PROTEINS: Structure, Function, and Bioinformatics 58:759–763 (2005)

STRUCTURE NOTE

Crystal Structure of the YgfY from *Escherichia coli*, a Protein that May Be Involved in Transcriptional Regulation

Kap Lim,[1] Victoria Doseeva,[1] Elif Sarikaya Demirkan,[1] Sadhana Pullalarevu,[1] Wojciech Krajewski,[1] Andrey Galkin,[1] Andrew Howard,[2] and Osnat Herzberg[1]*

[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*
[2]*Biological, Chemical, and Physical Science Department, Illinois Institute of Technology, Chicago, Illinois*

Introduction. The *ygfY* gene from *Escherichia coli* encodes an 88 amino acid protein of unknown function, and is a member of a large sequence family present in both, prokaryotes and eukaryotes. A Psi-Blast search of the nonredundant and environmental nonredundant databases revealed 249 family members.[1] Of these sequences, 152 are from environmental samples, mostly from a Sargasso Sea sample filtered to include only microbial cells.[2] The first cycles of Psi-Blast iteration yielded 78 homologous bacterial proteins, including 42 from the Sargasso Sea. Plant and fungal proteins emerged in the second cycle, and insect and mammalian sequences emerged in the third cycle. While the bacterial proteins have approximately the same size as YgfY (except for incomplete sequences from environmental samples), the eukaryotic proteins are at least double in size. None of the sequence homologues has known biochemical function. The cellular role is also unknown, except for a remote relationship (E-score = 10^{-4}) to a protein from *Saccharomyces cerevisiae*, YOL071W/EM15 (GI:6324501), which is required for sporulation and for transcriptional induction of the early meiotic-specific transcription factor IME1.[3] This 162-amino-acid residue yeast protein exhibits 25% identity to YgfY over a stretch of 60 amino acids.

The crystal structure of YgfY was determined at 1.2 Å resolution as a part of our structural genomics project (http://s2f.umbi.umd.edu), revealing a five-helix fold similar to that of the homologous protein NMA1147 from *Neisseria meningitidis* (30% identity) determined recently by NMR methods.[4] Yet, there are inconsistencies in the details of these two structures that reflect the different levels of accuracy of the two methods of structure determination. We propose that the functional region is different from the one proposed based on the NMR structure, and highlight a particularly important structural difference associated with the proposed activity center.

Materials and Methods. YgfY from *E. coli* was amplified using PfuTurbo DNA polymerase (Stratagene), genomic DNA, and 5′- and 3′-end primers. In addition to the native gene sequence, a sequence consistent with a thrombin cleavage site was introduced, and a *NdeI* restriction site was designed to convert the 6xHis-tagged construct into a construct coding for native protein. The PCR product was introduced into the pET100/D-TOPO expression vector by TOPO directional cloning procedure (Invitrogen). For wild-type protein production, the *E. coli* strain BL21 Star (DE3) was transformed with the recombinant plasmid. An expression screen showed that the His-tagged protein was soluble, whereas the native protein was insoluble. Wild-type protein was produced by growing the cells at 37°C in Super Broth medium supplemented by ampicillin (100 μg/mL). Once the cell culture reached A_{600} = 0.6, expression was induced by the addition of 1 mM IPTG, and after 3 h the cells were harvested by centrifugation. To prepare selenomethionine (SeMet) containing protein, the *E. coli* B834 (DE3) strain was transformed with the recombinant vector. The cells were grown at 30°C in minimal medium supplemented with ampicillin (50 μg/mL), SeMet, and 19 amino acids other than methionine. When the cell culture reached A_{600} = 0.5, 1 mM IPTG was added, and after 3 h the cells were harvested.

Cells were suspended in 20 mM Tris HCl (pH 8.0), 0.5 M NaCl, and 5 mM imidazole, and lysed by passage through a French press. The soluble fraction was loaded on Ni-NTA metal affinity column (Qiagen). 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 the 6xHis tag, human thrombin (Haemtech) was added at 1:2000 molar ratio and incubated overnight at 4°C in Tris HCl (pH 8.0), and 500 mM NaCl. Thrombin was removed by passing the protein mixture through benzamidine column (Amersham Biosciences). The cleaved and uncleaved protein and the N-terminal peptide were separated on a second Ni-NTA column. The protein was dialyzed against a buffer of 50 mM NaCl and 20 mM Tris HCl (pH 7.5), and further purified with a size-exclusion chromatography

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 11 August 2004; Accepted 19 August 2004

Published online 8 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20337

TABLE I. Statistics of MAD Data

Space group	$P2_12_12$				
Cell dimensions (Å)	$a = 43.4$, $b = 87.7$, $c = 41.1$				
MAD data statistics	Se-λ_1	Se-λ_2	Se-λ_3	Au-λ_1	Au-λ_2
Wavelength (Å)	0.9796	0.9795	0.9641	1.0374	1.0397
Resolution (Å)	2.0	2.0	2.0	2.0	2.4
Number observed reflections	101,603	100,861	102,943	110,093	66,552
Number unique reflections[a]	20,304	20,185	20,519	19,527	11,486
Completeness (%)[b]	98.5 (98.2)	98.1 (96.3)	99.0 (96.8)	98.1 (98.3)	98.9 (99.2)
R_{merge}[c]	0.067 (0.149)	0.063 (0.100)	0.062 (0.213)	0.078 (0.169)	0.079(0.300)
$\langle I/\sigma \rangle$	14.2	15.9	14.0	11.1	8.2

[a]For phasing, Bijvoet pairs were treated as independent reflections.
[b]The values in parentheses are for the highest resolution shell, 2.07–2.00 Å (2.51–2.40 Å for Au-λ_2).
[c]$R_{merge} = \Sigma_{hkl}[(\Sigma_j | I_j - \langle I \rangle|)/\Sigma_j | I_j |]$, for equivalent reflections (Bijvoet pairs separated).

column. Protein expression, thrombin cleavage, and sample purity were assessed by SDS-PAGE, and the molecular weight was determined by mass spectrometry. The oligomeric state of the protein was measured by analytical size-exclusion chromatography using a G75 Superdex column (Amersham Pharmacia), and by dynamic light scattering using a Dyna Pro instrument (Protein Solutions).

Crystals were obtained by the vapor diffusion method in hanging drops at room temperature. Both SeMet-containing and wild-type protein crystals were formed in 1.1 M sodium malonate solution buffered with 0.1 M Tris HCl (pH 8.0). Crystals belonging to space group $P2_12_12$ with two molecules in the asymmetric unit were obtained for both the SeMet and wild-type proteins (Table I), and were used for structure determination. For the wild-type protein, crystals were also obtained in space group $I222$ with one molecule in the asymmetric unit.

For diffraction data collection, crystals were flash-cooled at 100 K in the crystallization solution. Multiple wavelength diffraction (MAD) data of selenium- and gold-containing crystals (Table I) were collected on the IMCA-CAT 17-ID beamline at the Advanced Photon Source (Argonne National Laboratory, Argonne, Illinois). The beamline was equipped with an ADSC Quantum 210 Charge-Coupled Device (CCD) detector. The home facility was used to characterize crystals and to acquire the 2.0 Å diffraction data of the wild-type protein $I222$ crystal. X-rays were supplied by a Siemens rotating anode which was equipped with MAR345 image plate. Ultra high resolution data at 1.2 Å for a wild-type protein crystal were collected at IMCA-CAT. Data processing was performed using the computer programs HKL[5] and CrystalClear (Rigaku MSC).

The structure of YgfY was determined by combining selenium and gold MAD data from the $P2_12_12$ crystal form. The Se sites were identified using the computer program SHELXD,[6,7] but the phase quality was inadequate for map interpretation. Thus, a wild-type YgfY crystal was soaked with 1 mM $KAu(CN)_2$ for three days, and a two-wavelength diffraction data set was collected (Table I). Two gold sites were identified in a difference Fourier map calculated with phases derived from the Se data. Phase combination was carried out using the MLPHARE program[8] as implemented in CCP4.[9] The phases were improved by solvent flattening and noncrystallographic symmetry averaging using RESOLVE.[10] The resulting electron density map at 2.0 Å resolution was traced using the graphics program O.[11]

The wild-type structure was refined at 1.2 Å resolution, initially with the CNS program,[12] and then with the program REFMAC.[13] Unisotropic temperature factors were introduced towards the end of the refinement, and hydrogen atoms were included in the last cycle (the addition of hydrogen atoms reduced R_{free} by 1%). For the $I222$ crystal form, the structure was determined at 2.0 Å resolution by Molecular Replacement, using BEAST,[14] and then refined using CNS.

Structure analysis was carried out using the computer programs: Procheck[15] for analysis of geometry, and Molscript[16] and Raster3D[17,18] for depiction of structure.

Results and Discussion. Cleavage of the His-tag left three amino acids residues N-terminal to the authentic polypeptide (Gly-Ser-His). Of these, only the main chain of the histidine residue is seen in the electron density map. In the following discussion, we use a numbering scheme based on the authentic sequence with the first residue being Met1. The structure is well ordered, and all residues of the authentic native sequence are included in the model. For the 1.2-Å resolution structure, three side chains in one molecule and four side chains in the second molecule of the asymmetric unit were modeled with alternate conformations.

The structure of YgfY was determined at ultra high resolution (1.2 Å) in the $P2_12_12$ crystal form, and at moderately high resolution (2.0 Å) in the $I222$ crystal form (Table II). The packing of molecules in the two crystal forms is very similar and, since both were obtained under the same crystallization condition, it is unclear whether the change of space group was due to slight variation in the growth conditions or due to perturbations during flash-cooling. The molecules pack into dimers in both crystal forms, which differ only by a rotation of two degrees between the two molecules. The structure is essentially the same within the limit of the resolution. Exchange between $P2_12_12$ and $I222$ space groups was reported previously.[19,20]

In contrast to the dimeric packing in the crystals, dynamic light scattering measurements and analytical size-exclusion chromatography indicated that the protein exists in solution primarily in a monomeric form. Close inspection of the dimer interface revealed that despite the

TABLE II. Refinement Statistics

Space group	$P2_12_12$	$I222$
Cell dimensions (Å)	$a = 43.5$, $b = 87.1$, $c = 40.8$	$a = 40.4$, $b = 42.7$, $c = 84.9$
Wavelength (Å)	0.9795	1.5418
Unique reflections	43,601	5,225
Completeness (%)[a]	88.3 (51.4)	97.1 (95.5)
R_{merge}[a]	0.045 (0.308)	0.059 (0.328)
$\langle I/\sigma \rangle$	25.3 (4.2)	7.8 (2.0)
Refinement resolution (Å)	8.0–1.2	8.0–2.0
No. of reflections	43,438	4,592
No. of protein atoms	1,511	737
No. of water molecules	268	37
R_{cryst}[a,b]	0.171 (0.213)	0.188 (0.185)
R_{free}[a,c]	0.211 (0.265)	0.242 (0.297)
RMS deviation from ideal geometry		
Bond length (Å)	0.018	0.018
Bond angle (°)	1.5	1.6
Temperature factors (Å^2)		
Protein	14	28
H_2O	28	32
Ramachandran plot (%)[d]	92.4, 7.6, 0.0, 0.0	91.1, 8.9, 0.0, 0.0

[a]The values in parentheses are for the highest resolution shell (1.24–1.20 Å for $P2_12_12$ space group and 2.07–2.00 Å for $I222$ space group).
[b]$R_{cryst} = \Sigma_{hkl}||F_o| - |F_c||/\Sigma_{hkl}|F_o|$, where F_o and F_c are the observed and calculated structure factors, respectively.
[c]R_{free} is computed from randomly selected reflections omitted from the refinement (4,385 from $P2_12_12$ space group and 526 from $I222$ space group).
[d]Ramachandran plot categories are most favored, allowed, generously allowed, and disallowed.

suggestive packing, the surfaces do not match well and a layer of solvent molecules fills the gap. Moreover, the amino acid residues at the interface are not conserved in the YgfY sequence family. Thus, we propose that the biological unit is a monomer.

YgfY folds into a compact, five-α-helical bundle [Fig. 1(A)], similar to the fold of the sequence family member, NMA1147 from *Neisseria meningitides*, determined by NMR.[4] The closest SCOP fold classification to YgfY fold is of the cyclin-like domains, characterized by five α-helices with one α-helix (helix 3) surrounded by the other four,[21,22] helices 1–3 forming an antiparallel up-down cluster, and helices 4–5 running orthogonally to one another. In addition to the cyclin family, this super family also includes the transcription factor IIB (TFIIB) core domain and retinoblastoma tumor suppressor domains. These are all protein binding domains. Interestingly, helix 3 of YgfY and NMA1147 is exposed to solvent because the ensuing loop and helix 4 are positioned above the C-terminus of helix 3. As discussed below, we propose that the exposed surface of helix 3 is involved in the active center of the molecule.

The RMS deviation between super-positioned YgfY and NMA1147 α-carbon atom pairs is 2.2 Å. Although the two structures share the same overall fold, there are many differences in details. These include inefficient packing in the NMR structures that results in holes or pockets that are not observed in the X-ray structure, different side chain interactions including those between core residues, and a key conserved residue that forms a strikingly repulsive electrostatic arrangement in NMA1147 (Asp54 in NMA1147, Asp51 in YgfY), whereas in YgfY it forms a perfect ion pair with a conserved arginine (Arg18 in NMA1147, Arg15 in YgfY). We believe some of these discrepancies are due to errors in the NMR structure determination that may have arisen because of the relatively low number of inter-residue constraints (10 per amino acid residue).

A pair of residues, conserved in the sequence family, forms an ion pair, Arg15 and Asp51 [Fig. 1(B)]. The only exception is a sequence from *Arabidopsis thaliana*, where Asp51 is replaced by an asparagine. The same residues are conserved in the sequence of NMA1147 (Arg18 and Asp54). However, in the NMR structure, the Asp54 side-chain is buried and does not form electrostatic interactions to compensate for the charge burial [Fig. 1(C)]. The Arg15-Asp51 salt bridge is located at the center of a surface spanning the contact region of helices 1 and 3, the loop between helices 1 and 2, and the loop between helices 3 and 4. In addition to the residue involved in the salt bridge, several other residues tend to be conserved or conservatively replaced in the sequence family, including Met1 (an initial methionine is always conserved but in YgfY it is well defined and involved is specific interactions), Arg8, Trp11, Gly16, and Met17. At the edge of this region, Glu19, Asp21 are conserved and Phe55 is on the surface and conservatively replaced. In addition to forming a salt bridge, the precise positioning of Arg15 and Asp51 is assured by the interaction of Asp21 carboxyl group with the backbone amide of Arg15, by the interaction of the guanidinium group with the backbone carbonyl groups of residues 47 and 49, and by the carboxyl group of Asp51 interaction with the backbone amide of the invariant Gly16. This exquisite arrangement maintains the structural integrity of the loops that connect helices 1 and 2, and helices 2 and 4. The residue conservation and structure pattern support the notion that this region is likely to be important for function.

A region juxtaposed to the above surface, on the other side of Glu19 and Phe55, comprises a shallow depression that is equally balanced with polar and hydrophobic

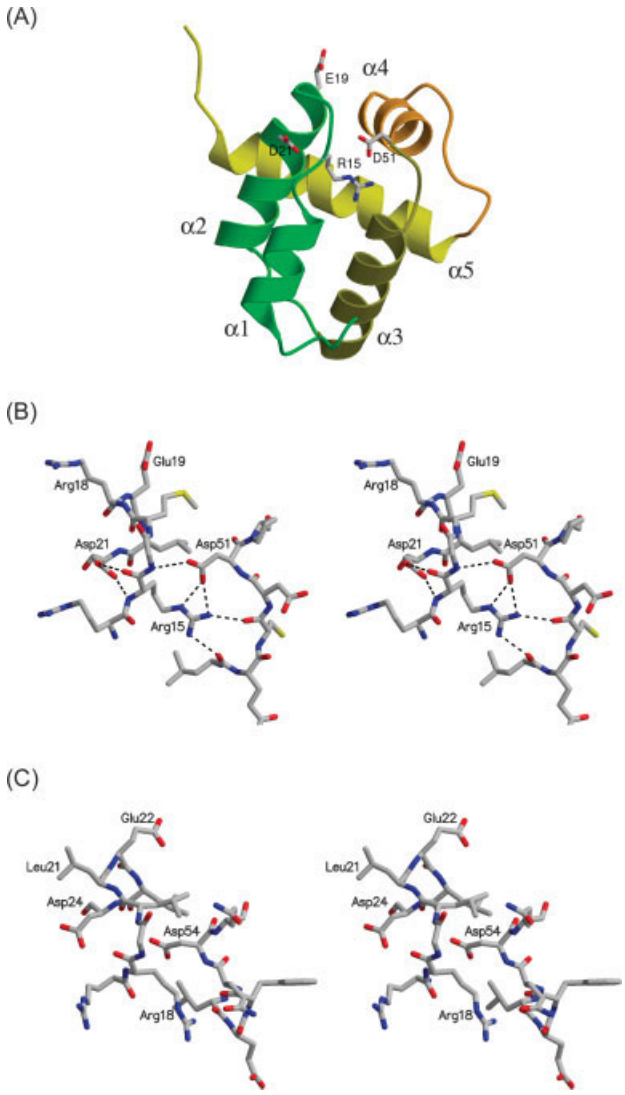

Fig. 1. Crystal structure of YgfY. **A:** Ribbon representation of the overall fold. Key residues conserved in the sequence family are displayed as stick models. **B:** Stereoscopic representation of the YgfY environment in the vicinity of the ion pair, Arg15 and Asp51. Atomic colors are used: carbon, gray; oxygen, red; nitrogen, blue; sulfur, yellow. **C:** Stereoscopic representation of the same environment within the NMR structure of NMA1147. The orientation is the same as in (B). Here, no salt bridge is formed and the carboxyl group of Asp54 (equivalent to Asp51 in YgfY) is buried without forming electrostatic interactions to compensate for charge burial.

residues. The residues associated with the depression are not conserved in the sequence family. In contrast, a deep pocket lined with hydrophobic residues is present in the structure of NMA1147. When the X-ray and NMR structures are compared, it emerges that the large pocket in the NMR structure is a consequence of inefficient packing that leads to the exposure of core residues and to a deep solvent-accessible pocket. Whereas Liu and colleagues implicated this region in function,[4] the X-ray structure does not support this conclusion.

Although the YgfY sequence family contains three conserved residues with functional groups that potentially could be involved in enzymatic reaction, their disposition does not resemble an arrangement typical of enzyme active sites. Moreover, the relatively flat nature of the associated surface suggests that it is more likely to be involved in protein–protein interaction. The structural analogy to the cyclin-like domains involved in protein–protein interactions supports this hypothesis, because often (although not always) a common fold indicates common general function. Yet, the interaction regions of the cyclins and TFIIB do not map to our proposed site of action in the YgfY family. Nevertheless, because the yeast protein homologue was shown to be involved in transcriptional induction of the early meiotic-specific transcription factor IME1,[3] it is tempting to speculate that TFIIB is the true functional analogue of YgfY, and YgfY family members are also associated with complexes that regulate transcription.

ACKNOWLEDGMENTS

We thank John Moult and Eugene Melamud for the use and help with their bioinformatics web site (http://s2f.carb.umbi.umd.edu) and the Structural Genomics team at CARB 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 codes of ygfY are 1X6I and 1X6J.

REFERENCES

1. 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;25:3389–3402.
2. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. Environmental genome shotgun sequencing of the Sargasso Sea. Science 2004;304:66–74.
3. Enyenihi AH, Saunders WS. Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae. Genetics 2003;163:47–54.
4. Liu G, Sukumaran DK, Xu D, Chiang Y, Acton T, Goldsmith-Fischman S, Honig B, Montelione GT, Szyperski T. NMR structure of the hypothetical protein NMA1147 from Neisseria meningitidis reveals a distinct 5-helix bundle. Proteins 2004;55:756–758.
5. Otwinowski Z, Minor W. Processing of x-ray diffraction data collected in oscillation mode. Meth Enzymol 1997;276:307–326.
6. Schneider TR, Sheldrick GM. Substructure solution with SHELXD. Acta Crystallogr D Biol Crystallogr 2002;58:1772–1779.
7. Uson I, Sheldrick GM. Advances in direct methods for protein crystallography. Curr Opin Struct Biol 1999;9:643–648.
8. Otwinowski Z. Maximum likelihood refinement of heavy atom parameters. Paper presented at the *Isomorphous replacement and anomalous scattering*, Warrington, UK; 1991.
9. CCP4. The CCP4 suite: programs for protein crystallography. Acta Crystallgr D Biol Crystallogr 1994;50: 760–763.
10. Terwilliger TC. Maximum-likelihood density modification using pattern recognition of structural motifs. Acta Crystallogr D Biol Crystallogr 2001:57;1755–1762.
11. Jones TA, Zou JY, Cowan SW, Kjeldgaard. Improved methods for

building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 1991:47;110–119.
12. Brünger 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.
13. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood methods. Acta Crystallog 1997:D53;240–255.
14. Read RJ. Pushing the boundaries of molecular replacement with maximum likelihood. Acta Crystallogr D Biol Crystallogr 2001:57; 1373–1382.
15. Laskowski RA, MacArthur MW, Moss DS, Thornton J. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
16. Kraullis PJ. A program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.
17. Bacon DJ, Anderson WF. A fast algorithm for rendering space-filling molecule pictures. J Mol Graph 1988;6:219–220.
18. Merritt EA, Bacon DJ. Raster3D: Photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.
19. Stahlberg J, Divne C, Koivula A, Piens K, Claeyssens M, Teeri TT, Jones TA. Activity studies and crystal structures of catalytically deficient mutants of cellobiohydrolase I from Trichoderma reesei. J Mol Biol 1996;264:337–349.
20. Shoham M, Yonath A, Sussman JL, Moult J, Traub W, Kalb AJ. Crystal structure of demetallized concanavalin A: the metal-binding region. J Mol Biol 1979;131:137–155.
21. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247: 536–540.
22. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.