Journal of Biomolecular NMR **28:** 397–400, 2004.
KLUWER/ESCOM

Letter to the Editor: Structure of the hypothetical protein At3g17210 from *Arabidopsis thaliana*[*]

Betsy L. Lytle[a,c], Francis C. Peterson[a,c], Kelly L. Kjer[a,c], Ronnie O. Frederick[b,c], Qin Zhao[b,c], Sandy Thao[b,c], Craig Bingman[b,c], Kenneth A. Johnson[b,c], George N. Phillips, Jr.[b,c] & Brian F. Volkman[a,c,**]

[a]*Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI 53226, U.S.A.;* [b]*Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706-1544, U.S.A.;* [c]*Center for Eukaryotic Structural Genomics, U.S.A.*

Received 12 August 2003; Accepted 17 October 2003

Key words: automated refinement, high-throughput, NMR, structural genomics

Biological context

Completion of the *Arabidopsis thaliana* genome sequence (The Arabidopsis Initiative, 2000) has fostered an array of large-scale research efforts directed on this model plant system, including an extensive gene knockout collection (Krysan et al., 1999; Sussman et al., 2000; Alonso et al., 2003), a series of functional genomics projects (Ausubel, 2002), and international structural genomics efforts (Yokoyama et al., 2000, Norvell and Machalek, 2000). Here we describe the structure of the hypothetical *Arabidopsis thaliana* protein At3g17210, determined by NMR and X-ray crystallography as part of a pilot project on the feasibility of high-throughput structure determination. Goals of this project also include the characterization of proteins of unknown function and identification of novel folds. At3g17210 has no identifiable sequence homology to characterized proteins, but adopts a novel dimeric fold that was only recently discovered in the structure of a bacterial monooxygenase (Sciara et al., 2003). Inspection of the structure has yielded no clear indication of its biological function, though a deep cavity is formed within each domain of the At3g17210 dimer. Interestingly, the At3g17210 sequence is 50% identical to a thermostable stress-responsive protein, SP-1, from *Populus tremula* (also known as Pop3 in other *Populus* species), though neither has significant homology to other stress-related proteins or the small heat-shock proteins (Wang et al., 2002). At3g17210 and its homologs may therefore represent a new class of stress-response proteins in plants.

Methods and results

The *Arabidopsis thaliana* gene encoding a hypothetical protein of 109 residues, At3g17210, was cloned from genomic DNA into the pET-15b vector (Novagen) for overexpression in *E. coli*. The expression vector was constructed using the restriction endonucleases *NdeI* and *BamHI* such that the final purified target protein would include three additional residues (NH_3^+-Gly-Ser-His-) after thrombin removal of an N-terminal hexahistidine affinity tag. A freshly-transformed culture using the Rosetta(DE3) strain (Novagen) was grown at 37 °C to an OD_{600} of 0.7 and induced with 1 mM IPTG for 6 h. Stable isotope labeling for NMR was achieved by production on M9 medium (Maniatis et al., 1986) containing ^{15}N (98%) ammonium chloride and ^{13}C (98%) glucose, supplemented with vitamins (Weber et al., 1992). The protein was purified by metal affinity chromatography and cleaved using ∼ 8 ng thrombin (Sigma) (1:10,000 molar ratio). Thrombin cleavage was irreversibly inhibited by a 5-fold excess of FPR-chloromethyl ketone (Bachem). A subsequent metal affinity chromatography step retained the uncut protein and cleaved tag peptide while allowing pure target protein to be recovered from the column flow-through fractions. A single [$U-^{15}N$, ^{13}C]-labeled sample was prepared at 1 mM protein concentration in 90% H_2O/10% D_2O containing 50 mM sodium chloride and 20 mM sodium phosphate buffer at pH 6.5. The 2D ^{15}N-^1H HSQC spectrum (Figure 1) contained ∼120 highly dispersed signals of uniform intensity, consistent with

[*]Coordinates for the NMR structure ensemble have been deposited in the PDB (accession 1Q53), and all time-domain data and resonance assignments have been deposited in the BMRB (accession 5843).

[**]To whom correspondence should be addressed. E-mail: bvolkman@mcw.edu

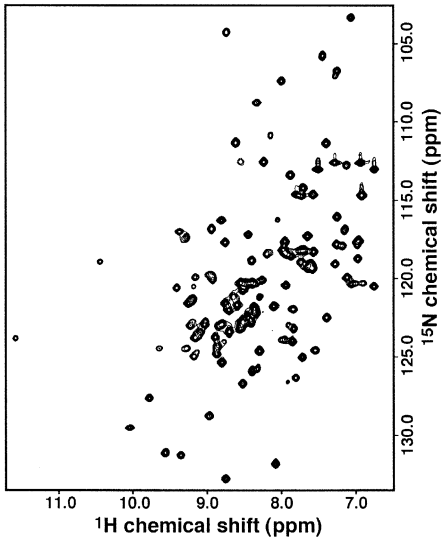

Figure 1. 2D ^{1}H-^{15}N HSQC spectrum of [U-^{15}N, ^{13}C] At3g17210 at 600 MHz and 25 °C.

the number (115) predicted from the amino acid sequence, suggesting that the protein was folded in a unique conformation suitable for high-resolution structural analysis.

All NMR data were acquired at 25 °C on a Bruker 600 MHz spectrometer equipped with a triple-resonance CryoProbe® and processed with NMRPipe (Delaglio et al., 1995). Backbone resonance assignments were obtained in a semi-automated manner using the program Garant (Bartels et al., 1996) with peaklists generated manually with XEASY (Bartels et al., 1995) from 3D HNCO, HNCACO, HNCA, HNCOCA, HNCACB, and CCONH spectra. Sidechain assignments were completed manually using 3D ^{15}N-edited TOCSY-HSQC, HCCH-TOCSY and a ^{13}C-edited NOESY-HSQC optimized for aromatic groups. Distance constraints were obtained from 3D ^{15}N-edited NOESY-HSQC and ^{13}C-edited NOESY-HSQC spectra (τ_{mix} = 80 ms). Each FID was the average of either 2 or 4 transients, and the total acquisition time for all NMR spectra was 180 h. All time-domain NMR data and chemical shift assignments have been deposited in BioMagResBank entry 5843.

Structure calculations were performed using the torsion angle dynamics program Cyana (Güntert et al., 1997). In addition to NOE-derived distances, backbone ϕ and ψ angle constraints were obtained from analysis of chemical shifts using the program TALOS (Cornilescu et al., 1999). Initial NOE assignments and a consistent tertiary fold were obtained for At3g17210 using the automated CANDID iterative refinement module of Cyana (Herrmann et al., 2002). Subsequent refinement was performed by manual editing of peaklists in XEASY to add or correct NOE assignments, followed by recalculation of structures in Cyana. At each stage, 100 structures were calculated using 10,000 steps of simulated annealing, and a final ensemble of 20 structures was selected based on Cyana target function values.

Database searches for sequence and structure homology performed at intermediate stages of refinement yielded no matches, but a VAST search performed on the nearly completed At3g17210 structure identified a protein with clear structural similarity (PDB ID 1LQ9). Interestingly, the x-ray structure of this protein, ActVA-Orf6 from *Streptomyces coelicolor*, is dimeric (Sciara et al., 2003), while the At3g17210 NMR structure had been refined as a monomer. We measured the translational self-diffusion coefficient (D_s) for At3g17210 (12.5 kDa monomer/ 25 kDa dimer) and a series of monomeric proteins with M_r from 8.6 to 25 kDa in solution conditions identical to those used for the structure determination. Of the model proteins, chymotrypsinogen (25, kDa, 0.98×10^{-6} cm^2 s^{-1}) had the D_s value closest to At3g17210 (0.93×10^{-6} cm^2 s^{-1}). These results and comparisons with D_s values obtained in other PFG diffusion studies of protein oligomerization suggest that At3g17210 is dimeric under these solution conditions.

Upon closer inspection, we discovered that the most persistent NOE violations clustered in two regions of the At3g17210 sequence that correspond to domain-swapped regions of the ActVA-Orf6 dimer. Reassignment of a series of these problematic restraints as intermolecular NOEs substantially improved the overall agreement, and refinement of At3g17210 was therefore completed in the context of the homodimeric assembly. The final structure was generated using 141 dihedral angle constraints and 2001 unique, non-trivial distance constraints for each monomer (1831 intramonomer, 170 intermonomer) derived from a total of 4152 assigned NOE crosspeaks. Structural statistics (Table 1) were obtained using Cyana, Molmol (Koradi et al., 1996) and Procheck-NMR (Laskowski et al., 1996). The ensemble of 20 structures (Figure 2A) and the experimental constraints were deposited in the PDB (accession code: 1Q53). The At3g17210 tertiary structure consists of a four-stranded antiparallel β-sheet and three α-helices, arranged in a β-α-β-β-α-α-β topology. Two sheets join at their edges to form an oblong β-barrel, flanked by

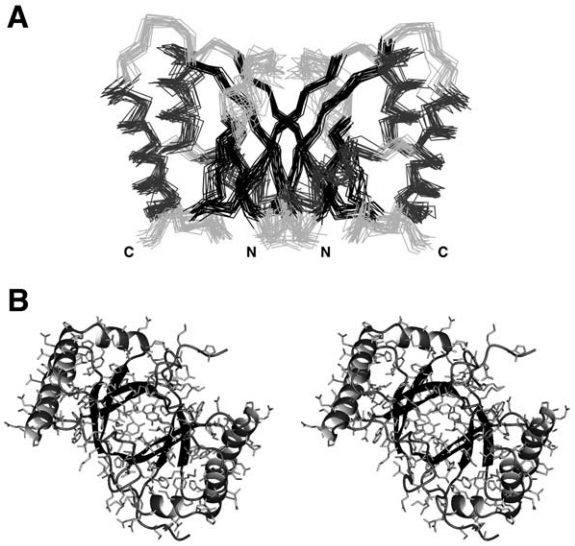

Figure 2. NMR structure of At3g17210. (A) Ensemble of 20 At3g17210 Cyana structures (C^{α} trace). Helices, β-strands and loops are shown in gray, black and light gray, respectively. Disordered residues of the N-terminus (1–8) are removed for clarity. (B) Stereoview of a representative conformer (nearest to the mean) with all sidechain heavy atoms shown. The ribbon diagram is rotated 90° about the horizontal axis relative to the family of structures.

four helices on each of the opposing faces (Figure 2B). C-terminal residues extending from the β4 strand of each monomer wrap around and connect with the β2 strand and α1 helix of the opposing monomer to form the dimer interface.

Discussion and conclusions

In a parallel effort, selenomethionine-labeled At3g17210 was crystallized in space group $P6_2$ and its structure solved to 1.9 Å with a four-wavelength MAD experiment around the Se(K) X-ray absorption edge (Bingman, C., Johnson, K. A. and Phillips, G. N. Jr., *in preparation*). An intimate dimer is situated around a crystallographic twofold axis. Solution and refinement of the NMR and crystal structures were completed independently and concurrently, and the results are compared with the ActVA-Orf6 crystal structure in Figure 3. All three structures display the same overall fold and secondary structure topology. Helices 2 and 3 arch over the β-sheet forming a deep cavity in each monomer that serves as the monooxygenase active site in ActVA-Orf6. While its functional role is unknown, the same cleft in At3g17210 is lined with hydrophobic groups and surrounded by the sidechains from His 13, Lys 20, Lys 74 and His 86. Along with

Table 1. Structural statistics for 20 At3g17210 NMR structures

Experimental constraints	Number
TALOS dihedral (ϕ and ψ)	141
NOE distance (total used for dimer structure)	
Long (intramolecular)	1052
Long (intermolecular)	340
Medium	844
Short	1030
Intraresidue	736
Total	4002
Ramachandran statistics (% of all residues)	
Most favored	81.1
Additionally allowed	14.4
Generously allowed	2.4
Disallowed	2.1
Violations	
Target function ($Å^2$)	0.67 ± 0.15
Upper limit violations	
Number > 0.1 Å	6 ± 2
Sum of violations (Å)	2.9 ± 0.5
Maximum violation (Å)	0.28 ± 0.04
Torsion angle violations	
Number > 2°	1 ± 1
Sum of violations (°)	17.9 ± 5.9
Maximum violation (°)	2.57 ± 1.19
Van der Waals violations	
Number > 0.2 Å	0 ± 0
Sum of violations (Å)	2.9 ± 0.5
Maximum violation (Å)	0.14 ± 0.03
Atomic r.m.s.d. to mean structure (Å)	
Dimer (residues 10-54,60-112, 210-254,260-312)	
Backbone	0.95 ± 0.13
Heavy atom	1.30 ± 0.11
Monomer A (residues 10-54,60-112)	
Backbone	0.83 ± 0.08
Heavy atom	1.20 ± 0.10
Monomer B (residues 210-254, 260-312)	
Backbone	0.83 ± 0.15
Heavy atom	1.21 ± 0.14

Asp 104, the only acidic group near the cavity opening, each of these basic residues is highly conserved among proteins from *Arabidopsis thaliana* with significant sequence homology to At3g17210 (Figure 3D). In contrast, none of the active site residues identified in the ActVA-Orf6 structure (Sciara et al., 2003) are conserved in At3g1720, suggesting that a common fold has evolved to support distinct functions in plants and bacteria.

In our analysis of At3g17210, we sought to establish a prototype system for high throughput NMR structure determination at the Center for Eukaryotic Structural Genomics. Data acquisition on a cryoprobe-equipped spectrometer provided a significant time reduction by eliminating the need for extensive signal averaging. We obtained correct backbone resonance

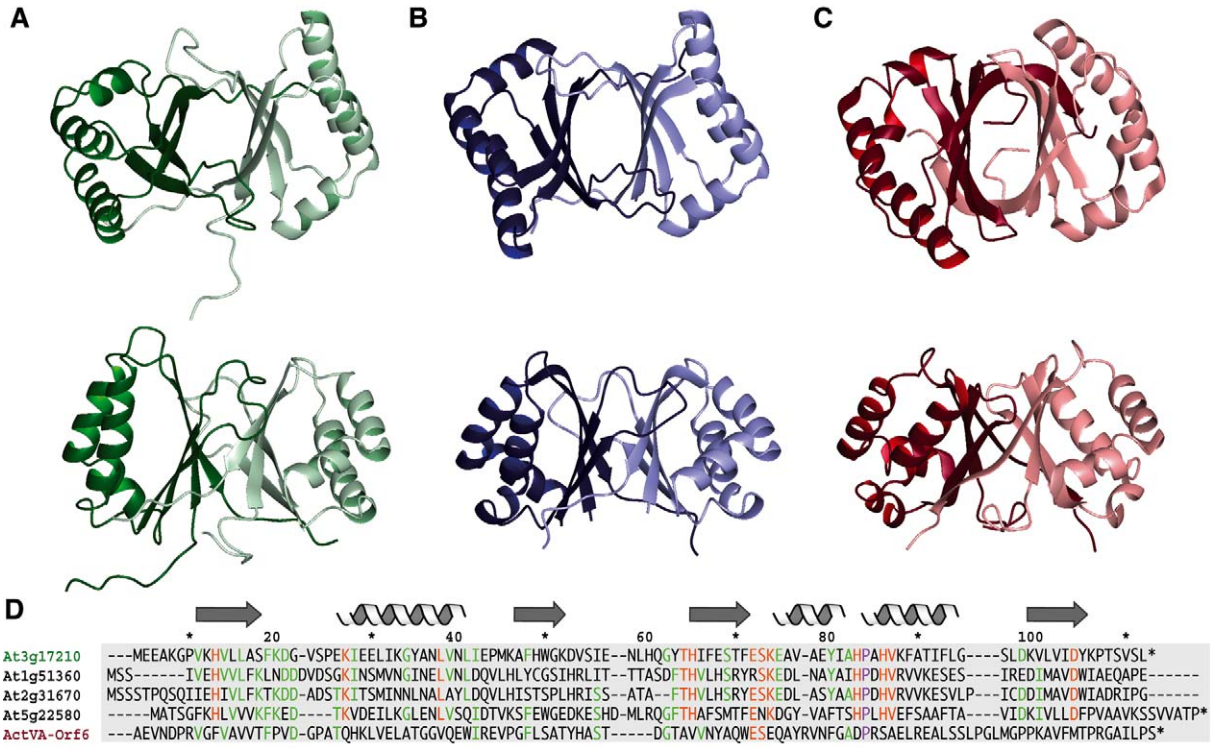

Figure 3. Structures of At3g17210 and ActVA-Orf6 are closely related despite low sequence homology. Ribbon diagrams of (A) the NMR (PDB ID 1Q53) and (B) X-ray crystal (PDB ID 1Q4R) structures of At3g17210; and (C) the ActVA-Orf6 crystal structure (PDB ID 1L9Q). The two orientations are related by a 90° rotation about the horizontal axis, and light and dark coloring distinguished the two monomers within each dimeric structure. (D) Multiple sequence alignments for At3g17210, ActVA-Orf6 and other hypothetical proteins in the *Arabidopsis* genome with sequence similarity to At3g17210. Conserved positions are colored green, orange and magenta for residues conserved in 3, 4 and 5 of the sequences. The level of sequence identity for the regions aligned with At3g17210 is 24% (At1g51360), 23% (At2g31670), 35% (At5g22580) and 50% (ActVA-Orf6). Asterisks identify C-terminal residues; the other sequences extend beyond the aligned domains.

assignments for 92% of At3g17210 residues using the backbone triple-resonance experiments listed above and the automated approach of Garant (Bartels et al., 1996). Automated NOE refinement with CANDID (Herrmann et al., 2002) was noteworthy for its ease of use and computational speed.

In summary, we report the structure of At3g17210, a protein of unknown function from *Arabidopsis thaliana* determined using a semi-automated assignment and refinement strategy. Knowledge of the novel three-dimensional structure of this protein will be coupled to the results of functional genomics efforts on this eukaryotic system, a combination that will yield important biochemical insights in the plant kingdom.

Acknowledgements

The authors gratefully acknowledge Adrian Hegeman for mass spectrometry analysis. The Center for Eukaryotic Structural Genomics is supported by NIH grant P50 64598 (John L. Markley, PI).

References

Alonso, J.M. et al. (2003) *Science*, **301**, 653–657.
Ausubel, F.M. (2002) *Plant Physiol*, **129**, 394–437.
Bartels, C. et al. (1996) *J. Biomol. NMR*, **7**, 207–213.
Bartels, C. et al. (1995) *J. Biomol. NMR*, **5**, 1–10.
Cornilescu, G. et al. (1999) *J. Biomol. NMR*, **13**, 289–302.
Delaglio, F. et al. (1995) *J. Biomol. NMR*, **6**, 277–293.
Güntert, P. et al. (1997) *J. Mol. Biol.*, **273**, 283–298.
Herrmann, T. et al. (2002) *J. Mol. Biol.*, **319**, 209–227.
Koradi, R. et al. (1996) *J. Mol. Graphics*, **14**, 51–55.
Krysan, P.J. et al. (1999) *Plant Cell*, **11**, 2283–2290.
Laskowski, R.A. et al. (1996) *J. Biomol. NMR*, **8**, 477–486.
Maniatis, T. et al. (1986) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Norvell, J.C. and Machalek, A.Z. (2000) *Nat. Struct. Biol.*, **7** (Suppl.), 931.
Sciara, G. et al. (2003) *Embo J.*, **22**, 205–215.
Sussman, M.R. et al. (2000) *Plant Physiol.*, **124**, 1465–1467.
The Arabidopsis Initiative (2000) *Nature*, **408**, 796–815.
Wang, W.X. (2002) *Plant Physiol.*, **130**, 865–875.
Weber, D.J. et al. (1992) *Proteins*, **13**, 275–287.
Yokoyama, S. et al. (2000) *Nat. Struct. Biol.*, **7** (Suppl.), 943–945.