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

Michel Vincentz 1,2*, Paulo S. Schlögl1, Luis Gustavo G. Corrêa1, Fabiana Kühne1 and Adilson Leite1

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 (Ciceri et al., 1999); defense mechanisms (Niggegeweg 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; Niggegeweg 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 an 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.

1Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, 13081-970 Campinas, SP, Brazil.
2Departamento 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 (A12gBZN). 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 families, 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 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 available 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
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...
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 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 ot 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-
cane EST genome project) database (http://www.sucest.lib.de) 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, http://bozeman.mbt.washington.edu/phrap.docs/phrap.html; -penalty -15 -bandwidth 14 -minscore 100 -shatter_greedy, Meidanis personal 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, Phylony 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; HPB1, 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, T04477; RITA1, BAB12406; GBF2, P42775; GBF3, P42776. Family X: HY5, BAA21116. Family XII: VIF1, AAF37279; POSF21, AAD26486. The accession numbers of the bZIP factors shown in Figure 4 are: LG2Zm (liguleless2), AAC39351; RF2A0s, AAC49832; O2Clj, S42493; O2Sh, CAA50642; O2Zm, A34800; OHP1Zm, QJ2147; OHP2Zm, QJ2148; REB0s, BAA36492; BLZ1Hv, T04477; RITA1Os, T03990. The accession numbers for RISBZ4Os and RISBZ5Os 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: IL8, SCCCL3001H09.g; VIII.1, SCVRZ3030A04.g; VIII.3, SCCCL4005C09.g; VIII.4, SCQSRT1036D12.g; XII.2, SCCCRT2002D06.g; XII.3, SCJFRZ2034D03.g; XII.4, SCAGHR1016H08.g; XII.5, SCFSFLR2016F09.g; XII.6, SCACRZ3034F03.g; XII.7, SCCLAM1064H01.g.

ACKNOWLEDGMENTS

This work was supported by grant from Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) / Auxílio à Pesquisa Nº 1999/02839-9. PSS and LGGC are supported by grants from FAPESP. We thank an anonymous referee for helpful comments.

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 genoma 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


