470–478 Nucleic Acids Research, 1994, Vol. 22, No. 3

Novel conserved sequence motifs in plant G-box binding proteins and implications for interactive domains

Iris Meier+ and Wilhelm Gruissem*
Department of Plant Biology, University of California, Berkeley, CA 94720, USA

Received August 27, 1993; Revised and accepted December 17, 1993 EMBL accession nos X74941–X74943 (incl.)

ABSTRACT

The G-box is a cis-acting DNA sequence present in several plant promoters that are regulated by diverse signals such as UV irradiation, anaerobiosis, abscissic acid and light. Several basic/leucine zipper (bZIP) proteins from different plant species have been identified as high affinity G-box binding proteins. Although their capability to enhance transcription has been demonstrated, their precise function in transcriptional activation is still unknown. We have isolated three cDNAs from young tomato fruit that encode bZIP G-box binding proteins (GBF4, GBF9 and GBF12). They bind to the G-box sequence in the tomato *rbcS1*, *rbcS2* and *rbcS3A* promoters. GBF9 binding resulted in a DNase I footprint identical to that obtained with tomato nuclear extract and different from the DNase I protection obtained with GBF4 and GBF12. The mRNAs of all three GBFs were most abundant in tomato fruit and seeds, moderately abundant in root and least abundant in leaves. Protein sequences outside of the bZIP domains were compared with the known GBFs from other plants and seven conserved motifs of seven to 35 amino acids length have been identified. Based on the presence of these motifs, three classes of GBFs can be defined that are conserved among plant species. GBF9, the predominantly expressed tomato GBF, is the first member of its class isolated from dicot plants. Three conserved motifs from two of the classes are highly hydrophilic and are predicted to be exposed on the surface of the proteins. These motifs likely define novel interactive domains in the different classes of GBFs that could provide a new tool to determine how distinct regulatory signals are transmitted through GBFs to activate transcription.

INTRODUCTION

Plant promoters activated by diverse stimuli such as UV irradiation, anaerobiosis, abscissic acid and light share a DNA sequence with the core motif CACGTG, called G-box [1], box II [2] or Em1a [3]. The DNA sequence element is required for full activity of the various promoters. Binding studies with nuclear extracts have shown that proteins interact with this sequence in different plant species [1, 4, 5]. Different cDNAs encoding proteins that bind to oligonucleotides containing the CACGTG sequence (hereafter called G-box) or closely related sequences have been cloned [3, 6, 7, 8, 9]. The G-box binding factors (GBFs) are members of a family of bZIP proteins that contain a basic DNA binding domain and a leucine-zipper motif [10, 11]. Outside of the bZIP domain and an N-terminal proline-rich region shared by some of these proteins, they show little similarity to each other. cDNAs for more than one type of G-box binding protein (GBF) have been isolated from different plants, e.g., three different GBFs have been identified that can bind the *rbcS* G-box in Arabidopsis [6]. There are three GBF-like proteins in parsley (*CPRFs*) that bind the box II element in the chalcone synthase promoter [8], and two GBF-like proteins from wheat (*HBPs*) interact with the related hexamer sequence element C-ACGTCA in the promoter of the histone H3 gene [12]. These results suggest that families of related bZIP proteins exist in most plants which recognize the CACGTG motif, most likely including those plants for which at present only one GBF has been reported [7].

Two lines of evidence demonstrate a general transcriptional activation function for G-box binding factors. First, the tobacco protein TAF-1 enhances transcription from a chimeric CaMV '–90' 35S promoter fused to six copies of the TAF-1 binding site GGTACGTGGC [7]. Second, fusion of the proline-rich N-terminal fragment of the Arabidopsis factor GBF1 to the DNA binding domain of the yeast GAL4 protein can activate transcription from a promoter containing the GAL4 binding site in mammalian cells [13]. These results support a model for GBF function in plants first established by cis-analyses, suggesting that GBFs have a general enhancing activity which is necessary for full activity of the respective promoters.

Considering the multiplicity of GBF-like proteins in higher plants, two questions emerge regarding their potential function in transcriptional activation. First, how is the transmission of the different regulatory signals linked to GBFs? And second, are all GBFs within one plant species functionally equivalent or do specific GBFs link the transcriptional machinery to different signal transduction pathways?

*To whom correspondence should be addressed

+Present address: Institut für Allgemeine Botanik, Angewandte Molekularbiologie der Pflanzen, Ohnhorststr. 18, D-2000 Hamburg 52, Germany

We are addressing the above questions in the context of the *rbcS* gene family in tomato which we use as a model system to dissect mechanisms by which plants establish spatial and temporal patterns of gene expression. Of the five *rbcS* genes in tomato, three (*rbcS1*, *rbcS2* and *rbcS3A*) contain a G-box motif in their promoters [14]. Run on transcription experiments have shown that *rbcS1*, *rbcS2* and *rbcS3A* are transcribed at high levels in leaves, where their transcription is light regulated [15]. In contrast, only *rbcS1* and *rbcS2* are transcribed at a significant level in young, developing tomato fruit [15]. The region of the G-box, however, is bound by a protein in all three promoters in both organs [16, 17]. Because GBF can act as a general transcriptional activator in short range interaction with the basic transcriptional machinery [7, 13], the above results suggest that a G-box binding protein in the *rbcS3A* promoter may be negatively regulated or that DNA binding is unproductive. A DNA sequence motif (F-box) immediately upstream of the G-box in the *rbcS3A* promoter binds specifically to a protein from young fruit nuclear extract but no binding is detected with leaf nuclear extract. This interaction results in a DNase I footprint that is contiguous with the G-box protection (Meier and Gruissem, unpublished results; [16, 17]). The immediate proximity of this DNA-binding site (F-box) could be of importance for the regulation of GBF function on the *rbcS3A* promoter in the young fruit.

We have cloned three different cDNAs for GBF-like proteins from young tomato fruit as part of our effort to understand how GBF function can be regulated in response to developmental and environmental signals. Among the different GBFs represented by the cDNAs, we have identified one protein which appears to interact specifically with the G-box in the *rbcS* promoters. All tomato GBFs have similar expression patterns, with their expression being highest in young fruit and seeds. Comparison of all plant GBFs reveals conserved amino acid sequence motifs outside of the bZIP and proline-rich domains which allows them to be grouped into different classes. We propose that these conserved motifs represent domains by which GBF activity can be regulated in response to different signals.

METHODS

DNAse I footprinting

RbcS promoter fragments were isolated from the respective plasmids and end-labeled as described by Manzara *et al.* [16]. Footprinting reactions contained 1 fmol ^{32}P labeled fragment and 4.5 μg *E. coli* extract. Binding was for 30 min. at room temperature. DNAse I reactions and gel-electrophoreses were performed according to Manzara *et al.* [16].

DNA sequence analysis

Double-stranded DNA was sequenced by the dideoxy chain termination method using a Sequenase 2.0 sequencing kit (Stratagene). Overlapping fragments were created either by subcloning of suitable restriction fragments into pBluescript (Stratagene) or by creating nested deletions, starting at the 5′ terminus of the cDNAs [18]. Nucleotide and amino acid sequence analysis was performed on a Macintosh computer using the MacVector program and the UWGCG program package [19]. The database search was done using the FASTA program.

Expression of fusion proteins in *E. coli*

pBluescript plasmids containing the cDNA inserts were rescued from λZAP II (Stratagene) and retransformed into *E. coli* XL1 blue (Stratagene). A 50 ml culture was grown at 37°C to $O.D._{600}$ = 0.2 and IPTG was added to a final concentration of 10 mM. Cultures were incubated at 37°C with vigorous shaking for 2 h and cells were harvested by centrifugation at 1600 g for 10 min. Cells were suspended in 2.5 ml 50 mM Tris−HCl, pH 7.5, 1 mM EDTA, 5 mM DTT, and 50 μg/ml PMSF, frozen in liquid N_2 and thawed at room temperature. Five hundred ml of 10 μg/ml lysozyme in 10 mM Tris−HCl, pH 8.0, was added and the samples were incubated for 15 min on ice. After centrifugation for 30 min at 20,000 g, 4°C the supernatants were dialyzed three times for 1 h against 20 mM Hepes, pH 7.6, 40 mM NaCl, 0.2 mM EDTA, 20% glycerol, and 1 mM DTT. The dialyzed protein extracts were aliquoted, frozen in liquid N_2 and stored at −80°C. The proteins are fusion proteins, containing 7 amino acids of β-galactosidase and 29 amino acids derived from the pBluescript polylinker at the N-terminus (Stratagene).

Preparation of nuclear extracts and mobility shift assays

Nuclear extracts were prepared from young, fully expanded leaves and 3 − 8 mm fruit as described by Manzara *et al.* [16]. Reactions for mobility shift assays contained 3.2 ng G-box DNA fragment (α-^{32}P end-labeled by filling in staggered ends with sequenase 2.0 enzyme), 1.2 μg of poly(dIdC), and 2 μg nuclear protein in 20 mM HEPES, pH 7.6, 40 mM NaCl, 1 mM EDTA, 20 % glycerol, 1 mM DTT in a total volume of 12 ml and were incubated for 30 min at room temperature. Reactions were loaded onto a 5% polyacrylamide gel (19:1 acrylamide to bisacrylamide in 0.5×TBE [18]. Electrophoresis was in 0.5×TBE at 10 V/cm. For the binding competition assays 320 ng unlabeled double-stranded oligonucleotides were added to the binding reactions.

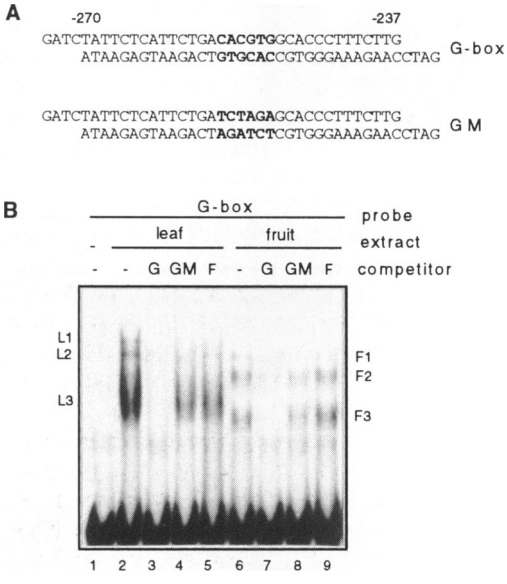

Figure 1. A G-box binding activity is present in young tomato fruit. **(A)** Sequence of the oligonucleotides G-box and GM used for the mobility shift assays. The central 6 bp of the G-box motif and the respective basepairs in the mutant are highlighted. The numbers refer to the position of the sequence with respect to the transcriptional start site of *rbcS 3A*. **(B)** Mobility shift assay of the G-Box oligonucleotide with nuclear extracts from tomato leaves and young fruit. G: G-box oligonucleotide, GM: GM-oligonucleotide, F: unrelated oligonucleotide F (GATCCTTTTAGGATG AGATAAGACTATTCTCATTCTGA). L1, L2, L3, F1, F2 and F3: protein-DNA complexes formed with leaf- and fruit-extract, respectively.

RNA isolation and RNA gel blot analysis

Total RNA was prepared from various tomato tissues as described [20]. Roots were harvested from hydroponically grown VFNT cherry LA1221 tomato plants. Leaf and young fruit tissue were harvested from greenhouse grown VFNT cherry LA1221 tomato plants. Leaves were young, fully expanded leaves, young fruit were 3−8 mm in diameter. Seeds were mature seeds. Ten mg of total RNA was separated on a 1.2 % agarose gel containing formaldehyde [18] and transfered to Hybond N membrane (Amersham) as described [20]. Prehybridization and hybridization were carried out in 0.5 M NaH_2PO_4/Na_2HPO_4, 7% SDS, 1 mM EDTA, pH 7.2, at 65°C. Filters were washed for 10 min at 65°C in 2×SSC, 0.2%SDS, followed by 3 washes of 10 min each at 65°C in 0.2×SSC, 0.2 % SDS. DNA fragments were labeled by random prime labeling using α^{32}P-dCTP and purified through push columns (Stratagene). DNA probes were the 1.2 kb, 1.2 kb and 1.0 kb EcoR1−Xho I fragments of pGB4, pGB9 and pGB12, respectively.

Screening of a cDNA expression library from tomato fruit

An amplified λZAP II cDNA expression library constructed with polyadenylated mRNA from 3−8 mm large VFNT cherry LA1221 tomato fruit was provided by Dr. Jonathon Narita. The library was screened for proteins interacting with the *rbcS3A* G-box sequence as described by Singh *et al.* [21] with the following modifications: Denaturation of the filterbound proteins was performed in 6M Guanidinium HCl in buffer B (20 mM HEPES pH 7.6, 40 mM NaCl, 0.2 mM EDTA, 1 mM DTT). The proteins were renatured by stepwise dilution in buffer B. The binding reaction was performed in buffer B supplemented with 0.25% non-fat dry milk, 10 μg/ml denatured salmon sperm DNA and 10 ng/ml DNA probe (10^6 cpm/ml), synthesized and labeled as described by Sambrook *et al.* [18]. The binding reaction was carried out for 4 h at 4°C. Filters were washed three times 10 min. in buffer B + 0.25 % non-fat dry milk at 4°C. Nine positive plaques were carried through four subsequent rounds of purification. Plaque pure clones were probed with both the *rbcS3A* G-box and the unrelated oligonucleotide A2 (GATCT-CAAAACCAACCTCAATCATACATTCATATCCTCTTCG). Four out of nine clones showed sequence specific binding to the *rbcS3A* G-box oligonucleotide and were characterized further.

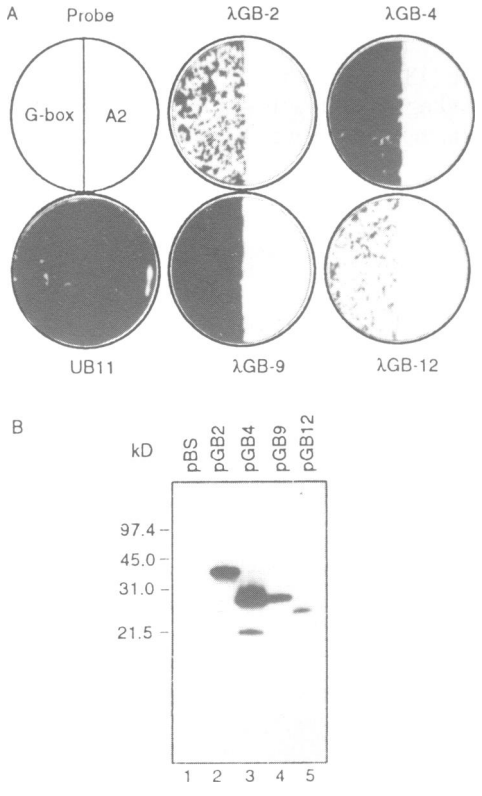

Figure 2. **(A)** Sequence specificity of the DNA-binding proteins encoded by λGB-2, λGB-4, λGB9 and λGB12. Filters containing protein expressed by the respective phage were cut in halves and incubated with either the G-box oligonucleotide or the oligonucleotide A2 as described in Methods. The phage UB11 expresses a non-specific DNA binding protein that was used as a control. **(B)** Expression of the G-box binding proteins in *E. coli*. Each lane contains 20 μg of protein that was separated on SDS-PAGE, transfered to nitrocellulose and incubated with the G-box oligonucleotide as described in Methods. The sizes of the molecular weight markers are indicated on the left. pBS: Protein extract from *E. coli* cells containing the vector alone.

'South-western' blot analysis

Twenty μg *E. coli* protein was separated on 12% SDS-PAGE, blotted onto nitrocellulose using an electroblotting chamber, denatured in 6 M Guanidinium-HCl in 20 mM HEPES pH 7.6, 40 mM NaCl, 0.2 mM EDTA, 1 mM DTT (buffer B) and renatured by stepwise dilution in buffer B as described for the library screen. The filter was blocked overnight in 5% non-fat dry milk in buffer B. The binding reaction was carried out in 10 ml buffer B + 0.25 % non-fat dry milk with 10^6 cpm/ml DNA probe synthesized and labeled as described by Sambrook *et al.* [18], and 10 μg/ml salmon sperm DNA at 4°C for 2 h. The filter was washed 3 times for 10 min. in buffer B + 0.25% non-fat dry milk. Autoradiography was for 2 h at −80°C with an intensifying screen.

Standard cloning techniques

Basic molecular biology techniues were carried out according to Sambrook *et al.* [18].

RESULTS

A GBF activity interacts with the *rbcS3A* G-box in young tomato fruit

Previous analysis of DNA-protein interactions in the tomato *rbcS* promoters has established a protection of the *rbcS1*, *rbcS2* and *rbcS3A* G-box sequence (16, 17). We confirmed by gel mobility shift analysis that the DNase I footprint over the G-box motif observed with young fruit nuclear extract indeed represents a G-box binding protein. Using an oligonucleotide containing the DNA sequence for the *rbcS3A* G-box (Figure 1A), protein-DNA complexes were formed with nuclear extracts from leaves and young fruit (Figure 1B). Leaf nuclear extract showed a diffuse retarded band (L3), possibly consisting of more than one complex, as well as two minor bands of reduced mobility (L1 and L2, lane 2). Only the L3 complex was specifically competed by a 100-fold excess of the G-box oligonucleotide (lane 3). Although the signal of L1 and L2 was reduced as well, their complete competition required higher concentrations of the G-box oligonulceotide (data not shown). A 100-fold excess of the mutated G-box oligonucleotide GM or the unrelated oligonucleotide F showed less or no competition for L3 (lanes 4 and 5). With fruit nuclear extract, three complexes were formed (F1, F2 and F3, lane 6). Only the fastest migrating complex F3

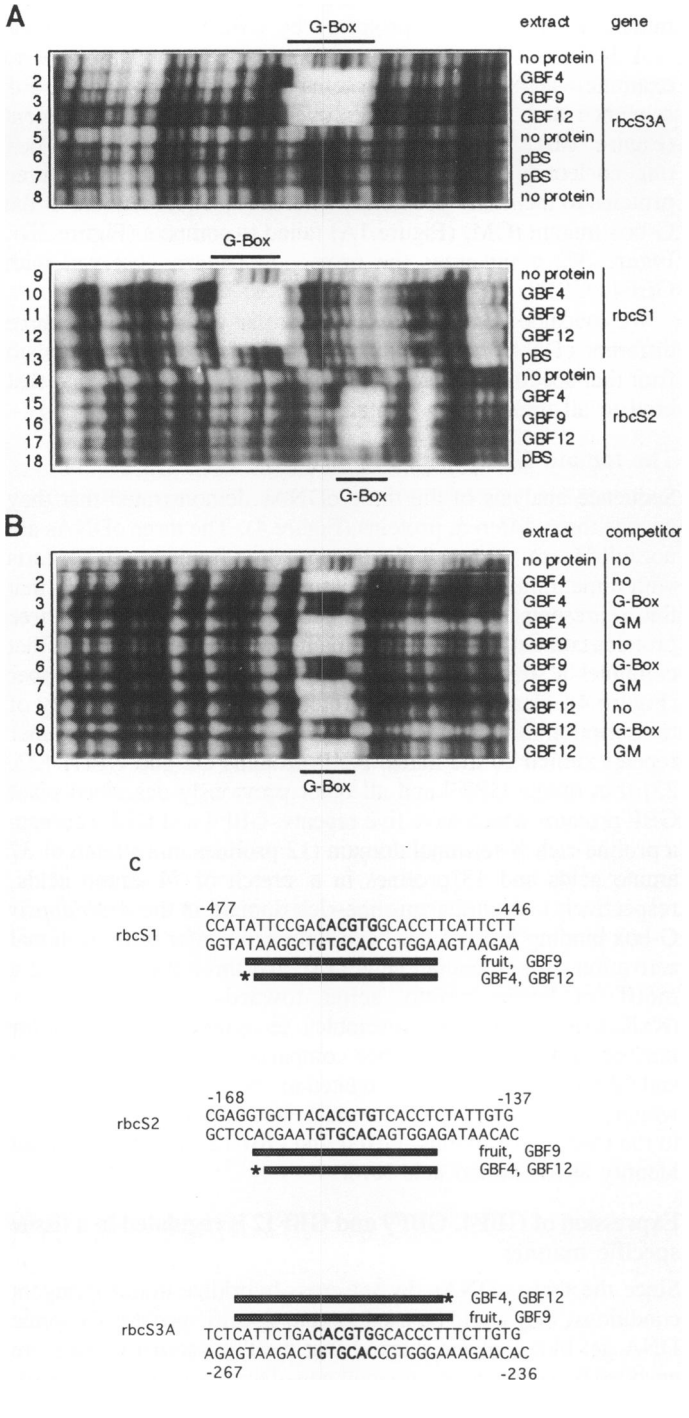

Figure 3. GBF4, GBF9 and GBF12 bind to the G-box in all three *rbcS* promoters and require the CACGTG motif for binding. **(A)** DNAse I footprinting of the *rbcS 1*, *rbcS2*, and *rbcS3A* promoters with protein extracts from the *E. coli* strains expressing GBF4, GBF9 and GBF12. The non-coding strand of *rbcS1* and *rbcS2* and the coding strand of *rbcS3A* are shown. The position of the G-box in the three promoters is indicated by black bars. Reactions contained 1 fmol of fragment. 4.5 μg of protein was added as indicated. pBS: Protein extract from cells carrying the vector only. **(B)** Sequence-specific competition of the binding of GBF4, GBF9 and GBF12 to the *rbcS1* G-box. Reactions were carried out with 1 fmol fragment and 4.5 mg protein. If indicated, a 5000 fold molar excess of either G-box or GM oligonucleotide (Figure 1) was added to the reactions. The noncoding strand is shown. The position of the G-box is indicated by a black bar. **(C)** Comparison of the sequences in the three promoters that are protected by GBF4, GBF9, GBF12 and young fruit nuclear extract [17]. Numbers indicate the distance from the transcriptional start site. Bars indicate the protected regions and stars the enhanced cleavage sites. The central CACGTG of the G-box is highlighted.

(and to a lesser extent F1 and F2) was competed specifically by the G-box oligonucleotide, but not by the GM or F oligonucleotides (lanes 7, 8 and 9). Thus, L3 and F3 represent specific G-box binding activites in leaf and fruit, respectively. The complex F3 differed from the respective complex detected with leaf extract (L3) in its mobility. The additional low mobility bands detectable with leaf (L1 and L2) and fruit nuclear extracts (F1 and F2) are likely complexes of multiple G-box binding activites interacting with the G-box oligonulceotide.

Isolation of cDNAs encoding G-box binding proteins from young tomato fruit

A cDNA expression library was screened for proteins capable of binding to the *rbcS3A* G-box oligonucleotide (Figure 1A) to identify cDNAs encoding for G-box binding proteins in young tomato fruit. Nine positive phage were purified to homogeneity. Plaque filter binding assays with the *rbcS3A* G-box oligonucleotide as well as the unrelated oligonucleotide A2 (see Methods) demonstrated that the four proteins encoded by the phage λGB-2, λGB-4, λGB-9 and λGB-12 bind specifically to the G-box oligonucleotide (Figure 2A). The other five phage clones encoded proteins which bind equally well to both

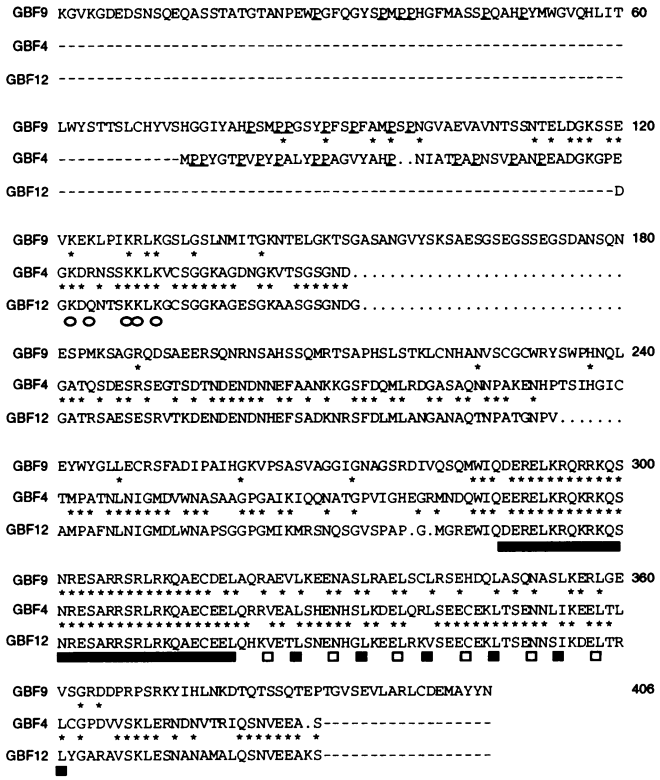

Figure 4. Amino acid sequences of GBF4, GBF9 and GBF12. Alignment of the amino acid sequences of the three tomato GBF proteins is shown. The proline residues in the amino-terminal proline-rich domains of GBF4 and GBF9 are underlined. The basic region and the leucine or isoleucine residues at positions 7 of the leucine zipper motifs are marked by black bars, and the hydrophobic residues at positions 4 of the leucine zipper motifs are marked by open bars. The basic amino acids in the predicted nuclear import signal are indicated by ovals. Stars indicate amino acid identities. Gaps introduced for optimal alignment are indicated by dots. Numbers correspond to the amino acid positions in GBF9. The EMBL accession numbers for DNA sequences of the cDNAs are X74941 (GBF12), X74942 (GBF4) and X74943 (GBF9).

sequences and are likely non-specific DNA-binding proteins (e.g. UB11, Figure 2A). The plasmids containing the cDNA inserts of λGB-2, λGB-4, λGB-9 and λGB-12 were rescued and protein extracts were prepared from *E. coli* cells containing the rescued plasmids pGB2, pGB4, pGB9 and pGB12 or the vector pBS alone. Filter binding assays were performed with the electrophoretically separated proteins and the labeled, concatamerized G-box oligonucleotide (see Methods). A G-box binding protein was detected in each of the bacterial extracts, with the sizes of 35 kD, 28 kD, 28 kD and 25 kD for pGB2, pGB4, pGB9 and pGB12, respectively. No specific binding was detected with an extract derived from cells carrying the vector only (Figure 2B). The approximately 21 kD protein detected with pGB4 is likely a shorter translation or proteolytic product of the protein encoded by the pGB4 cDNA.

Reciprocal filter hybridization of the four cDNA inserts showed that they represent three different DNA sequences because only pGB9 and pGB2 cross-hybridize under high stringency conditions (data not shown). Therefore, only the larger clone pGB2 was included in the following experiments. The proteins encoded by the cDNAs of pGB4, pGB12 and pGB2 were named GBF4, GBF12 and GBF9, respectively.

DNase I footprint analysis using the *E. coli* extracts was carried out to establish that the proteins encoded by the isolated cDNAs have binding properties similar to that of the protein detected in tomato nuclear extract. Figure 3A shows that GBF4, GBF9 and GBF12 expressed in *E.coli* protect a region centered over the G-box of the *rbcS3A* promoter similar to that observed with nuclear extract from young fruit [17]. In contrast to GBF9 which protected the exact number of cleavage sites as the nuclear extract (Figure 3A, lane 3), GBF4 and GBF12 showed a shorter protection at the 3′ end of the footprint, enhancing the cleavage of the last nucleotide protected by GBF9 (Figure 3A, lane 2, lane 4). No protection was observed with the *E. coli* extract from control cells transformed with the vector alone (Figure 3A, lane 6 and lane 7).

To investigate whether the proteins bind equally well to the G-box in the different sequence contexts of the *rbcS1* and *rbcS2* promoters, DNaseI footprint experiments with these promoter fragments were performed. As shown in Figure 3A, all three proteins protected the respective sequences in the two promoters.

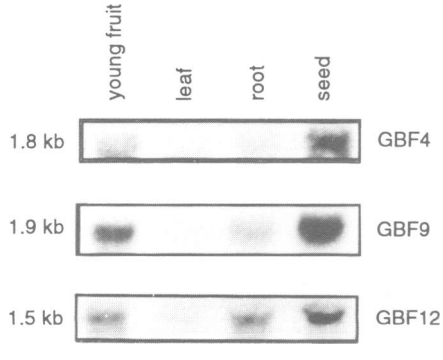

Figure 5. Expression pattern of GBF4, GBF9 and GBF12 mRNAs. Ten μg total RNA from seed, root, leaf and young fruit was hybridized to the cDNAs encoding GBF4, GBF9 and GBF12. The approximate size of the RNAs is indicated on the left. Equal loading of RNA was monitored with a DNA probe for ribosomal RNA (Wanner and Gruissem, 1991) in control hybridizations (not shown).

Again, GBF4 and GBF12 showed an enhancement of the outermost cleavage site protected by GBF9 and plant extract.

A DNAse I footprint competition experiment was done to examine whether the three proteins indeed recognize the G-box sequence and not a different, overlapping sequence element (Figure 3B). A 5000-fold excess of the *rbcS3A* G-box oligonucleotide successfully competed for the binding of the three proteins to the *rbcS1* promoter, whereas an equal amount of the G-box mutant (GM) (Figure 1A) failed to compete (Figure 3B). Figure 3C summarizes the protection pattern observed with GBF4, GBF9, GBF12 and fruit nuclear extract.

We conclude from these experiments that we have isolated three different cDNAs coding for proteins present in young tomato fruit that specifically interact with the G-box sequence and that bind to all three G-box containing *rbcS* promoters *in vitro* .

The tomato GBFs are bZIP proteins

Sequence analysis of the three cDNAs demonstrated that they encode three different proteins (Figure 4). The three cDNAs are not full-length. Although the open reading frame of GBF4 starts with a methionine, sequence comparison makes it unlikely that this represents the start of the protein (see below). All three proteins contain a conserved domain towards the C-terminus that combines a stretch of basic amino acids with a leucine-zipper (Figure 4), classifying them as members of the bZIP family of transcription factors [8]. GBF4 and GBF12 contain six heptamer repeats similar to the maize bZIP proteins O2 and OPH1 [22, 23], but unlike GBF9 and all other previously described plant GBF proteins which have five repeats. GBF4 and GBF9 contain a proline-rich N-terminal domain (12 prolines in a stretch of 37 amino acids and 13 prolines in a stretch of 74 amino acids, respectively). A similar proline-rich domain in the *Arabidopsis* G-box binding factor 1 has been shown to confer transcriptional activation in mammalian cells [13]. All three proteins share a motif of basic amino acids towards the N-terminus (KxKxxxKK/RLK) that resembles recognition sequences for nuclear import [24]. Sequence comparison revealed that GBF4 and GBF12 are more closely related to one another than to GBF9 which, outside of the bZIP domain, has no sequence similarity to the two other proteins. GBF4 and GBF12 show 62% overall identity at the amino acid level.

Expression of GBF4, GBF9 and GBF12 is regulated in a tissue specific manner

Since the three cDNAs do not cross-hybridize under stringent conditions, they could be used as gene specific probes. Genomic DNA gel blot analysis showed that the three tomato GBFs are encoded by single genes or small gene families (data not shown). RNA gel blot analysis was performed using total RNA from young fruit, leaf, root and seeds to determine their organ-specific expression (Figure 5). GBF4 and GBF9 detect mRNAs of approximately 1.8 kb and 1.9 kb in size, respectively, whereas GBF12 hybridizes to a smaller mRNA species of approximately 1.5 kb. The mRNAs from the three genes accumulate to their highest levels in seed and fruit. The mRNA levels are significantly lower in roots andsubstantially reduced in leaves. GBF9 showed the highest mRNA accumulation relative to GBF4 and GBF12.

Different classes of G-box binding proteins are conserved between species

To date, no function has been attributed to any part of G-box binding proteins outside of the bZIP domain and the proline-rich

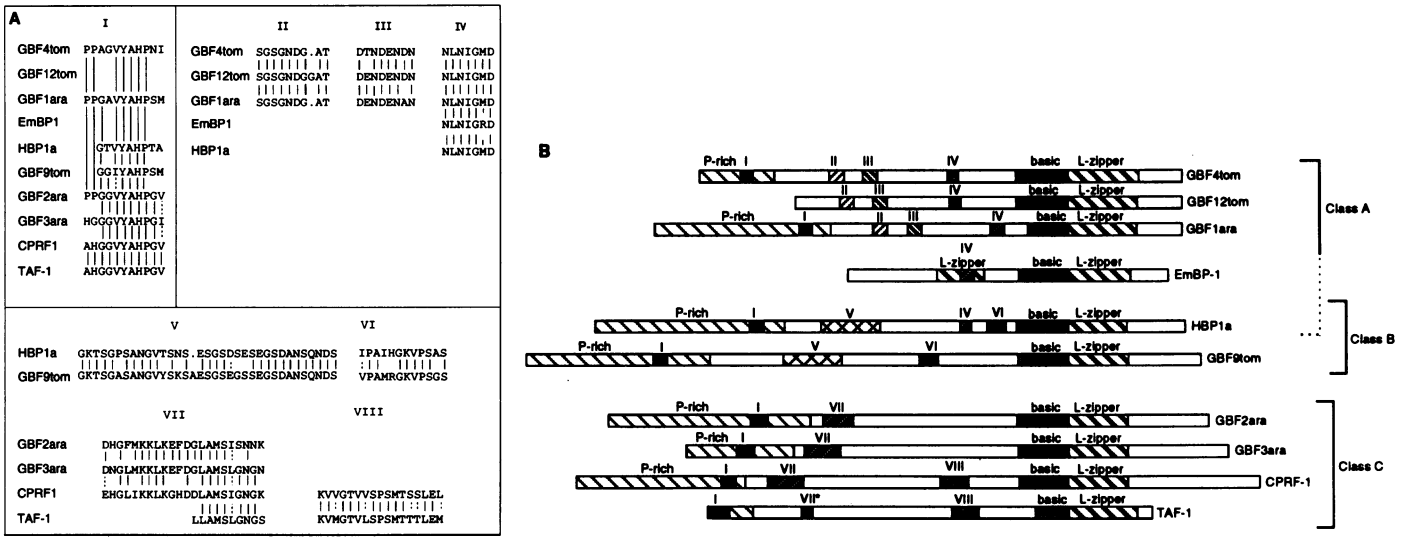

Figure 6. Conserved sequence motifs define different types of G-box binding proteins. **(A)** Peptide sequence motifs that are shared by two or more proteins are shown. Vertical bars indicate identical amino acids, broken vertical bars conservative substitutions. Gaps are indicated by dots. **(B)** Position of the conserved sequence motifs within the G-box binding proteins. The different proteins are aligned with respect to the junction between basic domain and leucine-zipper. The boxes indicate the position and size of the respective sequence element, the numbers refer to the numbering in (A). The grouping of the proteins in three different classes is indicated by brackets.

activation domain. We compared the sequence of the tomato GBFs with the sequence of all other plant GBFs to identify conserved amino acid sequence domains that could be of functional significance. These are GBF4, GBF9 and GBF12 from tomato (this paper), GBF1, GBF2 and GBF3 from *Arabidopsis* [6], TAF-1 from tobacco [7], CPRF-1, CPRF-2 and CPRF-3 from parsley [8] and HBP-1a [9] and EmBP-1 [3] from wheat. Plant bZIP proteins that do not have a high affinity to DNA sequence elements containing the CACGTG core sequence such as TGA-1a [25], HBP-1b [12] or O2 [22] were not included in this comparison.

Within the region between the proline-rich domain and the bZIP domain seven peptide motifs of seven to 35 amino acids were identified that are highly conserved between at least two different GBFs (Figure 6A). According to the presence of these motifs 10 of the 12 compared GBFs can be grouped into three distinct classes (class A, B, and C, Figure 6B). GBF4, GBF12 and GBF1 share motifs II, III, and IV, which are conserved in sequence as well as in spacing. The two monocot GBFs EmBP-1 and HBP-1a are related to this class by sharing motif IV. HBP-1a, however, is more closely related to the tomato protein GBF9. The two proteins share motifs V and VI, which are indicative of class B. Class C is represented by the two Arabidopsis proteins GBF2 and GBF3, CPRF-1 from parsley and TAF-1 from tobacco and is defined by the presence of motif VII or its shorter derivative VII*. Additionally, the two proteins CPRF-1 and TAF-1 share motif VIII. All GBFs that have been isolated as sufficiently long cDNA clones to contain sequence information for the proline-rich amino-terminus which contains a derivative of motif I with the core pentamer V/IYAHP. Sequence motifs II, III and V are characterized by a high percentage of polar amino acids and an acidic pI. The two G-box binding proteins CPRF-2 and CPRF-3 of parsley have neither similarity to any other GBFs nor to one another and were therefore not included in Figure 6B. Database searches revealed no significant similarities of motifs I through VIII to other known proteins.

We compared the position of the conserved motifs in the three tomato GBFs with the predicted secondary structure according to the combination of the Chou and Fasman and Robson and Garnier algorithms [26, 27] as well as the Antigenic Index surface probability calculation according to Jameson and Wolf [28] (Figure 7). There are no predicted α-helices and β-sheets spanning motifs II through VI. A long α-helical stretch spans the bZIP domain, consistent with the results from the crystal structure of the bZIP domain of the yeast protein GCN4 [10]. The Antigenic Index calculation predicts the basic region in all three proteins to be exposed on the surface (plotting above the graph axis), consistent with its function as the DNA binding surface. The leucine-zipper is predicted not to be exposed, consistent with it being a domain of hydrophobic interaction. The acidic motifs II and III in GBF12 as well as GBF4, and motif V in GBF9, are located within areas of the protein that are predicted to be surface-exposed. Motif IV in GBF4 and GBF12, and motif VI in GBF9, are predicted to be non-exposed.

The implications of this analysis are twofold. First, the conservation of different classes of GBFs between species suggests that each class has a distinct function with respect to the potential role in linking signal transduction chains to gene activation. Second, the results indicate a functional significance for a region of the protein other than the previously characterized activation and DNA binding/dimerization domains.

DISCUSSION

Differences in DNase I footprinting define specific binding

We have isolated cDNAs coding for three G-box binding bZIP proteins from young tomato fruit (GBF4, GBF9 and GBF12). The β-galactosidase fusion proteins of GBF4, GBF9 and GBF12

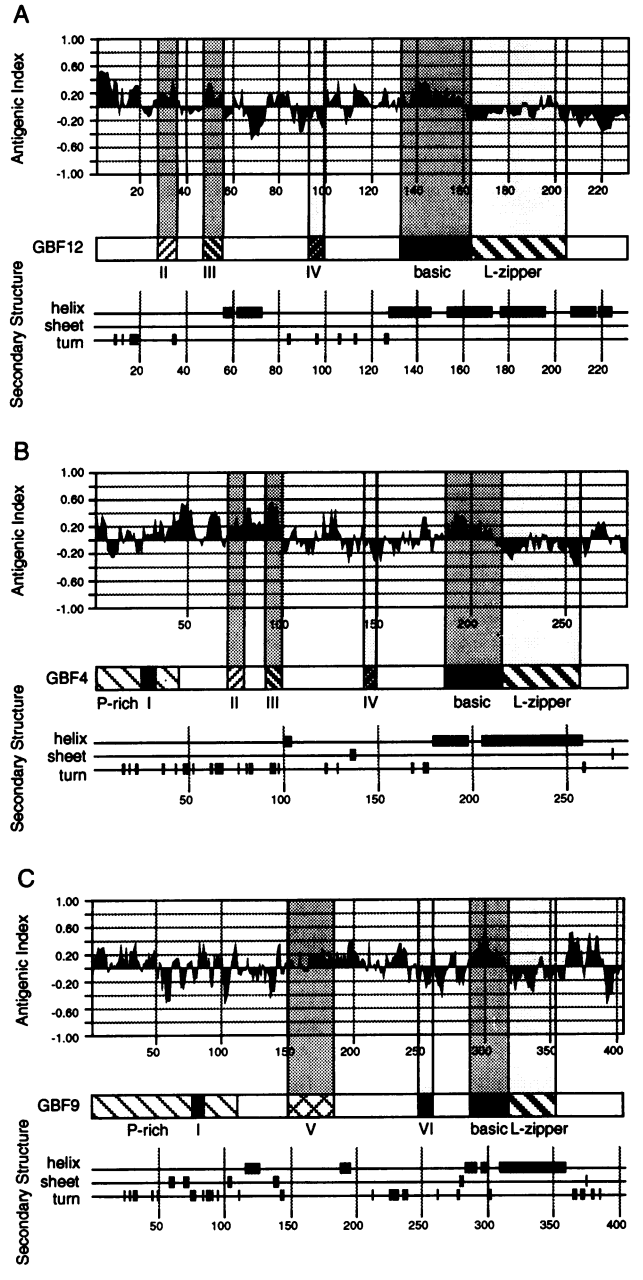

Figure 7. Structural features of the tomato GBFs. Antigenic Index analysis according to Jameson and Wolf [28] is shown in alignment with the schematic primary structure and the secondary structure prediction according to Chou and Fasman [26] and Robson and Garnier [27]. Boxed areas align the conserved peptide elements, the proline-rich amino-terminus and the basic and leucine-zipper domains (indicated by patternd boxes in the schematic primary structure in Figure 6), with the respective position in the Antigenic Index analysis. Shaded boxes represent regions that plot above the graph's zero x-axis and that are predicted to be exposed on the protein surface. (A) GBF12. (B) GBF4. (C) GBF9.

bind sequence-specifically to the G-box motif in the tomato *rbcS1*, *rbcS2* and *rbcS3A* promoters (Figure 3). The protection patterns obtained in DNaseI footprinting experiments with the recombinant proteins are similar for GBF4 and GBF12, but differ for GBF9. In all three promoters the footprint produced by GBF4 and GBF12 is characterized by the enhanced cleavage of one residue at the border of the protected region (Figure 3A and 3C). It is unlikely that this difference in protection pattern is the result of N-terminal truncated GBFs fused to β-galactosidase. For example, GBF4 and GBF12 share significant amino acid sequence similarities and produce identical DNaseI footprints, but GBF12 lacks the entire proline-rich N-terminal domain (Figure 4). Similarly, it has been shown for the yeast bZIP protein GCN4 that the isolated bZIP domain shows a binding pattern indistinguishable from the full length protein [10]. It is likely, therefore that the correlation between binding characteristics and sequence similarity indicates that GBF4/12 and GBF9 form structurally distinct complexes with the DNA. It has been reported previously that of the three *Arabidopsis* GBFs, GBF2 differs in its binding behavior from the two other proteins [6] suggesting that similar structurally (and functionally) distinct GBF-DNA complexes can be found in *Arabidopsis* as well.

The enhanced cleavage site detected with GBF4 and GBF12 in *E. coli* protein extracts was not observed with nuclear extracts from tomato [17]. Therefore, the DNAseI footprint pattern found for GBF9 resembles most closely the binding by proteins in the nuclear extract. This observation suggests that GBF9 comprises the largest fraction of active G-box binding factor that interacts with the *rbcS* promoters in tomato fruit nuclear extracts. Based on sequence comparison, the protein most closely related to GBF9 is the wheat histone H3 promoter binding protein HBP-1a. Tomato GBF9 is therefore the first member of this type of G-box binding protein isolated from dicot plants.

Expression pattern of tomato GBFs

The three tomato genes for GBF proteins show a similar expression pattern at the RNA level, with highest accumulation of mRNA in young fruit and mature seed, moderate expression in root and low expression in leaves (Figure 5). Of the three Arabidopsis GBFs, GBF1 and GBF2 mRNAs accumulate in light and dark grown leaves as well as in root, whereas GBF3 mRNA is detected only in dark grown leaves and root, indicating that expression of this gene is light-regulated [6]. TAF-1 mRNA is detected in tobacco root tissue, but not in leaf, stems [7] or mature seed [29]. This is consistent with a report that the high affinity TAF-1 binding site GCCACGTGGC enhances transcription of a GUS reporter gene in tobacco root but not in seed [29], thus making TAF-1 the best candidate for the active GBF protein that controls transcription in root, but not seed, of tobacco. The three tomato GBFs are similar to TAF-1 with respect to their high level of mRNA accumulation in root and low level in leaves. They differ from TAF-1, however, by their high mRNA levels in mature seed. It is therefore unlikely that one of the tomato GBFs is the functional equivalent of the tobacco TAF-1 protein.

The G-box of *rbcS1*, *rbcS2* and *rbcS3A* is protected by both leaf and young fruit nuclear extract as has been shown previously by DNase I footprinting [17]. The very low mRNA accumulation for GBF4, GBF9 and GBF12 in leaf therefore suggests that other, still unidentified GBFs are responsible for the protection seen in DNase I footprints in leaves as well as for the slower migrating complex detected in mobility shift assays using leaf nuclear extract (Figure 1). Alternatively, posttranscriptional processes could result in higher levels of GBF activity in leaves compared to the levels of mRNA accumulation from the respective genes. Different modifications or higher-order complex formation of the same factors in the two organs could then explain the different gel mobilities of the complexes.

Conserved sequence motifs define novel, potentially interactive domains in GBFs

The G-box DNA sequence motif has been identified in several promoters regulated by entirely different stimuli. In the wheat *Em* promoter two G-box like DNA sequences are located within a 75 bp fragment that activates transcription in response to ABA. Mutation of the upstream G-box (Em 1a) abolishes ABA mediated activation [3]. In the *CHS* promoter of parsley, a G-box DNA sequence is necessary for the activation by UV light, and the respective GBF was shown to bind to this sequence *in vivo* in response to UV light [2]. In the anaerobiosis-induced *Adh 1* promoter of Arabidopsis, a protein binds to a G-box DNA sequence *in vivo* [30, 31]. Mutagenesis of the G-box sequence results in a significant decrease in transient transcription from the *Adh-1* promoter [32]. Mutations in the G-box DNA sequence of the Arabidopsis *rbcS1-A* promoter leads to a significant decrease of promoter activity in transgenic tobacco plants [33].

In all of the above studies, an intact G-box is necessary for full promoter activity, which implies that binding of GBF is required for the activation of transcription. It is not known, however, if and how the different signals required for the activation of the respective promoter are transmitted through GBF. We have therefore examined the different plant GBF protein sequences for shared sequence determinants that could function as potential interactive or regulatory domains. The only domains within GBF proteins for which a function has been established are the proline-rich activation domain and the bZIP DNA-binding/dimerisation domain [11]. We have identified eight amino acid sequence motifs outside of these DNA-binding, dimerization and activation domains that are highly conserved between GBFs in different plant species (Figure 6). They are all located in the region between the proline-rich N-terminal domain and the bZIP domain. They are conserved not only with respect to their sequence but also with respect to their relative position within this domain, and therefore allow the grouping of GBFs into three classes. The evolutionary conservation of amino acid sequence motifs within proteins of a given class suggests that the members of the three classes represent functionally different proteins that could be regulated by different signal transduction pathways.

Of the conserved sequence motifs, motifs II and III of class A and motif V of class B GBFs (Figure 6A) are of particular interest because of their high proportion of acidic and polar residues. Acidic domains have been assigned functions as exposed surfaces in protein-protein interaction in a variety of transcription factors [34]. We have used the Antigenic Index calculation developed by Jameson and Wolf [28] to locate such exposed surfaces in GBF4, GBF9, and GBF12, and determined whether or not motifs II, III and V are located within these areas (Figure 7). This approach combines information from hydrophilicity, surface probability and backbone flexibility predictions along with the secondary structure predictions of Chou-Fasman and Robson-Garnier to produce a composite prediction of the surface contour of a protein [28]. The algorithm locates the basic domain in all three GBFs to the surface of the protein and the leucine-zipper to a non-exposed region of the protein, which is in good agreement with the known functions of the two domains in DNA-binding and hydrophobic interaction. Motifs II and III both in GBF4 and in GBF12 as well as motif V in GBF9 are located in domains that are predicted to be exposed to the surface (Figure 7). The position of motifs II and III in GBF12 coincides exactly with two small peaks of high probability of surface-exposure, whereas motifs II and III in GBF4 are located within a larger domain of surface exposure that spans both sequences. Motif V consists of an N-terminal part with low and a C-terminal part with high probability for surface exposure, consistent with the fact that the C-terminal half of this motif has a higher proportion of polar and acidic amino acids (Fig. 6A). The localization to the protein-surface of motifs II and III in class A GBFs as well as motif V in the class B GBF together with their acidic pI, make them potential target domains for regulatory signaling events specific to the different classes or regulatory functions of GBFs.

It is interesting to note that recently an *Arabidopsis* cDNA was cloned for a protein (GF14) that is associated with a protein-G-box complex, but that itself does not bind DNA [35]. Because the respective G-box binding protein of the complex has not been cloned yet, it is not clear whether this protein directly interacts with a GBF or with another potential component of the complex. If the protein indeed binds to GBF, it will be interesting to determine whether it discriminates between different types of GBFs and what the site of interaction is.

In tomato, the fruit-specific DNA-binding protein FBF recognizes a DNA-sequence immediately upstream of the G-box specifically in the inactive *rbcS3A* promoter. This results in a contiguous DNase I footprint between GBF and FBF (Meier and Gruissem, unpublished results, [16]. It is presently not known, however, if these two proteins physically interact, and if FBF down-regulates the activity of GBF.

The identification of a novel conserved domain in plant GBFs should provide a basis for mutational studies to identify their function. Furthermore, they might allow the identification of proteins that interact with GBFs outside of the leucine-zipper through one of the recently developed methods of functional cloning [36]. This may help to establish how different GBFs are linked to the diverse signal transduction mechanisms leading to gene activation in response to distinct stimuli.

ACKNOWLEDGEMENTS

We thank Drs. Jonathon Narita, Thianda Manzara and Râdiger Hell for helpful discussions and David Somers, Dr. Thianda Manzara and Dr. Susan Abrahamson for critical reading of the manuscript. The research was funded by a grant from the National Science Foundation and the NSF Center for Plant Development.

REFERENCES

1. Guiliano, G, Pichersky, E, Malik, V.S, Timko, M.P, Scolnik, P.A and Cashmore A.R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7089–7093.
2. Schulze-Lefert, P., Dangl, J.L., Becker-André, M., Hahlbrock. K. and Schulz, W. (1989) *EMBO J.*, **8**, 651–656.
3. Guiltinan, M.J., Marcotte, W.R. and Quatrano, R.S. (1990) *Science*, **250**, 267–271.
4. Staiger, D., Kaulen, H. and Schell, J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**,
5. De Lisle, A.J. , Ferl, R.J. (1990) *Plant Cell*, **2**, 547–557.
6. Schindler, U., Terzaghi, W., Beckmann, H., Kadesh, T. and Cashmore, A.R. (1992) *EMBO J.*, **11**, 1275–1289.
7. Oeda, K., Salinas, J. and Chua, N.-H. (1991) *EMBO J.*, **10**, 1793–1802.
8. Weisshaar, B., Armstrong, G.A, Block, A., da Costa e Silva, O. and Hahlbrock, K. (1991) *EMBO J.*, **10**, 1777–1786.
9. Tabata, T., Tasake, H., Takayama, S., Mikami, K., Nakatsuka, A., Kawata, T., Nakayama, T. and Iwabuchi, M. (1989) *Science*, **245**, 965–967.
10. Ellenberger, T.E., Brandl, C.J., Struhl, K. and Harrison, S.C. (1992) *Cell*, **71**, 1223–1237.

11. Vinson, C.R., Sigler, P.B. and McKnight, S.L. (1989) *Science*, **246**, 911–916.
12. Tabata, T., Nakayama, T., Mikami, K. and Iwabuchi, M. (1991) *EMBO J.*, **10**, 1459–1467.
13. Schindler, U., Menkens, A.E., Beckmann, H., Ecker, J.R. and Cashmore, A.R. (1992) *EMBO J.*, **11**, 1261–1273.
14. Manzara, T. and Gruissem, W. (1988) *Photosynth. Res.*, **16**, 117–139.
15. Wanner, L.A. and Gruissem, W. (1991) *Plant Cell*, **3**, 1289–1303.
16. Manzara, T., Carrasco, P. and Gruissem W. (1991) *Plant Cell*, **3**, 1305–1316.
17. Manzara, T., Carrasco, P. and Gruissem, W. (1993) *Plant Mol. Biol.*, **21**, 69–88.
18. Sambrook, J., Fritsch, E.F. and Maniatis T. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor.
19. Devereux, J., Haeberli, P. and Smithies O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
20. Piechulla, B., Pichersky, E., Cashmore, A.R. and Gruissem, W. (1986) *Plant Mol. Biol.*, **7**, 367–376.
21. Singh, H., Clerc, R.G. and LeBowitz, J.H. (1989) *Biotechniques*, **7**, 252–261.
22. Hartings, H., Maddaloni, M., Lazzaroni, N., Di Fonzo, F., Motto, M., Salamini, F. and Thompson, R. (1989) *EMBO J.*, **8**, 2795–2801.
23. Pysh, L.D., Aukerman, M.J. and Schmidt, R.J. (1993) *Plant Cell*, **5**, 227–236.
24. Dingwall, C. and Laskey, R.A. (1991) *Trends in Biol. Sci.*, **16**, 478–481.
25. Katagiri, F., Lam, E. and Chua N.-H. (1989) *Nature*, **340**, 727–730.
26. Chou, P.Y. and Fasman, G.D. (1978) *Ann. Rev. Biochem.*, **47**, 251–276.
27. Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) *J. Mol. Biol.*, **120**, 97–120.
28. Jameson, B.A. and Wolf, H. (1988) *Comput. Applic. in the Biosciences*, **4**, 181–186.
29. Salinas, J., Oeda, K. and Chua, N.-H. (1992) *Plant Cell*, **4**, 1485–1493.
30. McKendree, W.L., Paul, A.-L., DeLisle, A.J. andFerl, R.J. (1990) *Plant Cell*, **2**, 207–214.
31. Ferl, R.J. and Laughner, B.H. (1989) *Plant Mol. Biol.*, **12**, 357–366.
32. McKendree, W.L. and Ferl RJ. (1992) *Plant Mol. Biol.*, **19**, 859–862.
33. Donald, R.G.K. and Cashmore, A.R. (1990) *EMBO J.*, **9**, 1717–1726.
34. Mitchell, P.J. and Tijan R. (1989) *Science*, **245**, 371–378.
35. Lu, G., DeLisle, A.J., deVetten, N.C. and Ferl, R.J. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 11490–11494.
36. Guarente, L. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 1635–1637.