PROTEIN STRUCTURE REPORT

Structure of the B3 domain from *Arabidopsis thaliana* protein At1g16640

JEANETTE K. WALTNER, FRANCIS C. PETERSON, BETSY L. LYTLE, AND BRIAN F. VOLKMAN

Department of Biochemistry and Center for Eukaryotic Structural Genomics, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA

Abstract

A novel DNA binding motif, the B3 domain, has been identified in a number of transcription factors specific to higher plant species, and was recently found to define a new protein fold. Here we report the second structure of a B3 domain, that of the *Arabidopsis thaliana* protein, At1g16640. As part of an effort to 'rescue' structural genomics targets deemed unsuitable for structure determination as full-length proteins, we applied a combined bioinformatic and experimental strategy to identify an optimal construct containing a predicted conserved domain. By screening a series of N- and C-terminally truncated At1g16640 fragments, we isolated a stable folded domain that met our criteria for structural analysis by NMR spectroscopy. The structure of the B3 domain of At1g16640 consists of a seven-stranded β-sheet arranged in an open barrel and two short α-helices, one at each end of the barrel. While At1g16640 is quite distinct from previously characterized B3 domain proteins in terms of amino acid sequence similarity, it adopts the same novel fold that was recently revealed by the RAV1 B3 domain structure. However, putative DNA-binding elements conserved in B3 domains from the RAV, ARF, and ABI3/VP1 subfamilies are largely absent in At1g16640, perhaps suggesting that B3 domains could function in contexts other than transcriptional regulation.

Keywords: B3 domain; NMR; protein structure; structural genomics; bioinformatics

Ongoing efforts in structural genomics are changing the landscape of structural biology. Annotation of the *Arabidopsis thaliana* genome has fostered these advances, providing an excellent eukaryotic system from which to identify novel targets. NMR spectroscopy is an invaluable tool for high-throughput screening and protein structure determination in structural genomics, but production of sufficient numbers of tractable structural targets often represents a critical bottleneck. Improved methods to overcome problems such as low protein expression, insolubility, protein aggregation, and lack of foldedness are critical to the progression of high-throughput proteomics.

As part of a structural genomics effort directed at eukaryotic proteins, the At1g16640 protein was selected from the *A. thaliana* genome as a target likely to reveal novel structural information. At1g16640 was predicted in the Pfam database to contain a DNA-binding domain unique to higher plant species (Bateman et al. 2004). This motif, called the B3 domain (Pfam accession 02362), has been characterized in a number of plant transcription factors. The first B3 domains were identified in the proteins Abscisic Acid-Insensitive 3 (ABI3) from *Arabidopsis* and Viviparous 1 (VP1) from *Zea Mays* (Giraudat et al. 1992). Since then, many B3 domain-containing proteins have been classified functionally as factors re-

sponsive to abscisic acid and auxin, phytohormones that play critical roles in developmental processes such as plant growth and seed maturation (McCarty et al. 1989; Ulmasov et al. 1997). Three major classes of transcription factors containing B3 domains have been identified to date, including factors resembling ABI3 and VP1 (ABI3/VP1-like factors), proteins similar to the *Arabidopsis* protein, RAV1 (RAV-like family), and auxin response factors (ARFs) (Riechmann et al. 2000). These B3 domains bind to specific DNA sequences six base pairs in length (Suzuki et al. 1997; Ulmasov et al. 1997; Kagaya et al. 1999). The recognition sequences are conserved among members of the same family, but differ between the three identified families. When At1g16640 was selected as a target, no B3 domain structures were reported; since then the first B3 domain structure (from the *Arabidopsis* protein RAV1) was determined by NMR (Yamasaki et al. 2004). Based on ambiguous screening results by 2D NMR, the full-length At1g16640 protein (134 residues) was judged an unsuitable target for structure determination and dropped from the production pipeline. Promising aspects of the initial NMR data and the potential value of a B3 domain structure led us to consider methods for salvaging high-priority targets that may contain folded domains but fail due to aggregation, insolubility, or other problems caused by various portions of the protein.

In this report we show that a bioinformatic approach, combined with experimental screening of a modest number of expression constructs, can be used to rescue proteins that would be otherwise unsuitable for structure determination. Using this approach, we identified a stable, folded domain in the *A. thaliana* protein At1g16640, a structural genomics target rejected at the HSQC screening stage as a full-length protein. Inspection of the structure of the optimal At1g16640 B3 domain construct determined by NMR spectroscopy reveals a highly conserved tertiary fold.

Results and Discussion

Using high-throughput production methods for structural genomics, At1g16640 was selected from the *Arabidopsis* genome as a target, cloned, expressed in *Escherichia coli*, affinity purified, screened by 2D NMR and submitted for small-scale crystallization trials (Tyler et al. 2005a,b). The protein failed to crystallize and was evaluated as "HSQC$+/-$", based on the nonuniformity of peak intensities and chemical shift dispersion in ^{15}N–^{1}H HSQC spectrum (Tyler et al. 2005a), despite the presence of weak signals that clearly indicated the presence of a folded domain (Fig. 1A). While the full-length form of the At1g16640 protein (134 amino acids) was judged unsuitable for NMR structure determination, we hypothesized that a well-behaved domain could be identified through a combination of bioinformatic analysis and systematic screening of a panel of truncated proteins.

Domain identification and construct design

From the Pfam database (Bateman et al. 2004) we found that At1g16640 was predicted to contain a B3 DNA-binding domain (PF02362) encompassing residues 8–102. We hypothesized that truncation of residues extraneous to
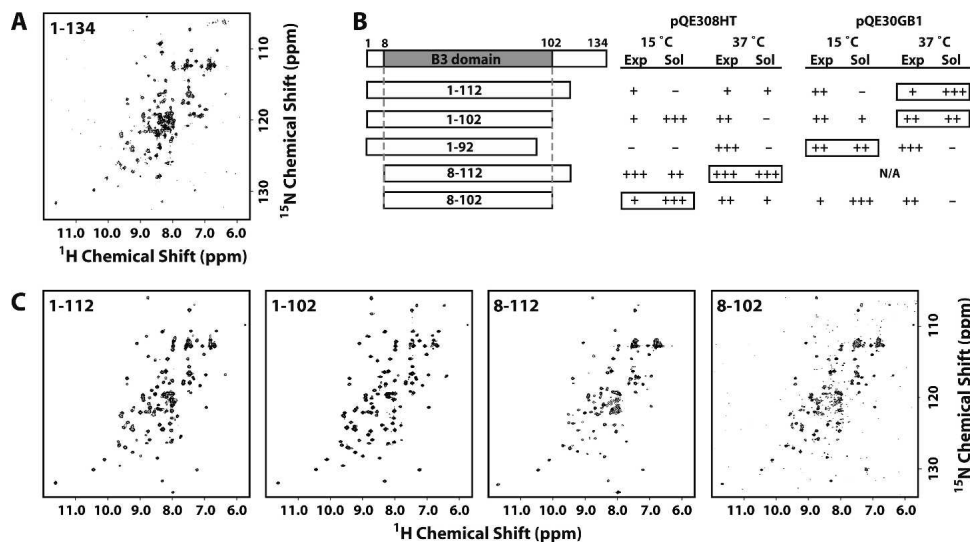

Figure 1. Expression and HSQC screening of At1g16640. (*A*) Two-dimensional ^{15}N–^{1}H HSQC spectra of full-length At1g16640 (residues 1–134). (*B*) Expression and solubility of At1g16640 truncations. (−) No expression or solubility; (+ + +) a high level of expression or solubility. Boxes indicate the optimal conditions chosen for production of each domain construct. (*C*) Two-dimensional ^{15}N–^{1}H HSQC spectra of At1g16640 truncations.
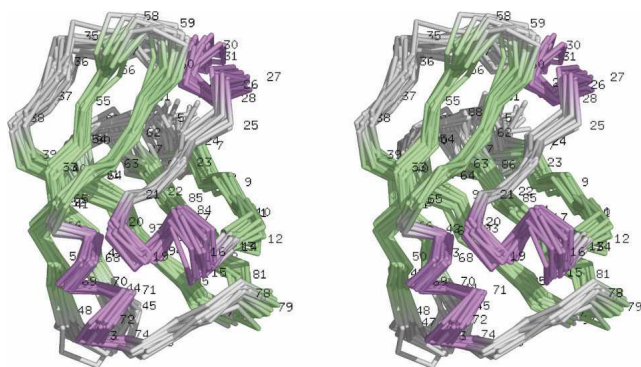
Figure 2. Structure of the At1g16640 B3 domain. The Cα-trace of the ensemble of 20 NMR structures is shown as a stereo image, with α-helices in violet and β-strands in green, produced using the program PyMOL (Delano 2002). For clarity, disordered residues of the N and C terminus (1–3 and 98–102) are not shown.

the predicted domain might improve the NMR spectrum, and designed a series of N- and C-terminal truncations of At1g16640. Using the boundaries of 8–102 as a guide, DNA fragments coding for residues 1–92, 1–102, 1–112, 8–102, and 8–112 of the protein were amplified by PCR and incorporated into two different plasmids for expression testing. Both expression vectors incorporate an N-terminal His-tag for affinity purification, and one includes the B1 Ig binding domain of protein G (GB1) as a solubility tag (Huth et al. 1997). Cloning into the GB1 vector failed for the 8–112 construct, so a total of nine expression constructs were evaluated.

Expression screening

We compared protein expression levels for each At1g16640 domain construct in *E. coli* at 15°C and 37°C (Fig. 1B). Total protein and the fraction of protein in the soluble cell lysate were assessed by SDS-PAGE and found to vary significantly. These differences arose not only between constructs, but were also dependent upon the expression vector used and the temperature at which the proteins were induced, with no obvious pattern.

Interestingly, removal of the N-terminal affinity tag with TEV protease was successful only for the fusion proteins that included the At1g16640 N terminus (1–92, 1–102, and 1–112), but not for the 8–102 or 8–112 versions. Structural results presented below reveal that Val 7 is the initial residue of the first β-strand of the B3 domain, suggesting that the cleavage site may have been sequestered by secondary structure in the N-terminally truncated constructs.

HSQC screening

A comparison of 2D ^{15}N–^{1}H HSQC spectra of the truncated At1g16640 constructs revealed significant differ-

ences (Fig. 1C). We evaluated the spectra based on the number of signals, chemical shift dispersion, and the uniformity of the peak intensities and linewidths. One sample, the 1–92 construct, precipitated heavily, precluding NMR analysis. The At1g16640 1–102 construct produced the best HSQC spectrum, with good peak dispersion and uniform peak intensity. Spectral features in the HSQC of the full-length protein (Fig. 1A) consistent with the presence of disordered residues and aggregation were eliminated. Thus, with a small set of constructs of At1g16640 designed around the predicted B3 DNA-binding domain, we isolated the folded portion of the protein and obtained a more uniform HSQC spectrum.

Structure determination

We determined the structure of the optimized At1g16640 B3 domain corresponding to residues 1–102 by NMR spectroscopy, using an automatic iterative NOE refinement method to obtain a consistent set of experimental constraints. The final NMR structure ensemble is shown in Figure 2, and structural statistics are summarized in Table 1. The structure reveals a compact seven-stranded β-barrel-like topology with a short α-helix near each end.

Table 1. *Structural statistics for 20 NMR structures*

| Constraint summary		
Dihedral angles from TALOS[19]	φ	70
	ψ	74
NOE	Long	564
	Medium	193
	Short	300
	Intraresidue	244
	Total	1445
Ramachandran from PROCHECK[24]	Most favored	87.07%
	Additionally allowed	11.42%
	Generously allowed	1.25%
	Disallowed	0.23%
Deviation from idealized geometry		
Bond lengths	RMSD (Å)	0.015
Torsion angle violations	RMSD (°)	1.3
Constraint violations		
NOE distance	Number > 0.5 Å	0 ± 0
	RMSD (Å)	0.017 ± 0.001
Torsion angle violations	Number > 5°	0 ± 0
	RMSD (°)	0.68 ± 0.12
Atomic RMSD to mean structure		
(residues 4–97) (Å)	Backbone	0.71 ± 0.09
	Heavy atom	1.28 ± 0.09
WHATCHECK[25] quality indicators	Z-score	−1.42 ± 0.35
	RMS Z-score	
	bond lengths	0.78 ± 0.02
	bond angles	0.70 ± 0.02
	Bumps	0 ± 0.00
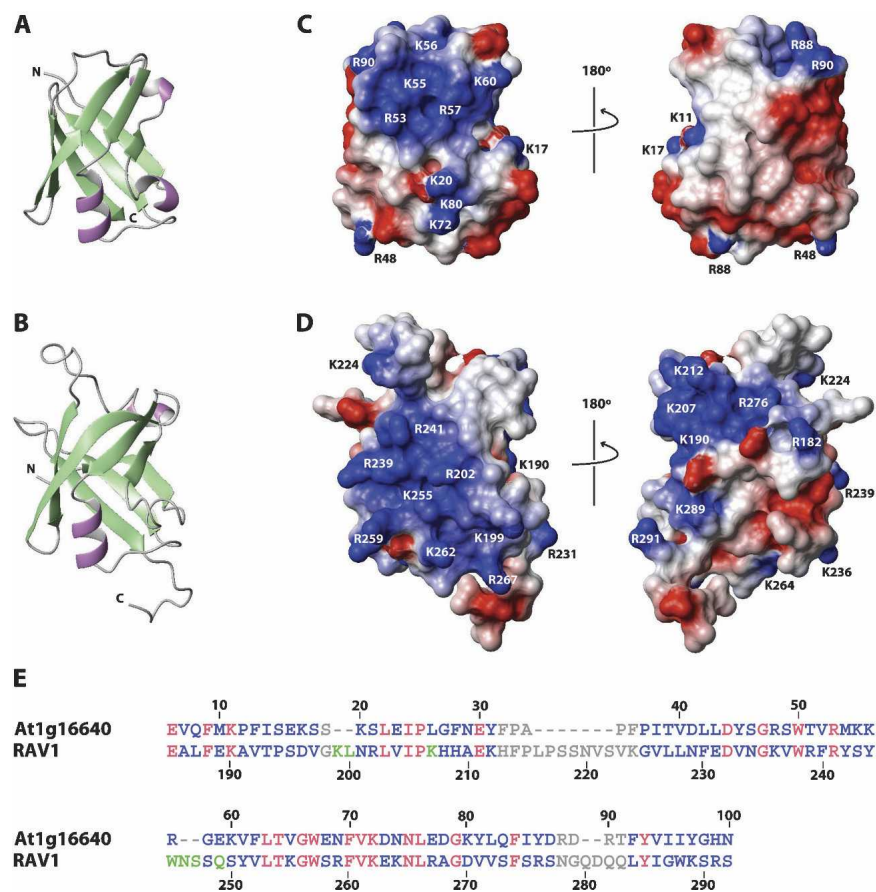
Figure 3. Comparison of the B3 domains of At1g16640 and RAV1. Ribbon diagrams of the B3 domains of (*A*) At1g16640 (residues 4–97; PDB code 1YEL) and (*B*) RAV1 (PDB code 1WID) are aligned in the same orientation. Contact surfaces showing electrostatic potentials (blue, positive; red, negative) of the B3 domains of (*C*) At1g16640 (residues 4–97) and (*D*) RAV1 are shown in the same orientation as the ribbon diagrams (*left*), and after a 180° rotation about the vertical axis (*right*). Structural representations were produced using MOLMOL (Koradi et al. 1996). (*E*) Sequences of At1g16640-B3 and RAV1-B3 aligned according to their structures using the program VAST (Gibrat et al. 1996). Residues in blue indicate structural alignment, residues in red indicate sequence identity, and residues in gray indicate regions of the proteins that do not align structurally (e.g., loop of variable length). Residues in green in RAV1 are those predicted to bind DNA (Yamasaki et al. 2004).

The B3 domain of At1g16640 thus adopts the same novel fold as that first observed in the recently reported RAV1 B3 domain structure (Yamasaki et al. 2004).

Sequence and structure comparison

Comparison of the B3 domain of At1g16640 and the previously determined structure of RAV1-B3 reveals significant structural homology (Fig. 3A,B). The conserved domains of these two proteins contain nearly identical tertiary structures (backbone RMSD ~2 Å), with the greatest differences restricted to three loops of variable length, which are longer in RAV1 than in At1g16640.

The electrostatic surface potentials of these two domains are less similar. The surface of the RAV1 B3 domain (Fig. 3D) contains two highly basic patches, within which specific residues have been shown to interact

with DNA through NMR titration experiments (Yamasaki et al. 2004), identified with green lettering in Figure 3E. Structural models of the B3 domains of ARF1 and ABI3 contain similar basic surfaces (Yamasaki et al. 2004). In contrast, At1g16640 (Fig. 3C) contains significantly fewer positively charged residues, which are clustered in a single basic region, adjacent to a dense patch of acidic residues on its surface. Potential DNA binding surfaces of At1g16640 are thus quite distinct from other classes of B3 domains.

In terms of amino acid sequence, At1g16640 is strikingly divergent from other classes of B3 domains, which display high sequence conservation. B3 domains within the ARF class are 72% identical on average. Likewise, RAV-like and ABI3/VP1-like proteins average 64% identity within their subfamilies. RAV1 and At1g16640, despite their structural similarity, share only 26% se-

quence identity. At1g16640, in fact, shows similarly weak homology to the other two classes, sharing only approximately 22% and 20% identity with ARFs and ABI3/VP1-like proteins, respectively (Poirot et al. 2004). Thus, based on sequence similarity, At1g16640 is unlikely to be categorized as a member of any of these B3 protein subfamilies.

Although At1g16640 appears to be quite distinct from the RAV, ARF, and ABI3/VP1 classes of B3 domains, these subfamilies represent only a fraction of B3-containing proteins. The B3 superfamily currently includes 363 members from various plant species, grouped into 16 distinct structural architectures based on their association with other conserved domain combinations (Bateman et al. 2004). Unlike most well-defined B3 proteins, At1g16640 contains only one identifiable domain. By comparison, RAV1 contains an additional DNA-binding motif of the AP2/ERF-type, and most ABI3/VP1- and ARF-like proteins contain additional protein interaction or dimerization domains (Yamasaki et al. 2004). Presumably, the accompanying domains contribute to the biological activity of these transcription factors and in their absence At1g16640 may function quite differently.

Conclusions

In determining the structure of the At1g16640 B3 domain, we have shown that bioinformatic analysis and 2D NMR screening of a small panel of truncated protein constructs can be used to salvage failed structural genomics targets. Our results present the second structure of a B3 domain and show that this novel fold is highly conserved among family members, despite relatively low sequence conservation. The At1g16640 protein has not been shown to bind DNA. Compared to RAV1 and other B3 proteins that bind DNA, At1g16640 has a less electropositive surface, lacks conserved putative DNA-binding residues and possesses no additional recognizable interaction domains. Thus, we hypothesize that At1g16640 may not participate in transcriptional regulation, but instead represents a distinct functional class of B3 domains.

Materials and methods

Cloning

Gene fragments were amplified by PCR from a plasmid containing the cDNA coding for full-length At1g16640 using DNA primers specific for various N- and C-terminal truncations as described in the Results section. The primers used also coded for 5′ BamHI and 3′ HindIII sites to facilitate ligation of the gene fragments into modified pQE30 vectors (Qiagen). The vectors, known as pQE308HT and pQE30GB1, both contained histidine affinity tags (His$_8$ in pQE308HT and His$_6$ in pQE30GB1) and a tobacco etch virus (TEV) protease cleavage site, while the latter also contained an insertion between the His-tag and TEV cleavage site coding for the B1 Ig binding domain of protein G (GB1). All expression constructs were verified by DNA sequencing.

Protein expression

Plasmids were transformed into E. coli strain SG13009[pREP4] (Qiagen) for expression. Cells were grown in 25 mL LB media containing 150 µg/mL ampicillin and 50 µg/mL kanamycin at 37°C until reaching a cell density of A$_{600}$ = 0.6. Isopropyl-β-D-thiogalactopyranoside was then added to a final concentration of 1 mM to induce expression of the proteins. Upon induction, the cultures were split into two equal parts and grown at both 37 and 15°C. One-milliliter samples were taken 2.5 and 5 h post-induction and 5 h and 24 h post-induction for the cultures at 37°C and 15°C, respectively. The samples were harvested, sonicated, and analyzed for protein expression and solubility by SDS-PAGE. After selecting the proper expression conditions for each construct, isotopically-labeled proteins were prepared for NMR by growing 1-L cultures of protein in M9 media containing ^{15}N-ammonium chloride and/or ^{13}C-glucose as the sole nitrogen and carbon sources, respectively.

Protein purification

Cells harvested from a 1-L culture were lysed using a French pressure cell and purified by metal affinity chromatography according to a previously published protocol (Lytle et al. 2004). Following purification, the protein solutions were each dialyzed into 2×4 L of 20 mM sodium phosphate at pH 7.0, 50 mM sodium chloride. The resulting purified proteins were then concentrated to 500 µL for analysis by NMR, and the identity and purity of the proteins were verified by SDS-PAGE.

NMR spectroscopy

NMR samples were prepared in buffers containing 20 mM sodium phosphate at pH 7.0, 50 mM sodium chloride, and 5% ^{2}H$_2$O. Soluble domain constructs were screened by ^{15}N–^{1}H HSQC using samples containing ~0.2–0.5 mM U-^{15}N protein, and the sample used for structure determination of At1g16640 1–102 contained ~1 mM U-^{13}C/^{15}N protein. All NMR data were acquired at 25°C on a Bruker 600 MHz spectrometer equipped with a triple-resonance Cryo-Probe and processed with NMRPipe software (Delaglio et al. 1995). The total acquisition time for all NMR spectra was ~280 h. Over 90% of the backbone ^{1}H, ^{15}N, and ^{13}C resonance assignments were obtained in an automated manner using the program Garant (Bartels et al. 1996), with peaklists from 3D HNCO, HNCACO, HNCA, HNCOCA, HNCACB, and CCONH spectra generated manually with XEASY (Bartels et al. 1995) or automatically with SPSCAN. Side chain assignments were completed manually from 3D HCCONH, HCCH-TOCSY, and ^{13}C(aromatic)-edited NOESY-HSQC spectra.

Structure determination

Distance constraints were obtained from 3D ^{15}N-edited NOESY-HSQC and ^{13}C-edited NOESY-HSQC spectra (τ_{mix} = 80 msec). Backbone φ and ψ dihedral angle constraints were generated

from secondary shifts of the $^1H^\alpha$, $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C'$, and ^{15}N nuclei using the program TALOS (Cornilescu et al. 1999). Structures were generated in an automated manner using the CANDID module of the torsion angle dynamics program CYANA (Herrmann et al. 2002), which produced an ensemble with high precision and low residual constraint violations that required minimal manual refinement. The 20 CYANA conformers with the lowest target function were subjected to a molecular dynamics protocol in explicit solvent (Linge et al. 2003) using XPLOR-NIH (Schwieters et al. 2003).

Accession numbers

Coordinates and restraints have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/) under PDB code 1YEL. All time-domain NMR data and chemical shift assignments have been deposited in BioMagResBank (http://www.bmrb.wisc.edu/) under BMRB entry 6464.

Acknowledgments

References

Bartels, C., Xia, T.-H., Billeter, M., Güntert, P., and Wüthrich, K. 1995. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **5:** 1–10.

Bartels, C., Billeter, M., Güntert, P., and Wüthrich, K. 1996. Automated sequence-specific NMR assignments of homologous proteins using the program GARANT. *J. Biomol. NMR* **7:** 207–213.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. 2004. The Pfam protein families database. *Nucleic Acids Res.* **32:** D138–D141.

Cornilescu, G., Delaglio, F., and Bax, A. 1999. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13:** 289–302.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. 1995. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6:** 277–293.

Delano, W.L. 2002. The PyMOL molecular graphics system. DeLano Scientific, San Carlos, CA.

Gibrat, J.F., Madej, T., and Bryant, S.H. 1996. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6:** 377–385.

Giraudat, J., Hauge, B.M., Valon, C., Smalle, J., Parcy, F., and Goodman, H.M. 1992. Isolation of the *Arabidopsis* ABI3 gene by positional cloning. *Plant Cell* **4:** 1251–1261.

Herrmann, T., Güntert, P., and Wüthrich, K. 2002. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319:** 209–227.

Huth, J.R., Bewley, C.A., Jackson, B.M., Hinnebusch, A.G., Clore, G.M., and Gronenborn, A.M. 1997. Design of an expression system for detecting folded protein domains and mapping macromolecular interactions by NMR. *Protein Sci.* **6:** 2359–2364.

Kagaya, Y., Ohmiya, K., and Hattori, T. 1999. RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. *Nucleic Acids Res.* **27:** 470–478.

Koradi, R., Billeter, M., and Wüthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14:** 51–55.

Linge, J.P., Williams, M.A., Spronk, C.A., Bonvin, A.M., and Nilges, M. 2003. Refinement of protein structures in explicit solvent. *Proteins* **50:** 496–506.

Lytle, B.L., Peterson, F.C., Qiu, S.H., Luo, M., Zhao, Q., Markley, J.L., and Volkman, B.F. 2004. Solution structure of a ubiquitin-like domain from tubulin-binding cofactor B. *J. Biol. Chem.* **279:** 46787–46793.

McCarty, D.R., Carson, C.B., Stinard, P.S., and Robertson, D.S. 1989. Molecular analysis of viviparous-1: An abscisic acid-insensitive mutant of maize. *Plant Cell* **1:** 523–532.

Poirot, O., Suhre, K., Abergel, C., O'Toole, E., and Notredame, C. 2004. 3DCoffee@igs: A web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res.* **32:** W37–W40.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., et al. 2000. Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* **290:** 2105–2110.

Schwieters, C.D., Kuszewski, J.J., Tjandra, N., and Clore, G.M. 2003. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160:** 65–73.

Suzuki, M., Kao, C.Y., and McCarty, D.R. 1997. The conserved B3 domain of VIVIPAROUS1 has a cooperative DNA binding activity. *Plant Cell* **9:** 799–807.

Tyler, R.C., Aceti, D.J., Bingman, C.A., Cornilescu, C.C., Fox, B.G., Frederick, R.O., Jeon, W.B., Lee, M.S., Newman, C.S., Peterson, F.C., et al. 2005a. Comparison of cell-based and cell-free protocols for producing target proteins from the *Arabidopsis thaliana* genome for structural studies. *Proteins* **59:** 633–643.

Tyler, R.C., Sreenath, H.K., Singh, S., Aceti, D.J., Bingman, C.A., Markley, J.L., and Fox, B.G. 2005b. Auto-induction medium for the production of [U-15N]- and [U-13C, U-15N]-labeled proteins for NMR screening and structure determination. *Protein Expr. Purif.* **40:** 268–278.

Ulmasov, T., Hagen, G., and Guilfoyle, T.J. 1997. ARF1, a transcription factor that binds to auxin response elements. *Science* **276:** 1865–1868.

Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Tomo, Y., et al. 2004. Solution structure of the B3 DNA binding domain of the *Arabidopsis* cold-responsive transcription factor RAV1. *Plant Cell* **16:** 3448–3459.