Published online 30 January 2007

Nucleic Acids Research, 2007, Vol. 35, No. 4 ***1145–1154***
doi:10.1093/nar/gkm001

DNA binding mechanism revealed by high resolution crystal structure of *Arabidopsis thaliana* WRKY1 protein

Ming-Rui Duan[1,2], Jie Nan[1], Yu-He Liang[1], Peng Mao[1,2], Lu Lu[1,2], Lanfen Li[1], Chunhong Wei[1,2], Luhua Lai[3], Yi Li[1,2]* and Xiao-Dong Su[1,*]

[1]The National Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing 100871, P.R. China, [2]Peking-Yale Joint Center for Plant Molecular Genetics and Agrobiotechnology, Beijing 100871, P.R. China and [3]College of Chemistry and Chemical Engineering, Peking University, Beijing 100871, P.R. China

Received August 11, 2006; Revised December 20, 2006; Accepted December 22, 2006

ABSTRACT

WRKY proteins, defined by the conserved WRKYGQK sequence, are comprised of a large superfamily of transcription factors identified specifically from the plant kingdom. This superfamily plays important roles in plant disease resistance, abiotic stress, senescence as well as in some developmental processes. In this study, the *Arabidopsis* WRKY1 was shown to be involved in the salicylic acid signaling pathway and partially dependent on NPR1; a C-terminal domain of WRKY1, *At*WRKY1-C, was constructed for structural studies. Previous investigations showed that DNA binding of the WRKY proteins was localized at the WRKY domains and these domains may define novel zinc-binding motifs. The crystal structure of the *At*WRKY1-C determined at 1.6 Å resolution has revealed that this domain is composed of a globular structure with five β strands, forming an antiparallel β-sheet. A novel zinc-binding site is situated at one end of the β-sheet, between strands β4 and β5. Based on this high-resolution crystal structure and site-directed mutagenesis, we have defined and confirmed that the DNA-binding residues of *At*WRKY1-C are located at β2 and β3 strands. These results provided us with structural information to understand the mechanism of transcriptional control and signal transduction events of the WRKY proteins.

INTRODUCTION

WRKY proteins comprise a large group of transcription factors identified specifically from the plant kingdom, ranging from lower plants such as mosses to higher plants (1,2). Members of the WRKY superfamily play important roles in a variety of developmental and physiological processes in plants. The most documented function for this superfamily of genes is their involvement in salicylic acid (SA) signaling and disease responses (3–8). SA is an important signal molecule in both local defenses and systemic acquired resistance (SAR). Endogenous accumulation or exogenous application of SA activates SAR and pathogenesis-related (PR) gene expression. In Arabidopsis, 49 out of 72 tested WRKY genes can be induced by pathogen infection or SA treatment (7), suggesting that a broad range of WRKY members could be involved in playing regulatory roles in disease resistance. This role of disease resistance was further confirmed by the finding of an over-presentation of the *cis*-element activated by the WRKY proteins, the W-box, within the promoters of a number of genes that are co-expressed in SAR (9,10). *Arabidopsis* NON-EXPRESSOR OF *PR1* (NPR1) has been demonstrated as a key protein in SAR signaling pathway (11). SAR and expression of *PR1* are prevented in *Arabidopsis npr1* mutants even after SA treatment (12). Previous reports also showed that over-expression of WRKY18 and WRKY70 in *Arabidopsis* resulted in increased resistance to pathogens and enhanced expression of *PR* genes (3,13).

*To whom correspondence should be addressed: Tel: +86 10 6275 9743; Fax: +86 10 6276 5669; Email: su-xd@pku.edu.cn; Yi Li. Tel: +86 10 6275 9690; Fax: +86 10 6275 4427; E-mail: liyi@pku.edu.cn
The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

In addition, WRKY genes are involved in abiotic stress, such as drought (14–17), cold (14,17,18), heat (15), salinity (17) and wounding (19); and in biosynthesis of anthocyanin (20) and starch (21). Some WRKY genes can also regulate developmental process such as embryogenesis (22), senescence (4,23), tricome development (20) as well as seed size (24).

There are more than 70 WRKY family members in *Arabidopsis* (1) and more than 100 in rice (25). The name of WRKY is derived from the conserved WRKY sequence motif which is present in all WRKY members. The WRKY domain is a stretch of about 60 amino acids with strictly conserved WRKYGQK sequence at its N-terminal followed by a putative novel zinc-binding motif with features of C–C–H–H or C–C–H–C (26–28). The existence of Zn^{2+} is crucial for the DNA-binding activity, which implies the importance of the putative zinc-binding motif (29,30). WRKY proteins can be classified into three groups according to the number of WRKY domains and zinc-binding pattern (26). Group I WRKY proteins are distinct from groups II and III by containing two WRKY domains. The zinc-binding motif of group III proteins is C–X_7–C–X_{23}–H–X–C, different from the common pattern of C-X_{4-5}-C-X_{22-23}-H-X-C of group I and II proteins. It has been reported that substitutions of the invariable WRKYGQK residues in the WRKY domain decreased the DNA-binding affinity, and any mutations of the conserved cysteine and histidine of the zinc-binding motif abolished the protein–DNA interaction (29).

The two WRKY domains in group I proteins play different roles in DNA-binding activities (27). It has been shown that the specific binding to W-box is mediated mainly by the C-terminal WRKY domain (27,30), whereas the function of N-terminal WRKY domain remains unclear. *Arabidopsis thaliana* WRKY1 protein (also known as ZAP1), the first identified WRKY transcription factor from Arabidopsis, is localized in Chromosome 2 as a single copy and belongs to group I WRKY family (30). The recognition of *At*WRKY1 with W-box mainly depends on the C-terminal WRKY domain while the N-terminal WRKY domain just slightly affect the protein–DNA interaction (30). The structural information of WRKY proteins and how they interact with W-box were lacking until recently when an NMR structure of C-terminal WRKY domain of *At*WRKY4, *At*WRKY4-C was determined (31). However, the domain division of the NMR sample seemed to be short and may not represent the structure of the whole domain. After the whole genome cloning of the transcription factors of *A. thaliana* (32), we have carried out functional and structural studies on *At*WRKY1.

In this report, we have showed that *At*WRKY1 can be induced by SA treatment and the induction was partially dependent on NPR1, suggesting that this protein may be involved in the defense response. Most importantly, we present the crystal structure of the *At*WRKY1-C at 1.6 Å and a model of protein–DNA complex from site-directed mutagenesis studies.

MATERIAL AND METHODS

Plant material and growth conditions

Wild-type *Arabidopsis* plants (ecotype Col-0) were grown in the soil at a growth room (23/20°C, 12-h-light/12-h-dark cycle). All the seeds were vernalized at 4°C for at least 2 days before placement in a growth environment. The non-functional NPR1 mutant *npr1–3* (12) and the transgenic line overexpressing NPR1 gene (33) were kindly provided by Dr Frederick M. Ausubel (Harvard Medical School, USA) and Dr Xin-Nian Dong (Duke University, USA), respectively.

SA treatment and northern blotting

Arabidopsis plants were sprayed with 2 mM SA diluted from stock solution (100 mM, adjusted to pH 6.5 with KOH). Leaves from different 4-week-old plants were harvested at indicated time points for total RNA preparation with RNAzol (Vigrous, Beijing) according to the manufacturer's instruction. Ten micrograms of total RNA was separated on 1.5% agrose–formaldehyde gels and transferred to a nylon membrane to hybridize with digoxigenin (DIG) labeled DNA probes. DIG DNA probes labeling, hybridization and detection were performed according to DIG Application Manual for Filter Hybridization (Roche, Penzberg, Germany). The following primer combinations were used to amplify DNA probes by PCR: *PR1*: forward 5′-ATGAATTTTACTGG CTATTCTCG-3′, reverse 5′-TTAGTATGGCTTCTCGT TCAC-3′; *At*WRKY1-C: forward 5′-ATGGCTGAGGT GGGAAAAGTTCTG-3′, reverse 5′-GCTTTGGGCAG GCTCTGTCTTGGG-3′. Equal loading was confirmed by staining the gel with ethidium bromide (EB).

Protein preparation and crystallization

The preparation and crystallization of the *At*WRKY1-C protein was described previously (34). Briefly, the cDNA fragment covering the C-terminal WRKY domain of *At*WRKY1 was amplified by polymerase chain reaction (PCR) from a full-length cDNA clone of *At*WRKY1 obtained from *Arabidopsis* (32) and cloned into the expression vector pET21aDEST. The protein was expressed in *E. coli* strain Rosetta and was purified to homogeneity in a two-step procedure of Ni^{2+} chelating and size exclusion chromatography. The mutant proteins were prepared similarly. The protein sample was kept in 20 mM Tris, 200 mM NaCl, pH 7.5 prior to crystallization. Crystals of the *At*WRKY1-C were obtained at 277 K from condition containing 1.2 M succinic acid, 0.1 M Tris–HCl pH 7.0 and 1% w/v PEG MME 2000, by using the hanging-drop vapor diffusion methods.

Data collection and structure determination

Two crystal forms were obtained from the same growth condition. One belongs to the space group $P2_1$, with diffraction to 2.5 Å resolution at home X-ray sources as published before (34), whereas the other crystal form

belongs to the space group $P3_221$, and diffracted to better than 1.6 Å at beamline 3W1A of Beijing Synchrotron Radiation Facilities (BSRF). Diffraction data were collected on a MAR165 CCD camera (MARresearch GmbH, Hamburg). The wavelength was set to 1.24 Å, at the higher energy side of the absorption peak of zinc. The data with anomalous signals were processed using the program DENZO and SCALEPACK (35). The statistics for the X-ray data collection and processing of the $P3_221$ crystal form are summarized in Table 1.

The zinc site was located using the program SHELXD (36) and the heavy atom parameters were refined using the program SOLVE (37). The initial phases were obtained by the program OASIS-2004 (38), then refined by using the program DM (CCP4) (39). Finally, ARP/warp (40) was used for the model auto-building.

Since the initial phases and auto-traced model were of good qualities, the subsequent refinement work could be easily carried out by using the program CNS (41). Manual adjustment of the structure was done with the program O (42), and a succinic acid as well as water molecules were added into the positive difference densities. The final model was refined by REFMAC5 (43).

Table 1. Data collection, phasing and refinement statistics (SAD)

Data collection		Refinement	
Space group	$P3_221$	Resolution (Å)	30–1.6
Cell dimensions		No. of reflections	10 536
a, *b*, *c* (Å)	45.55,45.55,68.96	R_{work}/R_{free} (%)	17.9/20.4
α, β, γ (°)	90,90,120	No. of atoms	727
Resolution (Å)	30–1.6 (1.66–1.6)	Protein	609
R_{sym} or R_{merge} (%)	9.1 (35.4)	Ligand/ion	8/1
$I/\sigma I$	22.9 (2.9)	Water	100
Completeness (%)	97.9 (95.5)	Average *B*-factors	20.95
Redundancy	5.6 (4.6)	RMSD[a] bond lengths (Å)	0.015
		RMSD[a] bond angles (°)	1.61

[a]RMSD in bond lengths and angles are the root-mean squared (RMS) deviations from ideal values.

Mutant design and site-directed mutagenesis

The design of mutants was based on a DNA–protein complex model between the W-box with the sequence 5′-ATCGTTGACCGAGTTGA-3′ and *At*WRKY1-C, which was generated from the known GCM–DNA complex (44) with the PDB ID 1ODH. GCM in the complex was replaced by *At*WRKY1-C with least square (LSQ) fitting between *At*WRKY1-C and GCM using the program O (42). The root mean-square standard deviations (RMSD) after LSQ was 2.3 Å for 68 Cα atoms. A B-DNA model for specific binding was prepared using the program CNS (41) and NAB (45) with the sequence 5′-ATCGTTGACCGAGTTGA-3′. DNA in the GCM–DNA complex was then replaced using the function of LSQ fit in XtalView (46).

The wild-type *At*WRKY1-C cDNA was used as a template for the mutagenesis primers. Site-directed mutagenesis by PCR was introduced into *At*WRKY1-C. DNA sequencing verified the introduction of the desired mutations and demonstrated that no unwanted mutations were present in the mutated protein sequences.

Electrophoretic mobility shift assays (EMSA)

The oligonucleotides 5′-ATCGTTGACCGAGTTGA-3′ and 5′-TCAACTCGGTCAACGAT-3′ containing the optimal binding site for *At*WRKY1 were annealed to form DNA duplex. The typical binding reaction (20 µl) contained 3 µg dsDNA, 20 mM HEPES/KOH (pH 7.2), 40 mM KCl, 1 mM EDTA, 0.5 mM DTT, 10% glycerol and 5 µg purified wild-type *At*WRKY1-C protein or mutant proteins. The binding reaction mixture was incubated at room temperature for 20 min and the complex was separated from the free duplex on a 12% non-denaturing polyacrylamide gel electrophoresis (PAGE) in 0.5× TBE at 70 V for 3.5 h. The gel was stained with EB. Images were captured using the fluorescence imaging system (Alpha Innotech Corporation, USA). A more 'standard' EMSA analysis was also performed. The standard binding reaction (20 µl) contained 0.5 µg of poly(dI-dC), 20 mM HEPES/KOH (pH 7.2), 40 mM KCl, 1 mM EDTA, 0.5 mM DTT, 10% glycerol, 0–180 ng purified proteins were added and 10 ng labeled double-stranded synthetic oligonucleotide, labeled with [γ-^{32}P]ATP using T4 polynucleotide kinase (NEB). For the competition assay, 100-fold molar excess of the unlabeled specific probe was added. DNA–protein complexes were allowed to form at room temperature for 30 min and then resolved on a 12% non-denaturing polyacrylamide gel in 0.5× TBE. The gel was dried and exposed for 2 h on X-ray films.

CD spectroscopic assays

All the CD spectra of the wild-type and mutant *At*WRKY1-C proteins were recorded on a Jobin Yvon CD 6 spectrometer (Longjumeau, France) at 298 K. The CD spectra of all proteins were recorded in phosphate buffer saline (PBS) buffer, pH 7.5. For near-UV–CD spectrum, a cell with a path length of 1 mm was used. Each spectrum was the average of four scans corrected by subtracting a spectrum of the buffer solution in the absence of proteins recorded under identical condition. Each scan in the range of 195–260 nm was obtained by taking data points every 0.5 nm with a 2-nm bandwidth and integration time of 1 s.

RESULTS AND DISCUSSION

*At*WRKY1 is involved in the SA signaling

SA is an important signal molecule in plant disease resistance. It has been previously shown that SA treatment can induce the expression of PR genes and activate systemic acquired resistance (SAR), which makes plants resistant against a spectrum of pathogens (47). To study the potential involvement of *At*WRKY1 in SA signaling pathway and its possible relation to NPR1, a key protein in SAR, we characterized the expression profile of *At*WRKY1 by RNA gel blot in wild-type *Arabidopsis* seedlings, *npr1-3* mutant (12) and NPR1-H line (33) in

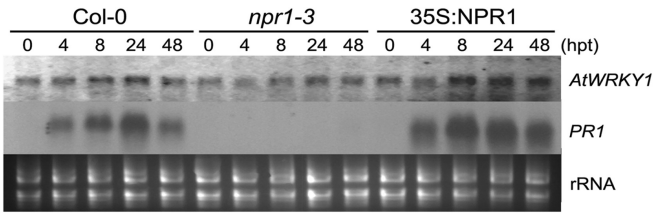

Figure 1. *AtWRKY1* is partially NPR1-dependent in SA pathway. Four weeks old wild type (Col-0), *npr1-3* and 35S::NPR1 *Arabidopsis* were sprayed with 2 mM SA and harvested at indicated time points. *PR1* transcripts were detected the same as *AtWRKY1* except for the difference of probe. The ethidium bromide stain of rRNA is shown for each lane to allow assessment of equal loading.

response to exogenous SA treatment. As an important marker of SAR, *PR1* was also examined to monitor the SA treatment procedure. In consistence with previous reports, *PR1* began to accumulate at 4 h post-treatment and the induction of *PR1* was abolished in *npr1-3* while increased in NPR1-H than that in wild type (Figure 1). The induced expression of *At*WRKY1 was observed after 8 h post-treatment and became obvious later than that of *PR1*. In *npr1-3*, *At*WRKY1 can still be expressed but at a lower level than that in wild-type plants. The opposite trend appeared in NPR1-H with higher expression level of *At*WRKY1 than in wild type (Figure 1). These data suggested that *At*WRKY1 was involved in SA signaling pathway and its induced expression was partially dependent on NPR1.

The nuclear localization of *At*WRKY1-C

As transcriptional regulators, WRKY proteins are shown to contain functional nuclear localization signal (NLS) and are targeted to nucleus (27,48). However, the NLS sequences of WRKY proteins are not well conserved, and may be distributed at different locations. The analysis of protein sequences of *At*WRKY1 revealed a potential NLS motif KRRKK between residues 273 and 277 near the C-terminal WRKY domain, which is a conserved feature for at least nine members of the group I WRKY proteins in *Arabidopsis* (Figure 2D). We have thus included this potential NLS motif in the *At*WRKY1-C sequence for this study. Introducing the fusion construct of *At*WRKY1-C with dimeric GFP (dGFP) into onion epidermal cells, the green fluorescent signal targeted specifically to the nucleus. However, the fluorescence of dGFP alone and fusion protein of *At*WRKY1-CΔNLS with dGFP were unevenly distributed throughout the cells (data not shown), which suggested that the potential NLS motif of *At*WRKY1-C is indeed responsible for the nuclear localization of *At*WRKY1.

Structure determination of *At*WRKY1-C crystal

Although in the same mother liquor and temperature, *At*WRKY1-C was crystallized in a different space group, $P3_221$, from the one published before (34). The structure was determined to 1.9 Å resolution by using the Single-wavelength Anomalous Dispersion (SAD) method. The

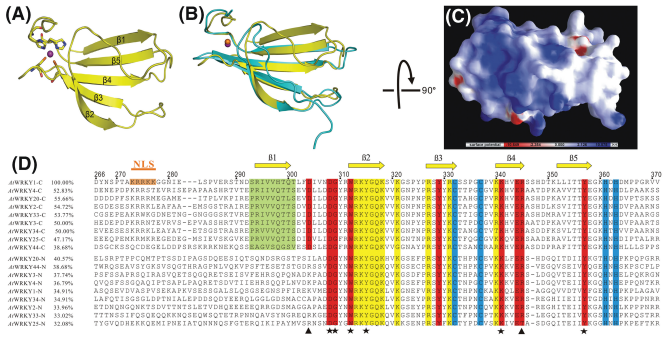

Figure 2. Structure of *At*WRKY1-C and multi-sequence alignment. (**A**) Ribbon representation of *At*WRKY1-C domain. The *At*WRKY1-C is composed of five β-strands (yellow ribbons), which are numbered from the N-terminus. The zinc ion is shown as a purple sphere and the zinc-coordinating residues are represented by sticks (yellow for C, red for O, blue for N, orange for S). (**B**) 3D-superimposition of the structure of *At*WRKY1-C and the best representative NMR structure of *At*WRKY4-C (model 15), using LSQ Fit in O. Macromolecular structures are shown by cartoons, with the *At*WRKY1-C colored in yellow and *At*WRKY4-C in cyan. The zinc ions are represented as spheres shown by magenta in *At*WRKY1-C and orange in *At*WRKY4-C. (**C**) Charge distribution on *At*WRKY1-C structure surfaces given by GRASP. Positive charges are represented by blue, negative charges are represented by red. (**D**) Structure-based sequence alignment of both N-terminus and C-terminus of the nine *Arabidopsis* WRKY proteins from group I. The zinc-coordinating residues are shown on blue background. Conserved residue elements for stabilizing the structure and recognizing DNA are drawn on red and yellow background, respectively. Residues of β1, highlighted in green but missed in *At*WRKY4-C structure, are rather conserved in all C-terminal domains of group I WRKY proteins. Residues marked by stars and triangles are scaffolds of two stable regions. Details of the interactions of the two regions are represented in Figure 3C and D.

average ratio of Bijovet pairs (<| Delta F|>/<F>) was 4.6%, in accordance with the high correlation coefficients (CC/all and CC/weak) 35.94/20.70, given by SHELXD (36) in zinc substructure locating. Phasing and density modification (DM) (39) were carried out consecutively using the software of OASIS-2004 (38) and DM (CCP4). With a figure of merit (FOM) value of 0.771 and good quality of density maps, 75 residues could be docked automatically by ARP/wARP (40). However, no density could be observed between residues 266 and 292 and between residues 369 and 371. The final model was refined manually to *R*-factor/*R*-free of 17.9 and 20.4%, respectively at 1.6 Å resolution (Table 1). The final model includes 76 residues corresponding to 293–368, one zinc ion, one succinic acid and 100 water molecules.

Description of the overall structure of *At*WRKY1-C

The *At*WRKY1-C structure is mainly composed of a five-stranded antiparallel β-sheet (β1, 294–300; β2, 312–318; β3, 327–332; β4, 340–345; β5, 352–358, see Figure 2D), with the disordered N-terminal 27 (including the potential NLS motif) and C-terminal three residues missing from the structure (Figure 2). WRKY protein's defining sequence, 'WRKYGQK', spans the entire β2 strand. Due to the long bridging loop (Thr301–Arg311) between β1 and β2 which covers one side of the β-sheet, the

structure of *At*WRKY1-C looks rather globular and stable (Figure 2A).

The crystal structure of *At*WRKY1-C differs from the NMR structure of a similar protein *At*WRKY4 C-terminal domain (*At*WRKY4-C, which shares about 53% sequence identity with *At*WRKY1-C) determined recently (31). As shown in Figure 2B and D, the β1 strand of *At*WRKY1-C was missing in the NMR *At*WRKY4-C model, thus the strands β2 to β5 in the crystal structure of *At*WRKY1-C corresponding to the strands β1–β4 of the NMR structure, and the RMSD corresponding to the similar regions of the two structures (*At*WRKY1-C residues 309–367, *At*WRKY4-C residues 411–469) is 1.5 Å, indicating a very similar structure for a common DNA-binding mechanism. Since the *At*WRKY4-C domain was constructed to start from VQTTS which is located in the middle of the β1 strand (Figure 2B and D with the alignment), the discrepancy in the organization of the β-sheet is most likely caused by the different ways of preparing the NMR and crystallization samples. In order to have a fair comparison, we have repeated and constructed the NMR samples in this work and compared the CD spectra for the two proteins, *At*WRKY4-C (Asp365–Ala469) similarly constructed as *At*WRKY1-C, and *At*WRKY4-CΔ1b (Val399–Ala469) similar to the NMR sample without the N-terminal β1 strand. The CD spectra showed similar pattern for zinc binding for both *At*WRKY4-C and *At*WRKY4-CΔ1b, and characteristics in both cases of a folded β-sheet structure. However, there was a reduction in the CD signal of *At*WRKY4-CΔ1b near the negative peak at 210–220 nm, indicating a reduction of β-sheet signal of the truncated protein (data not shown). We thus predict that the *At*WRKY4-C would have a similar globular structure like *At*WRKY1-C, had the *At*WRKY4-C sample prepared several residues longer at the N-terminus.

The conserved β1 and *B*-factor distribution of *At*WRKY1-C

As shown in Figure 2D, in addition to the conserved WRKY defining residues (in red, yellow and blue), the sequences in the β1 region (colored green) are well conserved in the C-terminal domains of the group I WRKY family members. We can thus conclude that all the C-terminal domains of the group I WRKY protein possess a very similar five-stranded β-sheet architecture as in *At*WRKY1-C.

Figure 3 describes the *B*-factor distribution over the *At*WRKY1-C structure. Figure 3A shows the *B*-factor plot against the residue numbers, and Figure 3B is the overall structure overlaid with the *B*-factors color coded as defined in Figure 3A. The regions with high *B*-factors (higher than 25) are in both the N- and C-terminal ends and on the loop between β2 and β3, this loop may be involved in conformational changes upon DNA binding. The middle part of the long loop connecting β1 and β2, and a few residues after β5 also showed slightly high *B*-factor between 20 and 25, whereas most of the structure are quite stable and ordered with an average *B*-factor of 21, particularly the zinc-binding site and the central β sheet are all with *B*-factors below 20.

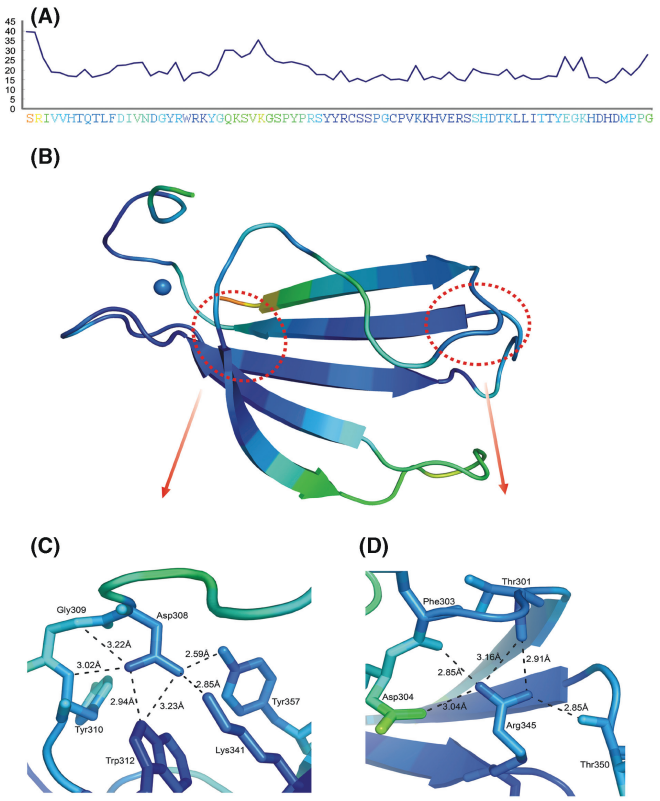

Figure 3. *B*-factor representation of *At*WRKY1-C crystal structure. All the objects are colored according to the *B*-factor values, from blue to red in the order of increasing *B*-factor. (**A**) Plot of *B*-factor against residues, with *B*-factor averaged over each residue. (**B**) The overall structure of *At*WRKY1-C in ribbon colored according to the *B*-factor values. The ends of the loop between strands β1 and β2 with low *B*-factor values are circled out by red rings for further amplification in 3C and 3D. Details of the interaction around these two terminals are described in 3C and 3D. (**C**) A D308-W312-K341 triad in the C-terminal of the loop between β1 and β2. Asp308 forms a well-defined salt-bridge with Lys341, extensively H-bonding with side chains of Trp312, Tyr357 and backbone of Tyr310. The three key residues Asp308, Trp312 and Lys341, components of D308-W312-K341 triad, and their hydrogen bonded residues Gly309, Tyr357 are marked by stars in Figure 2D. (**D**) The stable N-terminal end of the loop between β1 and β2. Arg345 is hydrogen bonded with the backbones of Thr301, Phe303 and Thr350. A potential salt-bridge between Arg345 and Asp304 was also observed as shown, and these two conserved residues are marked by triangles in Figure 2D.

The conserved salt-bridges and the D308-W312-K341 triad

Interestingly, both ends of the connecting loop between β1 and β2 show conserved pattern with ordered salt-bridge interactions and extensive hydrogen-bonding networks (Figure 3C and D). Figure 3C shows that in the C-terminal end of the loop, Asp308 forms a well-defined salt-bridge with Lys341, and Asp308 also forms extensive H-bonds with Tyr310, Trp312, Tyr357. All these residues are very well conserved among the WRKY family proteins, indicating a common function of stabilizing the domain structure. The pivotal residue at this constellation of residues is Asp308. The distance between Asp308 OD1 and Trp312 NE1 is 2.94 Å, and between Asp308 OD2 and Lys341 NZ is 2.85 Å. These three residues are with

the lowest *B*-factors, we thus name this constellation as D308-W312-K341 triad. In the NMR studies, the interactions around D308 could not be seen due to the undefined N-terminal of that structure, only the hydrophobic interaction of AtWRKY4-C W414-K443 (AtWRKY1-C residues W312-K341), which is part of the features of the previously defined triad constellation, has been mentioned (31). Figure 3D depicts another well conserved salt-bridge between Asp304 and Arg345 in the N-terminal end of the loop, and Arg345 is hydrogen bonded with the main chain of Thr301, Phe303 and Thr350.

The zinc-binding site in *At*WRKY1-C

One of the important features resolved by the high-resolution crystal structure of *At*WRKY1-C is the well-ordered zinc-binding site. As shown in the electron density map in Figure 4A, the high-resolution data of 1.6 Å made it possible to distinguish the zinc coordination with all the ligand atoms, and it was clearly shown that the Nδ of His361 and the Nε of His363, coordinate to the zinc ion (Figure 4A and B). The distances between the ligand atoms and the zinc ion are 2.06 Å (Zn–Nδ of His361), 2.07 Å (Zn–Nε of His363), 2.28 Å (Zn–Sγ of Cys332), 2.27 Å (Zn–Sγ of Cys337), all the distances are well within the ideal distances of a zinc-binding site in proteins (49).

The whole zinc-binding site is situated in the low *B*-factor regions (Figure 3), and the zinc ligand residues have almost the lowest *B*-factors in the structure, the zinc-binding site is right next to the low *B*-factor D308-W312-K341 triad described in the Figure 3C. Together with the D308-W312-K341 triad, the zinc-binding site is shaping up and stabilizing the *At*WRKY1-C structure, since the zinc binding has been shown to be crucial for the stability of WRKY proteins (31).

The sequence signature of this zinc-binding site, $C–X_4–C–X_{23}–H–X–H$, somewhat resembles a classic C2H2 profile (50), the feature of a zinc-finger. In the crystal structure, it has been clearly shown that the zinc ion is coordinated by two cysteines (Cys332, Cys337) and two histidines (His361, His363) (Figure 4), but structurally dissimilar to classic zinc fingers, instead the structure is similar to a sequence unrelated DNA-binding protein Drosophila GCM (glia cell missing) (44) (PDB ID 1ODH). In contrast with the classic C2H2 zinc finger, *At*WRKY1-C does not contain any helix, so it cannot bind DNA through the helix as most zinc-finger proteins do. The DNA-binding ability of *At*WRKY1-C is mediated through the beta-hairpin regions between β2 and β3, and this mode of DNA–protein interaction is similar to the insightful prediction by Church, Sussman and Kim long time ago for non-helical DNA binding (51). Previous work on DNA binding of WRKY family has also pointed out that the conserved WRKYGQK region located on β2 strand is important for DNA binding (29).

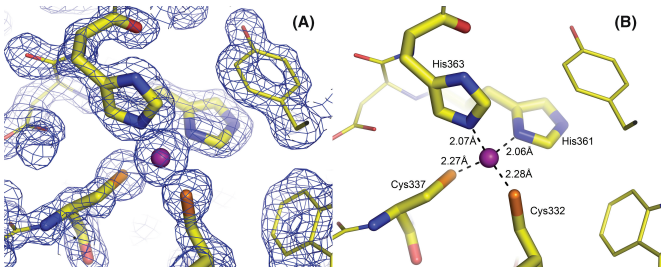

Figure 4. The zinc coordination environment. (**A**) $3F_o$-$2F_c$ electron-density map. The coordination of the zinc ion by Cys332, Cys337, His361 and His363 was shown. The map was prepared using CCP4 program suite and displayed with Pymol at the contour of 1.5σ level. The zinc ion is shown as a purple sphere and the zinc-coordinating residues are represented by stick (yellow for C, red for O, blue for N, orange for S, respectively). (**B**) The distances between the zinc ion and the coordinate residues. The orientation and colors are the same as shown in Figure 4A.

Conserved residues responsible for DNA binding

There are more than 70 members in the *Arabidopsis* WRKY family grouped into three major groups (26), nine members from the group I with two WRKY domains are selected for multiple sequence alignment based on secondary structure information (Figure 2D) (sequence data were downloaded from http://rarge.gsc.riken.go.jp/rartf/tf_info/fasta_p1/TF_9_P1.fasta). It was shown from the alignment that at least 18 residues are strictly conserved, including the four residues coordinating with the zinc ion, and three residues in the conserved D308-W312-K341 triad.

As shown in Figure 2D, more than 10 well-conserved residues are positioned in the region between β2 and β3 including the WRKYGQK motif. Particularly, the five consecutive residues RKYGQ, distinct components of the WRKY family scaffold, belong to the β2 strand; while the other five conserved residues located on β3 strand, in the pattern of PRSYYR/K. Considering the specific DNA sequence that all WRKY members could recognize, these conserved residues are good candidates for the specific DNA-binding process.

As shown in Figure 2C, the surface charge distribution of *At*WRKY1-C was uneven, with a calculated PI (point of iso-electricity) of 9.5. On one side, there were consecutive regions of positively charged residues, including Arg313, Lys314, Lys318, Lys321, Arg327 and Arg331, all belonging to the strictly conserved residues and located on β2 and β3 strands. Well-conserved Lys340 located at the beginning of strand β4 also belongs to this positively charged region. Since the WRKYGQK sequence falls nicely into this positively charged region and it has been shown that the WRKY motif is responsible to bind to the W-box (29), it can thus be deduced that the β2 and β3 strands are most likely to be involved in the process of DNA binding.

Identification and confirmation of the DNA-binding residues

After searching for similar structures of *At*WRKY1-C using the DALI server (52), a DNA–protein complex

structure, the Drosophila GCM (44) (PDB ID 1ODH, Z-score > 6.2) was returned. When the GCM protein was used to perform a least-square fitting with *At*WRKY1-C, the resulting RMSD between the aligned parts (68 residues) was 2.3 Å, a DNA complex model of W-box DNA and *At*WRKY1-C was thus constructed as shown in Figure 7A. By comparison of the potential DNA-binding residues (which are close to the DNA major groove as shown in the complex model), and previous work on the DNA-binding related residues on the WRKYGQK motif (29), the following mutants with a single residue change were constructed: R313E, K314A, K314R, Y315F, Y315R, G316F, Q317A, Q317K, R327A, R327E, Y330F, R331K, R331A and K340A. In Maeo *et al.*'s work (29), it has been shown that all residues in the WRKYGQK (corresponding to residues 312–318 in *At*WRKY1-C) motif are important for the DNA binding. Particularly for W312, K314, Y315, K318, a change to Ala for any of those residues would completely abolish the DNA binding. On the other hand, the R313, G316 and Q317 would only reduce the amount of DNA binding (29). We have thus constructed multiple mutants for R313, K314, Y315, G316 and Q317. Furthermore, the well-conserved residues R327, Y330, R331 and K340 are quite close to the DNA-binding sites, these residues were also selected for mutagenesis studies. Thirteen out of the fourteen mutants are located on strands β2 and β3. This region, particularly the connecting loop, showed flexible feature (high *B*-factor) in Figure 3A and B, indicating conformational changes might occur upon DNA binding. EMSA were used to test the DNA-binding abilities of these mutants with a fragment of 17-bp duplex DNA containing the W-box of TGAC as shown in Figure 5. The EMSA assay using ^{32}P-labeled probe according to standard protocol was also performed with similar results (not shown). To examine the specificity of the binding, a competition experiment was performed by using an excess amount of unlabeled probe of the same sequence. When a 100-fold excess of unlabeled oligonucleotides were included in the reaction, binding of *At*WRKY1-C to the labeled probe was abolished completely (data not shown). All the mutants could be readily expressed and purified as the wild-type *At*WRKY1-C, the circular dichroism (CD) spectra of these mutant proteins were used to monitor the correct folding of these proteins, thus the DNA-binding ability changes are not due to the structural change caused by the mutations, see Figure 6.

As shown in Figure 5, K314A, Y315R, G316F, R331A, R327A and R327E have completely abolished the DNA-binding abilities of those mutants, and the CD spectra in Figure 6 have shown those mutants were well-structured proteins. Thus the inabilities of DNA binding must be caused by the point mutations on the correct structural scaffold. The DNA-binding activity of the mutations R313E, K314R, Y315F and Q317K were reduced, whereas the mutations of Q317A, R331K, K340A and Y330F did not show any decrease of the DNA-binding ability, indicating that these mutations did not affect the DNA binding of *At*WRKY1-C.

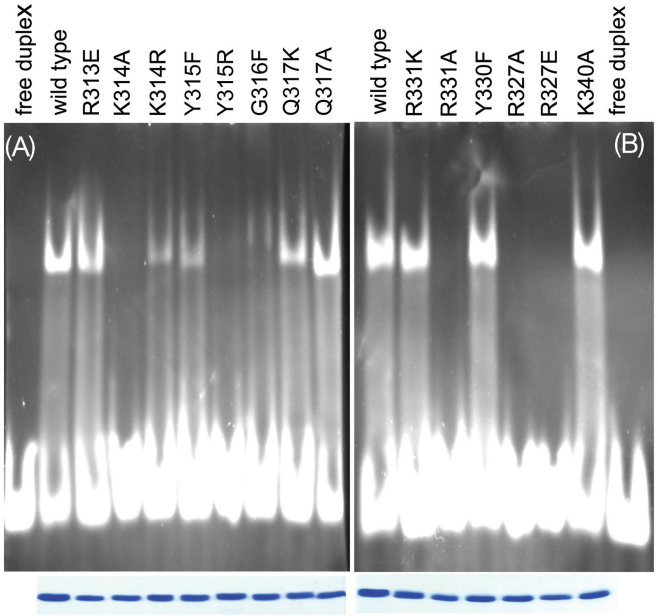

Figure 5. EMSA with the consensus W-box binding site and wild-type *At*WRKY1-C as well as mutant proteins. (**A**) Binding activity examination with mutations on β2. The binding reaction contained 5 μg purified proteins and 3 μg DNA duplex. The gel was stained with EB after the electrophoresis. The same amount of purified proteins were run on SDS–PAGE and stained with Coomassie Brilliant Blue to show the equal loading. (**B**) Binding activity examination with mutations on β3 and β4.

These results are in good agreement with the previous work (29), and conclusions from the NMR structure (31).

In the crystal structure as well as in the NMR structure, a kink was formed in the middle of β2 around Gly316. The formation of this kink was considered helpful for the protruding of Gly316 into DNA major groove deeply and attending specific binding. The assumption was confirmed by the mutation G316F, which has lost almost all the DNA-binding ability. The complete loss of DNA binding of the mutants R327E, R327A and R331A are of great interests, since these two positively charged residues localized on the β3 strand that has been shown to be crucial for the DNA binding for the first time by site-directed mutagenesis. Furthermore, these two arginine residues, which are separated by four residues, is consistent with Tateno's model of DNA binding at the major groove with a pair of β strands (53).

As a summary, all the mutation results further confirmed that strands β2 and β3 participate in DNA binding. It also suggested that residues Lys314, Gly316, Lys318 on β2, and Arg327, Arg331 on β3 are the key residues for DNA-specific recognition and substitution any one of them resulted in no specific binding. In addition, the distribution of these residues is i, $i+2$, $i+4$ on one strand and j, $j+4$ on the other strand. This distribution is in agreement with Tateno's model (53), which proposed that the binding of DNA to proteins with two β strands should happen on the DNA major groove in a convex manner. Based on such a model, it has been found that normally six consecutive base pairs are

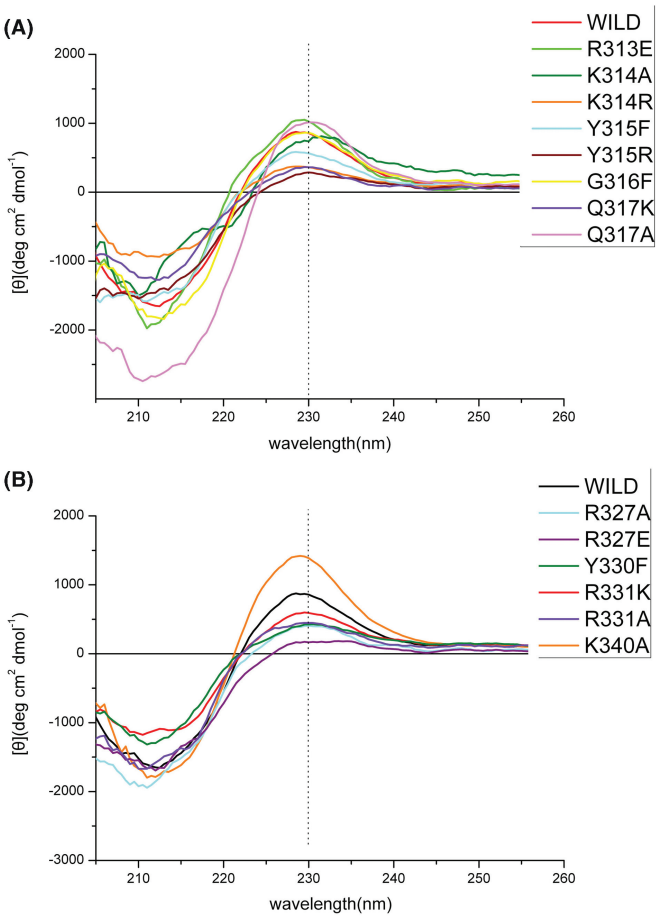

Figure 6. CD spectra. CD spectra of wild-type *At*WRKY1-C and the mutant proteins on β2 strand. The different proteins were distinguished from each other by colors. (A) CD spectra of wild-type *At*WRKY1-C and the mutant proteins on β3 and β4 strands.

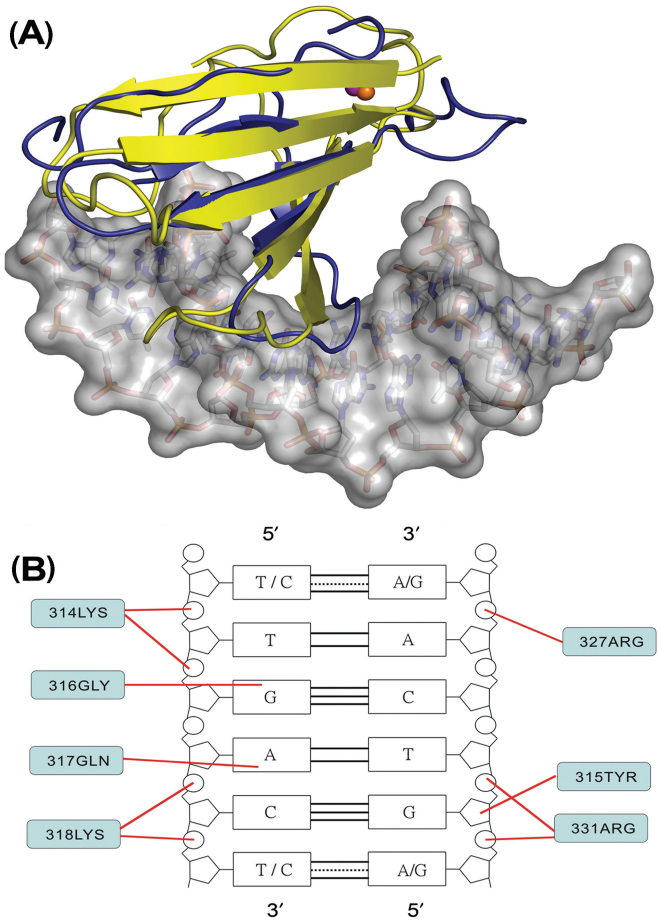

Figure 7. A model of the *At*WRKY1-C and W-box interaction. (**A**) Superimposition of *At*WRKY1-C and the overlapping structure of the Drosophila GCM domain using LSQ Fit in the program O. Macromolecular structures are shown by cartoon loops, with the GCM domain colored yellow, *At*WRKY1-C blue and zinc ions as orange and magenta spheres. The DNA double-helix is shown with sticks wrapped inside surface rendering. (**B**) The final schematic model of the specific interaction of *At*WRKY1-C with the W-box, combining all information obtained from the structural and mutagenesis experiments presented in this work.

involved in specific recognition, and the large-sized residues (not only charged), such as lysines at i and $i+4$, as well as arginines at j and $j+4$, could bridge the large gap separating the DNA.

The structure model of DNA–*At*WRKY1-C complex we made in this study has been based on the GCM–DNA structure (44), which is different from the Tateno's DNA-binding model (53) in details. The β-strands of the GCM protein adapted a previously undescribed binding mode to interact with DNA (44). Despite of the absence of sequence and topology similarity, *At*WRKY1-C's five-stranded β-sheet overlaps with the corresponding strands of GCM with unexpected high accuracy (Figure 7A), except for the β2–β3 region including the flexible loop, which will be subjected to conformational changes upon DNA binding. We thus refined our DNA–AtWRKY1-C complex according to the mutagenesis results, and the resulting model is quite similar to the GCM–DNA complex in topology. A schematic description of this refined model for AtWRKY1-C and DNA interactions is described as in Figure 7B. This figure summarizes all the residues that might interact with the W-box DNA, providing structural explanations for the function of the WRKY superfamily of transcription factors.

ACCESSION NUMBER

The Protein Data Bank accession number for the *At*WRKY1-C crystal structure discussed in this paper is PDB ID 2AYD.

ACKNOWLEDGEMENTS

We thank Dr Xinnian Dong for providing NPR1-H seeds, Dr Frederick M.Ausubel for npr1-3 seeds, Dr Biao Ding for pRTL2:dGFP (dimeric GFP); Ms Huanhuan Liang is acknowledged for her help with CD spectrum measurement. We also thank the beamline staff (Drs Yu-Hui Dong and Peng Liu) at the Beijing Synchrotron Radiation Facilities for help with data collection. This work was

supported by Key Basic Research Program (973) grant no: 2006CB101905 (to YL) and National Natural Science Foundation of China (NSFC) grants no. 30600095 (to MRD) and 30325012 (to XDS). Funding to pay the Open Access publication charge was provided by Peking University, P.R. China.

Conflict of interest statement. None declared

REFERENCES

1. Ulker,B. and Somssich,I.E. (2004) WRKY transcription factors: from DNA binding towards biological function. *Curr. Opin. Plant Biol.*, **7**, 491–498.
2. Zhang,Y. and Wang,L. (2005) The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evol. Biol.*, **5**, 1.
3. Chen,C. and Chen,Z. (2002) Potentiation of developmentally regulated plant defense response by AtWRKY18, a pathogen-induced Arabidopsis transcription factor. *Plant Physiol.*, **129**, 706–716.
4. Robatzek,S. and Somssich,I.E. (2002) Targets of AtWRKY6 regulation during plant senescence and pathogen defense. *Genes Dev.*, **16**, 1139–1149.
5. Yoda,H., Ogawa,M., Yamaguchi,Y., Koizumi,N., Kusano,T. and Sano,H. (2002) Identification of early-responsive genes associated with the hypersensitive response to tobacco mosaic virus and characterization of a WRKY-type transcription factor in tobacco plants. *Mol. Genet. Genomics*, **267**, 154–161.
6. Chen,C. and Chen,Z. (2000) Isolation and characterization of two pathogen- and salicylic acid-induced genes encoding WRKY DNA-binding proteins from tobacco. *Plant Mol. Biol.*, **42**, 387–396.
7. Dong,J., Chen,C. and Chen,Z. (2003) Expression profiles of the Arabidopsis WRKY gene superfamily during plant defense response. *Plant Mol. Biol.*, **51**, 21–37.
8. Kalde,M., Barth,M., Somssich,I.E. and Lippok,B. (2003) Members of the Arabidopsis WRKY group III transcription factors are part of different plant defense signaling pathways. *Mol. Plant Microbe. Interact.*, **16**, 295–305.
9. Maleck,K., Levine,A., Eulgem,T., Morgan,A., Schmid,J., Lawton,K.A., Dangl,J.L. and Dietrich,R.A. (2000) The transcriptome of Arabidopsis thaliana during systemic acquired resistance. *Nat. Genet.*, **26**, 403–410.
10. Turck,F., Zhou,A. and Somssich,I.E. (2004) Stimulus-dependent, promoter-specific binding of transcription factor WRKY1 to its native promoter and the defense-related gene PcPR1-1 in Parsley. *Plant Cell*, **16**, 2573–2585.
11. Cao,H., Bowling,S.A., Gordon,A.S. and Dong,X. (1994) Characterization of an Arabidopsis mutant that is nonresponsive to inducers of systemic acquired resistance. *Plant Cell*, **6**, 1583–1592.
12. Cao,H., Glazebrook,J., Clarke,J.D., Volko,S. and Dong,X. (1997) The Arabidopsis NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell*, **88**, 57–63.
13. Li,J., Brader,G. and Palva,E.T. (2004) The WRKY70 transcription factor: a node of convergence for jasmonate-mediated and salicylate-mediated signals in plant defense. *Plant Cell*, **16**, 319–331.
14. Mare,C., Mazzucotelli,E., Crosatti,C., Francia,E., Stanca,A.M. and Cattivelli,L. (2004) Hv-WRKY38: a new transcription factor involved in cold- and drought-response in barley. *Plant Mol. Biol.*, **55**, 399–416.
15. Rizhsky,L., Liang,H. and Mittler,R. (2002) The combined effect of drought stress and heat shock on gene expression in tobacco. *Plant Physiol.*, **130**, 1143–1151.
16. Pnueli,L., Hallak-Herr,E., Rozenberg,M., Cohen,M., Goloubinoff,P., Kaplan,A. and Mittler,R. (2002) Molecular and biochemical mechanisms associated with dormancy and drought tolerance in the desert legume Retama raetam. *Plant J.*, **31**, 319–330.
17. Seki,M., Narusaka,M., Ishida,J., Nanjo,T., Fujita,M., Oono,Y., Kamiya,A., Nakajima,M., Enju,A. *et al.* (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant J.*, **31**, 279–292.
18. Huang,T. and Duman,J.G. (2002) Cloning and characterization of a thermal hysteresis (antifreeze) protein with DNA-binding activity from winter bittersweet nightshade, Solanum dulcamara. *Plant Mol. Biol.*, **48**, 339–350.
19. Hara,K., Yagi,M., Kusano,T. and Sano,H. (2000) Rapid systemic accumulation of transcripts encoding a tobacco WRKY transcription factor upon wounding. *Mol. Gen. Genet.*, **263**, 30–37.
20. Johnson,C.S., Kolevski,B. and Smyth,D.R. (2002) TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. *Plant Cell*, **14**, 1359–1375.
21. Sun,C., Palmqvist,S., Olsson,H., Boren,M., Ahlandsberg,S. and Jansson,C. (2003) A novel WRKY transcription factor, SUSIBA2, participates in sugar signaling in barley by binding to the sugar-responsive elements of the iso1 promoter. *Plant Cell*, **15**, 2076–2092.
22. Lagace,M. and Matton,D.P. (2004) Characterization of a WRKY transcription factor expressed in late torpedo-stage embryos of Solanum chacoense. *Planta*, **219**, 185–189.
23. Hinderhofer,K. and Zentgraf,U. (2001) Identification of a transcription factor specifically expressed at the onset of leaf senescence. *Planta*, **213**, 469–473.
24. Luo,M., Dennis,E.S., Berger,F., Peacock,W.J. and Chaudhury,A. (2005) MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 17531–17536.
25. Xie,Z., Zhang,Z.L., Zou,X., Huang,J., Ruas,P., Thompson,D. and Shen,Q.J. (2005) Annotations and functional analyses of the rice WRKY gene superfamily reveal positive and negative regulators of abscisic acid signaling in aleurone cells. *Plant Physiol.*, **137**, 176–189.
26. Eulgem,T., Rushton,P.J., Robatzek,S. and Somssich,I.E. (2000) The WRKY superfamily of plant transcription factors. *Trends Plant Sci.*, **5**, 199–206.
27. Eulgem,T., Rushton,P.J., Schmelzer,E., Hahlbrock,K. and Somssich,I.E. (1999) Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. *EMBO J.*, **18**, 4689–4699.
28. Rushton,P.J., Torres,J.T., Parniske,M., Wernert,P., Hahlbrock,K. and Somssich,I.E. (1996) Interaction of elicitor-induced DNA-binding proteins with elicitor response elements in the promoters of parsley PR1 genes. *EMBO J.*, **15**, 5690–5700.
29. Maeo,K., Hayashi,S., Kojima-Suzuki,H., Morikami,A. and Nakamura,K. (2001) Role of conserved residues of the WRKY domain in the DNA-binding of tobacco WRKY family proteins. *Biosci. Biotechnol. Biochem.*, **65**, 2428–2436.
30. de Pater,S., Greco,V., Pham,K., Memelink,J. and Kijne,J. (1996) Characterization of a zinc-dependent transcriptional activator from Arabidopsis. *Nucleic Acids Res.*, **24**, 4624–4631.
31. Yamasaki,K., Kigawa,T., Inoue,M., Tateno,M., Yamasaki,T., Yabuki,T., Aoki,M., Seki,E., Matsuda,T. *et al.* (2005) Solution structure of an Arabidopsis WRKY DNA binding domain. *Plant Cell*, **17**, 944–956.
32. Gong,W., Shen,Y.P., Ma,L.G., Pan,Y., Du,Y.L., Wang,D.H., Yang,J.Y., Hu,L.D., Liu,X.F. *et al.* (2004) Genome-wide ORFeome cloning and analysis of Arabidopsis transcription factor genes. *Plant Physiol.*, **135**, 773–782.
33. Cao,H., Li,X. and Dong,X. (1998) Generation of broad-spectrum disease resistance by overexpression of an essential regulatory gene in systemic acquired resistance. *Proc. Natl. Acad. Sci. USA*, **95**, 6531–6536.
34. Duan,M.R., Ren,H., Mao,P., Wei,C.H., Liang,Y.H., Li,Y. and Su,X.D. (2005) Crystallization and preliminary X-ray analysis of the C-terminal WRKY domain of Arabidopsis thaliana WRKY1 transcription factor. *Biochim. Biophys. Acta*, **1750**, 14–16.
35. Otwinowski,Z. and Minor,W. (1997). In: Carter,C.W. and Sweet,R.M. (eds). *Methods in Enzymology* Academic Press, New York, Vol. 276, pp. 307–326.
36. Sheldrick,G.M., Dauter,Z., Wilson,K.S., Hope,H. and Sieker,L.C. (1993) The application of direct methods and Patterson interpretation to high-resolution native protein data. *Acta Crystallogr. D*, **49**, 18–23.

37. Terwilliger,T.C. (2003) SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol.*, **374**, 22–37.
38. Wang,J.W., Chen,J.R., Gu,Y.X., Zheng,C.D., Jiang,F. and Fan,H.F. (2004) Optimizing the error term in direct-method SAD phasing. *Acta Crystallogr. D*, **60**, 1987–1990.
39. Cowtan,K.D. and Main,P. (1993) Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints. *Acta Crystallogr. D Biol. Crystallogr.*, **49**, 148–157.
40. Perrakis,A., Morris,R. and Lamzin,V.S. (1999) Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.*, **6**, 458–463.
41. Brunger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.S., Kuszewski,J., Nilges,M. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D*, **54**, 905–921.
42. Jones,T.A. and Kjeldgaard,M. (1997) Electron density map interpretation. *Methods Enzymol.*, **277**, 173–208.
43. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D*, **53**, 240–255.
44. Cohen,S.X., Moulin,M., Hashemolhosseini,S., Kilian,K., Wegner,M. and Muller,C.W. (2003) Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. *EMBO J.*, **22**, 1835–1845.
45. Macke,T. and Case,D.A. (1998). Modeling unusual nucleic acid structures. In: Leontes,N.B. and SantaLucia,J.Jr. (eds). *Molecular Modeling of Nucleic Acids* American Chemical Society, Washington DC, pp. 379–393.
46. McRee,D.E. (1999) XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.*, **125**, 156–165.
47. Ryals,J.A., Neuenschwander,U.H., Willits,M.G., Molina,A., Steiner,H.Y. and Hunt,M.D. (1996) Systemic acquired resistance. *Plant Cell*, **8**, 1809–1819.
48. Robatzek,S. and Somssich,I.E. (2001) A new member of the Arabidopsis WRKY transcription factor family, AtWRKY6, is associated with both senescence- and defence-related processes. *Plant J.*, **28**, 123–133.
49. Simonson,T. and Calimet,N. (2002) Cys(x)His(y)-Zn2+ interactions: thiol vs. thiolate coordination. *Proteins*, **49**, 37–48.
50. Krishna,S.S., Majumdar,I. and Grishin,N.V. (2003) Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.*, **31**, 532–550.
51. Church,G.M., Sussman,J.L. and Kim,S.H. (1977) Secondary structural complementarity between DNA and proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 1458–1462.
52. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
53. Tateno,M., Yamasaki,K., Amano,N., Kakinuma,J., Koike,H., Allen,M.D. and Suzuki,M. (1997) DNA recognition by beta-sheets. *Biopolymers*, **44**, 335–359.