PROTEINS: Structure, Function, and Genetics 49:423–426 (2002)

STRUCTURE NOTE

Structure of HI1333 (YhbY), a Putative RNA-Binding Protein From *Haemophilus influenzae*

Mark A. Willis,[1] Wojciech Krajewski,[1] Vani Rao Chalamasetty,[2] Prasad Reddy,[2] Andrew Howard,[3,4] and Osnat Herzberg[1]*

[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*
[2]*The National Institute of Standards and Technology, Gaithersburg, Maryland*
[3]*Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois*
[4]*Illinois Institute of Technology, Chicago, Illinois*

Introduction. The structures of a number of small α/β RNA-binding proteins with diverse biological functions are known.[1] Their topologies and the locations of the RNA-binding sites vary considerably, consistent with the plasticity of RNA due to base pair mismatches, bulges, and loops. Yet the protein-binding surfaces can be recognized because they are enriched with positively charged residues that either form salt bridges with the negatively charged RNA or contribute favorably to the electrostatic environment. Protein regions that exhibit conformational flexibility are also good candidates for RNA-protein interactions because binding is usually accompanied by some mutual conformational adjustments.[1] We have determined the crystal structure of HI1333 (YhbY) from *Haemophilus influenzae,* a protein annotated as hypothetical in sequence databases. We propose that this protein and its close sequence relatives (25 in the nonredundant sequence database at the time of writing) comprise a new class of RNA-binding proteins.

Materials and Methods. The gene encoding HI1333 was amplified from *Haemophilus influenzae* KW20 genomic DNA and cloned into pRE1[2] for expression in *E. coli* MZ1. Cells were grown in LB media containing ampicillin (50 μg/mL) at 32°C until the A_{650} reached 0.4 and diluted with an equal volume of fresh LB media kept at 60°C, achieving 42°C where protein induction occurs. Cells were broken by passage through a French press and cell debris removed by centrifugation at 100,000 × g for 1 h. The protein was purified by a combination of DE-52 anion-exchange chromatography and cation exchange on a Shodex CM-2025 HPLC column. Fractions containing the protein were pooled, concentrated, and dialyzed against 20 mM $NaPO_4$, pH 7.0, 100 mM NaCl to achieve a final protein concentration of 18 mg/mL. The molecular weight of the protein was confirmed by MALDI-TOF mass spectroscopy, and dynamic light scattering indicated that the protein is monomeric in solution.

Crystals of HI1333 belonging to space group $P2_1$ (with cell dimensions of a = 30.6 Å, b = 51.9 Å, c = 59.1 Å, β = 102.2° and two molecules in the asymmetric unit) appeared in a few days at room temperature in hanging-drop vapor diffusion experiments using 5 μL of protein (18 mg/mL in 20 mM $NaPO_4$, pH 7.0, 100 mM NaCl), 1.5 μL 15% heptane-1,2,3-triol, and 3.5 μL well solution (29–30% PEG 4000, 100 mM Tris pH 8.0, 10 mM $CaCl_2$). Diffraction data at 100 K on cryoprotected crystals (using perfluoropolyether) oil MW = 2800 and 32% PEG 4000, 100 mM Tris, pH 8.0, 15 mM $CaCl_2$, 6.1% heptane-1,2,3-triol, 15% glycerol) were collected on the IMCA-CAT beamlines (17-ID and 17-BM) at the Advanced Photon Source (Argonne National Laboratory, Argonne, IL). In addition to the 1.37 Å native data, two MAD data sets were collected at 1.85 Å. One set for a platinum derivative was obtained by soaking crystals in cryosolution augmented with 2 mM K_2PtCl_4 for 3 days before flash-cooling, and the second set for a bromide derivative was obtained by soaking crystals for 90 s in a 1 M NaBr cryosolution.

All data sets were processed with the HKL Suite[3] and scaled by using SOLVE[4] (MAD data) and XPREP[5] (native data). Heavy atom sites were found by using SOLVE[4] and CNS.[6] The SOLVE phases derived from the combined MAD data sets were modified by the program RESOLVE[7] and used to produce a high-quality electron density map into which the two molecules in the asymmetric unit were built with use of the program O.[8] Initial refinement using all the native data from 25.9 to 1.37Å was performed with CNS, which was followed by refinement with SHELX-97[9] using anisotropic displacement parameter refinement. Native data processing statistics and refinement statistics are shown in Tables I and II, respectively.

Results and Discussion. HI1333 is a 99 amino acid α/β protein consisting of a four-stranded mixed β-sheet sandwiched between two helices on one side and one helix on the other [Fig. 1(a)]. The packing of molecules in the

Grant sponsor: National Institutes of Health; Grant number: P01 GM57890.

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

Received 24 June 2002; Accepted 26 June 2002

TABLE I. Data Processing Statistics

Space group	$P2_1$							
Cell dimensions								
Native	a = 30.6, b = 51.9, c = 59.1, β = 102.2							
Br	a = 30.8, b = 52.2, c = 58.2, β = 104.1							
Pt	a = 30.9, b = 52.4, c = 58.4, β = 104.4							
No. of molecules/asymmetric unit	2							
	Native	Br w1	Br w2	Br w3	Br w4	Pt w1	Pt w2	Pt w3
Wavelength (Å)	0.91840	0.91963	0.92017	0.91990	0.90632	1.07149	1.07196	1.05518
Resolution (Å)	1.37	1.85	1.85	1.85	1.85	1.86	1.86	1.86
No. of observations	246633	57935	58254	57300	57644	105273	105918	108119
Unique reflections[a]	38069	15418	15422	15412	15418	14820	14824	14868
Completeness (%)[b]	99.9 (100)	100 (99.9)	100 (100)	99.9 (99.5)	100 (99.7)	96.1 (62.4)	96.4 (65.3)	97.1 (72.5)
R_{sym} (I)[b,c]	0.031 (0.271)	0.051 (0.363)	0.044 (0.337)	0.058 (0.453)	0.055 (0.481)	0.067 (0.258)	0.067 (0.263)	0.064 (0.293)
⟨I/σ⟩	14.0	10.5	10.8	9.0	8.8	9.8	9.8	9.4
Anomalous and dispersive R-factors (%)[d]								
	Br w1	Br w2	Br w3	Br w4		Pt w1	Pt w2	Pt w3
Br w1	4.0	1.9	1.7	1.9	Pt w1	4.8	3.2	3.6
Br w2		2.6	1.7	2.8	Pt w2		4.3	3.6
Br w3			4.5	2.2	Pt w3			4.4
Br w4				3.7				

[a]Friedel pairs are treated as independent reflections for derivative data.
[b]Values in parentheses are for the highest resolution bin (1.42–1.37Å for native data, 1.92–1.85Å for Br data, and 1.92–1.86Å for Pt data).
[c]$R_{sym} = \Sigma_{hkl}[\Sigma_j|I_j - \langle I\rangle|)/\Sigma_j|I_j|]$.
[d]$R = [\Sigma_{hkl}(||F_{obs}| - |\Delta F_{calc}||)/\Sigma_{hkl}(|\Delta F_{obs}|)]$, where ΔF is the observed (obs) or calculated (calc) dispersive (off-diagonal elements) or Bijvoet difference (diagonal elements) of data used in the phasing routine (from low resolution to 1.86Å for Pt w1 and w3, 2.1 Å for Pt w2, 2.3 Å for Br w1 and w3, 2.5 Å for Br w2 and w4).

TABLE II. Refinement Statistics

Resolution (Å)	25.9–1.37
Wavelength (Å)	0.9184
Unique reflections (F > 0)	38052
Completeness (%)	99.9
No. of protein atoms	1554
No. of glycerol molecules	7
No. of water molecules	264
R_{work}[a]	0.138 (0.132)
R_{free}[a]	0.221 (0.213)
RMSD[b] from ideal geometry	
Bond lengths (Å)	0.011
Bond angle distances (Å)	0.030
Average B factor (Å^2)	
Molecule A	26
Molecule B	32
Glycerol	65
Water	42
Ramachandran plot (%)	
Most favored	92.4
Allowed	6.5
Generously allowed	0.0
Disallowed	1.2

[a]The crystallographic R-factor, $R = (\Sigma_{hkl}||F_{obs}| - k|F_{calc}||/\Sigma_{hkl}|F_{obs}|)$. R_{work} is calculated for the reflections used in refinement. R_{free} is calculated for a randomly selected 8% set of reflections not included in the refinement. The values in parentheses are for reflections with $F > 4\sigma(F)$.
[b]RMSD, root-mean-square deviation.

crystal is consistent with a monomeric protein. The two molecules in the asymmetric unit are quite similar with a main-chain atom root mean square deviation (RMSD) of 0.4 Å and an all-atom RMSD of 0.9 Å. The largest deviations occur in a region with high crystallographic temperature factors that involves two loops and a type 1′ reverse turn: residues 26–28 (between β1 and α2), residues 52–59 (between β2 and α3), and residues 78 and 79 (reverse turn between β3 and β4). These differences are attributable to crystal-packing interactions.

All residues except Glu46 fall into the most favored (92.4%) or additionally allowed (6.5%) regions of a Ramachandran plot.[10] The electron density for Glu46 is well defined, showing that the main-chain adopts a strained conformation ($\varphi = 72°$, $\psi = -57°$).

The distribution of electrostatic charges on the surface of the protein reveals a region rich in positively charged residues that includes the exposed β-sheet above α1 and continues around the edge of β3 to include a portion of α3 [Fig. 1(b)]. It is also this region of the protein that exhibits the best conservation of residues among HI1333 sequence relatives [Fig. 1(c)]. The extensive basic region and the correlation with amino acid conservation suggest that the function of HI1333 involves binding of nucleic acids.

Another surface region that contains conserved residues includes a number of hydrophobic residues and Glu46 (the sterically strained residue is seen as an isolated red spot in Figure 1(b). The residues involved define a small pocket. In the HI1333 sequence family, position 46 is occupied mostly by a glutamic acid and in a few cases by a glycine residue. The pocket would be larger with a glycine at this position. The backbone conformation of residue 46 and the amino acid conservation pattern suggest that this region is

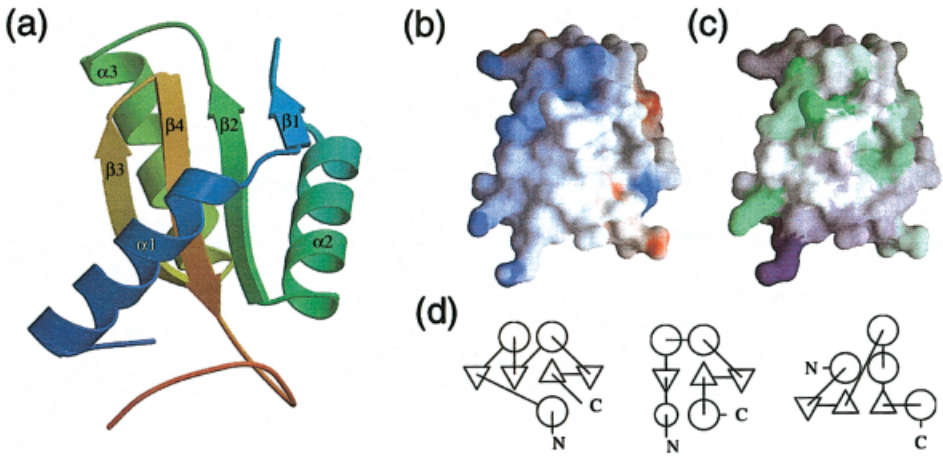

Fig. 1. Structure of HI1333. **a:** Ribbon diagram of HI1333 colored according to the progression of the polypeptide chain from blue (N-terminus) to red (C-terminus). **b:** Charge distribution on the surface of HI1333 in the same orientation as shown in (a). Colors are from blue to red for basic to acidic residues. The surface for the conserved Glu46 can be seen as a red spot to the lower right of center on the charge distribution plot. With the exception of the continuation of the basic region on the bottom left (β3) onto the back face of the protein, the face hidden from view shows fewer basic residues. **c:** Residue conservation distribution, colored from purple (no conservation) to green (high conservation), in the same orientation as in (a). The larger conserved patch correlates with the positively charged region, and the smaller patch defines a small pocket that includes Glu46 and several hydrophobic residues. This may indicate a region of protein-protein interactions. **d:** Topology diagrams. **Left:** HI1333 (1JO0), IF3-C (1TIG), and YhhP (1DCJ) (α1 is missing or not well defined in 1TIG and 1DCJ, respectively). **Center:** Type I KH domain (IKHM). **Right:** Type II KH domain, residues A186–A283 (IEGA). Helices are depicted as circles and β-strands are shown as triangles.

functionally important as well, possibly involved in protein-protein interactions.

HI1333 was selected as a target for a structural genomics project (http://s2f.carb.nist.gov/) because the homologous proteins were annotated hypothetical. Sequence analysis using Psi-Blast[11] on the nonredundant database shows now that two of the close relatives (E values after the first cycle of Psi-Blast of $2e^{-6}$ and $5e^{-6}$) have vaguely based annotations of RNA binding with no reference to experimental work. The third iteration cycle reveals remote homology to CRS1 from maize, a large protein involved in gene splicing.[12] The homology extends over 87 of the protein's 715 amino acids, with 22% sequence identity. Gene splicing implies interactions with RNA.

A number of small α/β RNA-binding protein structures share a similar protein architecture to HI1333 [Fig. 1(d)], although the actual combination of topology and surface charge distribution of HI1333 are unique. The closest structural homologues obtained by using DALI[13] are the C-terminal domain of the translation initiation factor IF3 (IF3-C)[14] and YhhP, a protein implicated in cell division.[15] However, the sequence similarity of the structurally aligned residues is low: 16% and 6% identical residues, respectively. Moreover, the pattern of conservation does not correlate with that of the HI1333 sequence family, and the most basic area of both IF3 and YhhP is on the opposite side of the protein compared with that of HI1333.

As the manuscript was prepared for publication, the coordinates of YhbY from *E. coli* were deposited in the PDB (1LN4, Ostheimer et al., to be published). This structure is very similar to HI1333, consistent with the high-sequence identity. The title of this entry states that the protein is a representative of a new class of RNA-binding proteins.

Another RNA-binding domain with an α/β fold is the KH domain characterized by a conserved VIGXXGXXI sequence. Although type I and type II KH domain proteins share some sequence homology with HI1333, their topologies are different,[16] and neither matches that of HI1333 [Fig. 1(d)]. The positively charged regions of the KH domains are centered on the GXXG loop and include the helices on either side of this loop along with the outer β-strand adjacent to this region. HI1333 and its sequence homologues have the GXXG found in KH domains, but they tend to replace the isoleucine after the second glycine with a charged or polar residue. The structure of HI1333 in this region also differs from that of the KH domain with a short segment in extended conformation replacing the short helical section of the KH domain. Despite these topological and structural differences, the conserved GXXG of HI1333 may, like the GXXG of the KH domain, be involved in binding RNA.

Finally, we note that the completed genome of *Arabidopsis thaliana*[17] contains many middle-size and large proteins that exhibit sequence homology spanning much of the HI1333 domain as seen in CRS1.[12] *Oryza sativa* (rice) contains such domains as well. We propose that these proteins contain RNA-binding domains, and therefore are involved in activities related to RNA. At this time, there are no HI1333 sequence homologues in mammalian genomes.

ACKNOWLEDGMENTS.

We thank John Moult and Eugene Melamud for their bioinformatics work on the structural genomics project.

We also thank the staff at the IMCA-CAT beamlines at the Advanced Photon Source (APS) for their help with 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. Protein Data Bank coordinates entry code: 1JO0.

REFERENCES

1. Draper DE. Themes in RNA-protein recognition. J Mol Biol 1999;293:255–270.
2. Reddy P, Peterkofsky A, McKenney K. Hyperexpression and purification of Escherichia coli adenylate cyclase using a vector designed for expression of lethal gene products. Nucleic Acids Res 1989;17:10473–10488.
3. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
4. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr 1999;55:849–861.
5. Sheldrick GM. XPREP Programm zur Datenanalyse: Universität Göttingen; 1997.
6. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
7. Terwilliger TC. Maximum-likelihood density modification. Acta Crystallogr D Biol Crystallogr 2000;56:965–972.
8. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for binding protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 1991;47:110–119.
9. Sheldrick GM, Schneider TR. SHELXL: high resolution refinement. Methods Enzymol 1997;277:319–343.
10. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. Procheck—a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
12. Till B, Schmitz-Linneweber C, Williams-Carrier R, Barkan A. CRS1 is a novel group II intron splicing factor that was derived from a domain of ancient origin. RNA 2001;7:1227–1238.
13. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
14. Biou V, Shu F, Ramakrishnan V. X-ray crystallography shows that translational initiation factor IF3 consists of two compact alpha/beta domains linked by an alpha-helix. Embo J 1995;14:4056–4064.
15. Katoh E, Hatta T, Shindo H, Ishii Y, Yamada H, Mizuno T, Yamazaki T. High precision NMR structure of YhhP, a novel Escherichia coli protein implicated in cell division. J Mol Biol 2000;304:219–229.
16. Grishin NV. KH domain: one motif, two folds. Nucleic Acids Res 2001;29:638–643.
17. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 2000;408:796–815.