PROTEINS: Structure, Function, and Bioinformatics 57:874–877 (2004)

STRUCTURE NOTE

X-Ray Structure of HI0817 From *Haemophilus influenzae*: Protein of Unknown Function With a Novel Fold

Andrey Galkin,[1] Elif Sarikaya,[1] Christopher Lehmann,[1] Andrew Howard,[2,3] and Osnat Herzberg[1]
[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*
[2]*Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois*
[3]*Biological, Chemical, and Physical Science Department, Illinois Institute of Technology, Chicago, Illinois*

Introduction. The enormous amount of sequence data is being derived from 160 completed genomes (http://www.cbs.dtu.dk/services/GenomeAtlas/index.php) and from over 170 genome sequencing projects in progress (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/InProgress.html). For many protein sequences, which are annotated as "hypothetical proteins," neither a function nor a three-dimensional structure is available. The determination of the structures and functions of these proteins offer the basis for important discoveries such as new antibacterial drug targets and a better understanding of cellular processes. Our experience with Structural Genomics of 43 such proteins reveals an enrichment of novel folds compared with the general set of proteins entering the Protein Data Bank (PDB).

As of this writing, the *Haemophilis influenzae* 20.3-kD protein, HI0817, remains a hypothetical protein. It is one of a 24-member protein family that includes 12 human pathogens, one bovine-, one insect-, and four plant pathogens. The remaining six family members, representing nonpathogenic organisms, exhibit lower sequence homology to HI0817 than the members of the pathogenic subgroup. All known HI0817 homologs are found in the γ-subdivision of Proteobacteria.

The amino acid composition of the HI0817 contains an unusually high number of acidic residues (35 aspartic and glutamic acids compared with 6 arginines and lysines) leading to a theoretical pI of 4. Here, we report the crystal structure of HI0817 (ygfB) at 1.95 Å resolution determined by multiwavelength anomalous dispersion (MAD) phasing from a crystal of a selenomethinoine-containing (SeMet) protein. More information about HI0817 may be obtained from our Structural Genomics web site: http://s2f.umbi.umd.edu.

Materials and Methods. *Cloning and mutagenesis.* The gene encoding HI0817 from *Haemophilis influenzae* Rd KW20 was amplified using PfuTurbo DNA polymerase (Stratagene, La Jolla, CA), genomic DNA, and 5′- and 3′-end primers. The PCR product was introduced into pET100/D-TOPO expression vector by the TOPO directional cloning procedure (Invitrogen, La Jolla, CA). The 182-residue protein contains 2 methionine residues, including the N-terminal methionine that is often processed during expression in *Escherichia coli* or disordered in the structure. To facilitate structure determination by the Multiwavelength Anomalous Diffraction method (MAD) exploiting the absorption edge of Se atoms of a SeMet protein, one additional methionine residue was introduced instead of leucine at position 8. The mutated gene was termed HI0817m. Recombinants were isolated from *E. coli* TOP10 strain (Invitrogen).

Production and purification of the SeMet protein. The *E. coli* strain B834(DE3) was transformed with the pET100/HI0817m recombinant plasmid encoding the HI0817m. Minimal medium was supplemented with SeMet and all 19 amino acids other than methionine. Cells were grown at 37°C to an A_{600} of 0.5, when 1 mM IPTG was added. The cells were collected by centrifugation and lysed by passage through a French press. The soluble fraction was applied on a Ni-NTA affinity column (Qiagen, Chatsworth, CA). Protein was eluted with 20 mM Tris-HCl (pH 8.0), 0.5 M NaCl, and 250 mM imidazole. To remove the N-terminal sequence containing a 6xHis tag, human thrombin (Haemtech) was added at a molar ratio of 1:500 and left for 4 h at room temperature in Tris-HCl (pH 8.0), and 150 mM NaCl. The protein was further purified by ion-exchange chromatography on Sourse 15 Q column (Amersham Biosciences, Arlington Heights, IL). Fractions containing protein were collected and dialyzed against 50 mM Tris-HCl (pH 7.5), and 50 mM NaCl. Finally, the protein was concentrated to 12 mg/mL, flash-cooled in liquid nitrogen, and stored in aliquots at −80°C. Protein integrity and purity were assessed by polyacrylamide gel electrophoresis in the presence of SDS. The oligomeric state was assessed by analytical size exclusion chromatography on an ÄKTA purifier 10 using a Superdex-75 HR 10/30

Grant sponsor: National Institute of Health; Grant number: PO1 GM57890.

*Correspondence to: Osnat Herzberg, Center for Advanced Research in Biotechnology, 9600 Gudelsky Drive, Rockville MD 20850. E-mail: osnat@carb.nist.gov

Received 17 June 2004; Accepted 21 June 2004

Published online 5 August 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20260

TABLE I. X-Ray Data Collection and Refinement Statistics

Space group	$P4_32_12$		
Cell dimension (Å)	a = b = 42.29, c = 198.5		
No. molecule in asymmetric unit	1		
% solvent	42		
Data collection	**Peak**	**Edge**	**Remote**
Wavelength (Å)	0.9795	0.9796	0.9686
Resolution range (Å)	20–1.95	20–1.95	20–1.95
No. observations	298,675	298,304	306,364
No. unique reflections	24,638	24,680	24,751
Completeness (%)[a]	98(92)	98(92)	98(95)
R_{merge}[b]	0.050(0.097)	0.053(0.103)	0.048(0.108)
Refinement statistics			
No. reflections	13,482		
No. residues	170		
No. water molecules	201		
R_{cryst}[c]	0.194		
R_{free}[d]	0.255		
RMS deviation			
Bonds (Å)	0.020		
Angles (°)	1.8		
Average B factor (Å^2)	37.9		
Ramachandran plot (%)			
Most favored	90.3		
Allowed	9.0		
Generously allowed	0.7		
Disallowed	0.0		

[a]The values in parentheses are for the highest resolution shell
[b]$R_{merge} = \Sigma_{hkl}[(\Sigma_j \mid I_j - \langle I \rangle \mid)/\Sigma_j \mid I_j \mid]$, for equivalent reflections (bijvoet pairs separated).
[c]$R_{cryst} = \Sigma_{hkl} \| F_o \mid - \mid F_c \| / \Sigma_{hkl} \mid F_o \mid$, where F_o and F_c are the observed and calculated structure factors, respectively.
[d]R_{free} is computed for 951 reflections that were randomly selected and omitted from the refinement.

column (Amersham Pharmacia Biotech) in 50 mM Tris-HCl (pH 7.0) and 150 mM NaCl.

Crystallization, data collection, and structure determination. Crystals were obtained at 4°C by the vapor diffusion method in hanging drops. The protein solution was mixed with an equal volume of mother liquor containing 1.4 M ammonium sulfate, 0.05 M potassium phosphate (pH 4.8), and 4% isopropanol, and equilibrated against the mother liquor reservoir. Crystal parameters are provided in Table I. Crystals were flash-cooled at 100 K in mother liquor to which glycerol was added to a final concentration of 25%. MAD data, exploiting the absorption edge of Se, were collected on the IMCA-CAT 17-ID beamline at the Advanced Photon Source (Argonne National Laboratory, Argonne, IL). The beamline was equipped with an ADSC Quantum 210 Charge-Coupled Device (CCD) detector. Data processing was performed using the HKL program suite[1] (Table I). The computer program SOLVE[2] was used to locate the selenium sites and to calculate the phases, and RESOLVE[2] was used for density modification. One hundred and thirty of the 182 amino acid residues were built automatically by RESOLVE, and the remaining residues were built manually with the program O.[3] Structure refinement was carried out using the CNS program.[4] Simulated annealing molecular dynamics cycles were followed by alternating cycles of positional and individual temperature factor refinement. Water molecules were added to the model using difference Fourier maps, and peak density $\geq 3\sigma$ as the acceptance criteria. Structure analysis was carried out using a set of computer programs: Procheck[5] for analysis of geometry, Pymol[6] for depiction of structure, and GRASP[7] for molecular surface area calculations.

Results and Discussion. *The structure.* The refined structure at 1.95 Å resolution includes 169 of the 182 amino acid residues. Missing from the model are residues that were not associated with interpretable electron density: three N-terminal residues remaining after thrombin cleavage (with the sequence Gly-Ser-His), the authentic N-terminal methionine residue, and 11 residues at the C-terminus (residues 172-182). The refinement statistics are provided in Table I.

The monomer (approximate dimensions $25 \times 35 \times 47$ Å) adopts a novel fold as determined by the programs DALI[8] and SSM.[9] The new fold consists of seven α-helices arranged into two domains. The N-terminal domain (helices I–IV) [Fig. 1(A)] contains a 4-helix bundle, reminiscent of the 4-helix bundle structure of myosine phosphate inhibitor CPI-17 (PDB code 1k5o), with root mean square (RMS) deviation between 85 aligned Cα atoms of 3.4Å. The C-terminal domain (helices V–VII) contains a 3-helix up-down bundle, reminiscent of the 3-helix C-terminal subdomain of DnaK substrate binding domain (PDB code 1dkz), with RMS deviation between 61 aligned Cα atoms of 3.2 Å. The sequence identity between the HI0817 fragments and CPI-17 or DnaK regions is very low. The two HI0817 domains are connected through an inter-domain crossover (residues 78–91) traversing the molecule in an extended conformation.

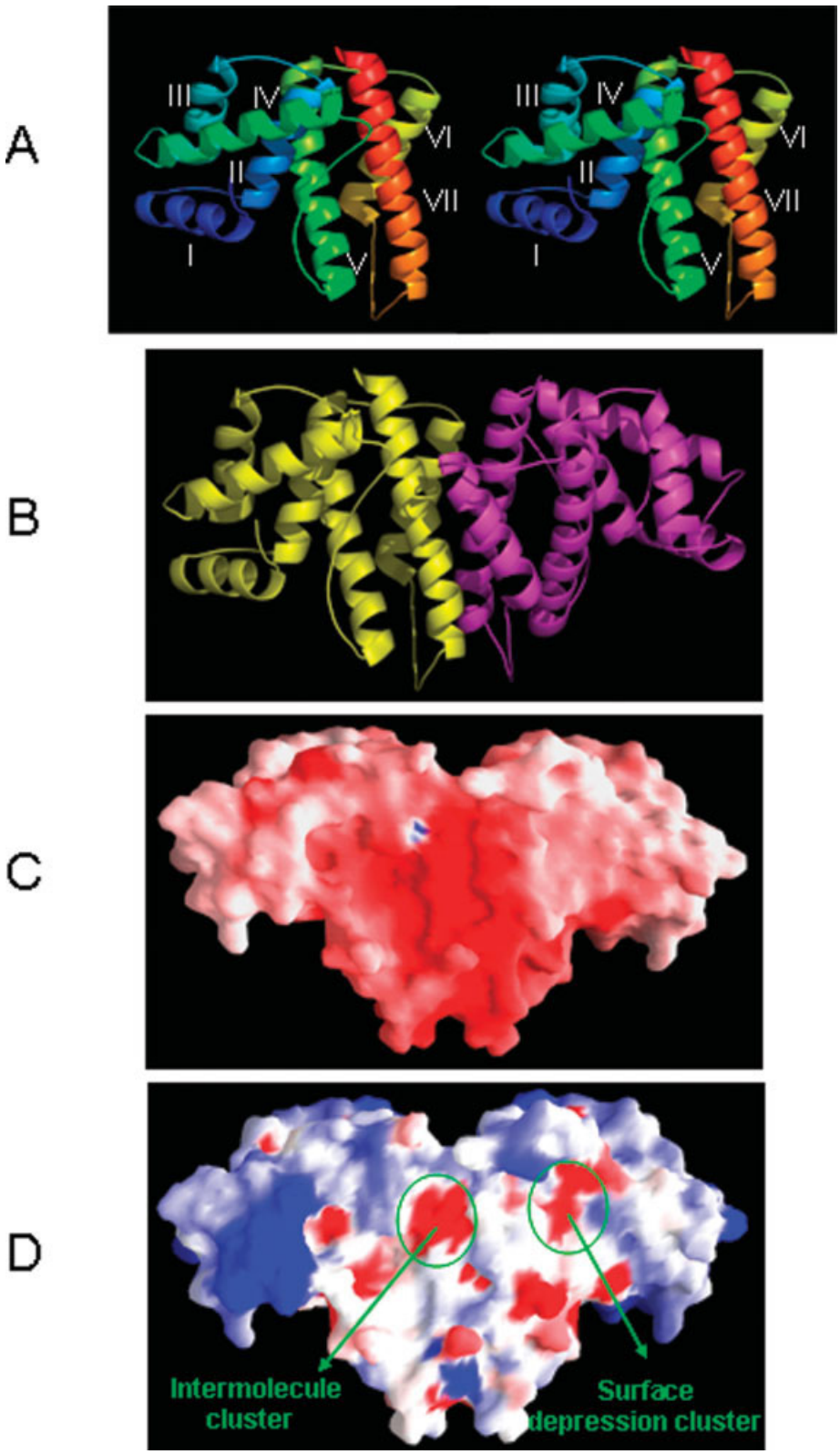

Fig. 1. Fold and structural features of HI0817. **A:** Stereoscopic view of the new fold. Helices I–IV form one domain and helices V–VII form a second domain. **B:** Ribbon diagram representation of the protein dimer. **C:** Molecular surface of HI0817 colored according to electrostatic potential. Negative and positive charged areas are colored in red and blue, respectively. **D:** Molecular surface of HI0817 colored according to an amino acid conservation scheme. High conservation regions and low conservation regions are colored in red and blue, respectively. Two clusters of conserved residues are highlighted.

The dimer and residue conservation. HI0817 associates into dimers both in solution and in the crystal [Fig. 1(B)]. Packing in the crystal is mediated by a crystallographic twofold axis. The surface of the dimer is highly enriched with negatively charged residues [Fig. 1(C)]. The dimer interface is formed through contacts of the C-terminal domains of the two monomers, particularly by close association of the helices VI and VII [Fig. 1(A)]. The buried intermolecular surface area per monomer is 985 Å^2, approximately 12% of the monomer surface area, and includes many protein-protein interactions seen in other oligomeric proteins.[10] The buried hydrophobic surface area is 289 Å^2, approximately 14% of the total hydrophobic surface area of a monomer. Strikingly, the interface includes extensive carboxyl-carboxylate interactions (488 Å^2 surface area, forming approximately 49% of total contact surface) with no compensating basic residues close by.

The residues at the dimer interface tend to be conserved in the HI0817 sequence family. They include three invariant residues, Glu123, Asp130, and Glu157, and two conservatively replaced residues, Tyr158 and Glu154 [Fig. 1(D)]. Such acidic interface is stable at the low pH of the crystals (4.8), but should be repulsive at physiological pH. We note that the analytical size exclusion experiments show that the dimer is the predominant form in solution at pH 7.0, and suggest that under physiological conditions the subunit interactions may be modified to include cations. Moreover, the specificity and conservation of the intersubunit interactions in the HI0817 sequence family are suggestive of a possible role for dimer-monomer transition upon binding to an unknown positively charged target molecule.

A cluster of four conserved residues, His25, Gly26, Trp102, and Phe106, is located in a depression formed at the domain interface of each monomer [Fig 1(D)]. The arrangement is suggestive of an anchoring site for tight helix-helix packing (helices II and V) [Fig. 1(A)] with very short inter-axial distance (the Cα-Cα distance between Gly26 and Phe106 is 4.3 Å). Sequence conservation of this site suggests that it is an essential structural and functional characteristic of the HI0817 sequence family.

It is worthwhile to mention one more conserved region of HI0817 protein family: the last C-terminal 4 residues, including the invariant His182. Although the fragment is not seen in the electron density map, the sequence conservation suggests an important functional role for these residues, and the flexibility may be a required feature correlated with the postulated interface reorganization.

Genome context. HI0817 is encoded by a gene located in close proximity to HI0816. The distance between the TAA stop codon of the HI0817 gene and the ATG start codon of HI0816 gene is only 12 nucleotides, indicating a potential operon organization and coordinated expression. Similar cluster organization (orientation and proximity of HI0817 and HI0816 gene homologs) is observed also in six other bacterial genomes. HI0816 exhibits 53% identity to proline aminopeptidase P (PepP). This enzyme was first isolated from *E. coli* and was subsequently characterized in many microbial mammalian and plant organisms. It releases the N-terminal residue of a peptide where the following residue is a proline. PepP is thought to play a role in the maturation of specific proteins and nascent polypeptides.[11] Although the physiological role of YgfB is unknown, the close association of the YgfB gene with the PepP gene in several pathogenic organisms indicates a potential involvement of the protein product in a peptide maturation process unique to pathogenicity.

Acknowledgments. We thank John Moult and Eugene Melamud for the use of and help with their bioinformatics web site (http://s2f.carb.umbi.umd.edu). We also thank Alexey Murzin and Alexey Teplyakov for stimulating discussions. We thank the staff of IMCA-CAT at the Advanced Photon Source for their help during data collection. The IMCA-CAT facility is supported by the companies of the Industrial Macromolecular Crystallographic Association, through a contract with IIT. Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Science, under contract W-31-109-Eng-38. The Keck foundation provided generous support for the purchase of X-ray equipment at CARB. The PDB coordinates entry code of HI0817 is 1IZM.

REFERENCES

1. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
2. Terwilliger TC. Automated structure solution, density modification and model building. Acta Crystallogr D Biol Crystallogr 2002;58:1937–1940.
3. Kleywegt GJ, Jones TA. Software for handling macromolecular envelopes. Acta Crystallogr D Biol Crystallogr 1999;55:941–944.
4. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
5. Laskowski RA, MacArthur MW. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
6. DeLano WL. The PyMOL user's manual. San Carlos, CA: DeLano Scientific; 2002.
7. Nicolls A, Sharp K, Honing B. Protein folding and association: insights from the interfacial and thermodynamics properties of hydrocarbons. Proteins 1991;11:281–296.
8. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
9. Krissinel E, Henrick K. Protein structure comparison in 3D based on secondary structure matching (SSM) followed by C alignment, scored by a new structural similarity function. In: Kungl AJ, Kungl PJ, editors. Proceedings of the 5th International Conference on Molecular Structural Biology, Vienna; 2003. p 88.
10. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
11. Matos J, Nardi M, Kumura H, Monnet V. Genetic characterization of pepP, which encodes an aminopeptidase P whose deficiency does not affect *Lactococcus lactis* growth in milk, unlike deficiency of the X-prolyl dipeptidyl aminopeptidase. Appl Environ Microbiol 1998;64:4591–4595.