Solution structure of the hypothetical protein TA0095 from Thermoplasma acidophilum: A novel superfamily with a two-layer sandwich architecture

Esther León, Adelinda Yee, Angel R. Ortíz, Jorge Santoro, Manuel Rico and M. Angeles Jiménez

Protein Sci. 2007 16: 2278-2286; originally published online Aug 31, 2007;
Access the most recent version at doi:10.1110/ps.072869607

PROTEIN STRUCTURE REPORT

Solution structure of the hypothetical protein TA0095 from *Thermoplasma acidophilum*: A novel superfamily with a two-layer sandwich architecture

ESTHER LEÓN,[1] ADELINDA YEE,[2] ANGEL R. ORTÍZ,[3] JORGE SANTORO,[1] MANUEL RICO,[1] AND M. ANGELES JIMÉNEZ[1]

[1]Instituto de Química Física Rocasolano, CSIC, 28006 Madrid, Spain
[2]Ontario Cancer Institute, Toronto, Ontario M5G 2M9, Canada
[3]Centro de Biología Molecular, CSIC, 28049 Madrid, Spain

(RECEIVED March 13, 2007; FINAL REVISION June 1, 2007; ACCEPTED June 24, 2007)

Abstract

TA0095 is a 96-residue hypothetical protein from *Thermoplasma acidophilum* that exhibits no sequence similarity to any protein of known structure. Also, TA0095 is a member of the COG4004 orthologous group of unknown function found in Archaea bacteria. We determined its three-dimensional structure by NMR methods. The structure displays an α/β two-layer sandwich architecture formed by three α-helices and five β-strands following the order β1-α1-β2-β3-β4-β5-α2-α3. Searches for structural homologs indicate that the TA0095 structure belongs to the TBP-like fold, constituting a novel superfamily characterized by an additional C-terminal helix. The TA0095 structure provides a fold common to the COG4004 proteins that will obviously belong to this new superfamily. Most hydrophobic residues conserved in the COG4004 proteins are buried in the structure determined herein, thus underlying their importance for structure stability. Considering that the TA0095 surface shows a large positively charged patch with a high degree of residue conservation within the COG4004 domain, the biological function of TA0095 and the rest of COG4004 proteins might occur through binding a negatively charged molecule. Like other TBP-like fold proteins, the COG4004 proteins might be DNA-binding proteins. The fact that TA0095 is shown to interact with large DNA fragments is in favor of this hypothesis, although nonspecific DNA binding cannot be ruled out.

Keywords: COG4004 orthologous group; DNA-binding protein; new fold; NMR; protein structure; structural genomics

Supplemental material: see www.proteinscience.org

Genome sequencing projects are providing us with the amino acid sequences of thousands of proteins, but understanding the biological role of these proteins requires knowledge of their structure and function. Protein structure prediction would greatly benefit this if all the protein folds were known. Therefore, one of the main objectives of structural genomics is to fill the protein structure space. The best candidates to exhibit a novel protein fold are proteins lacking sequence homologs of known structure. The 96-residue hypothetical protein TA0095 from the Archeon *Thermophilus acidophilum*, whose solution structure we report here, is among them. TA0095 cannot be related to any previously characterized protein, and its function is unknown. According to a BLAST search (http://www.ncbi.nlm.nih.gov/blast/), the sequence of TA0095 is 63% identical and 77% similar to

Reprint requests to: M. Angeles Jiménez, Instituto de Química Física Rocasolano, CSIC Serrano-119, 28006 Madrid, Spain; e-mail: majimenez@iqfr.csic.es; fax: 34-915642431.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.072869607.

the hypothetical protein TV0159 from *Thermoplasma volcanium*, and exhibits ~30% identity and 55% similarity to 12 other proteins (Fig. 1A) annotated as hypothetical proteins found in the genomes of Archaea bacteria. TA0095 and these proteins have been classified as members of the COG4004 orthologous group (Tatusov et al. 1997, 2003; Geer et al. 2002; von Mering et al. 2003). Determining the three-dimensional (3D) structure of TA0095 will contribute to complete the universe of protein folds, and will give us clues about the functional role of this conserved protein.

Results and Discussion

The NMR structure of TA0095

TA0095 protein is a monomer under the experimental conditions used for NMR spectra acquisition as confirmed on the basis of ^{1}H NH amide T_2 relaxation experiments. The structure calculated for TA0095 protein on the basis of distance and angular and orientation restraints is well defined (see RMSD values in Table 1), except for residues 1–22 belonging to the N-terminal histidine tag, which, as expected, are disordered and will be excluded from the following description. In the structures calculated solely using NOEs and dihedral constraints (data not shown), the C-terminal segment, 106–118 (helix α3), was less ordered than the rest of the protein (Table 1); in particular, its position relative to the rest of the protein was ill defined. Whether this result is a consequence of true flexibility was discarded based on relaxation parameters. Thus, the ^{1}H–^{15}N heteronuclear NOEs for residues 106–118 that are ~0.7 indicate the absence of flexibility in the C-terminal helix (see Supplemental Fig. SF5). Incorporation of RDC constraints during structure refinement gave a better defined C-terminal region, although still not as well as the rest of the protein structure (Fig. 1B; Table 1). Regarding the side chains, those of 50 non-Ala/Gly/Pro residues are well defined (χ1 angular RMSD ≤ ± 30°), 29 of them corresponding to solvent-buried residues (ASA ≤ 25%) (Fig. 1C).

The TA0095 structure (Fig. 1B–D) exhibits an α/β-topology with three helices spanning residues 36–46, 89–104, and 108–115, and five β-strands that extend to residues 24–26, 49–53, 56–60, 67–71, and 76–80, following the order β1-α1-β2-β3-β4-β5-α2-α3. The loops connecting secondary structure elements are not long, and most of them contain regular β-turns. They are of type I (84–87), II (61–64), II′ (53–56), and IV (29–32, 72–75), according to PROMOTIF analysis (Hutchinson and Thornton 1996). The structure consists of two layers, one formed by a five-stranded antiparallel β-sheet (β2-β3-β4-β5-β1), and the other by the three helices that are packed on the same face of the β-sheet. Although no hydrogen bonds were incorporated as restrictions in the NMR structure calculation, the TA0095 structure ensemble exhibits most of the characteristic hydrogen bonds between the CO of residue *i* and the NH of residue *i* + 4 for the three helices as well as the hydrogen bonds linking adjacent β-strands (Supplemental Table SM1; Supplemental Fig. SM3). Many of the hydrogen-bonded amide NH protons exchange slowly with solvent (Y26, I39, K40, R41, K42, I43, L46, F48, I58, A59, I69, K70, I76, V78, N79, F99, E101, L103, G104) (Supplemental Table ST1).

Side chains of buried residues, a total of 45 residues with ASA <25%, are located either at the outer β-sheet face (I58, S60, E68, K70, L77, N79) or between the two layers forming a compact hydrophobic core (Fig. 1C). Only nine of the buried residues are not hydrophobic; four of them are polar (S38, S60, N79, S107), and five are charged (K40, E68, K70, E101, E110). Based on the proximity between carboxyl, amino, and hydroxyl groups in TA0095 structures, the ammonium group of K40 might interact with the hydroxyl group of S52, the carboxyl group of E68 with the ammonium group of K70, probably forming a salt bridge, the carboxyl group of E101 with the hydroxyl group of S107, and the carboxyl group of E110 with the hydroxyl group of Y106 and maybe with the amino group of K29 as well. The E101/S107 and the E110/Y106-K29 interactions are probably involved in fixing the orientation of helix α3 (Fig. 1C).

Based on the exchange rates of most protected amide protons (those with log $P > 4$, 16 in total) (Supplemental Table SM1) at pH 6.5 and 298 K, the global free energy of unfolding (ΔG_u) of TA0095 can be evaluated to be 6.0 ± 0.6 kcal/mol^{-1}. The residues forming this slow-exchanging core of the protein encompass the three inner β-strands (β3, β4, and β5) and the second halves of helices α1 and α2 (see Supplemental Fig. SF6).

The TA0095 structure represents the common fold for orthologous COG4004 proteins

The secondary structure elements displayed by TA0095 structure coincide approximately with those predicted from the sequence as well as with the predictions for the other known COG4004 proteins, except for the β1-strand (Fig. 1A) and, in the case of the MK1058 protein from *Methanopyrus kandleri*, the helix α3. Hence, the NMR solution structure reported here for TA0095 can be considered as representative of the COG4004 ortholog group and could be used to model the structures of the other members of the group. It is instructive to analyze the location of the conserved residues within the TA0095 3D structure. The longest segment of conserved consecutive residues encompasses the loop linking helices α2 and α3 and most of helix α3, except in the case of the

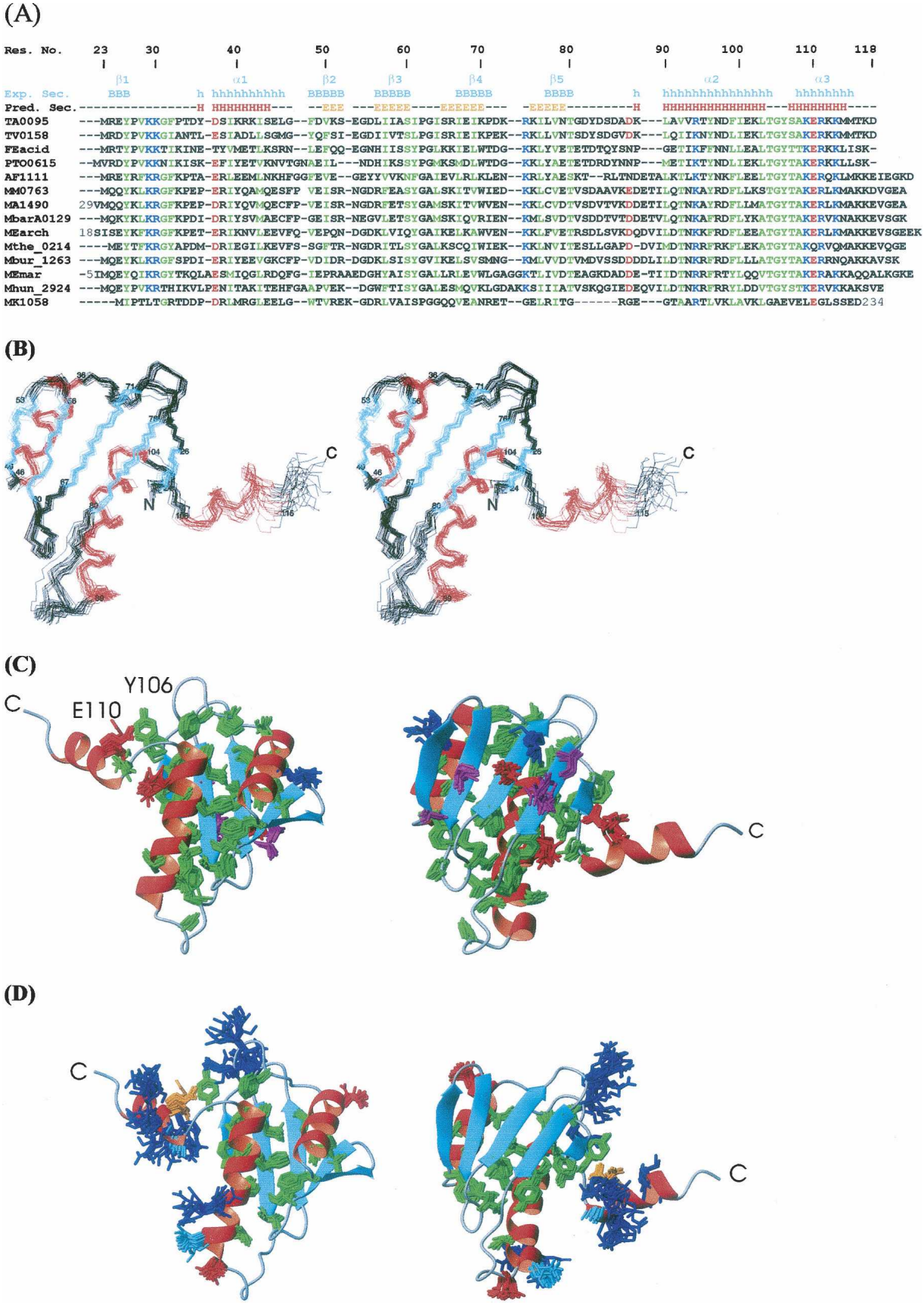

Figure 1. (Legend on next page)

MK1058, which, on the other hand, is the one showing less sequence similarity to all other proteins (Fig. 1) and the only one found as a domain of a larger protein. Apart from those regions, the less conserved segments and those where sequence insertions and deletions occur correspond generally to either loop regions or to helical N termini. A large number of conserved residues, mostly the hydrophobic ones (Y26, V28, F32, I39, I43, F48, V50, I57, I61, I64, I67, I69, I76, V78, V92, Y96, F99, I100, L103, T104, G105, Y106, E110), are buried in the TA0095 structure (<25% of solvent-accessible surface) (Fig. 1D), which is probably important in terms of structure stability. The fact that the sequence of the loop between helices α2 and α3 is identical in all the proteins, except for MK1058 (Fig. 1), indicates that it could be essential to maintain the singular position of helix α3, pointing toward the solvent and exhibiting contacts with other protein regions only at its N-terminal half. The proximity of the side chain of Y106 and that of the also conserved residue K29 in the TA0095 structure suggests that the interaction between these two residues, of the π-cation nature, might contribute to fix the position of helix α3 (Fig. 1C).

The TA0095 structure is the first known member of a novel superfamily

To find if TA0095 structure represents a novel fold, we searched for structural homologs by using DALI (http://www.ebi.ac.uk/dali/server; Holm and Sander 1993), CATHEDRAL (http://www.cathdb.info/cgi-bin/cath/CathedralServer.pl; Pearl et al. 2003), and MAMMOTH (Ortiz et al. 2002; Lupyan et al. 2005). By all the procedures, we found that the structure of segment 23–105 is structurally similar to α + β proteins belonging to the TATA-box-binding-protein fold (TBP-like fold, according to SCOP nomenclature) (Murzin et al. 1995). Thus, the DALI Z scores for two TATA-box-binding proteins, one from a yeast and the other from an Archaeon (pdb codes 1ytb and 1ais, respectively), are 6.2 and 6.0, respectively (DALI Z scores lower than 2.0 are not significant), and the corresponding RMSD values are 2.9 Å and 2.3 Å. However, the structure of TA0095 differs from all structures previously classified within that fold in the presence of an additional C-terminal helix, helix α3. Given this difference and the fact that helix α3 is a characteristic of the COG4004 orthologous proteins, the TA0095 structure reported here represents the first known member of a novel superfamily within the TBP-fold.

TA0095 and its orthologous COG4004 proteins might be DNA-binding proteins

Inferring function from protein structure is a subject of great interest in structural genomics, even when there is no straightforward relationship. Given the sequence similarities existing among the COG4004 proteins, which, at least up to now, have been found exclusively in archaeal genomes, they must play a common biological role. Assuming that TA0095 biological function probably requires its interaction with other biomolecules, examination of the TA0095 protein surface will provide us with insights into its function. In addition, we will take into account the solvent-exposed residues that are conserved in COG4004 proteins because these residues are the most likely to participate in their interactions with other biomolecules.

Two different regions are distinguished on the electrostatic surface of TA0095 structure: a large, positively charged patch encompassing the outer side of the helical layer and the interface between strand β1 and helix α3 (helix α3 is approximately in the middle of the patch) and a negatively charged cleft (Fig. 2A,B). Only two of the solvent-exposed residues conserved in COG4004 proteins have a negative charge, D37 and D88, and only the last one belongs to the negatively charged cleft. In contrast, all other

Figure 1. (*A*) Alignment of TA0095 amino acid sequence with those of their 13 known orthologs (TV0158, FEacid, and PTO0165 from *Thermoplasma volcanium*, *Ferroplasma acidarmanus*, and *Picrophilus torridus*, respectively, that together with *T. acidophilum* are classified as Thermoplasmatales; AF1111 from *Archeoglobus fulgidus*; MM0763, MA1490, MbarA0129, MEarch, Mthe_0214, Mbur_1263, MEmar, Mhun_2924 from the Methanomicrobia *Methanosarcina mazei*, *Methanosarcina acetivorans*, *Methanosarcina barkeri*, Methanogenic archaeon, *Methanosaeta thermophila*, *Methanococcoides burtonii*, *Methanoculleus marisnigri*, and *Methanospirillum hungatei*, respectively; and MK1058 from *Methanopyrus kandleri*), (Tatusov et al. 1997, 2003; Geer et al. 2002; von Mering et al. 2003). Their NCBI accession codes are NP_393573.1, NP_110677.1, ZP_00609935.1, YP_023393.1, NP_069940.1, NP_632787.1, NP_619056.1, YP_303694.1, CAJ38058.1, YP_842652.1, YP_565934.1, ZP_01390923.1, and YP_504335.1. The numbers at the ends of a sequence refer to additional nonaligned N- or C-terminal residues. Secondary structure elements predicted for TA0095 are shown at the *top* (http://www.compbio.dundee.ac.uk/~www-jpred/). (Blue) Conserved positively charged residues; (red) conserved negatively charged residues; (green) all other conserved amino acids. The numbering corresponds to the H_6-TA0095 sequence. (Cyan) α-Helices and β-strands present in the TA0095 structure determined by NMR. (*B*) Stereoview of the backbone atom superposition of the 20 lowest energy structures calculated for TA0095. Residues at the ends of helices and strands that are colored in red and blue, respectively, are numbered. (*C*) Ribbon representation in two different orientations showing the side chains for buried residues (ASA < 25%). (Green) Residues at the hydrophobic core; (magenta) those at the outer β-sheet face; except for charged residues, which, independently of their location, are blue if positive and red if negative. (*D*) TA0095 ribbon structure in the same orientations as panel *C* showing the side chains of residues conserved in the COG4004 orthologous proteins. (Green) Buried residues except for (orange) E110; solvent-exposed residues are in blue (K/R), red (D37), and cyan (noncharged residues).

Table 1. *Structural statistics for the ensemble of 20 lowest energy structures of TA0095*

	CYANA	CYANA + RDC
Distance restraints		
All	1433	1433
Intraresidue	436	436
Sequential	365	365
Medium range ($i - j$ <5)	231	231
Long range ($i - j$ >5)	401	401
Dihedral angle restraints		
All	186	186
φ	97	97
ψ	89	89
Residual dipolar couplings		
All	—	52
Axial	—	−9.3
Rhombic	—	−3.1
Maximum NOE violations per structure (Å)	0.20 ± 0.06	0.22 ± 0.06
Maximum dihedral angle violations per structure (°)	2 ± 1	3 ± 1
Maximum orientation violations per structure (H_2)	—	0.4 ± 0.4
Averaged pairwise RMSD		
Backbone atoms		
23–116	1.1 ± 0.4	0.9 ± 0.3
23–106	0.5 ± 0.1	0.6 ± 0.2
107–116	0.9 ± 0.4	0.7 ± 0.3
All heavy atoms		
23–116	1.8 ± 0.3	1.6 ± 0.2
23–106	1.3 ± 0.1	1.4 ± 0.1
107–116	1.8 ± 0.4	1.8 ± 0.4
Ramachandran plot (%)[a]		
Residues in most favored regions	86.2	86.7
Residues in additional allowed regions	13.8	13.3
Residues in generously allowed regions	0.0	0.0
Residues in disallowed regions	0.0	0.0

[a]Residues 1–22 corresponding to the histidine tag are excluded from these analyses.

solvent-accessible conserved residues, which are mostly positively charged (K29, K30, G31, R74, L90, R94, A108, K109, R111, K113; accessible surface >30%), lie within the positively charged patch (Fig. 2B). They belong to the first loop of the TA0095 chain (between strand β1 and helix α1), to the loop β4–β5 (R74), to the N-terminal region of helix α2 (L90, R94), and to helix α3 (Fig. 1A). This last helix and the loop β1–α1 are far apart in the protein sequence but spatially close in the TA0095 structure. All of them form a large, positively charged patch on the protein surface. Given the differences in residue conservation between positively and negatively charged surface regions, it seems more likely that TA0095 biological function requires binding to some negatively charged molecule, such as DNA, RNA, some lipids or small ligands, than to a positively charged ligand.

Although no typical DNA-binding motif was found by using the program Profunc (Laskowski et al. 2005a,b), since some of the proteins structurally similar to TA0095 are DNA-binding proteins, we considered it worthy to test whether TA0095 was able to bind to DNA. To that end, we recorded 1D ^{1}H-NMR spectra of TA0095 at increasing amounts of sheared salmon DNA. Upon DNA addition, the ^{1}H NMR signals of TA0095 broaden up to finally disappear at a large amount of sheared DNA (Fig. 2C). This result indicates that TA0095 is able to interact with DNA at least nonspecifically. Further work is needed to confirm if TA0095 binds DNA in a specific way. At the moment, it cannot be ruled out that TA0095 function would occur through binding to other negatively charged molecules, such as RNA or other small ligands.

Conclusion

TA0095 is an α/β protein with a two-layer sandwich architecture. The TA0095 structure can be considered as the first one determined of a novel superfamily within the TBP-like fold. An additional C-terminal helix is a unique characteristic of this new superfamily. The TA0095 structure represents a common fold for the COG4004 group of orthologous proteins that will also belong to this new superfamily.

This COG4004 orthologous domain is present in the genomes sequenced for archaeobacterias. However, no

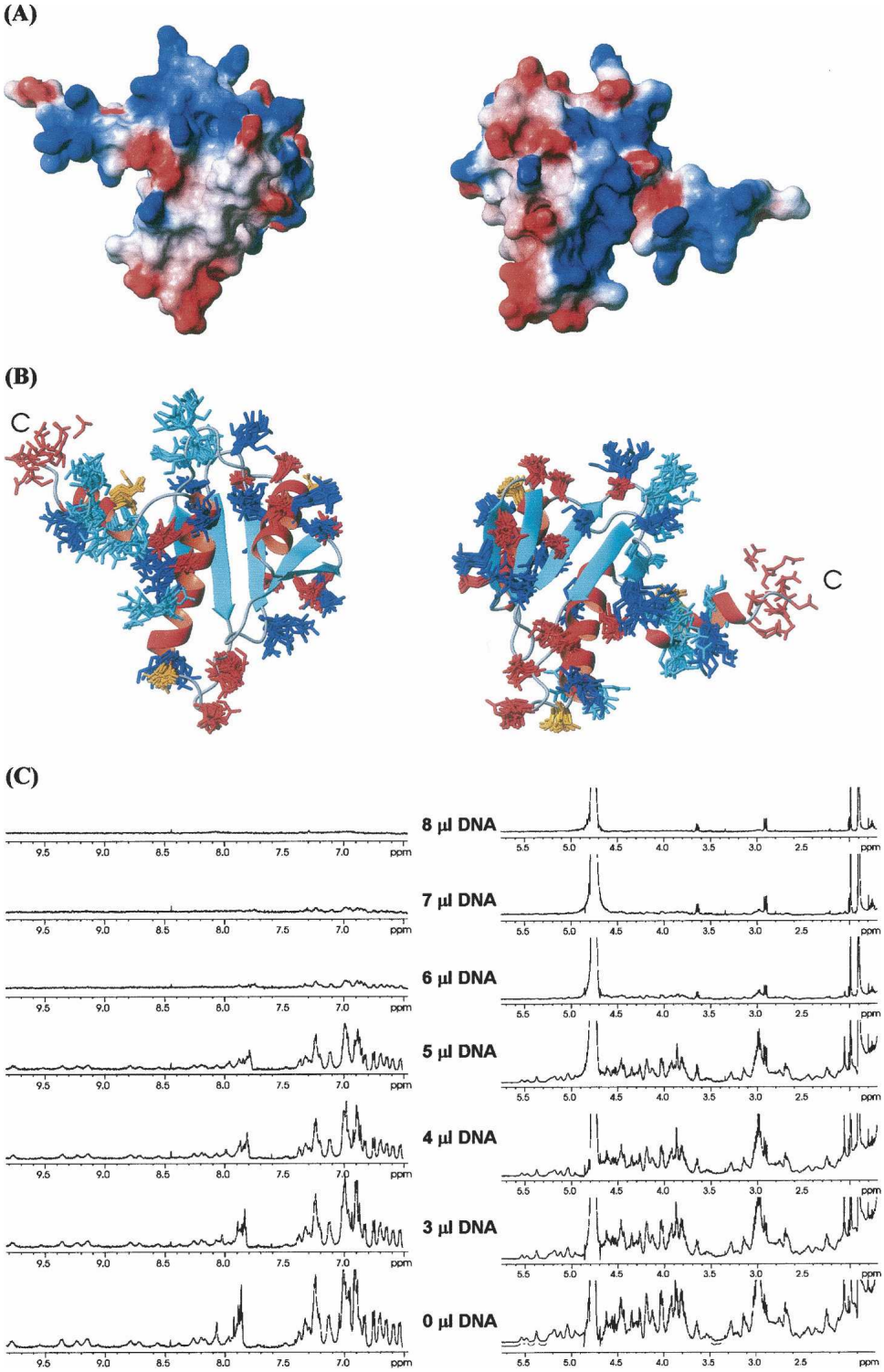

Figure 2. (*A*) TA0095 ribbon structure in two different orientations showing the side chains of charged residues. (Blue) K/R residues; (red) D/E. (Cyan) K/R residues conserved in the COG4004 orthologous proteins; (orange) conserved D/E. (*B*) Electrostatic surface of TA0095 structure with positive and negative charges shown in blue and red, respectively, in the same orientations as panel *A*. (*C*) 1D ^{1}H NMR spectra of 0.2 mM TA0095 with increasing amounts of a saturated solution of salmon sperm DNA (D_2O, pH 6.5, 25°C).

biological function has yet been found for them. Based on the structure determined for TA0095, that is, a large, positively charged patch, and structural similarity to DNA-binding proteins, we propose that the COG4004 proteins are DNA-binding domains. The fact that TA0095 binds to large DNA fragments is in favor of this hypothesis, though nonspecific DNA binding cannot be ruled out.

Materials and Methods

Expression and purification

Ta0095 was cloned in plasmid pET11b with an N-terminal $(His)_6$ tag and a Tobacco Etch Virus (TEV) protease cleavage site. The plasmid was transformed into the *Escherichia coli* BL21-Gold (DE3) strain for expression. For unlabeled samples, the cells were grown overnight in LB-(ampicillin + kanamicin) at 310 K from a single colony. Inoculation, at 1:100, was done in fresh medium, and the culture was grown until OD_{600nm} reached 0.6, at which point the cells were induced with 1 mM isopropyl β-D-thiogalactopyronoside (IPTG) and allowed to grow for 5 h to maximize expression levels. Then, the cells were harvested by centrifugation. For labeled samples, cells were grown in 1 L of Luria-Bertani medium at 310 K until OD_{600} reached 0.7. Then, the cells were pelleted, washed, and centrifuged for 10 min at 277 K and 6000*g*, using an M9 salt solution, containing no nitrogen and carbon sources. The cell pellet was resuspended in 250 mL of isotopically labeled M9 minimal media, prepared with ^{13}C-labeled glucose and ^{15}N ammonium chloride. After 1 h, protein expression was induced by addition of IPTG to 0.8 mM concentration, and the cells were allowed to grow overnight at 298 K. Then, the cells were harvested. Once harvested, the same protocol was followed for labeled and unlabeled samples. Harvested cells were resuspended in 5 mL of 20 mM Tris-HCl buffer (pH 8.0). After lysation by French press and centrifugation at 15,000*g*, the soluble fraction was submitted to Ni^{2+} affinity chromatography (Amersham Biosciences), and the affinity beads were washed three times with five column volumes of 500 mM NaCl, 20 mM Tris, and 20 mM imidazole (pH 8.0); the protein was eluted with five column volumes of the same buffer plus 500 mM imidazole. The purification was completed loading the sample in a HiLoad 26/60 Superdex 75 column in 0.1 M ammonium acetate (pH 7.0). The pool containing the recombinant protein was lyophilized.

NMR spectroscopy

Samples for NMR experiments were ~1 mM protein concentration in 0.5 mL of H_2O/D_2O (9:1 ratio by volume) or in pure D_2O at pH 6.5 in the case of ^{15}N-labeled and unlabeled TA0095 and ~0.5 mM protein concentration in 25 mM sodium phosphate buffer (pH 6.5), 0.45 mM NaCl, 10 mM DTT, and 20 μM $ZnCl_2$ in the case of the ^{13}C,^{15}N-labeled TA0095. pH was measured with a glass microelectrode and was not corrected for isotope effects. NMR experiments were recorded at 298 K in either a Bruker AV 600 MHz spectrometer equipped with a z-gradient cryoprobe or a Bruker AV 800 MHz spectrometer. The temperature of the NMR probe was calibrated using a methanol sample. 2D and 3D NMR spectra were acquired by using standard pulse sequences. NMR spectra were processed using the XWIN-NMR (Bruker Biospin) software and analyzed with Sparky (Goddard and Kneller 2007). Sodium 2,2-dimethyl-2-silapentane-5-sulphonate (DSS) was used as an internal chemical shift reference. The 0-ppm ^{13}C and ^{15}N δ-values were obtained indirectly by multiplying the spectrometer frequency that corresponds to 0 ppm in the ^{1}H spectrum, assigned to internal DSS reference by 0.251449530 and 0.101329118, respectively (Markley et al. 1998).

Sequential backbone assignments were accomplished using 3D HNCA, HNCACB, CBCA(CO)NH, and HN(CA)HA experiments (Sattler et al. 1999) acquired using the ^{13}C,^{15}N-labeled TA0095 sample and 3D HNHA spectra (Vuister and Bax 1993) recorded with a ^{15}N-TA0095 sample. Only the backbone amides of residues 5–10 belonging to the histidine tail were not assigned because of exchange broadening, as observed for histidine-tail residues in other proteins (Volpon et al. 2006). Side-chain resonances were assigned from ^{15}N-edited 3D 60-ms TOCSY and 80-ms NOESY spectra (Palmer et al. 1992) acquired with a ^{15}N-TA0095 sample, 2D homonuclear COSY, 60-ms TOCSY, and 100-ms NOESY experiments (Kumar et al. 1980; Bax and Davis 1985) recorded with unlabeled TA0095 in H_2O/D_2O 9:1 (v/v) and in D_2O and a ^{1}H–^{13}C HSQC recorded with the unlabeled sample in D_2O. The assigned chemical shifts have been deposited in the BioMagResBank (accession code: 15172; http://www.bmrb.wisc.edu).

Heteronuclear ^{1}H–^{15}N NOEs were determined from the ratio of two experiments with and without saturation as described previously (Farrow et al. 1995; Mulder et al. 1998).

Samples of oriented protein were prepared by adding previously formed liquid crystals to the lyophilized protein. The liquid crystals were prepared as described by Rückert and Otting (2000) using a weight percentage of 5% for the ratio of C8E5 [*n*-octyl-penta(ethylene glycol)] to water (H_2O/D_2O 9:1 [v/v]), and a final molar ratio of C8E5/*n*-octanol of 0.87 achieved by adding 1 μL aliquots of *n*-octanol. The pH was adjusted to 6.5. The solutions were biphasic at low alcohol concentrations and became instantaneously transparent and opalescent upon crossing the L_α phase boundary. Anisotropy of the oriented samples was verified from the doublet ^{2}H splitting of the solvent (Rückert and Otting 2000), 24 Hz at 293 K.

One-bond ^{1}H–^{15}N dipolar couplings were determined from the signal splitting in the nitrogen dimension in F1-coupled HSQC spectra acquired with a pulse sequence that separates the doublet components in the F1 dimension into two subspectra (Garcia-Mayoral et al. 2005). Spectra were acquired at 308 K for the isotropic medium (scalar contribution, $^1J_{NH}$) and 293 K for the anisotropic medium (scalar and dipolar contributions, $^1J_{NH} + {}^1D_{NH}$). The residual dipolar couplings were obtained from the difference in splitting between both conditions.

The NH exchange was followed by a tandem method. The exchange reaction was started by dissolving the lyophilized ^{15}N-TA0095 in D_2O at pH 6.5. Once transferred to a 5-mm NMR tube and shimmed, a series of consecutive 2D ^{1}H–^{15}N-HSQC experiments was run at 298 K over a 15-d period. The first 2D spectrum was recorded 13–16 min after dissolving the protein. 2D ^{1}H–^{15}N-HSQC spectra were acquired with 2048 complex data points in t_2 and with 256 t_1 increments with eight scans per increment. The acquisition time for each 2D ^{1}H–^{15}N-HSQC was ~45 min.

Hydrogen exchange rates were determined by fitting cross-peak volumes that were measured using the Sparky program to a first-order exponential decay:

$$I(t) = I(0)\exp(-k_{ex}t) \quad (1)$$

where I represents the volume of the cross-peak; $I(0)$ is the cross-peak volume at $t = 0$; k_{ex} is the experimental rate of hydrogen exchange, and t is the time in minutes. Data were fitted with the program Microcal Origin 6.0.

Hydrogen exchange data were analyzed within the EX2 exchange limit, the dominant mechanism of exchange for most proteins at moderate pH and temperature, where Equation 2 applies:

$$k_{ex} = K_{op} * k_{rc} \quad (2)$$

where K_{op} is the equilibrium constant for local transient opening of a hydrogen-bonded site; and k_{rc} is the intrinsic exchange rate constant for each amide proton, calculated as described (Bai et al. 1993). The equilibrium constant relates to ΔG_{op}, the structural free energy difference between the closed and open states of a hydrogen-bonded site, according to Equation 3:

$$\Delta G_{op} = -RT \ln K_{op} = -RT \ln(k_{ex}/k_{rc}) \quad (3)$$

with R being the gas constant and T the absolute temperature, and other parameters as defined above.

Then, the free energy of protein unfolding can be obtained from the exchange rates of the slow-exchanging amide protons by assuming that the exchange of the most protected amide protons occur via global unfolding (Englander et al. 1996). In this case, ΔG_{op} can be considered equal to ΔG_u (unfolding free energy variation). Hence, $\Delta G_u = \Sigma \Delta G_{op}(i)/i$, where i refers to the slow-exchanging amide protons. To account for the fact that the proline residues in the unfolded state do not have time to reach their isomeric equilibrium distribution during the exchange experiments, the ΔG_u value obtained was corrected by the effect of the four Pro residues present in TA0095 protein (P27, P33, P62, and P71; 0.53 kcal/mol^{-1} at 298 K) (evaluated as described by Bai et al. 1994).

Structure calculation

Distance constraints were derived from three different NOE spectra: a ^{15}N-edited 3D NOESY, and two 2D NOESY spectra, one recorded in H_2O/D_2O 9:1 (v/v) and the other in D_2O. ϕ and ψ angle restraints were obtained from $^1H_\alpha$, $^{13}C_\alpha$, and $^{13}C_\beta$ chemical shifts by using the TALOS program (Cornilescu et al. 1999). Structures were calculated using the program CYANA and the CANDID protocol for iterative automatic NOE assignment (Guntert et al. 1997). Experimental ^{1}H–^{15}N RDCs were introduced as a final step of the structure calculation within the program CYANA. The tensor parameters were obtained by best fitting of the RDC values (D^{ij}) to Equation 4:

$$D^{ij}(\theta,\phi) = D_a^{ij}\left[(3\cos^2\theta - 1) + \frac{3}{2}R\sin^2\theta\cos(2\phi)\right] \quad (4)$$

by using the unrefined TA0095 structure as input structural model, where D_a^{ij} is the axial component of the alignment tensor, R is its rhombicity, and θ and ϕ are the polar angles of the vector ij with respect to the tensor principal axis. Coordinates have been deposited in the Protein Data Bank (accession code: 2joi; http://www.rcsb.org/pdb/).

Structure analysis

Structures were analyzed by using MOLMOL (Koradi et al. 1996), PROCHECK/NMR (Laskowski et al. 1996), and PROMOTIF (Hutchinson and Thornton 1996). A side-chain torsion angle was considered as well defined when the root mean square deviation between values in the 20 calculated structures is $< \pm 30°$ or its order parameter is in the range 0.87–1.00 (the order parameter is zero for a totally random dihedral angle and one for a completely fixed conformation). The criteria for hydrogen-bond formation were: (1) proton-acceptor distance lower than 2.4 Å; and (2) the donor-acceptor angle lower than 35°.

DNA-binding assays

^{1}H 1D NMR spectra of 200 μM protein samples in D_2O at pH 6.5 were acquired at 25°C containing different amounts of sheared salmon sperm DNA (Sigma). Volume samples were 3-mm 200 μL in 3-mm NMR tubes. To prepare sheared DNA, salmon sperm DNA was dissolved in D_2O at saturation and sheared using a syringe with a 30-gauge, 8-mm hypodermic needle.

Electronic supplemental material

The Supplemental material includes a table listing NH/ND amide exchange data and six figures providing further NMR data, such as NOE summary, and ^{1}H–^{15}N heteronuclear NOEs.

Acknowledgments

We thank Dr. M.A. Treviño for help with protein preparation and Drs. J.M. Pérez-Cañadillas and C. González for suggestions on the DNA-binding assays. This work was supported by the Spanish MEC projects GEN2003-20642-C09-03 and BFU2005-01855/BMC and by the Ontario Research Development Challenge Fund and Genome Canada.

References

Bai, Y., Milne, J.S., Mayne, L., and Englander, S.W. 1993. Primary structure effects on peptide group hydrogen exchange. *Proteins* **17:** 75–86.

Bai, Y., Milne, J.S., Mayne, L., and Englander, S.W. 1994. Protein stability parameters measured by hydrogen exchange. *Proteins* **20:** 4–14.

Bax, A. and Davis, D.G. 1985. MLEV-17 based two-dimensional homonuclear magnetisation transfer spectroscopy. *J. Magn. Reson.* **65:** 355–360.

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.

Englander, S.W., Sosnick, T.R., Englander, J.J., and Mayne, L. 1996. Mechanisms and uses of hydrogen exchange. *Curr. Opin. Struct. Biol.* **6:** 18–23.

Farrow, N.A., Zhang, O., Szabo, A., Torchia, D.A., and Kay, L.E. 1995. Spectral density function mapping using ^{15}N relaxation data exclusively. *J. Biomol. NMR* **6:** 153–162.

Garcia-Mayoral, M.F., Pantoja-Uceda, D., Santoro, J., Martinez del Pozo, A., Gavilanes, J.G., Rico, M., and Bruix, M. 2005. Refined NMR structure of α-sarcin by ^{15}N–^{1}H residual dipolar couplings. *Eur. Biophys. J.* **34:** 1057–1065.

Geer, L.Y., Domrachev, M., Lipman, D.J., and Bryant, S.H. 2002. CDART: Protein homology by domain architecture. *Genome Res.* **12:** 1619–1623.

Goddard, T.D. and Kneller, D.G. 2007. *SPARKY 3*. University of California, San Francisco, CA.

Guntert, P., Mumenthaler, C., and Wüthrich, K. 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273:** 283–298.

Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233:** 123–138.

Hutchinson, E.G. and Thornton, J.M. 1996. PROMOTIF—A program to identify and analyze structural motifs in proteins. *Protein Sci.* **5:** 212–220.

Koradi, R., Billeter, M., and Wuthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14:** 29–32, 51–55.

Kumar, A., Ernst, R.R., and Wüthrich, K. 1980. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton—proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Commun.* **95:** 1–6.

Laskowski, R.A., Rullmannn, J.A., 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.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. 2005a. ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33:** W89–W93.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. 2005b. Protein function prediction using local 3D templates. *J. Mol. Biol.* **351:** 614–626.

Lupyan, D., Leo-Macias, A., and Ortiz, A.R. 2005. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* **21:** 3255–3263.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E., and Wüthrich, K. 1998. IUPAC-IUBMB-IUPAB inter-union task group on the standardization of databases of protein and nucleic acid structures determined by NMR spectroscopy. *Pure Appl. Chem.* **70:** 117–142.

Mulder, F.A.A., de Graaf, R.A., Kaptein, R., and Boelens, R. 1998. An off-resonance rotating frame relaxation experiment for the investigation of macromolecular dynamics using adiabatic rotations. *J. Magn. Reson.* **131:** 351–357.

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.

Ortiz, A.R., Strauss, C.E., and Olmea, O. 2002. MAMMOTH (MAtching Molecular Models Obtained from THeory): An automated method for model comparison. *Protein Sci.* **11:** 2606–2621.

Palmer, A.G., Cavanagh, J., Byrd, R.A., and Rance, M. 1992. Sensitivity improvement in three-dimensional heteronuclear correlation NMR spectroscopy. *J. Magn. Reson.* **96:** 416–424.

Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., and Orengo, C.A. 2003. The CATH database: An extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* **31:** 452–455.

Rückert, M. and Otting, G. 2000. Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J. Am. Chem. Soc.* **122:** 7793–7797.

Sattler, M., Schleucher, J., and Griesinger, C. 1999. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34:** 93–158.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41. doi: 10.1186/1471-2105-4-41.

Volpon, L., Young, C.R., Matte, A., and Gehring, K. 2006. NMR structure of the enzyme GatB of the galactitol-specific phosphoenolpyruvate-dependent phosphotransferase system and its interaction with GatA. *Protein Sci.* **15:** 2435–2441.

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.

Vuister, G.W. and Bax, A. 1993. Quantitative *J* correlation: A new approach for measuring homonuclear three-bond $J(H^{N}H^{\alpha})$ coupling constants in ^{15}N-enriched proteins. *J. Am. Chem. Soc.* **115:** 7773–7777.