Published in final edited form as:

Proteins. 2008 May 1; 71(2): 546–551. doi:10.1002/prot.21936.

Solution structure of At3g28950 from *Arabidopsis thaliana*

Norberto B. de la Cruz, Francis C. Peterson, and Brian F. Volkman*

Department of Biochemistry and Center for Eukaryotic Structural Genomics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin, 53226

Abstract

We determined the solution structure of At3g28950 from *A. thaliana*, a homolog of At5g39720, whose structure we solved earlier. The secondary structure of the 165-aa protein consists of a 5-strand anti-parallel β -barrel domain flanked by two α -helices and a 2-strand β -sheet; an additional free C-terminal α -helix extends into solution. Bioinformatic searches and analyses suggest that members of this growing set of structurally related proteins have been recruited to serve a wide variety of functions ranging from γ -glutamyl cyclotransferase activity, to participation in plant responses to chemical and biotic stimuli. Expression of a human homolog is elevated in bladder cancer tissues. Expression patterns for At3g28950 and its *Arabidopsis* paralogs suggest that each one evolved different a physiological role. The At3g28950 structure was solved as part of a structural genomics effort, and the results demonstrate how such a project can further understanding of genome evolution in addition to sequence-structure and structure-function relationships.

Introduction

Structural genomic efforts broaden our understanding of how nature adapts useful molecular motifs for diverse biological roles. Examples include the β -grasp domains used by ubiquitin and many related proteins 1, and the seven membrane-spanning helix motif employed by a large family of cell-surface receptors 2. Along with enriching the dataset for the analysis of sequence-structure and structure-function relationships, structural genomics projects are providing glimpses as to how such functions evolve within and across species. We previously determined the structure of At5g39720, an *Arabidopsis* protein containing a 5-stranded antiparallel β -barrel and an extended C-terminal α -helix 3. Our analysis indicated that At5g39720 is structurally related to proteins from *P. horikoshii*, *M. musculus*, and *E. coli*, which are homologous to enzymes involved in butirosin metabolism. It is also related to AIG2, an *Arabidopsis* protein expressed in response to bacterial infection 4, and is thus a member of the AIG2-like family. Upon Pfam reorganization, At5g39720 also emerged as the first representative structure for three protein families: AIG2-like, UPF0131, and ChaC (A. Bateman, personal communication). Here we report the structure of At3g28950, an At5g39720 homolog that is also from *Arabidopsis* and shares 60.7% identity, solved by NMR spectroscopy at the Center for Eukaryotic Structural Genomics (CESG; <http://www.uwstructuralgenomics.org>).

Although At3g28950's functional role is unknown, analysis of this new structure uncovered additional structural homologs including a human protein expressed in cancer cells, which further underscores the versatility of the motif. Moreover, analysis of the temporal and spatial gene expression patterns of all seven AIG2-like proteins from *Arabidopsis* suggests

*To whom correspondence should be addressed: Brian F. Volkman Medical College of Wisconsin Department of Biochemistry 8701 Watertown Plank Road Milwaukee, Wisconsin 53226 Tel: (414)-456-8400 Fax: (414)-456-6510 bvolkman@mcw.edu.

that family members, despite retaining structural similarities, have been recruited to very different physiological roles and thus evolved diversified functions.

Materials and Methods

Target selection

Protein At3g28950 was selected as part of the CESH effort to determine the structure of proteins with potentially unique folds. After screening for soluble expression and purification, the protein was placed in the pipeline for production, purification, and structure determination via NMR. Data associated with pipeline operations were managed using the SESAME LIMS 5.

Protein expression and purification

Samples of At3g28950 were prepared in [U - ^{13}C , ^{15}N]-labeled form according to the cell-based expression protocol described previously 6. Briefly, the protein was expressed with an N-terminal His₆ fusion tag in *E.coli* B834 in medium containing ^{15}N -NH₄Cl and ^{13}C -glycerol and ^{13}C -glucose, purified via Ni-IMAC affinity chromatography, treated with tobacco etch virus protease to remove the N-terminal tag, which was separated from the protein by gel filtration chromatography.

NMR Spectroscopy

All NMR data were acquired at 298 K on a Bruker 600 MHz spectrometer equipped with a triple-resonance CryoProbe and processed with NMRpipe software 7. The total acquisition time for all NMR spectra was ~260 h, and no protein degradation or unfolding was detected by comparison of one- and two-dimensional spectra acquired at the start and end of data collection. Heteronuclear ^{15}N - ^1H NOE values were determined from an interleaved pair of two-dimensional gradient sensitivity enhanced correlation spectra of [U - ^{13}C , ^{15}N]-At3g28950 using a 5 s ^1H saturation period and analyzed using CARG (Keller, 2004). About 85% of the backbone chemical shifts were assigned in an automated manner using the program Garant 8, with peaklists from three-dimensional HNCO, HNCACO, HNCOCA, HNCA, HNCACB, and CCONH spectra that were generated automatically with SPSCAN. Side-chain assignments were completed manually from three-dimensional HBHACONH, HCCONH, HCCH-TOCSY and ^{13}C (aromatic)-edited NOESY-HSQC spectra analyzed using XEASY 9.

Structure determination

A total of 1810 unique, non-trivial NOE distance constraints were obtained from three-dimensional ^{15}N -edited NOESY-HSQC, ^{13}C (aromatic)-edited NOESY-HSQC and ^{13}C edited NOESY-HSQC spectra ($\tau_{\text{mix}} = 80$ ms). A total of 189 backbone ϕ and ψ dihedral angle constraints were generated from secondary shifts of $^1\text{H}^\alpha$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}^\gamma$ and ^{15}N nuclei using the program TALOS 10. Structures were generated in an automated manner using the NOEASSIGN module of the torsion angle dynamics program CYANA 2.1 11. A starting model for the At3g28950 structure was created from the At5g39720 NMR structure (PDB ID 2G0Q) 3 using MODELLER 12. Automated calculations starting from random conformations or the homology model converged to the same folded structure, but use of the homology model resulted in a larger number of NOE constraints and higher precision for the resulting ensemble. An ensemble with high precision and low residual constraint violations was produced after a series of automated calculations and iterative manual refinements. The 20 CYANA conformers with the lowest target function were selected from a total of 100 and subjected to a molecular-dynamics protocol in explicit solvent 13 using XPLOR-NIH 14.

Results and Discussion

Structure of At3g28950

At3g28950 consists of 165 amino acid residues with a molecular weight of 18.5 kDa. The two-dimensional ^{15}N - ^1H HSQC spectrum (Fig. 1A) displayed a well-dispersed pattern of peaks with uniform intensity, consistent with a stable folded domain. Chemical shift assignments for the backbone and sidechains were 85% complete and NMR data were deposited in the BioMagResBank (entry 15295). The final ensemble of 20 conformers (PDB code: 2JQV) is shown in (Fig. 1B), and a summary of the experimental restraints and structural statistics is given in Table 1. Like At5g39720, the structure revealed a five-stranded antiparallel β -barrel (β 1- β 5- β 2- β 6- β 7) flanked by two α -helices and a β -sheet (β 3- β 4) (Fig. 1C). The C-terminus contains a 23-residue α -helix that protrudes from the molecule. Consistent patterns of ^{13}C and ^1H secondary shifts and $d_{\text{NN}}(i, i+1)$, $d_{\alpha\text{N}}(i, i+3)$ and $d_{\alpha\beta}(i, i+3)$ NOEs provide dihedral and distance constraints that define an α -helix for residues 123–146. However, high backbone atomic r.m.s.d. values (Fig. 1D) and the lack of long-range NOEs between the C-terminal helix and the rest of the molecule indicate that these residues are unconstrained relative to the rest of the protein. Heteronuclear ^{15}N - ^1H NOE values shown in Figure 1E reflect the relative rigidity of the secondary structure elements and disorder in the C-terminus, specifically beyond the untethered C-terminal α -helix. The NOE values of residues 111–116, which are each between 0.5 and 0.6, suggest that greater flexibility in this region allows the C-terminal helix to assume multiple orientations.

Structural Homologs

VAST 15 and FATCAT 16 searches returned a number of structural homologs most of which aligned with the first ~100 residues. At5g39720 aligns most closely with a sequence identity of 60.7% and an RMSD of 2.0 Å over 112 C^α atoms (Fig. 2A and B). The next highest match is the human protein C7orf24, whose crystal structure was recently solved by the CESG 17. This protein of unknown function displays 23.2% sequence identity with At3g28950 and alignment over 99 C^α atoms yielded an RMSD of 1.9 Å. C7orf24 was recently identified in a proteomic screen of bladder cancer, and the protein is expressed at elevated levels in bladder cancer relative to normal tissue 18. Several cancer cell lines also express C7orf24, and siRNA knock-down of protein levels produced an antiproliferative response, suggesting that this protein may be a suitable target for cancer therapy 18. Interestingly, the crystal structure of C7orf24 shows that the C-terminal helix packs against the β -barrel domain, in contrast to the solution NMR structures of At5g39720 and At3g38950 and the crystal structure of PH0828 in which the C-terminal helices extends away from the β -barrel domain 3. In the absence of solution NMR data it is uncertain whether the C7orf24 C-terminal helix associates specifically with the β -barrel domain or is a flexible element that adopts one particular orientation favored by crystal packing interactions.

Some of the lower hits in the search were previously reported as structural homologs of At5g39720 3. Designated in the Structural Classification of Proteins (SCOP) as BtrG-like, these include 1vkb (97 aligned C^α atoms, sequence identity: 17.5%, RMSD: 2.2 Å) from *M.musculus* 19, 1xhs (81 aligned C^α atoms, sequence identity: 27.2%, RMSD: 2.5 Å) from *E.coli* 20, and 1v30 (81 aligned C^α atoms, sequence identity: 24.7%, RMSD: 1.8 Å) from *Pyrococcus Horikoshii* 21. According to SCOP, each of these proteins is related to BtrG, which is part of the butirosin-biosynthetic gene cluster in *B. circulans*. Butirosin is an aminoglycoside antibiotic with a unique (S)-4-amino-2-hydroxybutyrate (AHBA) side chain that protects the antibiotic from several common resistance mechanisms. Structural homology searches also indicated some resemblance of At3g28950 to HSV POL 22, a DNA polymerase from HSV, and a cytochrome c oxidase from *T. thermophilus* 23.

Because the functions of the closest homologs are unknown, it is difficult to infer the biochemical roles that At3g28950 plays; however, the recurrence of its motif across a wide range of biological kingdoms is an indication of its versatility. Whereas family members are used in prokaryotes for the biosynthesis of antibiotics, they have likely been recruited to a different set of functions in metazoans.

The AIG2-like family

Another indication of the versatility of the AIG2-like fold is the molecular biology of At3g28950's paralogs in *Arabidopsis* (Table 2): At5g39720, whose structure we solved earlier, AIG2, At3g28940, At5g39730, At4g31310, and At2g24390. These are members of the AIG2-like family, which is designated in Pfam as PF06094 and is anchored by AIG2, a protein that exhibits RPS2- and avrRpt2-dependent induction early after infection with *P. syringae* 4. The six proteins exhibit 25–60% sequence identity to At5g39720.

To glean clues as to their functions, we used Genevestigator (Zimmermann et al. 2004) to compare the spatial and temporal gene expression patterns of seven AIG2-like family members (Fig. 2C). Whereas At3g28950 is constitutively expressed, AIG2 expression rises and peaks around the time of early inflorescence (in period 5) then drops for the rest of the lifespan. At5g39720 expression is silent except during period 8 at the completion of flowering and seed maturation, while At5g39730 has constitutive expression which doubles in period 8. At3g28940 has the most pronounced temporal variation in gene expression peaking at the time of rosette development (period 4) and flower production (period 6) with a drastic decline during period 5. Finally, At2g24390 is virtually silent while At4g31310 has constitutive expression similar to At3g28950.

Across various tissues, At5g39730 and at At4g31310 are the most widely expressed, followed by At3g28950, then AIG2 and At3g28940 (Table 2). At5g39720 is confined largely to seeds. Of the seven, AIG2, having been initially discovered as a plant defense related gene, is not surprisingly the most responsive to chemical and biological stimuli along with At3g28940.

These divergent expression profiles are interesting given that the AIG2-like family probably arose from one ancestral gene that was copied during the several rounds of genome duplication and consolidation 24. The products of daughter genes were then recruited to different physiological roles; sequence differences reflect the adaptation of a common structure to suit individualized functions, along with their respective gene expression control circuits. It is notable that At5g39720 and At5g39730 reside within 10 kbp of each other on chromosome five but display very different expression profiles over the lifetime of the plant and in various tissues. Likewise, AIG2, At3g28940 and At3g28950 all reside within a 60 kbp segment of chromosome three. However, AIG2 and At3g28940 levels are very sensitive to external, particularly biotic, stimuli, in contrast with the constitutively expressed At3g28950 gene.

In addition to new data for analysis of sequence-structure, and structure-function relationships, structural genomics studies expose a new facet of our understanding of genome evolution. Structures of the AIG2 family of proteins illustrate the manner in which nature adapts useful structural motifs to distinct biological functions, even when their specific biochemical roles remain undefined.

References

1. Raiborg C, Slagsvold T, Stenmark H. A new side to ubiquitin. Trends Biochem Sci. 2006; 31(10): 541–544. [PubMed: 16901703]

2. Harmar AJ. Family-B G-protein-coupled receptors. *Genome Biol.* 2001; 2(12):3013.
3. Lytle BL, Peterson FC, Tyler EM, Newman CL, Vinarov DA, Markley JL, Volkman BF. Solution structure of *Arabidopsis thaliana* protein At5g39720.1, a member of the AIG2-like protein family. *Acta Crystallograph Sect F Struct Biol Cryst Commun.* 2006; 62(Pt 6):490–493.
4. Reuber TL, Ausubel FM. Isolation of *Arabidopsis* genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes. *Plant Cell.* 1996; 8(2):241–249. [PubMed: 8742710]
5. Zolnai Z, Lee PT, Li J, Chapman MR, Newman CS, Phillips GN Jr. Rayment I, Ulrich EL, Volkman BF, Markley JL. Project management system for structural and functional proteomics: Sesame. *J Struct Funct Genomics.* 2003; 4(1):11–23. [PubMed: 12943363]
6. Tyler RC, Sreenath HK, Singh S, Aceti DJ, Bingman CA, Markley JL, Fox BG. 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.* 2005; 40(2):268–278. [PubMed: 15766868]
7. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR.* 1995; 6(3):277–293. [PubMed: 8520220]
8. Bartels C, Billeter M, Güntert P, Wüthrich K. Automated Sequence-specific NMR Assignments of Homologous Proteins using the program GARANT. *J Biomol NMR.* 1996; 7:207–213.
9. Bartels C, Xia T-H, Billeter M, Güntert P, Wüthrich K. The Program XEASY for Computer-Supported NMR Spectral Analysis of Biological Macromolecules. *J Biomol NMR.* 1995; 5:1–10. [PubMed: 7881269]
10. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR.* 1999; 13(3):289–302. [PubMed: 10212987]
11. Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol.* 2002; 319(1):209–227. [PubMed: 12051947]
12. Eswar, N.; Webb, B.; Marti-Renom, MA.; Madhusudhan, MS.; Eramian, D.; Shen, M.; Pieper, U.; Sali, A. *Current Protocols in Bioinformatics.* John Wiley & Sons, Inc.; 2006. Comparative Protein Structure Modeling Using Modeller.
13. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. Refinement of protein structures in explicit solvent. *Proteins.* 2003; 50(3):496–506. [PubMed: 12557191]
14. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson.* 2003; 160(1):65–73. [PubMed: 12565051]
15. Gibbs AC, Kondejewski LH, Gronwald W, Nip AM, Hodges RS, Sykes BD, Wishart DS. Unusual beta-sheet periodicity in small cyclic peptides. *Nat Struct Biol.* 1998; 5(4):284–288. [PubMed: 9546219]
16. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics.* 2003; 19(suppl. 2):ii246–ii255. [PubMed: 14534198]
17. Bae, E.; Wesenberg, GE.; Phillips, GN., Jr; Bitto, E.; Bingman, CA.; Center for Eukaryotic Structural Genomics (CESG). Protein Data Bank. 2006. (www.rcsb.org/pdb/home) PDB ID: 2I5T, DOI: 10.2210/pdb2i5t/pdb
18. Kageyama S, Iwaki H, Inoue H, Isono T, Yuasa T, Nogawa M, Maekawa T, Ueda M, Kajita Y, Ogawa O, Toguchida J, Yoshiki T. A novel tumor-related protein, C7orf24, identified by proteome differential display of bladder urothelial carcinoma. *Proteomics - Clinical Applications.* 2007; 1(2):192–199. [PubMed: 21136669]
19. Joint Center for Structural Genomics (JCSG). Wilson, IA. Protein Data Bank. 2004. (www.rcsb.org/pdb/home) PDB ID: 1VKB, DOI: 10.2210/pdb1vkb/pdb
20. Aramini, JM.; Huang, YJ.; Swapna, GVT.; Paranjli, RK.; Xiao, R.; Shastry, R.; Acton, TB.; Cort, JR.; Kennedy, MA.; Montelione, GT.; Northeast Structural Genomics Consortium (NESG). Protein Data Bank. 2004. (www.rcsb.org/pdb/home) PDB ID: 1XHS, DOI: 10.2210/pdb1xhs/pdb
21. Tajika, Y.; Sakai, N.; Yao, M.; Watanabe, N.; Tanaka, I. Protein Data Bank. 2003. (www.rcsb.org/pdb/home) PDB ID: 1V30, DOI: 10.2210/pdb1v30/pdb

22. Liu S, Knafels JD, Chang JS, Waszak GA, Baldwin ET, Deibel MR Jr, Thomsen DR, Homa FL, Wells PA, Tory MC, Poorman RA, Gao H, Qiu X, Seddon AP. Crystal structure of the herpes simplex virus 1 DNA polymerase. *J Biol Chem.* 2006; 281(26):18193–18200. [PubMed: 16638752]
23. Soulimane T, Buse G, Bourenkov GP, Bartunik HD, Huber R, Than ME. Structure and mechanism of the aberrant ba(3)-cytochrome c oxidase from thermus thermophilus. *EMBO J.* 2000; 19(8): 1766–1776. [PubMed: 10775261]
24. Ziolkowski PA, Blanc G, Sadowski J. Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome. *Nucleic Acids Res.* 2003; 31(4): 1339–1350. [PubMed: 12582254]
25. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 2003; 31(13):3497–3500. [PubMed: 12824352]
26. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W, GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol.* 2004; 136(1):2621–2632. [PubMed: 15375207]

Figure 1.

(A) ^{15}N - ^1H HSQC spectrum of At3g28950. Representative resonance assignments are indicated by residue number and one-letter amino acid code. (B) Stereoview of the ensemble of 20 At3g28950 conformers. Secondary structure elements are colored cyan (helix) or magenta (β -sheet). (C) Ribbon diagram of At3g28950 with α -helices and β -strands shown in cyan and magenta, respectively. A portion of the flexible C-terminal helices (white) is omitted for clarity. Residue ranges for secondary structure elements include: $\alpha 1$ (19–27), $\alpha 2$ (71–81), $\alpha 3$ (122–147), $\alpha 4$ (152–157), $\beta 1$ (9–13), $\beta 2$ (32–39), $\beta 3$ (42–43), $\beta 4$ (54–55), $\beta 5$ (61–70), $\beta 6$ (86–95) and $\beta 7$ (101–109). (D) Backbone RMSD values for At3g28950 ensemble and (E) ^{15}N - ^1H heteronuclear NOE values are plotted as a function of amino acid sequence.

Figure 2.

(A) At3g28950 (red) and structural homologs At5g39720 (blue) and C7orf24 from *Homo sapiens* (PDB code 2I5T). (B) ClustalW 25 alignment of sequences for At3g28950, At5g39720 and human C7orf24. (C) Gene expression level for members of the AIG2-like family as a function of *Arabidopsis* growth and developmental periods: 0) 0.0–5.9 days, 1) 6.0–13.9 days, 2) 14.0–17.9 days, 3) 18.0–20.9 days, 4) 21.0–24.9 days, 5) 25.0–28.9 days, 6) 29.0–35.9 days, 7) 36.0–44.9 days, 8) 45.0–50.0 days 26.

Table 1

Structural statistics for 20 At3g28950 conformers

Constraint summary		
Dihedral angles from TALOS	ϕ	97
	ψ	92
NOE	long	668
	medium	319
	short	317
	intraresidue	506
	total	1810
Ramachandran from PROCHECK (%)	Most favored	85.9 ± 2.0
	Additionally allowed	11.0 ± 1.5
	Generously allowed	1.7 ± 1.0
	Disallowed	1.4 ± 1.0
Deviations 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)	0.022 ± 0.0013
	Maximum distance violation	0.353
Torsion angle violation	Number > 5°	0 ± 0
	RMSD (°)	0.79 ± 0.29
	Maximum angle violation (°)	7.2
Atomic RMSD to mean structure (9–44,60–80,85–111)	Backbone	0.63
	Heavy atom	1.05
WHATCHECK quality indicators	RMS Z-scores	
	Bond lengths (Å)	0.76 ± 0.002
	Bond angles (°)	0.75 ± 0.003
	1 st generation packing	-1.31 ± 0.18
	Bumps	0.00 ± 0.00
Lennard-Jones potential	KJ/Mol	$-3,279 \pm 124$

Table 2

Inducibility and tissue specific expression of AIG2-like family members.

Gene	Inducibility	Tissue location
At3g28950	Constitutively expressed	Floral organs, leaves, stems, and roots
At5g37920	Expressed only in seed, no induction by chemical or biotic stimuli	Seed
AIG2	Induced by biotic and chemical treatments including: <i>B. cinerea</i> , <i>P. infestance</i> , <i>P. syringae</i>	Leaves and stems
At3g28940	Induced by biotic and chemical treatments including: <i>B. cinerea</i> , <i>P. infestance</i> , <i>E. cichoracearum</i>	Leaves, root, stems
At5g39730	Constitutive expression	All organs
At4g31310	Generally repressed by biological stimuli	All organs, not seeds
At2g24390	Generally repressed by biological and chemical stimuli	Mainly in leaves