Phylogenetic relationships between *Arabidopsis* and sugarcane bZIP transcriptional regulatory factors

Michel Vincentz [1,2]*, Paulo S. Schlögl[1], Luis Gustavo G. Corrêa[1], Fabiana Kühne[1] and Adilson Leite[1]

Abstract

We built a complete and non-redundant database of bZIP transcriptional regulatory factors from the *Arabidopsis* reference genome. These *Arabidopsis* bZIP factors were ordered into thirteen families of evolutionary related proteins and this classification was used to identify and organize sugarcane cDNAs encoding bZIP proteins. We also show how this classification should help in defining putative clusters of orthologous groups of higher plant bZIP regulators and briefly discuss the expected benefits of this procedure to efficiently characterize sugarcane bZIP transcriptional regulators.

INTRODUCTION

Growth and development of all organisms largely relies on appropriate regulation of gene expression. Differential gene expression mainly occurs through the control of transcription initiation rates by transcriptional regulatory factors. These factors are usually defined as sequence-specific DNA binding proteins that recognize regulatory sequences in the promoter of a gene and are capable of modulating transcription (Holstege and Young, 1999; Kornberg, 1999 and Singh, 1998). Transcriptional regulators can be grouped into families (or super families) of related proteins according to the structural or primary sequence similarities of their DNA binding domain (Riechmann *et al.*, 2000; Wingender *et al.*, 2000).

The basic leucine zipper (bZIP) transcriptional regulatory factors have been described in all eukaryotes. Their DNA binding domain consists of a region rich in basic amino acids that binds to DNA and a so-called leucine zipper that consists of several heptad repeats of hydrophobic residues and which causes dimerization. The X-ray structure of the yeast GCN4 bZIP domain complexed to DNA target sites has shown that the bZIP is completely α-helical in structure. The two leucine zippers are packed in a coiled-coil structure for dimerization, while the basic regions of the dimer fits into the major groove of the half-sites of the target DNA (Hurst, 1995).

Genetic, molecular and biochemical studies indicate that the bZIP factors of higher plants are important regulators of plant specific processes such as fotomorphogenesis (Osterlund *et al.*, 2000); organ development (Walsh *et al.*, 1997; Chuang *et al.*, 1999); cell elongation and morphogenesis (Yin *et al.*, 1997; Fukazawa *et al.*, 2000); control of nitrogen to carbon balance during seed development (Cice-ri *et al.*, 1999); defense mechanisms (Niggeweg *et al.*, 2000; Zhang *et al.*, 1999); sucrose signalling (Rook *et al.*, 1998) and the response to hormones (Choi *et al.*, 2000; Finkelstein et Lynch, 2000; Uno *et al.*, 2000; Niggeweg *et al.*, 2000) and light (Schindler *et al.*, 1992; Wellmer *et al.*, 1999).

With the sequencing of the *Arabidopsis thaliana* (*Arabidopsis*) genome, a possible complete higher plant gene index was described (The *Arabidopsis* Genome Initiative, 2000). This repertoire of genes is likely to be representative of all higher plant genes that carry out essential functions and it therefore constitutes a invaluable reference data set which will help to better understand the evolution of cellular and developmental processes of higher plants.

Within this context, we initiated a comprehensive characterization of higher plant bZIP factors and we describe here, the generation of a probable complete and non redundant set of 72 bZIP factors encoded by the reference *Arabidopsis* genome (see also Riechmann *et al.*, 2000). A phylogenetic classification of this set of factors was established using conditions that were used previously to assess the phylogenetic relationships of 50 higher plant bZIP factors (Vettore *et al.*, 1998). We show how this classification has allowed us to efficiently characterize sugarcane expressed sequence tags (ESTs) encoding bZIP proteins and illustrate how this classification can be used to identify putative clusters of orthologous groups of higher plant bZIP factors including sugarcane bZIP genes. It is expected that defining such clusters should be useful in rationalizing the systematic characterization of higher plant bZIP proteins and more specifically sugarcane bZIPs.

---

[1]*Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, 13081-970 Campinas, SP, Brazil.*
[2]*Departamento de Genética e Evolução, IB, Universidade Estadual de Campinas, 13081-970 Campinas, SP, Brazil.*
Send correspondence to Michel Vincentz. E-mail: mgavince@obelix.unicamp.br.

RESULTS AND DISCUSSION

Phylogenetic classification of *Arabidopsis* bZIP transcriptional regulatory factors

A complete and non-redundant set of *Arabidopsis* bZIP factors was built from the NCBI GenBank and protein databases and MIPS MATDB accessions. The amino acid sequences of the bZIP domain of four accessions were further edited based on amino acids sequences alignments (BAB02051; AAD23721; T06089 and AAF67360) and one new putative bZIP protein not yet annotated at MATDB or GeneBank was identified (At2gBZN). Three proteins with a truncated basic region or leucine zipper were not included in our database, the total number of proteins in our database being 72.

The evolutionary relationships between the members of our *Arabidopsis* bZIP proteins collection was evaluated by phylogenetic analysis of the aligned amino acids sequences of their bZIP domain (Figure 1). The unrooted tree inferred from neighbor-joining analysis of the bZIP domain data set is shown in Figure 2. Based on the branching pattern, the tree was resolved into thirteen families. Most of the families show moderate to strong bootstrap support. Concerning families VI and VII, which are poorly resolved, we noticed that all members of these two families, as well as the genes of families IV and V, form a group of bZIP genes without introns. We also noticed that all members of several families share partially identical exon-intron gene organization (data not shown), supporting the pattern of clustering defined here. Finally, the bZIP protein AAG51519 does not fit into any of the Familie, although we included it into Family X based on its blastp best hit with proteins of Family X.

Index of sugarcane bZIP factors

The ordered set of *Arabidopsis* bZIP regulators was used to efficiently detect and classify sugarcane contigs encoding bZIP transcriptional regulators. In a first step, one or two query sequences consisting of full-length protein sequence of each of the 13 *Arabidopsis* bZIP families (Figure 2) were utilized to screen the SUCEST database, candidate sugarcane contigs being selected based on the presence of at least one conserved protein motif among several members of each *Arabidopsis* bZIP family. In a second step, selected sugarcane contigs were included into one of the *Arabidopsis* families according to their blastp best hit. Our strategy allowed us to identify 121 sugarcane contigs encoding candidate bZIP transcription factors. The pattern of

Basic Region	Leucine Zipper

KRQKRKQSNRESARRSRLRKQAECEQLQQRVESLSNENQSLRDE

Figure 1 - Amino acid sequence of the bZIP domain used in the phylogenetic analysis of *Arabidopsis* bZIP factors. The sequence is from the *Arabidopsis* GBF1 regulator. Leucines of the leucine zipper are underlined.

distribution of the sugarcane contigs among the 13 *Arabidopsis* families is shown in Figure 3. No sugarcane contig related to Families IV and XIII were detected. The interpretation of this pattern is not straightforward but we suggest that it may reflect the number of genes included in each *Arabidopsis* family and/or the expression level of sugarcane genes related to each of these families.

Putative clusters of orthologous groups of monocot and dicot bZIP factors

To further characterize the sugarcane bZIP factors we initiated a comparative analysis to identify Putative Clusters of Orthologous Groups (PCOG) of higher plants bZIP factors. A Cluster of Orthologous Group (COG) consists of individual orthologous genes or orthologous groups of paralogs from several completely sequenced genomes (Tatusov *et al.*, 1997). The term ortholog refers to homologous genes that have been created by a speciation event, *i.e.* are versions of the same gene in different organisms, and paralogs are homologous genes that result from a duplication event within a genome (Tatusov *et al.*,1997 and Thornton and DeSalle, 2000). Orthologs usually retain the same function, whereas paralogs can explore new functions. An important consequence of defining COGs is that it allows to predict with some confidence the structure and function of uncharacterized members of the COG.

To detect PCOGs of bZIP factors of higher plants, we built a data set consisting of all monocot and dicot bZIP protein sequences avalaible in GenBank plus the reference database formed by the 13 *Arabidopsis* bZIP families (Figure 2). The neighbor-joining distance method (Saitou and Nei, 1987) was used to identify the PCOGs. Several of the situations we encountered are illustrated in Figure 4. A simple PCOG consisting of individual putative orthologs which includes the maize regulator *Liguleless2* is shown in Figure 4A. The simplest interpretation of this PCOG is that the *Arabidopsis* AAF22906 and the sugarcane II.8 proteins are functionally related to the maize regulator *Liguleless2* involved in maize leaf development (Walsh *et al.*, 1997).

Several PCOGs with more complex relationships between members are shown in Figures 4B and 4C. For instance, PCOG 1 of family XII (Figure 4C) can be described as an orthologous group of two *Arabidopsis* and two sugarcane paralogs. Cluster 1 of family VIII (Figure 4B) is even more complex. It consists of a putative group of *Arabidopsis* /monocot orthologs (*Arabidopsis* AAF67360, maize OHP1, rice REB, barley BLZ1 and the sugarcane VIII.4 proteins) and one group of monocot orthologs (maize, *Coix* and *Sorghum* Opaque2 regulators).

We noticed that some *Arabidopsis* bZIP factors are encoded by genes that are part of two co-linear genomic sequences formed by several highly similar genes. Such proteins are therefore likely to be paralogs that originated with the large-scale chromosomal duplications that formed the
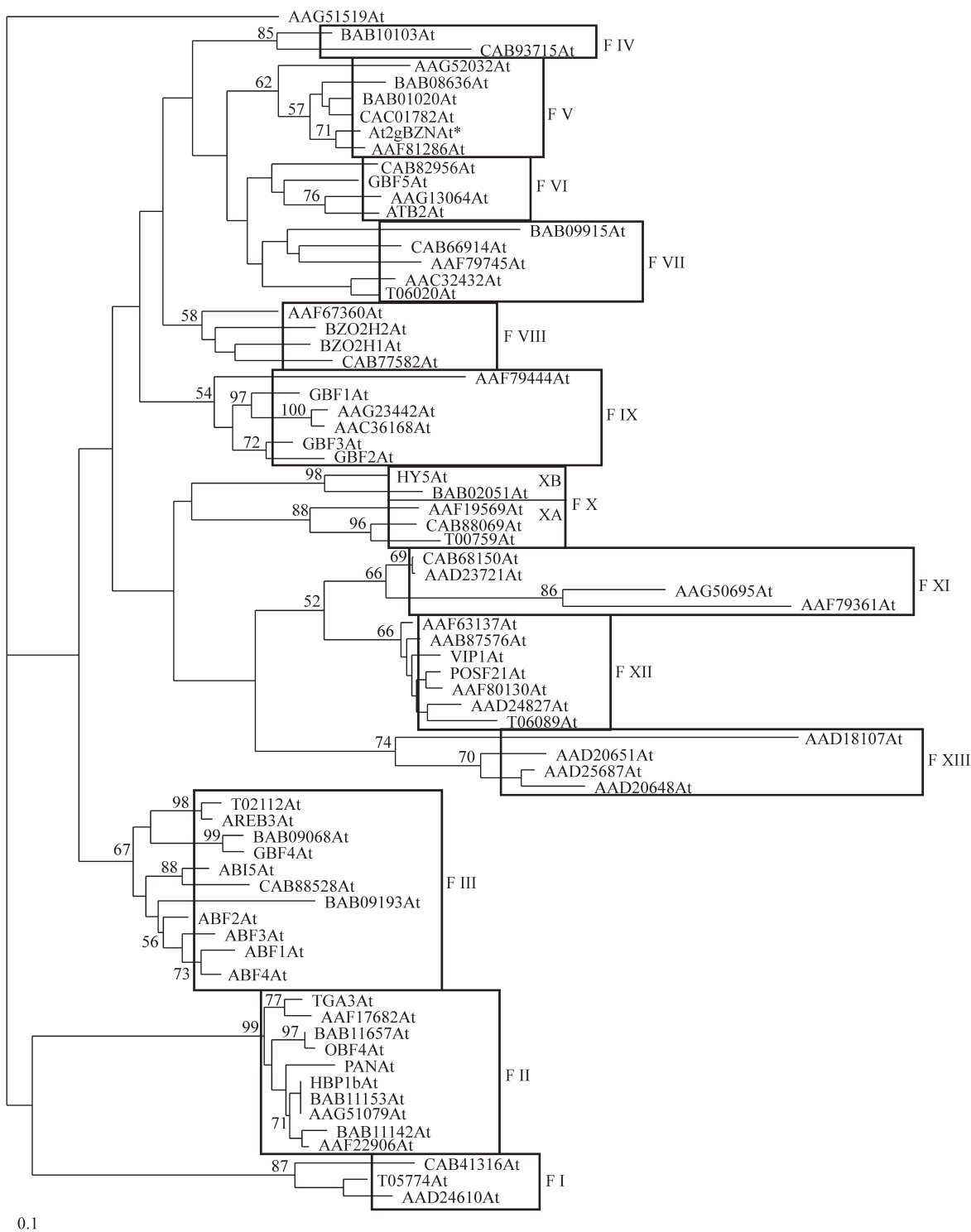
Figure 2 - Phylogenetic relationships of the complete set of *Arabidopsis* bZIP transcription factors. Unrooted tree of the *Arabidopsis* bZIP factors inferred from a neighbor-joining analysis of bZIP domain amino acid sequences data set (Figure 1). The tree was organized into thirteen families (F I to F XIII). Bootstrap of 1000 replicates is indicated as percentages along the branches when higher than 50%. In most case the proteins are identified by the accession number. Accession numbers of proteins with a name are given in Materials and Methods. At2gBZNAt* is a bZIP protein not yet annotated. The scale bar corresponds to 0.1 estimated amino acid substitution per site.

Arabidopsis genome (The *Arabidopsis* Genome Initiative, 2000; Vision *et al*., 2001). For instance, POSF21 and AAF80130 in PCOG 3 of Family XII (Figure 4C), are encoded by genes that are part of two co-linear segments of at least six genes on chromosome II and I, respectively (Result not shown). These two Arabidopsis bZIPs paralogs are
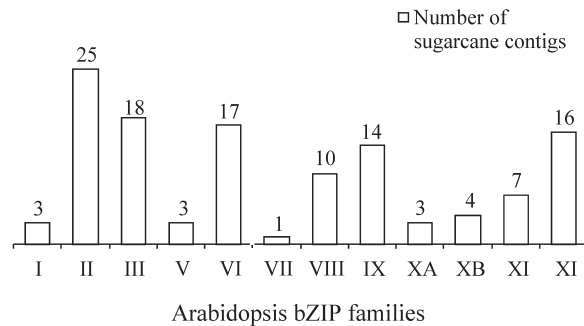
Figure 3 - Distribution of sugarcane contigs encoding bZIP factors among the thirteen Arabidopsis Families defined in Figure 2.

closely related to the rice RF2A that seems to be important for differentiation of leaf cells (Yin *et al.*, 1997). It remains to be shown whether or not that they are functionally related to RF2A and also to what extent they are redundant.

The polyploid origin of the sugarcane genome (Daniels and Roach, 1987) may prevent us distinguishing sugarcane paralogs from allelic forms of the same locus. However, this complexity should not hamper our ability to reach reasonable conclusions about the clustering pattern and functional inference. For example, it is difficult to infer whether or not the two sugarcane contigs XII.4 and XII.6 in PCOG 2 (Figure 4C) are two alleles of the same gene or not, while contig XII.7 could be a corresponding paralog (Figure 4C). However, a clear orthologous relationships between these three sugarcane bZIP proteins and the *Arabidopsis* protein VIP1 can be proposed (Figure 4C).

Based on the strategy described in this paper, we are now organizing all higher plant bZIP factors into PCOGs and hope to use this information to further characterize sugarcane bZIP transcriptional regulators.

MATERIALS AND METHODS

The non redundant data set of *Arabidopsis* bZIP factors was obtained through iterated searches of the GenBank and protein database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) and the Munich information center for protein sequences (MIPS) *Arabidopsis thaliana* database (MATDB, http://www.mips.biochem.mpg.de/proj/thal/) using different known bZIP query sequences and the blastp and tblastn programs (Altschul *et al.*, 1990) at the NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) and the MIPS (http://mips.gsf.de/ proj/thal/db/search/search_frame.html) servers. Additionally, with the recent publication of the *Arabidopsis* genome (The *Arabidopsis* Genome Initiative, 2000), a key word search was also performed at MATDB (v211200).

Sugarcane contigs (a contig or cluster is a consensus sequence derived from several overlapping and highly similar ESTs sequences) coding for bZIP proteins were detected by using *Arabidopsis* full-length bZIP protein sequence as query sequences to screen the SUCEST (sugar-
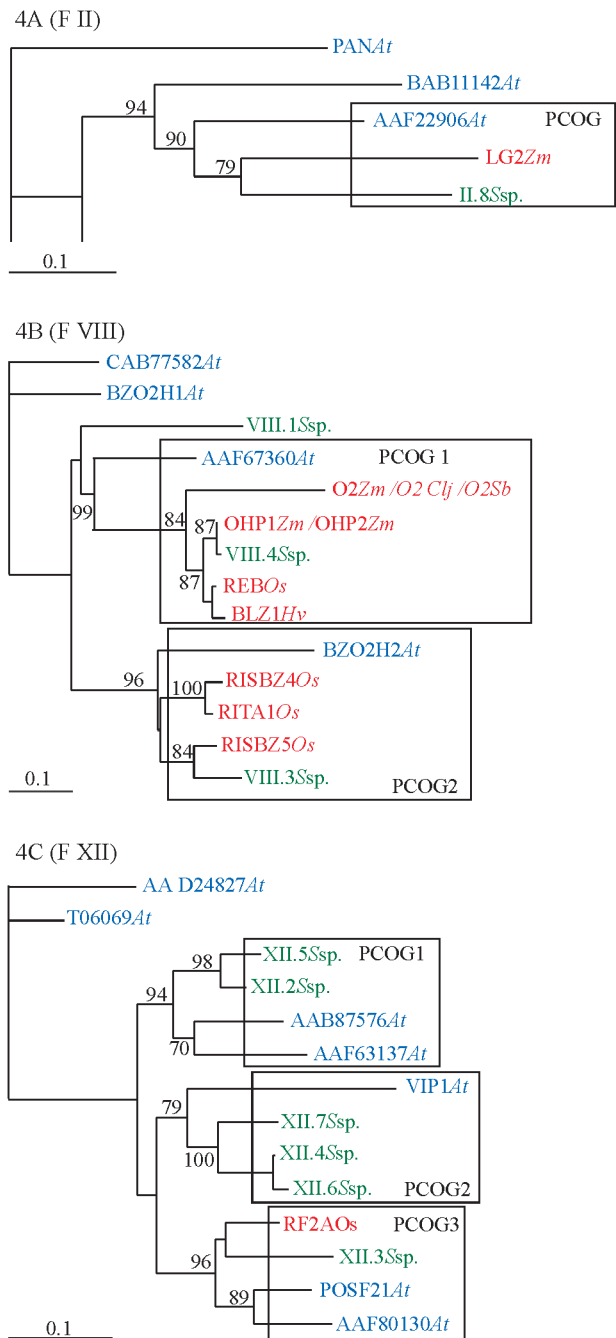

Figure 4 - Identification of Putative Clusters of Orthologous Groups of monocot and dicot bZIP factors (PCOG). Aligned amino acid sequence data set of monocot and dicot bZIP proteins belonging to the reference Arabidopsis Family II (F II; 4A), Family VIII (F VIII; 4B) and XII (F XII; 4C) were built and the corresponding unrooted trees inferred from a neighbor-joining analysis of these data sets are shown. Only part of the tree corresponding to Family II is shown (4A). Arabidopsis proteins are shown in blue, monocot proteins are shown in red and sugarcane proteins are shown in green. Accession numbers are given in Materials and Methods. Species abbreviations are: *At*, *Arabidopsis thaliana*; *Clj*, *Coix lacryma-jobi*; *Hv*, *Hordeum vulgare*; *Os*, *Oryza sativa*; *Sb*, *Sorghum bicolor*; *S*sp, *Saccharum* sp. and *Zm*, *Zea mais.* Bootstrap support is indicated as percentage of 1000 replicates. The scale bar corresponds to 0.1 estimated amino acid substitution per site.

cane EST genome project) database (http://www.sucest. lbi.dcc) with the locally available tblastn program. Sugarcane contigs consists of one to several overlapping and highly similar EST reads assembled with the PHRAP program (P.Green, htpp://bozeman.mbt.washington.edu/ phrap.docs/phrap.html; -penalty -15 -bandwidth 14 -minscore 100 -shatter_greedy, Meidanis personnal communication).

Protein sequences were aligned with the CLUSTALX program (Thompson *et al.*, 1997). Amino acid sequence data was analyzed by the neighbor-joining method (Saitou and Nei, 1987) using the NEIGHBOR program (PHYLIP, Phylogeny Inference Package version 3.57c; Felsenstein, 1993) and PAM distances (Dayhoff *et al.,* 1978), obtained with the PRODIST program (PHYLIP). Bootstrap assessment of tree topology in neighbor-joining analysis was performed with the SEQBOOT program (PHYLIP). Trees were displayed with the TREEVIEW program (Page, 1996). DNA sequence analysis was carried out with the DNASIS program (Pharmacia). Motifs conserved among members of each *Arabidopsis* bZIP family (Figure 2) were detected with the help of the MEME program (Bailey and Elkan, 1994; http://meme.sdsc.edu/meme/website/).

The accession numbers of the *Arabidopsis* bZIP proteins shown in Figure 2 are: Family II: PAN, AAD49979; OBF4, CAA49524; HBP1b, BAB11154; TGA3, S46523. Family III: ABF4 (AREB2), AAF27182; ABF3, AAF27181; ABF1, AAF27179; ABF2 (AREB1), AAF27180; AREB3, BAB12406; GBF4, P42777; ABI5, AAD21438. Family VI: GBF5, AAG17474; ATB2,T05279. Family VIII: BZO2H1, AAG25727; BZO2H2, AAG25728. Family IX: GBF1, P42774; GBF2, P42775; GBF3, P42776. Family X: HY5, BAA21116. Family XII: VIP1, AAF37279; POSF21, AAD26486. The accession numbers of bZIP factors shown in Figure 4 are: LG2*Zm* (*liguleless2*), AAC39351; RF2A*Os*, AAC49832; O2*Clj*, S42493; O2*Sb*, CAA50642; O2*Zm*, A34800; OHP1*Zm*, JQ2147; OHP2*Zm*, JQ2148; REB*Os*, BAA36492; BLZ1*Hv*, T04477; RITA1*Os*, T03990. The accession numbers for RISBZ4*Os* and RISBZ5*Os* sequences are not yet available but can be found in Onodera *et al.* (2001). The accession numbers at SUCEST for the Sugarcane contigs (clusters) shown in Figure 4 are: II.8, SCCCCL3001H09.g; VIII.1, SCVPRZ3030A04.g; VIII.3, SCCCCL4005C09.g; VIII.4, SCQSRT1036D12.g; XII.2, SCCCRT2002D06.g; XII.3, SCJFRZ2034D03.g; XII.4, SCAGHR1016H08.g; XII.5, SCSFLR2016F09.g; XII.6, SCACRZ3034F03.g; XII.7, SCJLAM1064H01.g.

ACKNOWLEDGMENTS

NOTE ADDED IN PROOF

Since we submitted this article for publication the protein At2gBZN in Figure 1 was annotated as At2g04038 at MATDB.

RESUMO

Construímos um banco de referência não redundante de fatores de regulação da transcrição do tipo bZIP a partir de dados do genôma de *Arabidopsis thaliana*. Os fatores bZIP de *Arabidopsis* foram ordenados em treze famílias de proteínas evolutivamente relacionadas e essa classificação foi usada para organizar os cDNAs de cana de açúcar que codificam proteínas bZIP. Além disso, mostramos que essa classificação poderá ser útil para definir "Putative Clusters of Orthologous Groups" de reguladores bZIP de plantas superiores.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and **Lipman, D.** (1990). Basic Local Alignment Tool. *J. Mol. Biol. 215*: 403-410.

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

Choi, H., Hong, J., Há, J., Kang, J. and **Kim, S.Y.** (2000). ABFs, a family of ABA-responsive element binding factors. *The Journal Biol. Chem. 275*: 1723-1730.

Chuang, C-F., Running, M.P., Williams, R.W. and **Meyerowitz, E.M.** (1999). The Perianthia gene encodes a bZIP protein involved in the determination of floral organ number in *Arabidopsis thaliana*. *Genes and Development 13*: 334-344.

Ciceri, P., Locatelli, F., Genga, A., Viotti, A. and **Schmidt, R.J.** (1999). The activity of the maize Opaque-2 transcriptional activator is regulated diurnally. *Plant Physiol. 121*: 1321-1327.

Daniels, J. and **Roach, B.T.** (1987). Taxonomy and evolution. In *Sugarcane improvement through breeding* (Heinz, D.J., ed.). Elsevier Press, Amsterdam, pp. 7-87.

Dayhoff, M.O., Schwartz, R.M. and **Orcutt, B.C.** (1978). In *Atlas of protein sequence and structure,* v. 5, Suppl. 3 (Dayhoff, M.O., ed.). National Biochemical Research Foundation, Silver Spring, MD: 345-352.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA.

Finkelstein, R.R. and **Lynch, T.J.** (2000). The *Arabidopsis* Abscisic acid response gene ABI5 encodes a basic leucine zipper transcription factor. *The Plant Cell 12:* 599-609.

Fukazawa, J., Sakai, T., Ishida, S., Yamaguchi, I., Kamiya, Y. and **Takahashi, Y.** (2000). Repression of shoot growth, a bZIP transcriptional activator, regulates cell elongation by controlling the level of gibberellins. *The Plant Cell 12*: 901-915.

Holstege, F.C.P. and **Young, R.A.** (1999). Transcriptional regulation: contending with complexity. *Proc. Natl. Acad. Sci. USA 96*: 2-4.

Hurst, H. (1995). Transcription factor 1: bZIP proteins. *Protein Profile 2*: 105-168.

Kornberg, R.D. (1999). Eukayotic transcriptional control. *Trends in Genetics 15*: 46-49.

Niggeweg, R., Thurow, C., Kegler, C. and **Gatz, C.** (2000). Tobacco transcription factor TGA2.2 is the main component of as-1-binding factor ASF-1 and is involved in salicylic acid-and auxin-inducible expression of as-1-containing target promoters. *The Journal Biol. Chem. 275*: 19897-19905.

Osterlund, M.T., Wei, N. and **Deng, X.W.** (2000). The roles of photoreceptor system and the COP1-targeted destabilization of HY5 in light control of *Arabidopsis* seedling development. *Plant Physiol. 124*: 1520-1524.

Page, R.D.M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences 12*: 357-358.

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

Rook, F., Gerrits, N., Kortstee, A., van Kampen, M., Borrias, M., Weisbeek, P and **Smeekens, S.** (1998). Sucrose-specific signalling represses translation of the *Arabidopsis* ATB2 bZIP transcription factor gene. *The Plant Journal 15*: 253-263.

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

Schindler, U., Menkens, A.E., Beckmann, H., Ecker, J.R. and **Cashmore, A.R.** (1992). Heterodimerization between light-regulated and ubiquitously expressed *Arabidopsis* GBF bZIP proteins. *EMBO J. 11*: 1261-1273.

Singh, K.B. (1998). transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiol. 118*: 1111-1120.

Tatusov, R.L., Koonin, E.V. and **Lipman, D.J.** (1997). A genomic perspective on protein families. *Science 278*: 631-637.

The *Arabidopsis* Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature 408*: 796-815.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and **Higgins, D.G.** (1997). The CLUSTALX windows interface: flexible strategies for multiple alignment aided by quality analysis tools. *Nucl. Acids Res. 25*: 4876-4882.

Thornton, J. and **DeSalle, R.** (2000). Gene family evolution and homology: genomics meets phylogenetics. *Annu. Rev. Genomics Hum. Genet. 1*: 41-73.

Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinozaki, K. and **Yamaguchi-Shinozaki, K.** (2000). *Arabidopsis* basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proc. Natl. Acad. Sci. USA 97*: 11632-11637.

Vettore, A.L., Yunes, J.A., Cord Neto, G., da Silva, M.J., Arruda, P. and **Leite, A.** (1998). The molecular and functional characterization of an Opaque2 homologue gene from *Coix* and a new classification of plant bZIP proteins. *Plant Mol. Biol. 36*: 249-263.

Vision, T.J., Brown, D.G. and **Tanksley, S.D**. (2001). The origin of genomic duplications in *Arabidopsis. Sience 290*: 2114-2117

Walsh, J., Waters, C.A. and **Freeling, M.** (1997). The maize gene liguleless2 encodes a basic leucine zipper protein involved in the establishment of the leaf blade-sheath boundary. *Genes and Development 11*: 208-218.

Wellmer, F., Kircher, S., Rügner, A., Frohmeyer, H., Schäfer, E. and **Harter, K.** (1999). Phosphorylation of the parsley bZIP transcription factor CPRF2 is regulated by light. *The Journal of Biol. Chem. 274*: 29476-29482.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I. and **Matys, V.** (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res. 28*: 316-319.

Yin, Y., Zhu, Q., Da, S., Lamb, C. and **Beachy, R.** (1997). RF2a, a bZIP transcriptional activator of the phloem-specific rice tungro bacilliform virus promoter, function in vascular development. *EMBO J. 16:* 5247-5259.

Zhang, Y., Fan, W., Kinkeman, M., Li, X. and **Dong, X.** (1999). Interaction of NPR1 with basic leucine zipper protein transcription factor that bind sequences required for salicylic acid induction of the PR-1 gene. *Proc. Natl. Acad. Sci. USA 96*: 6523-6528.