used in the calculation the duplication event dates to 15 million years. These dates place the duplication events after the divergence of the maize and millet lineages, which are estimated to be approximately 25 million years ago (Gaut and Clegg, 1991). Estimates based on the grass chs sequence data place the duplication events in the rice and rye lineages at approximately one million and 18 million years ago, respectively.

**Table 2.** Times of duplication of the two chs genes within each grass species using $5 \times 10^{-9}$ synonymous substitution per site per year as the rate of change and Jukes-Cantor corrected synonymous rates of substitution values.

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Time of duplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maize C2</td>
<td>Maize whp</td>
<td>17 million years ago</td>
</tr>
<tr>
<td>Millet 1</td>
<td>Millet 2</td>
<td>13 million years ago</td>
</tr>
<tr>
<td>Millet 1 exon 2</td>
<td>Millet 2 exon 2</td>
<td>15 million years ago</td>
</tr>
<tr>
<td>Rice 1</td>
<td>Rice 2</td>
<td>1 million years ago</td>
</tr>
<tr>
<td>Rye 3</td>
<td>Rye 2</td>
<td>18 million years ago</td>
</tr>
</tbody>
</table>

**Myb gene characterization.** Multiple screenings of the genomic library of pearl millet (Pennisetum glaucum) yielded 16 clones that hybridized with the maize c1 cDNA probe. All 16 clones were identical in sequence and contained a characteristic myb domain. Hereafter this gene is referred to as the millet myb 1 gene. The coding region is separated into 3 exons. Exon 1 (133 bp) codes for the 5’ end of the R2 subdomain of the myb motif. Exon 2 (130 bp) codes for the 3’ end of the R2 subdomain and the 5’ end of the R3 subdomain. Exon 3 (379 bp) codes for the rest of the R3 subdomain and the C-terminal end of the protein. The two introns have lengths of 107 and 125 bp, respectively; intron 1 interrupts a glycine residue whereas intron 2 interrupts a lysine residue that is adjacent to the second tryptophan site in the R3 subdomain. The placement of both of these introns in the millet myb 1 gene is exactly the same as for two of the maize myb genes, c1 and pl (Paz-Ares et al., 1987; Cone et al., 1993). A putative TATA box is evident in the 5’ flanking region.

**Amino Acid Sequence.** An amino acid alignment of the R2 and R3 subdomains characteristic of myb genes, including millet myb1, are represented in Figure 4. The alignment includes myb sequences from a select group of monocots and dicots, one gymnosperm (spruce), one bryophyte (moss) and one animal (chicken). The conserved tryptophan, glycine and arginine residues of the myb subdomains are indicated by an asterisk with the location of the first tryptophan site in the R3 subdomain represented by an arrow. These sites are separated by 18–19 amino acids in all cases. There is also a conserved domain surrounding the last tryptophan site (outlined by a box), that is thought to be involved in forming the contact site between myb proteins and DNA (Martin and Paz-Ares, 1997). These characteristics are found in all myb genes across three kingdoms (Martin and Paz-Ares, 1997). The millet myb1 gene identified in this study also contains these characteristics. The presence of a promoter region with a putative TATA box, and complete coding sequence with appropriate in frame start and stop codons, indicates that millet myb 1 is functional.

**Myb gene family phylogeny.** Nucleotide alignments of the myb domain of 72 genes from plants, animals and fungi were used to generate a neighbor-joining tree based on Kimura 2-parameter distance estimates. Only sites 1 and 2 of each codon were used in the distance estimates because most third position sites are saturated. The estimated phylogram is presented in Figure 5. The bootstrap values for 500 replicates are indicated. This tree has branches that are supported for clades that contain both monocot and dicot myb sequences. For example, the millet myb 1 gene, along with two barley and a maize gene, clusters with a tomato myb gene and several Arabidopsis genes with 82% bootstrap support. Several of the rice
Fig. 4. Amino acid alignment for select myb genes. Nucleotide sequences from GenBank, translated with Genetic Data Environment (GDE) and aligned with Clustal W (Thompson et al., 1994). Asterisks denote conserved tryptophan, glycine and arginine residues. The arrow indicates a change from a tryptophan to another amino acid. The conserved domain surrounding the last tryptophan site is outlined by a box.
Fig. 5. Neighbor-joining distance phylogram, using nucleotide bases 1 and 2 for the myb gene family sequences in select monocots, dicots, gymnosperms, bryophytes and fungi with animal sequences used as outgroups. The phylogram was constructed using PAUP (Swofford, 1998) with Kimura 2-parameter algorithms, as described in the Materials and Methods. Branch points with support greater than 70% bootstrap values are indicated.
genes cluster with tomato or Arabidopsis. The spruce gene (a gymnosperm) falls within the tree as does the moss (a bryophyte) sequence. This tree suggests that the dates of divergence of some of the myb gene lineages are older than the divergence of angiosperms.

C-Terminal Motifs. Figure 5 was generated using only the myb domains because there appears to be no homology in the C-terminal end for most of these genes. However, some sequences do show regions of similarity that may correspond to transactivation domains. Further analysis of the C-terminal end of the cluster of ten plant myb genes, including millet myb1, revealed a conserved motif of 14 amino acids. This motif has five absolutely conserved amino acids and three amino acids that have only one sequence with a conservative replacement. This motif is shown in Figure 6. The branch that leads to the clade that contains these sequences is supported with a bootstrap value of 77 and is based on myb domains; the C-terminal end is not included in the alignments used to generate the trees. Two Arabidopsis myb sequences were found that contained this same motif and this clade is located further down on the tree. No other sequences used in the myb tree analysis appeared to have this motif. As a contrast, the C-terminal end of another clade of myb genes represented by maize c1, maize pl and rice Y152 contain two C-terminal motifs that are different than the motif represented above. These motifs are not present in millet myb1 or any other plant myb gene represented in the myb tree analysis.

The millet myb1 sequence clusters with maize zm38 which has been characterized as an inhibitor of gene transcription in genes involved in anthocyanin production. Based on this clustering, it can be speculated that millet myb1 may also have an inhibitory function possibly on anthocyanin structural genes.

**DISCUSSION**

The salient facts of our investigation of chs gene family evolution are (1) the discovery of a greater multiplicity of
chs genes in grass genomes than was previously thought; (2) evidence that the recruitment of new duplicate chs genes occurs at a remarkably rapid rate in grass genomes; and (3) evidence that novel chs genes are functional and must therefore have adaptive value for the plant. The major conclusions from our analysis of myb gene family evolution are: (1) despite a high level of redundancy of myb genes in plant genomes most lineages so far described trace back to duplication events early in seed plant evolution; and (2) a novel amino acid motif is identified with the millet myb gene that is conserved in evolution and may therefore have a functional role. We begin by discussing the evidence supporting these conclusions. We then consider the broader question of why genetic redundancy is a major adaptive strategy of plants.

Patterns of chs gene recruitment. The neighbor-joining trees generated using nucleotide sequence data suggest that there are more copies of chs genes in the grass genome than had been previously reported. To date only two complete sequences are available from each of maize, barley, rice and rye. Data from the rbcL genes of the chloroplast genome have been used to estimate times of divergence of many of the grass genera; Zea (maize) and Pennisetum (millet) diverged approximately 25 million years ago, while Hordeum (barley) Zea, and Oryza (rice) diverged approximately 50 million years ago (Duvall and Morton, 1996). The divergence times require that orthologous chs genes from millet should cluster with orthologous maize chs genes. Instead, the neighbor-joining trees show the millet chs genes cluster outside of the clade that includes the two copies of the maize, rye, and rice chs genes and the barley chs 1 gene with a bootstrap value of 100. This suggests that the millet chs genes are not orthologous to the main group of chs genes in grasses. The barley chs 2 gene also falls outside of the main clade of grass genes on a strongly supported branch (bootstrap value of 96), likewise suggesting that this gene is not orthologous to the other group of grass chs genes. These data lead us to infer that additional chs-like genes will be discovered in the genome of maize and other grass species.

To further test the possibility of additional chs-like genes in the grass genome, Pioneer Hybrid performed a search of their maize expressed sequence tag library (containing in excess of 100,000 sequence tags) for chs-like tags at our request. This search revealed at least two additional unique chs-like tags in the maize genome. Finally, as already noted in the introduction, Southern blot analysis of barley indicated that there may be as many as seven chs genes in the barley genome (Christiansen et al., 1998).

The phylogenetic analysis of the grass family suggests a relatively recent duplication pattern of chs genes following the separation of the lineages that lead to each genus in the sample. Thus for example, the maize c2 and whp genes are believed to be the result of an allotetraploid event that occurred somewhere around 15 million years ago (Gaut and Doebley, 1997), after the maize lineage had separated from the Pennisetum lineage. The general picture is one of a highly plastic gene family that experiences new gene recruitment events every 15 to 25 million years. This is consistent with earlier estimates by Clegg et al., (1998) that indicate that the chs gene family recruits new copies at a relatively rapid rate.

Novel chs genes appear to be functional. The derived amino acid sequences of the millet chs genes contain sequence motifs that are conserved in monocot, dicot, and gymnosperm chs genes. These motifs include a cysteine residue at position 167 (in our alignment) known to be essential for chs activity (Lanz et al., 1991; Tropf et al., 1995). The two millet chs sequences also contain the chalcone synthase/stilbene synthase signature at the carboxyl end of the protein. All the necessary components for functionality are found in both millet chs genes. The genes are of reasonable length and the intron site in both genes is in the same position as in all known active chs gene sequences. There is also a translational start site that corresponds to a reading frame that translates to an amino acid sequence characteristic of CHS proteins. The high level of sequence conservation and its retention over a long period of evolutionary time strongly argues that the two novel millet genes code for functional enzymes.

Novel chs-like genes have conserved 5' regulatory sites. The 5' untranslated region of the two millet genes appear to have the necessary components for transcriptional activity. The TATA box sequence (TATATAT) has been shown to bind strongly with TATA-binding protein in vitro which in turn can enhance transcription rates (Hoopes et al., 1998). The same TATA sequence exists in the maize genes c2 and whp (Franken et al., 1991). Barley chs 1 (Rohde, et al., 1991) and rye chs 3 (Haussuhl et al., 1996) also have a known TATA box sequence (TATATAT) (Hoopes et al., 1998).

Within the 5' untranslated region for which we have sequence data, there are a number of conserved regulatory motifs. The millet genes do not appear to have a complete G-Box domain in the promoter region, however they both have the core motif of the G-Box. The two millet genes also have an identifiable CHS Box, as well as two H-Boxes, myb and myc binding domain, and a TACPyAT motif.

Evidence of inter-locus exchange. Although there is not a significant difference in nonsynonymous change between exons 1 and 2 of the duplicated millet genes, there is a difference in synonymous changes between the two exons. There are no synonymous changes in exon 1 between the two millet genes, yet there are 33 in exon
2. A contingency table test of the null hypothesis of proportional synonymous differences in both exons was rejected at the 2.5% level which suggests that a gene conversion or recombination event occurred subsequent to the duplication of these two genes. Further support for an exchange event can be seen from an analysis of the intron region 5’ to the 58 bp repeat where no nucleotide differences are observed, but following the repeat 10 differences are observed. The repeat region may mark one boundary of the putative exchange event. Moreover, the putative interlocus exchange event appears to have occurred recently in evolutionary time because there are no synonymous changes in exon 1 between the two millet chs genes. Evidence for interlocus exchange has also been detected in the alcohol dehydrogenase gene family in grasses (Gaut et al., 1999) so occasional exchanges among duplicated loci appear to be an aspect of plant genome evolution.

Rates of amino acid replacement substitution are accelerated in recently duplicated chs genes. Amino acid replacement rate acceleration is observed in maize and barley duplicate chs genes. Acceleration of evolutionary rate is also observed in other recently recruited chs genes. Thus for example, the Ipomoea chs A, B, and C genes exhibit accelerated rates of amino acid substitution relative to the older chs D lineage (Durbin et al., 1999). In addition, accelerated rates of protein evolution are observed in the grass adh 2 lineage relative to the adh 1 lineage (Gaut et al., 1999). Acceleration of protein evolution following duplication is often observed presumably because the newly duplicated protein is no longer under rigid functional constraint. The period following a duplication event is thought to permit a wider exploration of the space of potential protein structures than would be the case for a non redundant gene product with a well circumscribed catalytic function. Nevertheless, the long term retention of functional duplicate genes strongly implies that the novel protein structures have adaptive utility for the plant.

The millet myb gene appears to be functional. The predicted reading frame for myb1 encodes a myb domain at the NH₂-terminus and a probable transactivation domain at the C-terminal end. The myb domain of millet myb1 contains tryptophan residues that are characteristically found in subdomains, R2 and R3 of all known myb regions. In R2 all tryptophan residues are evenly spaced, 19 residues between each and in R3 the tryptophan residues are separated by 18 residues as found in other plant myb genes (Martin and Paz-Ares, 1997; Williams and Grotewold, 1997). These tryptophan residues are thought to align three alpha helices that form into a hydrophobic core for each subdomain (Frampton et al., 1991). Also identified in all plant myb genes is a change of the first tryptophan residue in the R3 subdomain to another phenolic amino acid (Martin and Paz-Ares, 1997; Williams and Grotewold, 1997). In the case of millet this residue is a phenylalanine.

Other conserved amino acids found in both subdomains and also identified in millet myb1, are the glycine residues at sites 21 and 74 and the arginine residues at sites 48 and 99. These amino acids are thought to form salt bridges in each domain in animal myb proteins, which contributes to tertiary folding of the hydrophobic core of the entire protein (Frampton et al., 1991). Another conserved amino acid motif is found surrounding the last tryptophan of the R3 subdomain. This string of amino acids is thought to be involved in stabilizing the helix-turn-helix tertiary structure of myb proteins (Martin and Paz-Ares, 1997).

Conserved C terminal amino acid motifs. Comparison of the C-terminal end of plant myb proteins may suggest similarities within clades and possible relatedness of function. Sequence analyses of the C-terminal amino acids of the translated millet myb1 gene identified a highly conserved and evidently ancient motif of fourteen amino acids, just downstream of the myb domain. Seven of the amino acids in this motif were identified earlier in barley myb 1 and myb 2 and snapdragon myb genes (308, 315, and 330) (Jackson et al., 1991). Our analysis indicates that the probable conserved domain is in fact larger and includes four amino acids upstream and three downstream of the region identified by Jackson et al. (1991). Seven of the fourteen amino acids are absolutely conserved (two have one change) and three more have conservative functional substitutions. This motif, when used in a BLAST search (Altschul et al., 1990), shows identity only with other myb proteins and all of these, except two Arabidopsis accessions, are in the same clade (bootstrap support value 77), which includes both dicot and monocot sequences. The two Arabidopsis myb sequences (AbThMYB4 and AbThMYB3321) that also contain the same motif are in a different clade, which is placed further down in the neighbor-joining tree. The high level of conservation found in the C-terminal motif evidently spans more than 200 million years of evolutionary time, the estimated time of divergence of monocots and dicots (Wolfe et al., 1989), which strongly suggests that this is an important functional domain.

Age of the myb gene family in plants. The phylogram based on the first and second codon positions of the myb domain suggests that the myb gene family is old. Indeed many of the clades of myb genes resolved in Figure 5 have clusters containing both monocot and dicot myb gene sequences. For example, the clade that contains millet myb1 includes not only other monocot representatives, (barley myb1 and myb2, maize myb38 and rice myb1), but
also has a dicot myb gene sequence (tomato myb27) (bootstrap support value 75). This clade also joins another clade (bootstrap support value 74) that includes other dicot myb gene sequences, (Arabidopsis mybY49 and mybM4E13). Further support for an ancient diversification of the plant myb gene family is shown in the clade that includes a bryophyte myb gene sequence (bootstrap support value 78) with a monocot (rice myb1402), and two dicot sequences (tomato myb16 and Arabidopsis myb20). Finally, note the placement of a gymnosperm myb gene sequence (spruce) within the middle of the phylogram again suggesting ancient gene duplication events. This type of clustering suggests that many of the gene duplications that define the myb gene family occurred early in land plant evolution.

Why is genetic redundancy a common feature of plant genomes? Plant genomes appear to exploit genetic redundancy as an adaptive strategy. Perhaps this derives from the fact that the individual plant must survive a wide range of environments both in a temporal sense (because of seasonal variation) and in the sense that major organ systems must be adapted to very different conditions ranging from subsurface environments (roots) to the exposed environments of the terrestrial atmosphere (leaves). It seems plausible that the exact enzymatic conditions that are optimal under one set of conditions may not be optimal under a different set of environmental conditions. The duplication and evolutionary specialization of important plant enzymes may thus be adaptive. Similarly the ability to deploy a vast combinatorial set of gene products through differential gene expression may further add to the plants ability to cope with environmental variation. Thus multiple regulatory cassettes that enhance the biochemical and metabolic flexibility of the plant may have been favored over the long course of terrestrial evolution. The patterns of chs and myb gene family evolution appear to be consistent with this speculation. What seems clear beyond dispute is that the multiple redundant copies of these genes do have adaptive value for the plant.

Implications for plant genomics. The research program of genomics has largely been based on the categorization of similar sequences into broad groupings. Thus for example, chs genes fall into a simple classification based on sequence similarity, but beneath this classification is the potential for substantial specialization in function that is not evident without more detailed analyses. It seems clear that evolutionary analyses have much to offer to the science of genomics. A comparative analysis can provide important circumstantial evidence on the adaptive value of redundant gene copies. Comparative analyses can also highlight conserved regions of a protein or non-coding sequence motifs associated with gene regulation that are likely to be functionally important by virtue of their sequence conservation. Based on these considerations it seems likely that the next phase in the exploitation of genomic data will rest on more detailed comparative analyses.

This research was partially supported by the Alfred P. Sloan Foundation. We thank Pioneer Hybrid International for sharing information about chs redundancy in their EST library. We also thank Dr. Karen Cone (University of Missouri, Columbia, MO) for providing the c1 clone from maize and we also thank Dr. Udo Wienand (Max Planck Institute, Zuchtungsforsch, Cologne, Germany) for providing the c2 clone from maize.

REFERENCES
Gene family evolution


