The hypothetical protein Atu4866 from *Agrobacterium tumefaciens* adopts a streptavidin-like fold

---

XUANJUN AI,^{1,2} ANTHONY SEMESI,³ ADELINDA YEE,³ CHERYL H. ARROWSMITH,³ WING-YIU CHOY,¹ AND SHAWN S.C. LI^{1,2}

¹Department of Biochemistry, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario N6A 5C1, Canada

²Siebens-Drake Medical Research Institute, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario N6A 5C1, Canada

³Ontario Centre for Structural Proteomics, University Health Network, Toronto, Ontario M5G 2C4, Canada

(RECEIVED August 9, 2007; FINAL REVISION September 29, 2007; ACCEPTED October 1, 2007)

Abstract

Atu4866 is a 79-residue conserved hypothetical protein of unknown function from *Agrobacterium tumefaciens*. Protein sequence alignments show that it shares $\geq 60\%$ sequence identity with 20 other hypothetical proteins of bacterial origin. However, the structures and functions of these proteins remain unknown so far. To gain insight into the function of this family of proteins, we have determined the structure of Atu4866 as a target of a structural genomics project using solution NMR spectroscopy. Our results reveal that Atu4866 adopts a streptavidin-like fold featuring a β -barrel/sandwich formed by eight antiparallel β -strands. Further structural analysis identified a continuous patch of conserved residues on the surface of Atu4866 that may constitute a potential ligand-binding site.

Keywords: β -barrel; NMR; protein structure; structural genomics; streptavidin; *Agrobacterium tumefaciens*

Supplemental material: see www.proteinscience.org

Agrobacterium tumefaciens is a soil bacterium belonging to the Rhizobiaceae family. It infects plants through wound sites by transferring a defined segment of DNA to the host that is then integrated into the plant's genome (Smith and Townsend 1907). Such infection causes changes in the growth hormone levels of the host, leading to the formation of tumors. The genome of the natural genetic engineer *A. tumefaciens* C58 was sequenced by Nester and his coworkers in 2001 (Wood et al. 2001). This 5.67-Mb genome consists of a circular and a linear chromosome, a plasmid named pAtC58, and a tumor-

inducing plasmid named pTiC58. Even though >5400 proteins are predicted to be encoded by the *A. tumefaciens* genome, only $\sim 64\%$ of them have had the functions putatively assigned. Structural and functional exploration of the remaining proteins is essential for a complete understanding of the mechanism by which this plant pathogen carries out its action.

The *Atu4866* gene (also known as *AGR_L_42p*) encodes for a 79 amino-acid protein of unknown function. Prediction by gene neighborhood analysis suggests that *Atu4866* may be related to *AGR_L_39p*, a gene encoding for a ketosteroid isomerase-related protein, and *mocA*, the product of which is predicted to be an oxidoreductase similar to aryl-alcohol dehydrogenase or a nicotinamide adenine dinucleotide phosphate (NADP)-dependent aldo/keto reductase (STRING server) (von Mering et al. 2003). Even though sequence alignments indicate that 20 other hypothetical proteins of bacterial origin share $\geq 60\%$ sequence identities with *Atu4866*, none of them has had

---

Reprint requests to: Shawn Li, University of Western Ontario, Siebens-Drake Medical Research Institute, 1400 Western Road, London, ON N6G 2V4, Canada; e-mail: sli@uwo.ca; fax: (519) 661-3175; or Wing-Yiu Choy, University of Western Ontario, Siebens-Drake Medical Research Institute, 1400 Western Road, London, ON N6G 2V4, Canada; e-mail: jchoy4@uwo.ca; fax: (519) 661-3175.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.073167408>.

Table 1. Structural statistics of the 20 lowest-energy NMR structures

Distance restraints	
All	908
Intraresidue	296
Sequential ($ i - j = 1$)	220
Medium range ($1 < i - j \leq 4$)	97
Long range ($ i - j > 4$)	295
Dihedral angle restraints	
All	86
ϕ	43
ψ	43
Deviation from idealized geometry	
Bonds (Å)	0.0025 ± 0.00006^a
Angles (°)	0.4615 ± 0.0060^a
Impropers (°)	0.320 ± 0.012^a
PROCHECK-NMR ^b	
Most favored regions (%)	76.0 ± 1.9^b
Additional allowed regions (%)	19.8 ± 2.1^b
Generously allowed regions (%)	1.4 ± 1.4^b
Disallowed regions (%)	2.8 ± 0.8^b
Pairwise atomic RMSD (Å)	
Backbone atoms (residues 3–79)	0.33 ± 0.11
Heavy atoms (residues 3–79)	1.02 ± 0.10
Average number of violations per structure	
NOE violations (>0.2 Å)	0
Dihedral violations ($>2^\circ$)	0
van der Waals violations (<1.6 Å)	2.1

^a Averaged deviations for the 20 lowest-energy structures evaluated by CNS v1.1 (Brunger et al. 1998).

^b Calculated using PROCHECK-NMR (Laskowski et al. 1996).

alignments (Supplemental Fig. S1) show that residues comprising the functional γ -loop of the DAP C domain are not conserved in Atu4866, implying that Atu4866 and DAP may possess different functions in spite of their structural similarity.

Another structural homolog of Atu4866 is a mutant of streptavidin (PDB 1SWU), which has the highest similarity (Z-score 5.9) among the single-domain proteins identified by DALI. Using the secondary structure-matching approach (Krissinel and Henrick 2004), 20 avidin or avidin-related proteins with high structural similarity to Atu4866 were identified. These proteins all have remarkably similar biotin-binding properties and the same tertiary fold as the core of streptavidin (Pazy et al. 2003). Despite the apparent structural similarity between Atu4866 and avidin/streptavidin, several tryptophan residues of streptavidin that are critical for biotin binding (Livnah et al. 1993) are missing in Atu4866 (Supplemental Fig. S2), suggesting that Atu4866 is unlikely to be a bona fide biotin-binding protein.

Identification of potential ligand-binding site

Figure 3A shows the sequence alignments of Atu4866 and 20 other hypothetical proteins with $\geq 60\%$ sequence

identities. The “conservation” index, calculated based on the method proposed by Livingstone and Barton (1993), reflects the residue-specific conservation of physiochemical properties in the sequence alignments. The result shows that 25 amino acid residues are highly conserved (i.e., with conservation values ≥ 9) in this group of proteins. As depicted in Figure 3, B and C, 11 of these highly conserved residues and six conserved residues (with conservation values ranging from 5 to 8) form a large, continuous surface of the Atu4866 structure. This set of residues, enriched in aromatic amino acids, is all located at one end of the β -barrel. Further, a cleft formed mainly by a subset of aromatic and hydrophobic residues (i.e., Ile16, Tyr26, Ala36, Tyr50, His70, and Phe75) is found at the mouth of the β -barrel. We speculate that this continuous surface formed by a set of conserved residues can be a ligand-binding site of Atu4866.

In summary, we have used NMR spectroscopy to elucidate the structure of Atu4866 from *Agrobacterium tumefaciens*. The protein shows high structural similarities to the C domain of DAP and streptavidin, which both adopt a streptavidin-like fold (Murzin et al. 1995). Meanwhile, Atu4866 shows low structural similarity with other superfamilies of streptavidin-like fold. Therefore, we conclude that Atu4866 is likely a new member of this fold, but with distinct function. In addition, based on the structural information extracted and the result of sequence

Figure 2. (A) Stereoview of the 20 lowest-energy structure ensemble of Atu4866. The numeric labels denote residue numbers. (B) Ribbon representation of Atu4866 in the same orientation as in A and after a 90-degree rotation about the vertical axis. The secondary structure elements are labeled sequentially from the N to C terminus. (Magenta) β -strands, (salmon) loops. This figure was generated using PyMOL (DeLano Scientific).

alignments, we have identified a potential ligand-binding site on the surface of Atu4866. Since gene neighborhood analysis indicates that the functions of the two neighbor-

ing genes of *Atu4866*, *AGR_L_39p* and *mocA*, are related to steroid metabolism, we speculate that *Atu4866* may also be involved in similar biological processes. Further

Figure 3. (A) Sequence alignment of *Atu4866* and 20 other hypothetical proteins identified by BLAST search. Multiple sequence alignment was performed using Clustal W (Thompson et al. 1994), and the diagram was generated using Jalview (Clamp et al. 2004). “Conservation” is a numerical index calculated by Jalview based on the method proposed by Livingstone and Barton (1993). (Diamonds) Fully conserved (or identical) residues, (+) highly conserved residues with conservation between 9 (light blue shading), and fully conserved (deep blue shading). (B) A ribbon representation of the lowest-energy *Atu4866* structure. (Sticks) Side chains of conserved residues, (deep green) highly conserved residues, (lime green) conserved residues. (C) Surface representation of the *Atu4866* structure in the same orientation and color scheme as in B. The figure was generated using PyMOL (DeLano Scientific).

investigation in this direction should help establish the function of Atu4866.

Materials and Methods

Protein expression and purification

The *Atu4866* gene was subcloned from genomic DNA into a pET-derived vector with N-terminal hexa-histidine followed by a TEV cleavage site (with the sequence MGSS(H)6SSGREN-LYFQG). The protein was expressed in *Escherichia coli* BL21-Gold (DE3), harboring a plasmid (pMgk) encoding three transfer RNAs for rare *E. coli* codons. Cells were grown in 0.5 L of M9 minimal media containing ^{13}C -glucose and $^{15}\text{NH}_4\text{Cl}$ as the sole carbon and nitrogen sources and supplemented with ZnCl_2 , thiamine, and biotin. The cells were grown at 37°C to an OD600 of 1.0, and protein expression was induced with 1 mM isopropyl beta-D-thiogalactoside. The temperature was then reduced to 15°C, and the cells were allowed to grow overnight before harvesting. Frozen cell pellets were thawed in 500 mM NaCl, 20 mM Tris, 5 mM imidazole (pH 8), and lysed by sonication. The proteins were then purified by nickel affinity chromatography. After loading the sample onto the nickel column, the Ni^{2+} beads (Qiagen) were washed three times with five column volumes of 500 mM NaCl, 20 mM Tris, 30 mM imidazole (pH 8), and the proteins were eluted with five column volumes of 500 mM imidazole in the same buffer. The purified proteins were concentrated and exchanged into the final NMR buffer containing 10 mM sodium acetate, 300 mM NaCl, 10 mM Zn^{2+} , 10 mM DTT, 0.01% NaN_3 , 1× inhibitor mixture (Roche Molecular Biochemicals), and 1 mM benzamidinium at pH 5.0. The 21-residue His-tag was retained in the NMR sample. The monomeric form of Atu4866 in solution was confirmed by size-exclusion chromatography.

NMR spectroscopy and structure calculations

All NMR data were recorded at 25°C on a Varian UnityINOVA 600 MHz spectrometer equipped with a triple resonance probe with xyz-gradient. HNCACB, CBCA(CO)NH, C(CO)NH, HNCO, HCCH-TOCSY, HC(CO)NH, and HNHA experiments were recorded for backbone and side chain assignments using the standard pulse sequences provided by Varian's Biopack. (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE (Yamazaki et al. 1993) experiments were recorded for the aromatic side chain assignments. Aliphatic and aromatic ^{13}C NOESY-HSQC and ^{15}N NOESY-HSQC, with mixing times of 100 ms, 100 ms, and 120 ms, respectively, were recorded for the NOE assignments. All data sets were processed using NMRPipe (Delaglio et al. 1995), and the spectra were analyzed using the program NMRView (Johnson and Blevins 1994).

Structure calculations were performed using CNS v1.1 (Brunger et al. 1998). Dihedral angle restraints were derived from the ^{15}N , $^1\text{H}_\alpha$, $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$, and $^{13}\text{C}'$ chemical shifts using the programs TALOS (Cornilescu et al. 1999). The geometries of the 20 lowest-energy structures (selected from 200 calculated structures) were analyzed using the program PROCHECK-NMR (Laskowski et al. 1996).

Atomic coordinates of Atu4866 and the experimental restraints used for the structure calculation have been deposited in the Protein Data Bank (PDB 2JMB). The chemical shift assignments have been deposited in BioMagResBank (<http://www.bmrb.wisc.edu>) with the accession code 15012.

Electronic supplemental materials

Supplemental materials include the structure-based sequence alignments of Atu4866 with D-aminopeptidase and avidin/streptavidin.

Acknowledgments

This work was supported by grants from the Ontario Research Development Challenge Fund (to S.S.C.L. and C.H.A.), Canadian Institutes of Health Research (CIHR) (to S.S.C.L.), and a CIHR resource Grant.

References

- Bompard-Gilles, G., Remaut, H., Villeret, V., Prange, T., Fanuel, L., Delmarcelle, M., Joris, B., Frere, J.M., and Van Beeumen, J. 2000. Crystal structure of a D-aminopeptidase from *Ochrobactrum anthropi*, a new member of the "penicillin-recognizing enzyme" family. *Structure* **8**: 971–980.
- Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, L.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr.* **D54**: 905–921.
- Clamp, M., Cuff, J., Searle, S.M., and Barton, G.J. 2004. The Jalview Java alignment editor. *Bioinformatics* **20**: 426–427. doi: 10.1093/bioinformatics/bfg430.
- 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.
- Holm, L. and Sander, C. 1995. Dali: A network tool for protein–structure comparison. *Trends Biochem. Sci.* **20**: 478–480.
- Johnson, B.A. and Blevins, R.A. 1994. NMRView: A computer-program for the visualization of and analysis of NMR data. *J. Biomol. NMR* **4**: 603–614.
- Krissinel, E. and Henrick, K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr.* **D60**: 2256–2268.
- Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R., and Thornton, J.M. 1996. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**: 477–486.
- Livingstone, C.D. and Barton, G.J. 1993. Protein sequence alignments: A strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**: 745–756.
- Livnah, O., Bayer, E.A., Wilchek, M., and Sussman, J.L. 1993. Three-dimensional structures of avidin and the avidin–biotin complex. *Proc. Natl. Acad. Sci.* **90**: 5076–5080.
- Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536–540.
- Pazy, Y., Eisenberg-Domovich, Y., Laitinen, O.H., Kulomaa, M.S., Bayer, E.A., Wilchek, M., and Livnah, O. 2003. Dimer–tetramer transition between solution and crystalline states of streptavidin and avidin mutants. *J. Bacteriol.* **185**: 4050–4056.
- Smith, E.F. and Townsend, C.O. 1907. A plant-tumor of bacterial origin. *Science* **25**: 671–673.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4680.
- von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. 2003. STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**: 258–261.
- Wood, D.W., Setubal, J.C., Kaul, R., Monks, D.E., Kitajima, J.P., Okura, V.K., Zhou, Y., Chen, L., Wood, G.E., Almeida Jr., N.F., et al. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* **294**: 2317–2323.
- Yamazaki, T., Forman-Kay, J.D., and Kay, L.E. 1993. 2-dimensional NMR experiments for correlating $^{13}\text{C}\beta$ and $^1\text{H}\delta/\epsilon$ chemical shifts of aromatic residues in ^{13}C -labeled proteins via scalar couplings. *J. Am. Chem. Soc.* **115**: 11054–11055.