A genomic approach to elucidating grass flower development

Marcelo C. Dornelas[1]* and Adriana P.M. Rodriguez[2]

Abstract

In sugarcane (*Saccharum* sp) as with other species of grass, at a certain moment of its life cycle the vegetative meristem is converted into an inflorescence meristem which has at least two distinct inflorescence branching steps before the spikelet meristem terminates in the production of a flower (floret). In model dicotyledonous species such successive conversions of meristem identities and the concentric arrangement of floral organs in specific whorls have both been shown to be genetically controlled. Using data from the Sugarcane Expressed Sequence Tag (EST) Project (SUCEST) database, we have identified all sugarcane proteins and genes putatively involved in reproductive meristem and flower development. Sequence comparisons of known flower-related genes have uncovered conserved evolutionary pathways of flower development and flower pattern formation between dicotyledons and monocotyledons, such as some grass species. We have paid special attention to the analysis of the MADS-box multigene family of transcription factors that together with the APETALA2 (AP2) family are the key elements of the transcriptional networks controlling plant reproductive development. Considerations on the evolutionary developmental genetics of grass flowers and their relation to the ABC homeotic gene activity model of flower development are also presented.

INTRODUCTION

Grasses, like all higher plants, develop by the elaboration of populations of undifferentiated stem cells (meristems) which give rise to all the organs of the plant. The shoot apical meristem (SAM) is the ultimate source for all aerial structures of the plant, including flowers. The identity of a meristem is defined by the types of structures it produces, and in sugarcane the shoot meristem undergoes several distinct transitions in identity during the life of the plant. A major transition occurs when the vegetative meristem ceases leaf production and is converted to an inflorescence meristem, which produces intermediate meristems that give rise to floral meristems which finally produce the organs of the flower.

The grasses are members of the monocotyledonous subclass of the angiosperms and produce highly modified wind-pollinated flowers (Schmidt and Ambrose, 1998; McSteen *et al.*, 2000). As in many other grasses, sugarcane flowers (florets) are organized into units called spikelets, each consisting of a pair of sterile bracts (glumes) enclosing a fixed number of florets (Moore, 1987). The regulation of floret number per spikelet is a prime determinant of spikelet architecture among members of the grass family. Sugarcane spikelets are determinate (*i.e.* have a fixed number of florets) and produce only one floret, arranged at a defined position on an axis called the rachilla (Moore, 1987), while other determinate grasses (*e.g.* maize) produce two. Some other species of grass (*e.g.* wheat) are indeterminate and produce spikelets with a variable number of florets (Schmidt and Ambrose, 1998).

The spikelet itself is but one component of the complex branched inflorescence of sugarcane. In contrast to dicotyledonous plants such as those from *Arabidopsis thaliana*, in which flowers are produced directly by the apical and lateral meristems (Hempel and Feldman, 1994), there are at least two distinct inflorescence branching steps in sugarcane before the spikelet meristem terminates in the production of the floret. These extra branching steps allow for greater morphological diversity among the grasses.

In contrast to the differences seen in inflorescence development, morphological studies have shown that the formation of the florets of grasses and the flowers of dicotyledons are similar (Greyson, 1994). As typical dicotyledons, which form four whorls of floral organs (sepal, petal, stamen and carpel), the sugarcane floral meristem also forms four different types of organs *i.e.* the palea, lodicule, stamen and carpel (Moore, 1987). The leaf-like palea and the reduced lodicule are thought to be analogous to the sepal and petal, respectively (McSteen *et al.*, 2000). In model plants, the identity of the floral organs is determined mainly by the combination of differentially expressed transcription factors belonging either to the MADS-box (for a review, see Theissen *et al.*, 2000) or to the APETALA2 (AP2, Jofuku *et al.*, 1994).

Although the evolutionary conservation of the mechanisms underlying flower development are well-known (Baum, 1998), an understanding of the molecular basis of

---

[1]*USP/Escola Superior de Agricultura "Luiz de Queiroz", Dep. Ciências Florestais, Av. Pádua Dias 11, 13418-900 Piracicaba, SP, Brazil.*
[2]*USP/Centro de Energia Nuclear na Agricultura, Lab. Biotecnologia Vegetal, Av. Centenário 303, 13400-970 Piracicaba, SP, Brazil.*
Send correspondence to M.C. Dornelas. E-mail: mcdornel@carpa.ciagri.usp.br.

grass flower development is just at its beginning. The availability of data from the extensive Expressed Sequence Tag (EST) sequencing project being developed by the Brazilian Sugarcane EST Project (SUCEST, http://sucest.lbi.dcc.unicamp.br/en/) and the evolutionary conservation of flower development have encouraged our attempts to identify most of the genes known to be involved in the control of flower development by using the SUCEST database.

In this article we concentrate on the analysis of the MADS-box multigene family of transcription factors that together with the APETALA2 (AP2) family are the key elements of the transcriptional networks which control plant reproductive development. Considerations on the evolutionary developmental genetics of grass flowers and their relation to the ABC model of flower development are also presented.

MATERIAL AND METHODS

Nucleotide and amino acid sequences of MADS-box and AP2 homologues

The 60-aminoacid-long MADS-box domain (Theissen *et al*., 2000) contained in the N-terminal region of the proteins and the 59-residue-long AP2 motifs (Jofuku *et al*., 1994) were obtained from data on *Arabidopsis* (GenBank database http://www.ncbi.nlm.nih.gov/entrez/ query.fcgi) and sugarcane (SUCESTdatabase http://sucest.Lbi.dcc.unicampbr/en/). The complete set of *Arabidopsis* MADS gene and protein sequences and the available sugarcane cluster sequences (old Phrap clustering) were obtained using the BLAST algorithm (Altschul *et al.,* 1990) and BLOSUM62 scoring matrix, with a threshold of e > 10^{-5} for positive hits. Accession numbers are given in the legends or can be obtained at http://www.mpiz-koeln.mpg.de/mads/.

Comparison of the amino acid sequences of MADS-box and AP2 homologues

The amino acid sequences corresponding to the MADS-box or AP2 domains (Theissen *et al*., 2000; Jofuku *et al*., 1994) were aligned using the CLUSTALX (Thompson *et al.,* 1994) software and the alignments were corrected by hand. The neighbor-joining (Saitou and Nei, 1987) trees and Bootstrap calculations (1000 replicates) were made using the MEGA software (http://www.megasoftware.net).

Phylogenetic analysis of the MADS-box and AP2 gene family

The nucleotide sequences corresponding to the coding region of the MADS and AP2 domains (see above) were aligned using the CLUSTALX (Thompson *et al.*, 1994) software. For the maximum parsimony analyses using PAUP software (version 3. I., Swofford, 1990), the third base of each codon was eliminated because of the redundancy of the genetic code. The 'branch and bound' and the 'exhaustive' search methods of the PAUP software were used. Double weight was attributed to transversions and simple weight to transitions. One hundred replicates were used for generating the bootstrap majority-rule consensus tree. Neighbor-joining and maximum likelihood trees were also constructed using the MEGA package.

RESULTS AND DISCUSSION

Most of our results on the data mining of genes involved in sugarcane flower development can be obtained at http://sucest.lbi.dcc.unicamp.br/private/mining-reports/ EF/EF-mining.htm.

The sugarcane MADS-box family

The large number of studies in recent years has culminated in the insight that inflorescence and flower development in higher flowering plants are determined by a network of regulatory genes, organized in a hierarchical fashion (for a review, see Theissen *et al.,* 2000). Most of these genes belong to the MADS-box family of transcription factors. Accordingly, the MADS-box genes have played an important role in the origin and evolution of flower development (Baum, 1998; Theissen *et al.,* 2000).

We have identified at least 21 sugarcane MADS-box genes (Figure 1), belonging to most of the known MADS-box gene groups. In model dicotyledons and monocotyledons, these groups are thought to perform distinct biological functions and play different roles inside the genetic circuitry that controls the different steps of meristem conversion, from vegetative to floral fates (Teissen *et al.,* 2000).

Close to the top of the hierarchy of flower-regulatory genes are the genes for 'late-flowering' and 'early-flowering' which are triggered by environmental factors, such as day length, light quality and temperature. These genes, including the FLOWERING LOCUS F (FLF, Sheldon *et al.*, 1999), SHORT VEGETATIVE PHASE (SVP; Hartmann *et al.*, 2000), and AGL20 (Borner *et al*., 2000) genes, mediate the switch from vegetative to reproductive development, perhaps by activating meristem identity genes. We have identified putative sugarcane homologues to all these genes (Figure 1).

Meristem identity genes control the transition from vegetative to inflorescence meristems and from inflorescence to floral meristems. Within floral meristems, cadastral genes set the boundaries of floral organ identity gene functions, thus defining the different floral whorls. In the model dicotyledonous plant *Arabidopsis*, floral organ identity genes specify the organ identity within each whorl of the flower by activating a set of yet unknown genes. In the classical ABC model (Coen and Meyerowitz, 1991), three
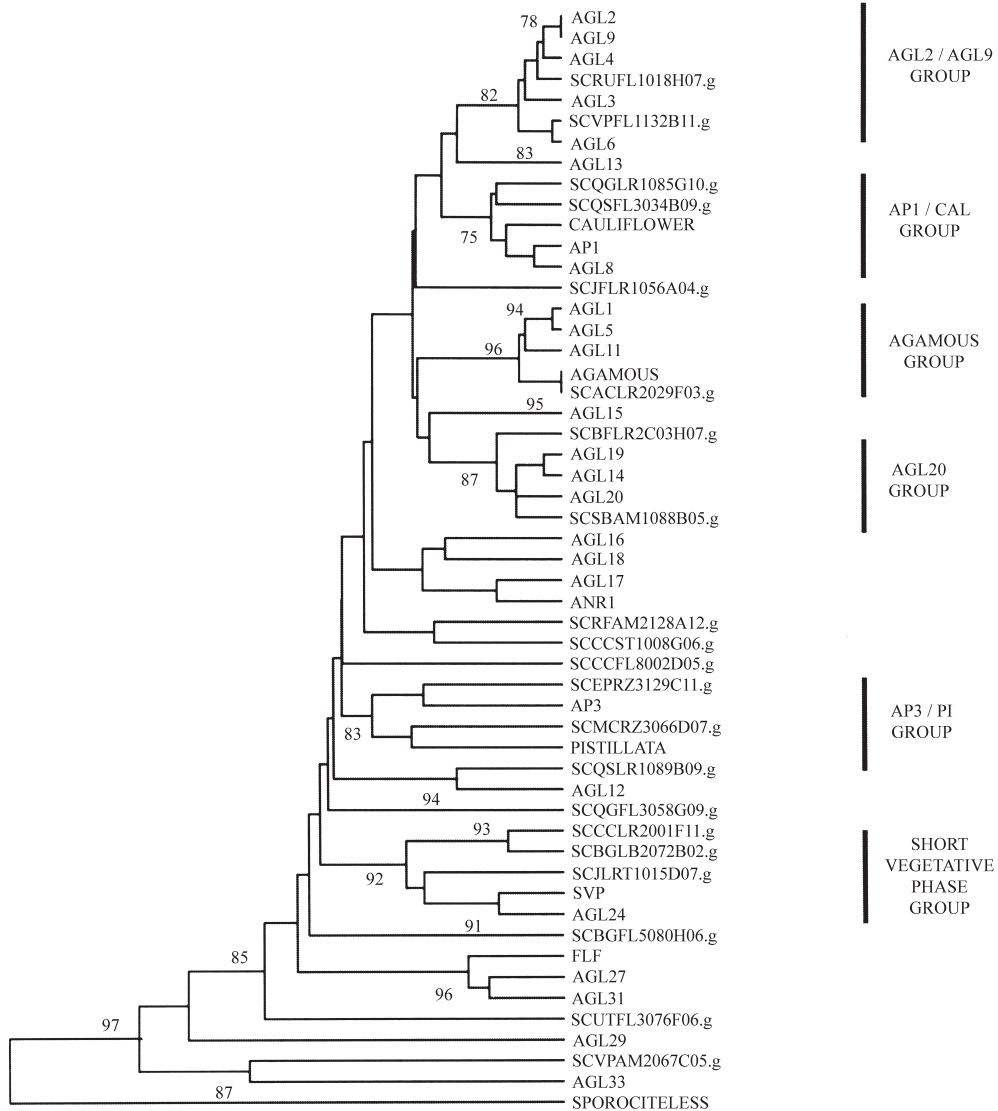
Figure 1 - Phylogenetic relationships of the sugarcane MADS-box transcription factors and their putative homologues from *Arabidopsis*. The most parsimonious consensus tree resulted from the branch-and-bound algorithm. Trees with similar topology were also obtained using a distance matrix and neighbor-joining. Only Bootstrap values (1000 replicates) higher than 75% are shown. For reference numbers and references to the gene sequences used, see Methods.

classes (A, B and C) of homeotic gene activities have been proposed. For any one of the four flower whorls, expression of the A class genes (the APETALA1 gene) on their own specifies sepal formation, while the combination of A and B class genes (the APETALA3 and PISTILLATA genes) specifies the development of petals and the combination of B and C class genes leads to stamen formation. Expression of C class genes (AGAMOUS gene) alone determines the development of carpels.

We have identified sugarcane putative homologues for all the ABC class genes (Figure 1). Although some monocotyledonous homologues of ABC class genes have been cloned (reviewed by McSteen *et al.*, 2000), this is the first report of grass homologues for all the MADS-box

sub-family groups. These groups are maintained even when only the grass MADS sequences are added to the analysis (Figure 2).

Gene expression patterns and phylogenetic studies have suggested orthologous relationships between a number of the maize MADS-box genes and their dicotyledonous relatives (Theissen *et al.*, 1996, 2000; Schmidt and Ambrose, 1998). Such studies have suggested that the *ZAG1* gene is a maize orthologue of the *Arabidopsis* C-function organ identity gene AGAMOUS. A putative null allele of the *ZAG1* gene was identified as carrying a mutator insertion (Mena *et al.*, 1996). In this case the ABC model would have predicted a loss of both reproductive organ development and floral meristem determinacy, but
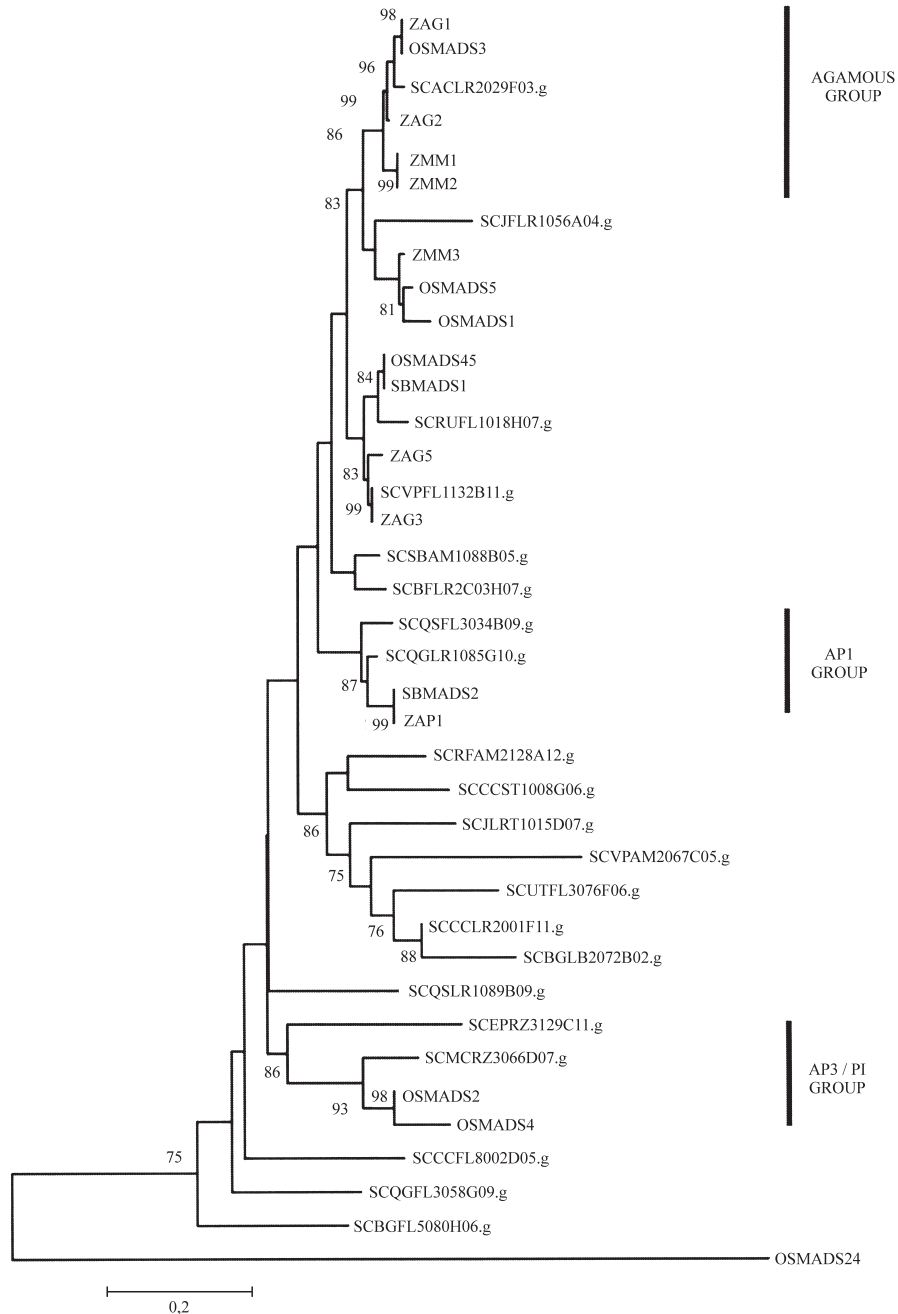
Figure 2 - Phylogenetic relationships among grass MADS-box transcription factors. The most parsimonious consensus tree had its branches scaled to the corresponding genetic distances. Only Bootstrap values (1000 replicates) higher than 75% are shown. For reference numbers and references to the gene sequences used, see Methods.

only loss of floral meristem determinacy was expressed in the phenotype, with supernumerary carpels being reiterated within the mutant *zag1* florets. The failure to affect stamen and carpel identity, in spite of the fact that *in situ* hybridization analyses indicated that *ZAG1* is expressed in the primordia of both reproductive organs, suggests a partial redundancy in C-function activity in maize (Mena *et al.*, 1996). By analogy to the MADS-box of the ZAG1 gene, this observation led to the identification and cloning of a closely related cDNA, the *ZMM2* gene (Theissen *et al.*, 1995). The pattern of *ZMM2* expression, along with its sequence and map position suggested it is a duplicate gene with activities that are non-identical, but partially overlapping, to those of *ZAG1*, *ZAG1* being more highly expressed in carpels and *ZMM2* more highly expressed in stamens. This pair of duplicate genes appears more ancient than most duplicate gene pairs in maize, and it has been suggested that the *ZAG1/ZMM2* duplication predates the divergence of

grass species (Schmidt and Ambrose, 1998, Theissen *et al.*, 2000). For a number of grass species, therefore, it is likely that C-function activity is shared by two genes, although C-function activity in rice is undertaken by a single gene, *OSMADS3* (Kang *et al.,* 1998) and we have found only a single putative homologue for the C-function in sugarcane (compare Figure 1 with Figure 2), which is closely related to *OSMADS3* and *ZAG1*. As we have identified putative MADS homologues for all of the MADS family groups, and the tree topology based only on the sugarcane nucleotide sequences (Figure 3) is quite similar to that observed in Figure 1, we assume that we have identified if not all, then at least most of the sugarcane MADS genes.

It is likely that the C-function is not duplicated in sugarcane, as well as in rice and that ZAG1 and OSMADS, as well as SCACLR2029F03.g, are the true orthologues of AGAMOUS. We therefore suggest that the grass lineage where the C-function duplication occurred may be the origin of unisexual flowers in monoecious grasses (*e.g.* maize) where this duplication is observed. The formation of unisexual flowers is thought to have evolved independently in many plant taxons to promote allogamy (Stebbins, 1957; Le Roux and Kellogg, 1999).

The sugarcane AP2 family

The APETALA2 (AP2) homeotic gene of *Arabidopsis*, has several functions in flower, seed, and ovule development (Jofuku *et al.* 1994; Modrusan *et al.* 1994). In addition to its role in determining floral organ identity, AP2 affects the regulation of floral meristem identity. For example, double mutants of the weak *ap2-1* allele with floral meristem identity mutants such as *leafy* or *apetala1* produce more inflorescence side branches in place of flowers (Bowman *et al.* 1993). Also, weak *ap2* alleles under short day-length conditions cause the formation of tertiary floral shoots in the axils of transformed sepals (Schultz and Haughn, 1993). The AP2 gene belongs to a large plant-specific gene family, called the EREB/AP2 family (Riechmann *et al.*, 2000). The two sub-families (AP2 and EREB) are characterized by the presence of two contiguous AP2 motifs in the AP2 sub-family proteins and only one in the EREB subfamily (Klucher *et al.*, 1996).

We have identified 44 sugarcane clusters that have a similarity coefficient (e-value) greater than 10^{-5}, when compared to either the first (r1) or second (r2) AP2 motifs (Figure 4). Since we made BLAST searches in the SUCEST database using only the two *Arabidopsis* AP2 motif sequences, the number of clusters in the EREB sub-family is probably underestimated.

The two AP2 sub-families (AP2 and EREB) have been distinguished by the presence of two contiguous AP2 motifs in the AP2 sub-family proteins and only one in the EREB subfamily (Klucher *et al.*, 1996). However, the comparison of the sugarcane homologues with the complete
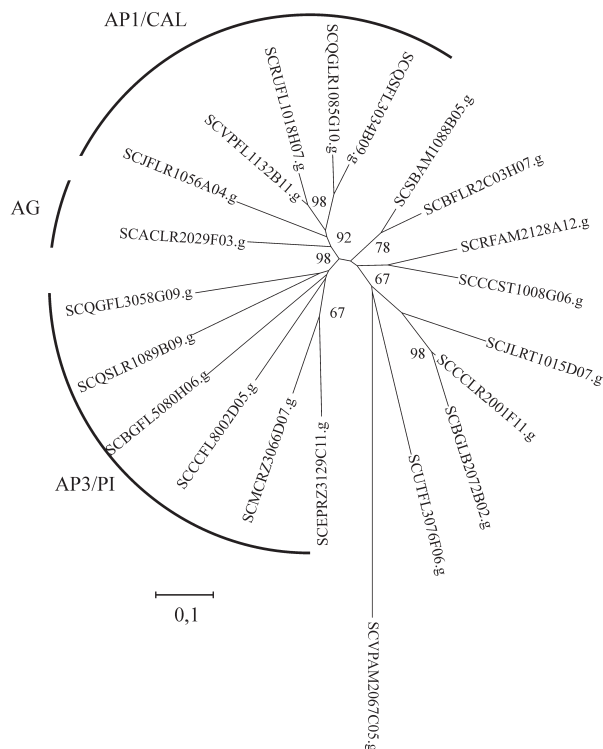

Figure 3 - Phylogenetic relationships among sugarcane nucleotide sequences coding for the MADS-box domain. The most parsimonious consensus tree had its branches scaled to the corresponding genetic distances and is shown in its radial form. Only Bootstrap values (1000 replicates) higher than 75% are shown.

Arabidopsis AP2 family has revealed that there are single-motif proteins that are more closely related to the AP2 sub-family than to the EREB group (Figure 4). This has also been observed when only the sugarcane sequences were analyzed.

The presence of sugarcane and *Arabidopsis* sequences in the same clad (Figure 4) would suggest that the divergence of the AP2 family into two distinct groups predates the divergence of monocotyledons and dicotyledons, estimated to have occurred 135-300 million years ago (Purugannan *et al.*, 1995; Theissen *et al.*, 1996). This observation is also valid for the divergence of the two tandem repeats found in the AP2 sub-family, suggesting that the last common ancestor of monocotyledons and dicotyledons already had two AP2-subfamily genes, one belonging to the AP2-group and the other to the ANT-group.

The divergence of the AP2 family into two distinct sub-families and the presence of two diverging groups in the AP2-subfamily are also observed when sugarcane nucleotide sequences alone are taken into account (Figure 5), suggesting that we have identified most, if not all, of the sugarcane genes belonging to the AP2 sub-family.

Numerous AP2 homologues have been identified in both monocotyledons and dicotyledons (Jofuku *et al.*, 1994; Ohme-Takagi and Shinshi 1995; Moose and Sisco, 1997; Chuck *et al.*, 1998). Mutants for the AP2-like gene
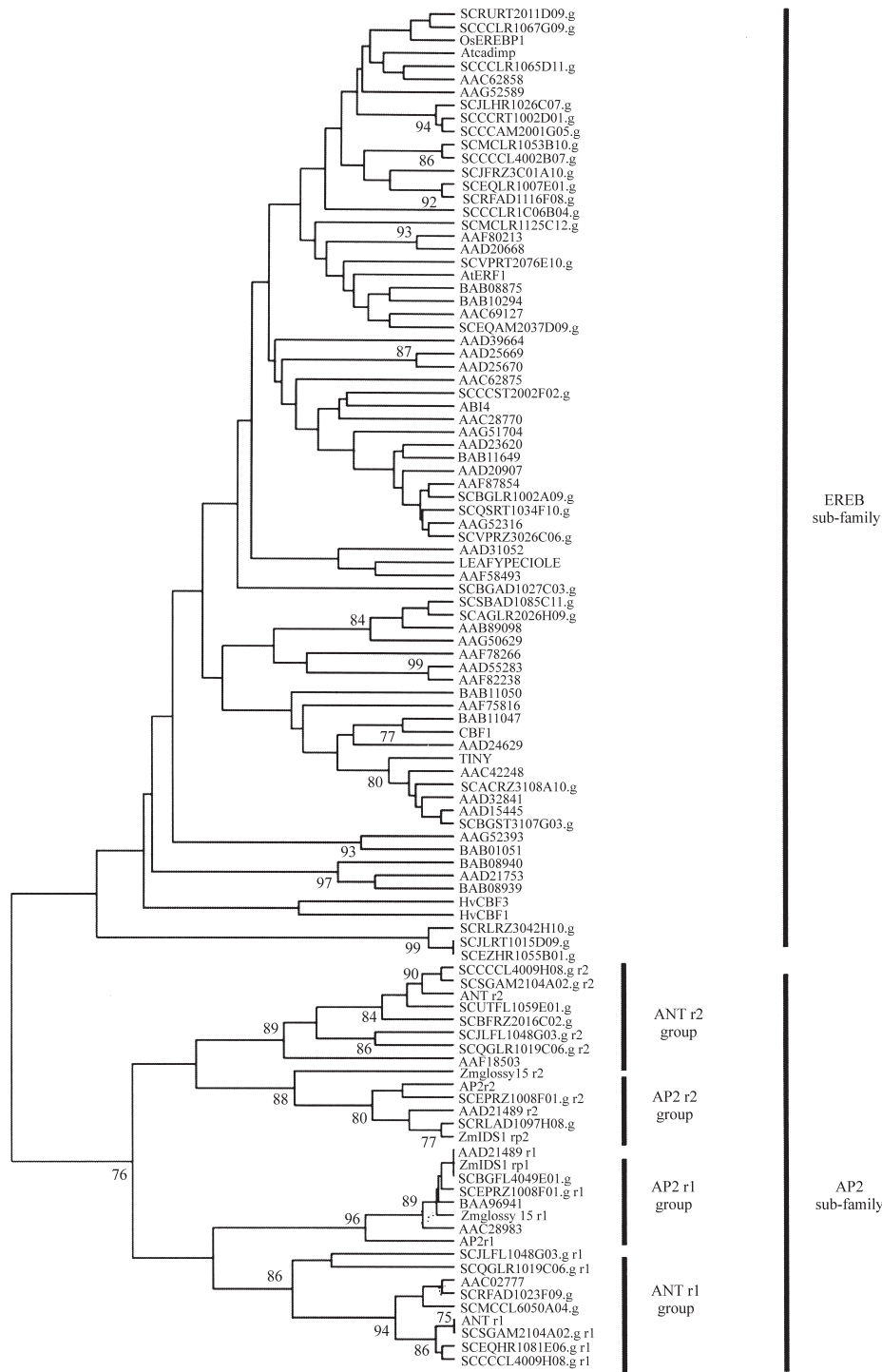
Figure 4 - Phylogenetic relationships of the sugarcane AP2 transcription factors and their putative homologues from *Arabidopsis*. The most parsimonious consensus tree resulted from the branch-and-bound algorithm. Trees with similar topology were also obtained using a distance matrix and neighbor-joining. Only Bootstrap values (1000 replicates) higher than 75% are shown. The r1 and r2 labels indicate the two tandem AP2 domains of a same protein that were analyzed separately. For reference numbers and references to the gene sequences used, see Methods.

AINTEGUMENTA are defective in ovule development (Elliot *et al.* 1996; Klucher *et al.* 1996). The *GLOSSY15* gene of maize has recently been shown to be an AP2-like gene that functions to repress adult leaf characters in juve-

nile leaves (Moose and Sisco, 1997). It thus seems that the analysis of the expression patterns of sugarcane AP2 homologues could help in elucidating the role of these homologues in flower development.
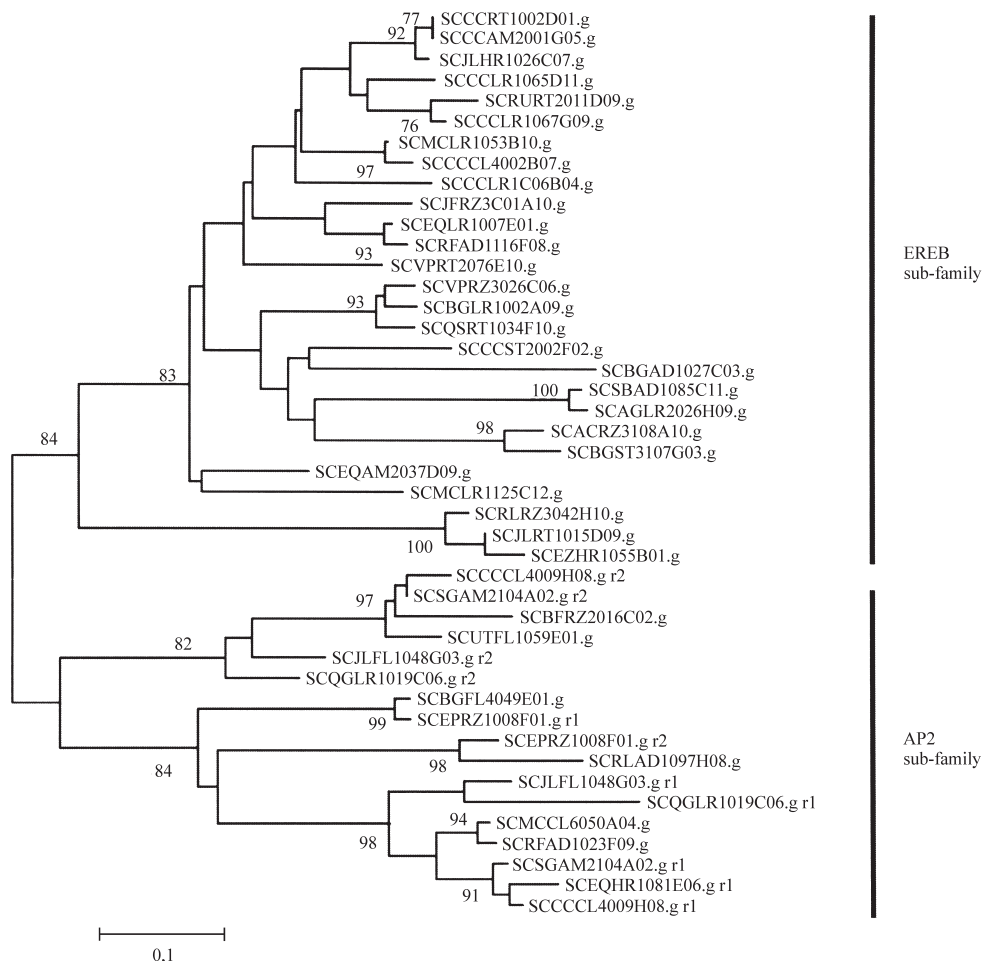
Figure 5 - Phylogenetic relationships among sugarcane nucleotide sequences coding for the AP2 domain. The most parsimonious consensus tree had its branches scaled to the corresponding genetic distances. Only Bootstrap values (1000 replicates) higher than 75% are shown.

RESUMO

Como na maior parte das gramíneas, num determinado momento do seu ciclo de vida, o meristema vegetativo da cana-de-açúcar é convertido em meristema reprodutivo. Em cana-de-açúcar há pelo menos duas conversões meristemáticas distintas entre a indução para o florescimento e a formação do florete. Em espécies dicotiledôneas modelo, a conversão sucessiva das identidades dos meristemas, bem como o arranjo concêntrico de órgãos florais são controlados geneticamente. Todos os genes e/ou proteínas sabidamente envolvidos no desenvolvimento floral foram anotados e identificados no banco de dados do SUCEST (Sugarcane EST Project). Comparações de seqüências entre genes reconhecidamente envolvidos no controle do desenvolvimento floral revelaram a conservação evolutiva entre os mecanismos de formação do padrão de desenvolvimento floral entre mono- e dicotiledôneas, bem como entre as gramíneas. Nossos estudos se concentraram na análise das famílias multigênicas dos fatores de transcrição do tipo MADS-box e AP2, uma vez que estes têm um papel importante na regulação do desenvolvimento reprodutivo vegetal. Também são apresentadas considerações sobre a genética evolutiva do desenvolvimento das flores de gramíneas e sua relação com o modelo ABC do desenvolvimento floral.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and **Lipman, D.J.** (1990). Basic local alignment search tool. *J. Mol. Biol. 2*15: 403-410.

Baum, D.A. (1998). The evolution of plant development. *Curr. Opin. Plant Biol. 1*: 79-86.

Borner, R., Kampmann, G., Chandler, J., Gleissner, R., Wisman, E. and **Apel, K., Melzer, S.** (2000). A MADS domain gene involved in the transition to flowering in *Arabidopsis. Plant J.* 24: 591-599.

Bowman, J.L., Alvarez, J., Weigel, D., **Meyerowitz, E.M.** and **Smyth, D.R.** (1993). Control of flower development in

Arabidopsis thaliana by *APETALA1* and interacting genes. *Development 119*: 721-743.

Chuck, G., Meeley, R. and **Hake, S.** (1998). The control of maize spikelet meristem fate by the APETALA2-like gene indeterminate spikelet1. *Genes Dev. 12*: 1145-1154.

Coen, E.S. and **Meyerowitz, E.M.** (1991). The war of whorls: genetic interactions controlling flower development. *Nature 353*: 31-37.

Elliot, R.C., Betzner, A.S., Huttner, E., Oakes, M.P. and **Tucker, W.Q.J.** (1996). AINTEGUMENTA, an APETALA2-like gene of *Arabidopsis* with pleiotropic roles in ovule development and floral organ growth. *Plant Cell 8*: 155-168.

Greyson, R.I. (1994). *The development of flowers*. Oxford University Press, Oxford.

Hartmann, U., Hohmann, S., Nettesheim, K., Wisman, E., Saedler, H. and **Huijser, P.** (2000). Molecular cloning of SVP: a negative regulator of the floral transition in *Arabidopsis*. *Plant J. 21*: 351-60.

Hempel, F.D. and **Feldman, L.J.** (1994). Bi-directional inflorescence development in Arabidopsis thaliana: Acropetal initiation of flowers and basipetal initiation of paraclades. *Planta 192*: 276-286.

Jofuku, K.D., van den Boer, B.G.W., van Montagu, M. and **Okamuro, J.K.** (1994) Control of Arabidopsis flower and seed development by the homeotic gene APETALA2. *Plant Cell 6*: 1211-1225.

Kang, H.G., Jeon, J.S., Lee, and **An, G.** (1998). Identification of class B and class C floral organ identity genes from rice plants. *Plant Mol. Biol. 38*: 1021-1029.

Klucher, K.M., Chow, H. Reiser, L. and **Fisher, R.L.** (1996). The AINTEGUMENTA gene of Arabidopsis required for ovule and female gametophyte development is related to the floral homeotic gene APETALA2. *Plant Cell 8*: 137-153.

Le Roux, L.G. and **Kellogg, E.A.** (1999). Floral development and the formation of unisexual spikelets in the Andropogoneae (Poaceae). *Am. J. Bot. 86*: 354-366.

McSteen, P., Laudencia-Chingcuanco, D. and **Colasanti, J.** (2000). A floret by another name: control of meristem identity in maize. *Trends in Plant Sci. 5*: 61-66.

Mena, M., Ambrose, B.A., Meeley, R.B., Briggs, S.P., Yanofsky, M.F. and **Schmidt, R.J.** (1996) Diversification of C-function activity in maize flower development. *Science 274*: 1537-1540.

Modrusan, Z.L., Reiser, K.A., Feldman, R.L., Fischer, R. and **Haughn, G.W.** (1994). Homeotic transformation of ovules into carpel-like structures in Arabidopsis. *Plant Cell 6*: 333-349.

Moore, P.H. (1987). Anatomy and Morphology. In: *Sugarcane Improvement Trough Breeding; Developments in Crop Science*, v. 11 (Heinz, D.J., ed.). Elsevier, New York, pp. 85-142.

Moose, S. and **Sisco, P.** (1997). Glossy15, an Apetala2-like gene from maize that regulates leaf epidermal cell identity. *Genes Dev. 10*: 3018-3027.

Ohme-Takagi, M. and **Shinshi, H**. (1995). Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell 7*: 173-182.

Purugganan, M.D., Rounsley, S.D., Schmidt, R.J. and **Yanofsky, M.F.** (1995) . Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics 140*: 345-356.

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

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

Schmidt, R.J. and **Ambrose, B.A.** (1998). The blooming of grass flower development. *Curr. Opin. in Plant Biol. 1*: 60-67.

Schultz, E.A. and **Haughn, G.W.** (1993). Genetic analysis of the floral initiation process (FLIP) in *Arabidopsis*. *Development 119*: 745-765.

Sheldon, C.C., Burn, J.E., Perez, P.P., Metzger, J., Edwards, J.A., Peacock, W.J. and **Dennis, E.S.** (1999). The FLF MADS box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *Plant Cell 113*: 445-58.

Stebbins, G.L. (1957). Self-fertilization and population variability in higher plants. *American Naturalist 91*: 337-354.

Swofford, D.L. (1990). PAUP: Phylogenetic Analysis using Parsimony, Version 3.1. University of Illinois, Champaign, IL.

Theissen, G., Strater, T., Fischer, A. and **Saedler, H.** (1995). Structural characterization, chromosomal localization and phylogenetic evaluation of two pairs of AGAMOUS-like MADS-box genes from maize. *Gene 156*: 155-166.

Theissen, G. Kim, J.T. and **Saedler, H**. (1996). Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. *J. Mol. Evol. 43*: 484-516.

Theissen, G. Becker, A., Di Rosa, A., Kanno, A.,Kim, J.T., Münster, T., Winter, K.U. and **Saedler, H.** (2000). A short history of MADS-box genes in plants. *Plant Mol. Biol. 42*: 115-149.

Thompson, J.D., Higgins, D.G. and **Gibson, T.J.** (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res. 22*: 4673-4680.