X-ray crystal structures of the conserved hypothetical proteins from Arabidopsis thaliana gene loci At5g11950 and AT2g37210

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · DECEMBER 2006

Impact Factor: 2.63 · DOI: 10.1002/prot.21166 · Source: PubMed

PROTEINS: Structure, Function, and Bioinformatics 65:1051–1054 (2006)

STRUCTURE NOTE

X-ray Crystal Structures of the Conserved Hypothetical Proteins From *Arabidopsis thaliana* Gene Loci At5g11950 and At2g37210

Won Bae Jeon, Simon T. M. Allard, Craig A. Bingman, Eduard Bitto, Byung Woo Han, Gary E. Wesenberg, and George N. Phillips Jr.*

Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin, USA

Introduction. The gene loci At5g11950 and At2g37210 from *Arabidopsis thaliana* encode highly conserved hypothetical proteins with unknown function. The gene products were annotated as putative lysine decarboxylase (LDC)-like proteins by genome analysis. The crystal structure of At5g11950 was determined by the single-wavelength anomalous dispersion method with an R factor of 15.9% (R_{free} = 21.3%) at 2.15 Å resolution, and the structure of At2g37210 was solved using molecular replacement with an R factor of 18.1% (R_{free} = 23.4%) at 1.95 Å resolution. The crystal structure of At5g11950 includes two monomers in the asymmetric unit, and the monomeric structure shows an α/β protein fold comprising eight α-helices and seven β-strands. The structure of At2g37210 is almost identical to that of At5g11950. The fully and highly conserved Arg98 and PGGxGTxxE motif, respectively, were identified by sequence alignment with members of the LDC-like proteins and were mapped onto the structure of At5g11950. A possible active site was suggested based upon the analysis of the location of invariant residues and the consensus motif on the structures of At5g11950 and At2g37210. The Center for Eukaryotic Structural Genomics (CESG) focuses on technology and methodology development for high-throughput X-ray or NMR structure determination of proteins from eukaryotic organisms.[1] The goals of this project also include the identification of new or unique protein folds and characterization of proteins of unknown structure or function. Through a process of selecting targets that have no close amino acid sequence relationship to those in Protein Data Bank (PDB),[2] CESG selected two open reading frames, At5g11950 and At2g37210, from *Arabidopsis thaliana* for structural characterization. These two genes encode highly conserved hypothetical proteins with molecular weights of 23.8 and 23.6 kDa, respectively. The biological functions of the At5g11950 and At2g37210 genes in *A. thaliana* are not yet established. Based on sequence similarities, the protein products of At5g11950 and At2g37210 are annotated as lysine decarboxylase (LDC)-like proteins; however, no indication of the basis for this annotation can be found. In *A. thaliana*, at least 11 hypothetical proteins are annotated as LDC-like proteins by genome analysis.[3] No biochemical evidence supporting this annotation is available. Here, we report the X-ray crystal structures of the proteins from *A. thaliana* gene loci At5g11950 and At2g37210 and describe the structural context of the characteristic motif of this protein family.

Methods. The genes were cloned[4] and proteins were expressed[5] and purified[6] by standard CESG protocols. Crystals of Se–Met-labeled At5g11950 were grown by the hanging drop method, from a 10 mg/ml protein solution in Buffer A (5 m*M* BisTris, 50 m*M* NaCl, 3.1 m*M* NaN_3, 0.3 m*M* TCEP, pH 6.0) mixed with an equal volume of well solution containing 13% (w/v) MePEG 2000, 280 m*M* KNO_3, 100 m*M* MOPS (pH 7.0 at 293 K). Crystals were cryoprotected by placing them serially in well solutions supplemented with increasing concentrations of ethylene glycol, up to a final concentration of 25% (v/v) ethylene glycol. Single-wavelength diffraction data were collected from Se–Met-labeled At5g11950 using an APS 1 detector on beamline 19-BM SBC-CAT at the Advanced Photon Source, Argonne National Laboratory. The data were integrated and scaled using the HKL2000 suite.[7] Localization of the Se positions, phasing, and phase improvement were performed with SOLVE[8] and RESOLVE[9] programs. The initial model was built using the automatic tracing procedure as implemented in ARP/wARP[10] and refined to 2.15 Å using Refmac5.[11]

Grant sponsor: NIH National Institute of General Medical Sciences; Grant numbers: P50 GM064598, U54 GM074901.

*Correspondence to: George N. Phillips Jr., Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706.
E-mail: phillips@biochem.wisc.edu

Received 12 May 2006; Revised 28 June 2006; Accepted 3 July 2006

Published online 17 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21166

TABLE I. Summary of Crystal Parameters, Data Collection, Phasing, Refinement and Model Statistics

	At5g11950	At2g37210
Space group	C2	$P2_12_12_1$
Unit Cell Parameters:		
a, *b*, *c* (Å)	121.6, 80.4, 50.7	53.4, 66.8, 98.6
α, β, γ (°)	90.0, 103.0, 90.0	90.0, 90.0, 90.0
Data collection and phasing Statistics		
Energy (keV)	12.659	12.700
Wavelength (Å)	0.97940	0.97625
Overall resolution range (Å)	35.45–2.15 (2.23–2.15)	46.96–1.95 (2.02–1.95)
Number of reflections (measured/unique)	177865/25771	282807/26424
Completeness (%)	99.9 (100.0)	99.8 (98.6)
R_{merge} (%)[a]	15.6 (38.4)	10.0 (51.8)
Redundancy	6.9 (6.2)	10.7 (4.8)
Mean $I/\sigma(I)$	6.27 (2.40)	15.01 (1.76)
Mean FOM of phasing	0.27	
MR correlation coefficient (MOLREP)		0.283
Refinement and model statistics		
Resolution range	35.45–2.15	46.98–1.95
Number of reflections (total/free)	25679/1300	25025/1336
R_{cryst}[b] (R_{free}[c])	0.159 (0.213)	0.181 (0.234)
RMSD bonds (Å)	0.020	0.021
RMSD angles (°)	1.939	1.712
ESU[d] based on *R*-free (Å)	0.168	0.157
Average *B* factor (Å^2)	31.14	29.22
Number of water molecules	319	222
Ramachandran plot, residues in		
Most favorable region (%)	95.8	96.2
Additional allowed region (%)	4.2	3.8
Generously allowed region (%)	0.0	0.0
Disallowed region (%)	0.0	0.0

[a] $R_{\mathrm{merge}} = \sum_h \sum_i |I_i(h) - \langle I(h)\rangle| / \sum_h \sum_i I_i(h)$, where $I_i(h)$ is the intensity of an individual measurement of the reflection and $\langle I(h)\rangle$ is the mean intensity of the reflection. Values in parentheses are for the highest resolution shell.

[b] $R_{\mathrm{cryst}} = \sum_h ||F_{\mathrm{obs}}| - |F_{\mathrm{calc}}|| / \sum_h |F_{\mathrm{obs}}|$, where F_{obs} and F_{calc} are the observed and calculated structure–factor amplitudes, respectively.

[c] R_{free} was calculated as R_{cryst} using 5.0% of the randomly selected unique reflections that were omitted from structure refinement.

[d] ESU, Estimated overall coordinate error.

Crystals of Se–Met-labeled At2g37210 were grown by the hanging drop method, from a 10 mg/ml solution in Buffer A (see above) mixed with an equal volume of well solution containing 22% (w/v) MePEG 2000, 84 m*M* $MgSO_4$, 100 m*M* BisTris (pH 6.5 at 296 K). Crystals were cryoprotected by placing them serially in well solutions supplemented with increasing concentrations of ethylene glycol, up to a final concentration of 20% (v/v) ethylene glycol. Single-wavelength diffraction data were collected using a MAR 225 detector on beamline 22-BM SER-CAT at the Advanced Photon Source, Argonne National Laboratory. The data were processed using the HKL2000 suite.[7] The structure was solved by molecular replacement using MOLREP[12] and the structure of At5g11950 as the phasing model. The structure of At2g37210 was refined to 1.95 Å using Refmac5.[11]

Results and Discussion. Table I summarizes data collection, phasing, refinement, and model statistics. Coordinates for the crystal structures and diffraction data have been deposited in the PDB under the accession codes 1YDH and 2A33 for At5g11950 and At2g37210, respectively. The monomeric structure of At5g11950 shows an α/β protein fold comprising eight α-helices and seven β-strands (β1α1β2α2β3α3β4α4β5α5β6α6α7β7α8) [Fig. 1(A)]. The central feature of this domain is the β-sheet formed by seven parallel β-strands surrounded by the eight α-helices. On one side of the central β-sheet are helices α1, α2, α3, and α8, and on the other side are helices α4, α5, α6, and α7. The tertiary structure of the At2g37210 monomer is almost identical to that of the At5g11950 monomer, with a 0.9 Å root mean square deviation (rmsd) and 71% identity over 167 aligned Cα positions. The only major difference between the structures is an absence of the short α3 helix in the At2g37210 structure. The loop that spans residues 82–89 was highly disordered in the electron density map, and thus, was not built into the final At2g37210 structure.

Two At5g11950 subunits associate to form a tight dimer in the crystalline asymmetric unit [Fig. 1(A)]. The dimer buries 1806 Å^2 surface area of each monomer and includes 12 hydrogen bonds. The interface between the two monomers is mostly hydrophobic (68%) and stabilized by contact of helices α5 and α6. The CASTp server[13] was used to search for pockets or cavities on the surface of At5g11950. It found a cleft with a surface area of 623 Å^2 and a volume of 322 Å^3. The bottom of the cleft is defined by residues from strands β1, β2, and β3 and a loop between β3 and α3, and the wall of the cleft is formed mostly by residues from helices α4 and α5.

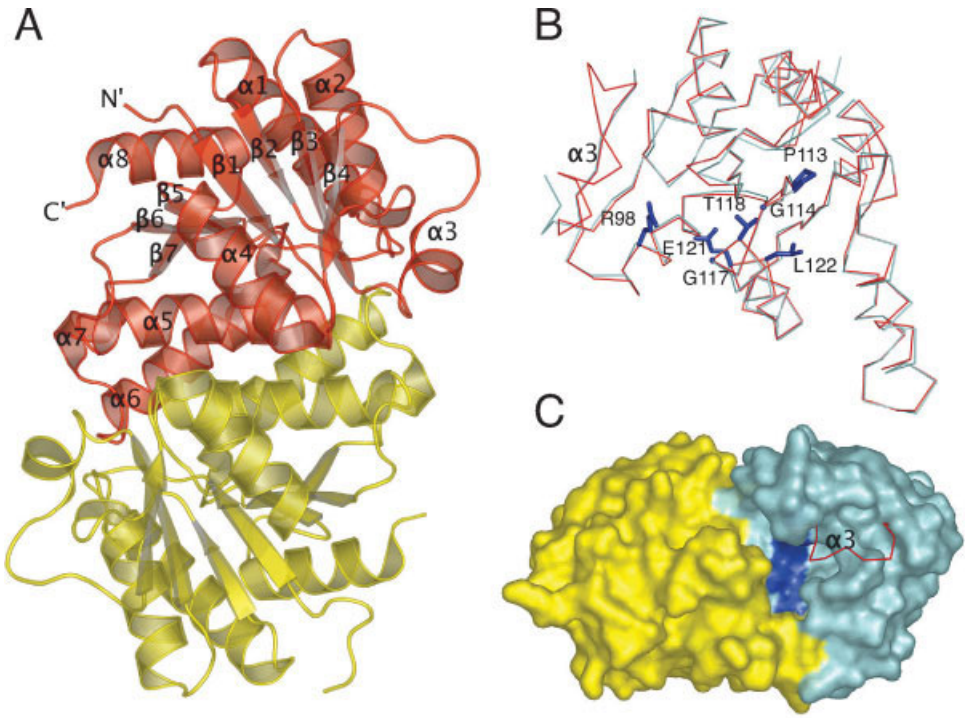

Fig. 1. (**A**) Ribbon diagram of the At5g11950 dimer with each monomer colored separately. (**B**) An overlay of Cα traces of At5g11950 (red) and At2g37210 (cyan). The residues corresponding to helix α3 and neighboring loops adopt a defined structure within At5g11950 (red). The same region is disordered in the At2g37210 structure (cyan). A fully conserved residue, Arg98, and the PGGxGTxxE motif, which is highly conserved among the LDC like proteins,[15] were mapped onto a Cα trace of the At5g11950 structure. (**C**) Surface view of the two monomers (yellow and cyan) in the At2g37210 dimer. The dark blue area represents the conserved motif PGGxGTxxE and residue Arg98, located at the bottom of a cavity. The red line highlights the residues corresponding to helix α3 of At5g11950. These residues are positioned near the entrance to the cavity and are disordered in the crystal structure of At2g37210, suggesting that they may be involved in controlling access of a substrate to the putative active site. The figures were generated using PyMol.[21]

The coordinates of At5g11950 and At2g37210 were analyzed by the DALI server[14] to find structurally similar proteins in the PDB. DALI returned 8 and 19 structural homologs for At5g11590 and At2g37210, respectively, all with a *Z* score over 7.0. The best match to both At5g11590 and At2g37210 was TT1887 (PDB 1WEH), with *Z* scores of 19.1 and 20.4, rmsd's of 2.4 and 2.1 Å and 23 and 25% identity, respectively, for the two proteins whose structures we determined. TT1887 is a hypothetical protein from *Thermus thermophilus* Hb8, which is also annotated as an LDC-like protein.[15] However, its true biological activity is still unknown. The second and third top matches were nucleoside 2-deoxyribosyltransferase (PDB 1F8X, *Z* score 7.5 and 7.8, rmsd 3.9 and 3.4 Å, 12 and 15% identity)[16] and UDP–*N*-acetylglucosamine 2-epimerase (PDB 1F6D, *Z* score 7.2 and 8.1, rmsd 3.0 and 2.9 Å, 12 and 12% identity),[17] respectively. Although they share an apparent structural similarity, both At5g11950 and At2g37210 may have a different function because they do not contain the active site residues of those two enzymes. Interestingly, At5g11950 and At2g37210 are also structurally similar to the negative transcriptional regulator NmrA (PDB 1K6I, Z score 6.7 and 7.6, rmsd 3.5 and 3.3 Å, 8 and 7% identity). NmrA is involved in the signaling pathway of nitrogen metabolite repression in various fungi.[18]

A VAST search[19] found 18 and 14 structural neighbors of At5g11950 and At2g37210, respectively, all with a VAST score over 13.0. Among these structures, four top neighbors with VAST scores greater than 17, rmsd values less than 2.0 Å and a sequence identity of over 23% were annotated as putative LDCs: YvdD (PDB 1T35) from *Bacillus subtilis*, Tm1055 (PDB 1RCU) from *Thermotoga maritima*, TT1465 (PDB 1WEK) and TT1887 (PDB 1WEH) from *T. thermophilus* Hb8.[15] All four structures display an α/β protein fold and contain 6, 7, or 8 α-helices flanking a central β sheet in a similar location to the At5g11950 and At2g37210 structures.

An FFAS03 search[20] confirmed that the four putative LDCs identified by the VAST server share distant sequence homology to At5g11950 and At2g37210, with FFAS03 scores below −49.8 and sequence identity just over 21%. Based upon the crystal structures and sequence homology searches, the protein fold of At5g11950 and At2g37210 was classified as part of the LDC family, pfam03641; however, at present there is no biochemical evidence to support this annotation. Structural analysis

with At5g11950 revealed that the consensus motif PGGxGTxxE[15] is within helix α5 and constitutes part of a cleft [Fig. 1(B)]. In addition, conserved residues Arg98, Thr118, and Glu121 are positioned at the bottom of the cleft, which is created by the β-sheet and helices α4 and α5 in each monomer [Fig. 1(C)]. With these findings, we speculate that the invariant residues and consensus motif are functionally important for biological activity, perhaps forming part of a catalytic site.

Acknowledgments. Data were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-BM beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions may be found at www.ser-cat.org/members. html. Use of the Argonne National Laboratory Structural Biology Center beamlines at the Advanced Photon Source, was supported by the U.S. Department of Energy, Office of Energy Research, under Contract No. W-31-109-ENG-38. Special thanks goes to all members of the CESG.

REFERENCES

1. Wrobel RL, Bingman CA, Jeon WB, Song J, Vinarov DA, Frederick RO, Aceti DJ, Sreenath HK, Zolnai Z, Vojtik FC, Bitto E, Fox BG, Phillips GN, Markley JL. Structural proteomics. In: Finnie C, editor. Plant proteomics. Oxford: Blackwell, in press.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
3. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res 2003;31:224–228.
4. Thao S, Zhao Q, Kimball T, Steffen E, Blommel PG, Riters M, Newman CS, Fox BG, Wrobel RL. Results from high-throughput DNA cloning of Arabidopsis thaliana target genes using site-specific recombination. J Struct Funct Genomics 2004;5:267–276.
5. Sreenath HK, Bingman CA, Buchan BW, Seder KD, Burns BT, Geetha HV, Jeon WB, Vojtik FC, Aceti DJ, Frederick RO, Phillips GN, Fox BG. Protocols for production of selenomethionine-labeled proteins in 2-L polyethylene terephthalate bottles using auto–induction medium. Protein Expr Purif 2005;40:256–267.
6. Jeon WB, Aceti DJ, Bingman CA, Vojtik FC, Olson AC, Ellefson JM, McCombs JE, Sreenath HK, Blommel PG, Seder KD, Burns BT, Geetha HV, Harms AC, Sabat G, Sussman MR, Fox BG, Phillips GN. High–throughput purification and quality assurance of Arabidopsis thaliana proteins for eukaryotic structural genomics. J Struct Funct Genomics 2005;6:143–147.
7. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
8. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr 1999;55:849–861.
9. Terwilliger TC. Maximum likelihood density modification. Acta Crystallogr D Biol Crystallogr 2000;56:965–972.
10. Perrakis PJ, Harkiolaki M, Wilson KS, Lamzin VS. ARP/wARP and molecular replacement. Acta Crystallogr D Biol Crystallogr 2001;57:1445–1450.
11. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by maximum likelihood method. Acta Crystallogr D Biol Crystallogr 1997;53:240–255.
12. Vagin A, Teplyakov A. MOLREP: an automated program for molecular replacement. J Appl Crystallogr 1997;30:1022–1025.
13. Binkowski TA, Naghibzadeh S, Liang J. CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 2003;31:3352–3355.
14. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
15. Kukimoto-Niino M, Murayama K, Kato-Murayama M, Idaka M, Bessho Y, Tatsuguchi A, Ushikoshi-Nakayama R, Terada T, Kuramitsu S, Shirouzu M, Yokoyama S. Crystal structures of possible lysine decarboxylases from *Thermus thermophilus* HB8. Protein Sci 2004;13:3038–3042.
16. Armstrong SR, Cook WJ, Short SA, Ealick SE. Crystal structures of nucleoside deoxyribosyltransferase in native and ligand-bound forms reveal architecture of the active site. Structure 1996;4:97–107.
17. Campbell RE, Mosimann SC, Tanner ME, Strynadka NC. The structure of UDP-*N*-acetylglucosamine 2-epimerase reveals homology to phosphoglycosyl transferases. Biochemistry 2000;39:14993–15001.
18. Stammers DK, Ren J, Leslie K, Nichols CE, Lamb HK, Cocklin S, Dodds A, Hawkins AR. The structure of the negative transcriptional regulator NmrA reveals a structural superfamily which includes the short-chain dehydrogenase/reductases. EMBO J 2001;20:6619–6626.
19. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
20. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 2005;33:W284–W288.
21. DeLano WL. The PyMol molecular graphics system. San Carlos, CA: DeLano Scientific; 2002. Available at http://www.pymol.org.