Published in final edited form as:

Proteins. 2011 July ; 79(7): 2146–2160. doi:10.1002/prot.23035.

Crystal structure of a metal-dependent phosphoesterase (YP_910028.1) from *Bifidobacterium adolescentis*: Computational prediction and experimental validation of phosphoesterase activity

Gye Won Han^{1,2}, Jaeju Ko³, Carol L. Farr^{1,2}, Marc C. Deller^{1,2}, Qingping Xu^{1,4}, Hsiu-Ju Chiu^{1,4}, Mitchell D. Miller^{1,4}, Jana Sefcikova⁵, Srinivas Somarowthu⁵, Penny J. Beuning⁵, Marc-André Elsliger^{1,2}, Ashley M. Deacon^{1,4}, Adam Godzik^{1,6,7}, Scott A. Lesley^{1,2,8}, Ian A. Wilson^{1,2,*}, and Mary Jo Ondrechen^{5,*}

¹Joint Center for Structural Genomics

²Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

³Department of Chemistry, Indiana University of Pennsylvania, Indiana, PA 15705, USA

⁴Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

⁵Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115, USA

⁶Center for Research in Biological Systems, University of California, San Diego, La Jolla, CA 92093, USA

⁷Program on Bioinformatics and Systems Biology, Sanford-Burnham Medical Research Institute, La Jolla, CA 92037, USA

⁸Protein Sciences Department, Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA

Abstract

The crystal structures of an unliganded and adenosine 5'-monophosphate (AMP) bound, metal-dependent phosphoesterase (YP_910028.1) from *Bifidobacterium adolescentis* are reported at 2.4 Å and 1.94 Å, respectively. Functional characterization of this enzyme was guided by computational analysis and then confirmed by experiment. The structure consists of a PHP (Polymerase and Histidinol Phosphatase, Pfam: PF02811) domain with a second domain (residues 105–178) inserted in the middle of the PHP sequence. The insert domain functions in binding AMP, but the precise function and substrate specificity of this domain is unknown. Initial bioinformatics analyses yielded multiple potential functional leads, with most of them suggesting DNA polymerase or DNA replication activity. Phylogenetic analysis indicated a potential DNA polymerase function that was somewhat supported by global structural comparisons identifying the closest structural match to the alpha subunit of DNA polymerase III. However, several other functional predictions, including phosphoesterase, could not be excluded. *THEMATICS*, a computational method for the prediction of active sites from protein 3D structures, identified

*Corresponding authors: Ian A. Wilson, The Scripps Research Institute, 10550 N. Torrey Pines Rd., BCC206, La Jolla, CA 92037, USA, wilson@scripps.edu, Tel: (858) 784-2232, Fax: (858) 784-2980. Mary Jo Ondrechen, Department of Chemistry and Chemical Biology, Northeastern University, 360, Huntington Ave., 102 HT, Boston, MA 02115, USA, mjo@neu.edu, Tel: (617) 373-2856, Fax: (617) 373-8795.

potential reactive residues in YP_910028.1. Further analysis of the predicted active site and local comparison with its closest structure matches strongly suggested phosphoesterase activity, which was confirmed experimentally. Primer extension assays on both normal and mismatched DNA show neither extension nor degradation and provide evidence that YP_910028.1 has neither DNA polymerase activity nor DNA proofreading activity. These results suggest that many of the sequence neighbors previously annotated as having DNA polymerase activity may actually be misannotated.

Keywords

Functional annotation; structural genomics; phosphoesterase; THEMATICS; active site prediction

INTRODUCTION

YP_910028.1 from *Bifidobacterium adolescentis* (strain ATCC 15703 / DSM 20083) was selected for crystallographic characterization because it is a member of a family of proteins that are over-represented in the human gut microbiome.

B. adolescentis is a gram positive bacterium which colonizes the human gut intestinal tract days after birth. It is particularly prevalent in breast fed infants¹ and its numbers remain steady until late adulthood when its population declines.² Members of the genus *Bifidobacteria* are reported to have probiotic activity³ and are widely used in the food industry often as bio-milks and bio-yoghurts.⁴ Reported probiotic effects in humans include: inhibition of carcinogenesis, re-establishment of normal gut flora after antibiotic treatment, production of anticholesteremic compounds, increased calcium resorption, destruction of anti-nutrition factors, increased vitamin synthesis and protein predigestion⁵. Little is known about the structure and function of *B. adolescentis* proteins and only eleven structures, the two structures (PDB IDs: 3e0f, 3o0f) presented here and nine others (PDB IDs: 3onq, 3cym, 3cpg, 3luy, 3fjy, 2gdu, 2gdv, 1r7a and 3i8b), are available from the Protein Data Bank (PDB).

Initial bioinformatics analyses of the YP_910028.1 amino-acid sequence yielded multiple potential functions. Phylogenetic analysis indicated a potential DNA polymerase or DNA replication function. However, a different prediction emerged from a local 3D structure analysis at the predicted active site, as described herein. THEMATICS (Theoretical Microscopic Anomalous Titration Curve Shapes)^{6,7} is a computational method for the identification of potential catalytic and binding residues based solely on the local environment of residues in the structure. THEMATICS computes the microscopic theoretical titration curves for all ionizable residues to identify sets of residues with unusual proton binding characteristics, defined as a spatial cluster of two or more such residues. This method accurately predicted active sites in a set of 170 experimentally characterized enzymes.⁸ It also has been used to classify members of the DJ-1 superfamily into functional subfamilies⁹ and to provide confirmation, or evidence against, putative annotations of proteins of unknown function.¹⁰ THEMATICS analysis and subsequent comparison of potential active site residues, based on local structural alignment at the predicted active site, strongly suggests phosphoesterase activity for YP_910028.1. Phosphoesterase activity as well as the absence of DNA polymerase and DNA proofreading activity were both confirmed by experiment.

Here we report the functional assignment of metal-dependent phosphoesterase activity to YP_910028.1, based on theoretical predictions coupled with analysis of its unliganded (Apo) and ligand (AMP) bound crystal structures and subsequent experimental

confirmation. The Apo-YP_910028.1 and AMP-YP_910028.1 crystal structures were determined to 2.4 Å and 1.94 Å, respectively, using the semi automated, high-throughput pipeline of the Joint Center for Structural Genomics (JCSG; <http://www.jcsg.org>)¹¹ as part of the NIGMS Protein Structure Initiative (PSI; <http://www.nigms.nih.gov/Initiatives/PSI/>).

MATERIALS AND METHODS

Protein production and crystallization

Clones were generated using the Polymerase Incomplete Primer Extension (PIPE) cloning method.¹² The gene encoding YP_910028.1 (GenBank: YP_910028, GI:gi119026183; Swiss-Prot: A1A2L3) was amplified by polymerase chain reaction (PCR) from *B. adolescentis* ATCC 15703 genomic DNA using *PfuTurbo* DNA polymerase (Stratagene) and I-PIPE (Insert) primers that included sequences for the predicted 5' and 3' ends. The expression vector, pSpeedET, which encodes an amino-terminal tobacco etch virus (TEV) protease-cleavable expression and purification tag (MGSDKIHSHHHHHENLYFQ/G), was PCR amplified with V-PIPE (Vector) primers. V-PIPE and I-PIPE PCR products were mixed to anneal the amplified DNA fragments together. *E. coli* GeneHogs (Invitrogen) competent cells were transformed with the I-PIPE / V-PIPE mixture and dispensed on selective LB-agar plates. The cloning junctions were confirmed by DNA sequencing.

Expression was performed in a selenomethionine-containing medium at 37°C with suppression of normal methionine synthesis.¹³ At the end of fermentation, lysozyme was added to the culture to a final concentration of 250 µg/mL, and the cells were harvested and frozen. After one freeze/thaw cycle, the cells were homogenized in lysis buffer (50 mM HEPES pH 8.0, 50 mM NaCl, 10 mM imidazole, 1 mM Tris(2-carboxyethyl)phosphine-HCl (TCEP)) and the lysate was clarified by centrifugation at 32,500 × g for 30 minutes. The soluble fraction was passed over nickel-chelating resin (GE Healthcare) pre-equilibrated with lysis buffer, the resin washed with wash buffer (50 mM HEPES pH 8.0, 300 mM NaCl, 40 mM imidazole, 10% (v/v) glycerol, 1 mM TCEP), and the protein eluted with elution buffer (20 mM HEPES pH 8.0, 300 mM imidazole, 10% (v/v) glycerol, 1 mM TCEP). The eluate was buffer exchanged with TEV buffer (20 mM HEPES pH 8.0, 200 mM NaCl, 40 mM imidazole, 1 mM TCEP) using a PD-10 column (GE Healthcare), and incubated with 1 mg of TEV protease per 15 mg of eluted protein. The protease-treated eluate was passed over nickel-chelating resin (GE Healthcare) pre-equilibrated with HEPES crystallization buffer (20 mM HEPES pH 8.0, 200 mM NaCl, 40 mM imidazole, 1 mM TCEP) and the resin was washed with the same buffer. The flow-through and wash fractions were combined and concentrated for crystallization trials to 19 mg/mL by centrifugal ultrafiltration (Millipore). Protein concentrations were determined using the Coomassie Plus assay (Pierce).

The oligomeric state of the YP_910028.1 protein in solution was determined using a 1 × 30 cm² Superdex 200 size exclusion column (GE Healthcare)¹² coupled with miniDAWN (Wyatt Technology) static light scattering (SEC/SLS) and Optilab differential refractive index detectors (Wyatt Technology). The mobile phase consisted of 20 mM Tris pH 8.0, 150 mM NaCl, and 0.02% (w/v) sodium azide. The molecular weight was calculated using ASTRA 5.1.5 software (Wyatt Technology).

Proteins were crystallized using the nanodroplet vapor diffusion method¹⁴ with standard JCSG crystallization protocols.^{11,15} Apo-YP_910028.1 crystals were grown using sitting drops composed of 200 nL protein solution mixed with 200 nL of crystallization solution and equilibrated against a 50 µL reservoir at 277 K for 25 days prior to harvest. The crystallization reagent was composed of 0.02 M CaCl₂, 30 % (v/v) 2-methyl-2,4-pentanediol (MPD), 0.1 M sodium acetate pH 4.6. No further cryoprotectant was added to the crystal.

Initial screening for diffraction was carried out using the Stanford Automated Mounting system (SAM)¹⁶ at the Stanford Synchrotron Radiation Lightsource (SSRL, SLAC National Accelerator Laboratory, Menlo Park, CA). The diffraction data were indexed in hexagonal space group P6₅.

The YP_910028.1 protein solution was supplemented with 3 mM AMP, and AMP-YP_910028.1 crystals were grown by vapor diffusion against a reservoir solution consisting of 24 % (v/v) MPD, 0.02 M CaCl₂ and 0.1 M sodium acetate pH 4.71, with drops composed of 150 nL of protein and 150 nL of reservoir solution. Crystals grew at 277 K within 3 days. Crystals were cryo-protected in reservoir solution supplemented with 10 % (v/v) ethylene glycol prior to flash cooling in liquid nitrogen.

Data collection, structure solution, and refinement

SAD (Single-wavelength anomalous diffraction) data were collected at SSRL on beamline BL11-1 from a single Apo-YP_910028.1 crystal at a wavelength corresponding to the selenium K absorption peak (λ_1) of a SAD experiment. The dataset was collected at 100 K with a MARMOSAIC 325 CCD detector (Rayonix) using the *Blu-Ice* data collection environment.¹⁷ The data processing and structure solution were carried out using an automatic structure solution pipeline developed at the JCSG. The SAD data were integrated and reduced using *XDS* and scaled with the program *XSCALE*^{18–20}. Selenium sites were located with *SHELXD*²¹. Phase refinement and automatic model building were performed using *autoSHARP*²² and *wARP*.²³

Data for the AMP-YP_910028.1 structure were collected at the Advanced Light Source (ALS, Berkeley, CA) on beamline 5.0.3. The structure of AMP-YP_910028.1 was determined by molecular replacement (MR) as implemented by *PHASER*^{24,25}, using the Apo-YP_910028.1 coordinates (PDB ID: 3e0f) as the search model.

Model completion and refinement for both structures were performed with *MIFIT* (<http://code.google.com/p/mifit/>), *COOT*²⁶ and *REFMAC5*²⁷. *CCP4* programs were used for other calculations and data manipulation.²⁸ The final structure of Apo-YP_910028.1 is shown in Figure 1. Data reduction and refinement statistics are summarized in Table 1.²⁹

The presence of zinc and iron was confirmed by an X-ray fluorescence excitation scans of the Apo-YP_910028.1 crystal, where a zinc K α emission peak was found at approximately 8531 eV (256 counts), and an iron K α emission peak near 6400 eV (288 counts), in addition to the selenium K α emission peak near 11208 eV (1560 counts). In order to identify the metal types at the individual zinc and iron sites from anomalous difference Fourier maps, X-ray datasets were collected ~500 eV above and below K absorption edges of zinc and iron elements on SSRL beamline 9-2 (Supplemental Table S1). These data were integrated with *MOSFLM*³⁰ and scaled with *SCALA*²⁸. *MAPMAN*³¹ was used to analyze the anomalous difference Fourier maps. The large change in anomalous peaks above and below the zinc edge for the Zn-301 (3.6/0.7) site and above and below the iron edge for the Fe-302 (4.6/1.9) and Fe-303 (7.5/1.6) metal sites clearly identify the metals (Supplemental Tables S2–S3).

Structure validation and deposition

The quality of the crystal structures was analyzed using the JCSG Quality Control server (<http://smb.slac.stanford.edu/jcsg/QC>). This server analyses the data using a variety of validation tools including *AutoDepInputTool*³², *MolProbity*³³, *WHATIF* 5.0.³⁴, *RESOLVE*³⁵ and *MOLEMAN2*³⁶. Protein quaternary structure analysis was performed using the *PISA* server.³⁷ Figure 1B was adapted from an analysis using *PDBsum*³⁸ and all others were prepared with *PyMOL* (The *PyMOL* Molecular Graphics System, Version 1.3, Schrödinger, LLC.). Atomic coordinates and experimental structure factors for

YP_910028.1 at 2.40 Å (Apo) and 1.94 Å (AMP bound) resolution have been deposited in the PDB and are accessible under PDB IDs: 3e0f and 3o0f, respectively.

Thermofluor assay

Thermofluor binding assay mixtures consisted of 12 μM purified YP_910028.1 supplemented with 200 μM ligand and 17× Sypro Orange (Invitrogen) in 150 mM NaCl, 20 mM HEPES pH 7.9. The temperature was ramped on an iCycler iQ5 real-time PCR system (Bio-Rad) from 20°C to 100°C at a rate of 1°C/min. Melting temperatures were determined by identifying the maximum derivative (d(RFU)/dt) of the curve.

Phosphoesterase assay

Phosphoesterase activity was determined by measuring the conversion of para-nitrophenyl phosphate (pNPP) to para-nitrophenol. Under alkaline conditions, this reaction yields an intense yellow-colored product, which can be quantified spectrophotometrically at 405 nm. Phosphoesterase activity was assessed using two independent preparations of YP_910028.1 in triplicate reactions (125 μl final volume) that were initiated by addition of 0–3 μM YP_910028.1 to a solution containing pNPP in 50 mM Tris-HCl pH 7.5, 100 mM NaCl, 10 mM MgCl₂, 1 mM TCEP buffer and incubating for 30 minutes at room temperature (23°C). The reactions were stopped with addition of NaOH to a final concentration of 0.6 M and the activity was determined from spectrophotometric absorbance at 405 nm and expressed as nmol P_i released per minute. Assays were performed in parallel (background subtracted) with negative controls.

DNA polymerase and DNA degradation assays

DNA polymerase activity was assayed by primer extension of matched primer:template DNA, as described.³⁹ Three different protein concentrations (40, 200, and 2000 nM) were assayed in reaction buffer containing 30 mM HEPES, pH 7.5, 4% (v/v) glycerol, 100 μg/mL BSA, 1 mM DTT, 1 mM ATP, 50 μM dNTP, 20 nM single-stranded DNA binding-protein (SSB). The salt and divalent cation concentration was varied as follows: 40 mM KCl, 3.0 mM MgSO₄ (low salt) at room temperature; and 80 mM KCl, 7.5 mM MgSO₄ (high salt) at room temperature or at 37°C. Time points were taken at 0, 15, 30, 50, and 80 or 82 minutes. Klenow fragment of DNA pol I (1 U, New England Biolabs) was included as a positive control under the following conditions: 40 mM KCl with 3.0 mM MgSO₄ or 80 mM KCl with 7.5 mM MgSO₄ at room temperature with 0, 1, and 5 minutes time points. Reactions were quenched with 50 mM EDTA in 82% (v/v) formamide containing bromophenol blue and xylene cyanol, analyzed on a 10% polyacrylamide gel, and imaged on a Storm phosphorimager (GE Healthcare).

Assays to determine hydrolytic proofreading activity on matched versus mismatched DNA were carried out at 37°C in a reaction buffer containing 30 mM HEPES, pH 7.5, 20 mM NaCl, 10 mM MgSO₄, 4% glycerol, 100 μg/mL BSA, 1 mM DTT, 1 mM ATP, with or without 100 mM dNTPs, with 200 nM or 2.0 μM YP_910028.1. Reactions were also carried out as described, but without ATP, and with and without 1 mM ZnSO₄. Time points were taken at 0, 15, 30, and 60 minutes. Reactions were quenched with 50 mM EDTA in 82% (v/v) formamide containing bromophenol blue and xylene cyanol, analyzed on a 12% polyacrylamide gel, and imaged on a Storm phosphorimager (GE Healthcare). The following DNA molecules were used for the template and primer in both assays: template DNA 5'-ggttactcagatcaggcctgcgaagacctgggcgtccgctgcagctg tactatcatatgc-3', primer DNA 5'-gcatatgatagtacagctgcagccggacgccx-3', where x = dC (matched) or x = dA, dG, or dT (mismatched).

RESULTS

Overall structure

The crystal structure of Apo-YP_910028.1 from *B. adolescentis* (strain ATCC 15703/DSM 20083) was determined in space group P6₅ at 2.4 Å resolution using the SAD method (PDB ID: 3e0f). Data collection, model, and refinement statistics are summarized in Table I. The final model of Apo-YP_910028.1 includes one protein molecule (residues 8 – 293), one Zn, two Fe, two tetrahedrally-shaped ligands, which were modeled as phosphate ions (see below), one acetate and 103 water molecules in the asymmetric unit. Residues 0 (the glycine residue left after TEV cleavage of the expression and purification tag), 1–7 and 294–300 were not visible in the electron density. No electron density was observed for the side chains of E64, E118, E159, Q207, R208, R210, R250, K264 and R293. The presence of zinc and iron was confirmed by X-ray fluorescence excitation and wavelength scans. The metal identity at each of the zinc and iron sites was confirmed based on anomalous difference Fourier map comparisons with data collected above and below the iron and the zinc K absorption edges. The electron density and coordination geometry at the phosphate sites are consistent with either phosphate or sulfate. Either ion could have co-purified with the protein as neither phosphate or sulfate was specifically added during purification or crystallization. The assignment of these ions as phosphates in the model is based on the role of phosphate in other proteins in the PHP family with characterized functions and from subsequent functional analysis described herein (see below). The Matthews' coefficient (V_m)⁴⁰ for YP_910028.1 is 2.52 Å³/Da, and the estimated solvent content is 51.2%. The Ramachandran plot produced by MolProbity⁴¹ shows that 97.2 % of the residues are in favored regions with no outliers.

The crystal structure of AMP-YP_910028.1 was also determined in space group P6₅ to a resolution of 1.94 Å using MR methods (PDB ID: 3o0f, Table I). The crystal of AMP-YP_910028.1 was isomorphous with the APO-YP910028.1.crystal. The final model of AMP-YP_910028.1 includes one protein molecule (residues 7 – 293), one Zn, two Fe, one tetrahedrally-shaped ligand that is modeled as phosphate, two ethylene glycol molecules, one AMP and 183 water molecules in the asymmetric unit. No electron density was observed for the side chains of residues E8, E64, E118, E139, E159, Q207, R208, R210, R250, K264 and R293. The Ramachandran plot produced by MolProbity⁴¹, shows that 97.9 % of the residues are in favored regions with no outliers.

YP_910028.1 is composed of seven β-strands (β1–β7), ten α-helices (H1–4, H6–10 and H12), and two ₃₁₀-helices (H5 and H11) (Figs. 1A and 1B). The total β-sheet, α-helical, and ₃₁₀-helical content is 15.7%, 38.5%, and 6%, respectively. Analytical size exclusion chromatography (data not shown) and analysis of the crystallographic packing using the PISA server³⁷ suggests that the monomer, as observed in the crystal structure, is the relevant biological state in solution.

The monomer of YP_910028.1 is composed of two domains (highlighted in blue and green in Fig. 1A). The larger domain consists of a 7-stranded α/β barrel and is a member of the PHP family (Pfam PF02811) and comprises residues 8–104 and 179–293 (green in Fig. 1A). The second smaller domain corresponds to an insertion in the PHP domain and consists of residues 105–178 (blue in Fig. 1B). The insert domain contains mainly α helices with orthogonal bundle architecture and adopts a cyclin-A topology by CATH classification.⁴² The predicted active site, described below, is in a deep cleft between the PHP domain and the insert domain and is occupied by the Zn, Fe and phosphate ions. The AMP binding site sits on top of the active site cleft and is composed largely of residues from the insert domain.

Functional inferences from the structure

Upon determination of the YP-910028.1 apo structure, several structural features were observed that provided some information about potential biochemical functions; namely the presence of a PHP domain and a trinuclear metal binding site. The PHP domain seen in YP-910028.1 has also been observed in DNA polymerase III, X family polymerases, histidinol phosphate phosphatase, and numerous uncharacterized proteins⁴³. Therefore, the presence of a PHP domain in YP-910028.1 cannot be used as a direct indicator of specific biochemical function.

The two iron and one zinc observed in the YP_910028.1 structure suggest the general vicinity of the likely active site, but again do not point to any one specific function. It is further noted that it is possible that the bound metal ions found *in vivo* in *B. adolescentis* could be different from those in the protein that we expressed in *E. coli*. Trinuclear metal binding sites are observed in a variety of proteins, including *E. coli* endonuclease IV⁴⁴, some phosphatases^{45,46}, *B. cereus* phospholipase C⁴⁷, terpenoid synthases⁴⁸, and *trans*-dienelactone hydrolase from *P. reineke*⁴⁹. Therefore, the presence of a trinuclear metal binding site in YP-910028.1 is also not a direct indicator of biochemical function. Furthermore, the set of six residues that coordinate the two iron ions in YP_910028.1: E74, H85, and H202 for Fe-302 and H17, H19, E74, and D260 for Fe-303, according to a PDBSiteScan⁵⁰ analysis, are best matched to the corresponding Fe-coordinating residues of N-acetylglucosamine-6-phosphate deacetylase from *B. subtilis*⁵¹: E136, H202, and H223 for Fe-1398 and H63, H65, E136, and D281 for Fe-1397.

Thus the observation of such structural features at the apparent active site of YP-910028.1 points to a variety of potential biochemical function(s). Thus, more sophisticated analysis of these local structural features was required, as implemented in this study by THEMATICS.

Sequence analysis

WU-BLAST2⁵² searches against the YP_9100928.1 sequence retrieved sequences of PHP domain containing proteins, all annotated as uncharacterized proteins or putative phosphoesterases. The Joined Assembly of Function Annotations (JAFa)⁵³ metasever was also utilized to investigate other potential function predictions and seek a consensus prediction. JAFa queries using the YP_9100928.1 sequence resulted in only one functional prediction originating from InterProScan⁵⁴, which predicted DNA polymerase activity. The Berkeley Phylogenomics Group server (BPG)⁵⁵ also yielded a prediction of DNA replication activity, albeit inferred from automated annotation. Phylogenetic analysis of YP_9100928.1 using a neighbor joining method placed it among putative DNA polymerases, although one close neighbor in the same subfamily was listed as a putative metal-dependent phosphoesterase. Very similar results were obtained using maximum likelihood and maximum parsimony methods. Therefore, the sequence analysis results all suggest either phosphoesterase or DNA polymerase activity, but it is important to note that the closest aligned sequences and neighbors in the phylogenetic tree only had putative functional annotations.

Structural similarity search

The closest structural similarity identified with the Secondary Structure Matching (SSM)⁵⁶ server was to another Structural Genomics protein, TM0559, a putative metal-dependent phosphoesterase from *Thermotoga maritima* also determined by the JCSG (PDB ID: 2anu). However, this similarity was relatively weak with a Q score of 0.23 and a Z score of 6.7. The next closest structure, 3-keto-L-gulonate 6-phosphate decarboxylase (PDB ID: 1q6l), was more remote, with a Q score of 0.13 and a Z score of only 3.1.

The closest structural alignment, as implemented by the Dali⁵⁷ server, was to the DNA polymerase III alpha subunit from *Thermus aquaticus*⁵⁸ (DNA pol, PDB ID: 2hpi; Z score = 17). The next-highest Z scores were (in decreasing order) TM0559 from *Thermotoga maritima* (PDB ID: 2anu, Z score = 15.5), YcdX protein from *E. coli*⁵⁹ (PDB ID: 1m68, Z score = 12), histidinol phosphate phosphatase from *Thermus thermophilus*⁴⁶ (HPP; PDB ID: 2yz5, Z score = 10.6), ribonuclease P protein component 3 from *Pyrococcus horikoshii*⁶⁰ (RNase P; PDB ID: 2czv, Z score = 9.2); isoaspartyl dipeptidase from *E. coli*⁶¹ (IADP, PDB ID: 1po9, Z score = 8.8), and N-acetylglucosamine-6-phosphate deacetylase from *E. coli*⁶² (NAPD, PDB ID: 2p50, Z score = 8.5). These structural alignments are summarized in Table II with their respective Z scores, root mean squared deviation (RMSD) values, and sequence identity scores. Superpositions of these structures are shown in Figure 2.

Interestingly, all of the structures shown in Table II and Figure 2, except for the protein component of RNase P (PDB ID: 2czv; *n.b.* in bacteria, RNase P has two components, a catalytic RNA, and a polypeptide chain), contain one or more divalent metal ions in the putative active site. For example, DNA pol (PDB ID: 2hpi), IADP (PDB ID: 1po9), NAPD (PDB ID: 2p50) and YcdX (PDB ID: 1m68), all contain Zn ions in the active site. Others contain Zn plus Cl anions (TM0559, PDB ID: 2anu), or Fe and PO₄ (HPP, PDB ID: 2yz5) in the active site. Note that the Zn-501, Fe-502, Fe-503 and PO4-2001 ions of the HPP structure superimpose well with the corresponding Zn-301, Fe-302, Fe-303 and PO4-304 ions of YP_910028.1 to within 0.75 Å. In comparison, the overall structural superposition of YP_910028.1 and HPP is relatively poor with a 3.8 Å RMSD for 166 aligned Cα atoms.

Analysis of the active site

Coordination of the metal ions in the predicted active site is depicted in Figures 3A and 3B). A Zn ion numbered as atom 301 (Zn-301) is coordinated by the side chains of D24, H49, H262, and a phosphate ion. An Fe ion, Fe-302, is tetrahedrally coordinated by E74, H85, H202 and a phosphate ion. A second iron, Fe-303, is observed to have square pyramidal coordination to H17, H19, E74, D260, and a phosphate ion, where H19 occupies the axial position. It is possible that a sixth ligand, such as a water molecule, occupies the remaining axial position to form an octahedral configuration, but no such axial ligand could be resolved in the structure. This pattern of histidine, aspartate and glutamate residues is particularly well conserved among homologues of YP_910028.1, suggesting that they are essential to the function of this protein family. Analysis of the surface clefts in the YP_910028.1 structure, using the CastP server⁶³, revealed a deep cavity with a volume of 1717 Å³ at the interface between the PHP domain (residues 8–104 and 179–293) and the insert domain (residues 105–178) (cavity highlighted by magenta surface in Fig. 3B). The putative active site is located in the lower part of this cleft and is occupied by the Zn, Fe and PO₄ ions.

Nucleotide binding site

Within the cleft observed in the YP_910028.1 apo structure (near PO4-305), our THEMATICS computational analysis (see below) predicted several functionally important residues, including R109, R112, R145, P146, F165, and Y176. These residues are located close to the metal binding center (NH2 of R112 forms the closest interaction, a 3.09 Å hydrogen bond with O4 of PO4-305), suggesting these residues potentially form a binding site for a co-factor or larger substrate molecule containing a phosphate moiety. It was, therefore, decided to screen a library of compounds to test for possible binding of a co-factor or substrate. Thermofluor screening revealed stabilization of the melting temperature (T_m) of YP_910028.1 by all of the adenine containing nucleotides and dinucleotides in the library (Fig. 3C). The ligands that increased the T_m of YP910028.1 (57°C) were adenosine 5'-triphosphate (ATP, ΔT_m = 8.0°C), adenosine 5'-diphosphate (ADP, ΔT_m = 11.0°C),

adenosine 5'-monophosphate (AMP, $\Delta T_m = 5.5^\circ\text{C}$) and flavin adenine dinucleotide (FAD, $\Delta T_m = 3.0^\circ\text{C}$). These data then inspired subsequent co-crystallization experiments in search of a YP_910028.1 structure with bound nucleotide or dinucleotide.

The crystal structure of AMP-bound YP_910028.1 (AMP-YP_910028.1) was determined to 1.94 Å resolution (Fig. 3D). Structural superposition of Apo-YP_910028.1 and AMP-YP_910028.1 show that the two structures are very similar (RMSD = 0.3 Å, 286 C α atoms aligned). The structural changes are restricted to an outward movement in the position of the R208 C α at the entrance of the ligand binding pocket and slight variation in the rotamers of R109, R145 and F165 that contact the AMP.

AMP contact residues include R109, R112, R145, P146, R161, S162, F165, G265 and K266 (Fig. 3D). R109, R112 and R145 coordinate the phosphate group of AMP via a series of hydrogen bonds. Hydrophobic portions of F165 and K266 stack against the adenine ring of the AMP and R161, S162 and the backbone carbonyl of G265 hydrogen bond with the nitrogen atoms of the adenine moiety of AMP. Six of these contact residues (all except R161, S162, and G265) were also predicted computationally by the POOL method (see below).

The nucleotide binding site occupies the top half of the active site cleft, and all AMP interacting residues, with the exception of G265 and K266, come from the insert domain. In contrast, all of the *THEMATICS*-predicted catalytic residues (Fig. 3B and Table III) occupy the bottom half of the active site cleft near the location of Zn-301, Fe-302 and Fe-303 and PO4-304 ions and all of these residues belong to the PHP domain.

THEMATICS analysis

THEMATICS^{6,7} calculations were performed on the structure of Apo-YP_910028.1 and on the top structural “matches” (listed in Table II) in order to locate the catalytic site of YP_910028.1 and attempt to identify a potential biochemical function. Residues were selected according to the method of Wei *et al.*⁸ *THEMATICS*-predicted reactive residues for YP_910028.1 and its closest structural matches are shown in Table III. The location of these residues on the YP_910028.1 structure is shown in Figure 3B.

Note that in Table III, RNase P (PDB ID: 2czv) does not have any similarity to the query protein in the region corresponding to the predicted active site of YP_910028.1, while NAPD (PDB ID: 2p50) only has two similar residues and IADP (PDB ID: 1po9) is similar in four aligned positions. DNA pol (PDB ID: 2hpi) gave a much better match, with similar residues in six aligned positions. HPP (PDB ID: 2yz5) gave the best match of all proteins in the set, with very close agreement of the *THEMATICS*-predicted residues. If aspartate and glutamate residues are regarded as chemically equivalent, then the *THEMATICS*-predictions for YP_910028.1 and HPP correspond exactly (E230 in YP_910028.1 aligns with D189 in HPP). The three weakest *THEMATICS*-matches, RNaseP, NAPD, and IADP, were eliminated as possible functional leads, since the composition of their predicted active sites differs significantly from that of YP_910028.1.

The YP_910028.1, DNA Pol, and HPP active sites were further analyzed and superposition of the predicted active site residues of DNA pol (PDB ID: 2hpi) and YP_910028.1 (PDB ID: 3e0f) is shown in Figure 4A, and of HPP (PDB ID: 2yz5) and YP_910028.1 in Figure 4B. Two residues in the predicted active site of YP_910028.1 are not conserved in DNA polymerase (Y74 of DNA pol corresponds to T76 of YP_910028.1 and C145 of DNA pol corresponds to H202 of YP_910028.1; non-conserved residues are identified by dashed circles in Fig. 4A). Superposition of the predicted active site of HPP and YP_910028.1 results in the best match (D189 of HPP corresponds to chemically equivalent E230 of

YP_910028.1, identified by dashed circle in Fig. 4B). While DNA pol, HPP, and YP_910028.1 have very similar metal binding sites and, thus, have structurally equivalent residues in common in their active sites, the difference of two residues between DNA pol and YP_910028.1 suggest a difference in function. On this basis, it was predicted that YP_910028.1 is a phosphoesterase and not a polymerase. To verify these predictions, YP_910028.1 was tested experimentally for phosphoesterase activity and for DNA polymerase and DNA proofreading activity (see below).

In an effort to gain more information about the observed crystallographic features in the region around the active site of the apo structure, a technique that expands the capabilities of THEMATICs was explored. Partial Order Optimum Likelihood (POOL)⁶⁴ is a machine learning methodology that predicts functionally important residues using THEMATICs and other input data. POOL uses THEMATICs data to predict all 20 amino-acid types. Furthermore, POOL improves the accuracy of catalytic residue predictions and can also accept additional types of input, including geometric, sequence conservation, or phylogenetic information. POOL generates a list of all residues, rank-ordered according to their functional importance. The input⁶⁵ for the POOL calculations on YP_910028.1 consisted of the coordinates for the apo structure and the output data from THEMATICs, INTREPID⁶⁶, and structure-only ConCavity⁶⁷. Residues in the top 8% of the POOL rank-ordered list are designated as potentially functionally important residues. Table IV shows the POOL-predicted residues calculated for the YP_910028 apo structure. These 23 residues are listed in rank order and their contacts with metal ions, AMP, or catalytic residues are tabulated. The nine highest-ranking residues include all nine metal coordinating residues in the active site region. The 14 next-highest ranking residues for the apo structure include eight residues, T76, R109, R122, R145, P146, F165, Y176 and K266, that the AMP-bound structure later showed to be in direct contact with AMP. Four additional residues, S83, D205, R208 and H133, are in direct contact with one or more metal-coordinating residues; these four residues are located in the second sphere around the active site. Thus, the POOL predictions on the YP_910028.1 apo structure identified both the (already known) metal binding residues and also the AMP binding residues (later confirmed by the AMP-bound structure).

Phosphoesterase assay and inhibitor binding studies

Phosphoesterase activity assays (Fig. 5A) show that YP_910028.1 hydrolyzes the phosphate group from *p*-nitrophenyl phosphate (pNPP) in a concentration dependent manner with activities ≥ 200 -fold over the three negative controls (non-phosphatase proteins purified from the same bacterial host line and expression vector using identical protocols). YP_910028.1 is an effective phosphoesterase, although not as active as Calf-intestinal phosphatase (CIP) which requires an order of magnitude less enzyme to hydrolyze pNPP at the same rate.

Both of the negative control proteins (YP_508449.1 and YP_754858) showed meager phosphoester hydrolysis (Fig. 5). Both sodium pyrophosphate and sodium molybdate (serine-threonine phosphatase and acid/phosphoprotein phosphatase inhibitors, respectively) showed clear thermal shifts in Thermofluor assays of YP_910028.1, stabilizing the T_m by 3°C (data not shown). Other generic phosphatase inhibitors were also tested, including imidazole (alkaline phosphatase inhibitor) and sodium tartrate (acid phosphatase inhibitor), and both had little or no effect on the T_m of YP_910028.1 (data not shown). The metal-chelator EDTA destabilized the T_m of YP_910028.1 by 4°C (data not shown).

DNA polymerase and degradation assays

YP_910028.1 was tested for DNA polymerase activity using a range of different salt and divalent metal concentrations and different temperatures. We did not detect DNA

polymerase activity in the YP_910028.1 preparation under any conditions tested (Fig. 6A). The Klenow fragment of DNA pol I was included as a positive control and was shown to be active.

In order to test whether YP_910028.1 possesses DNA proofreading activity, and can hydrolyze mismatched DNA ends, we incubated YP_910028.1 with DNA containing matched and mismatched ends under a range of conditions. We did not detect hydrolysis of any of the DNA constructs tested (Fig. 6B).

DISCUSSION

The crystal structure of YP_910028.1 revealed a deep cleft at the interface between the PHP and the insert domain, which contains Zn, Fe and PO₄ ions. In the putative active site, as predicted by *THEMATICS*, the Zn and Fe ions are coordinated by a cluster of histidine, aspartate and glutamate residues (H17, H19, D24, H49, E74, H85, H202, D260 and H262). These residues are conserved among homologues of YP_910028.1, suggesting that they are essential to the function of YP_910028.1. All of these predicted residues lie within the PHP domain suggesting that the active site shares features common to other PHP containing proteins.

The *THEMATICS*-predicted active site for YP_910028.1 matched best with that of HPP, suggesting phosphoesterase activity, although other bioinformatics analyses suggested otherwise. The present results demonstrate that common active site features may not be obvious from simple global structural or sequence alignments, but more detailed analysis of the active site region is required. Specifically, the predicted set of active site residues of YP_910028.1 shows very good spatial overlap with the active site residues of HPP, although the best overall 3D structural match is with DNA Pol III. Through the use of these local structural alignments at the predicted active site, we were able to predict correctly the biochemical function of YP_910028.1.

The AMP-bound YP_910028.1 structure provides strong evidence that the predicted active site is extended into the insert domain, which possesses the ability to bind AMP (and potentially larger nucleotides or dinucleotides). Examples of phosphoesterases that bind AMP have been reported previously, including AKAP18,⁶⁸ and it has been proposed that these regulatory proteins may act as AMP sensors, thereby coupling intracellular AMP levels with various phosphorylation and dephosphorylation processes. It is also possible that the bound AMP that we observe is a cognate or surrogate ligand and further biochemical studies are needed to identify the specific substrates, products and function of YP_910028.1.

While both phylogenetic analysis and structural comparison suggested DNA polymerase and replication activity, the experimental evidence described here does not support such a functional annotation. A number of sequence homologues of YP_910028.1 are annotated as having DNA replication activity. Based on the *THEMATICS* predictions and local structural comparison of the *THEMATICS*-predicted active site to known phosphoesterases, our YP_910028.1 structure (PDB ID: 3e0f) was, therefore, annotated as a putative, metal-dependent phosphoesterase whose activity was later confirmed experimentally. Moreover, after active site comparison together with *THEMATICS* predictions of YP_910028.1 homologues, phosphoesterase activity is now predicted for two additional, uncharacterized proteins, TM0559 (PDB ID: 2anu) from *Thermotoga maritima* and YcdX (PDB ID: 1m68) from *E. coli*.

It is well known that gut bacteria fulfill a host of beneficial functions in the human body, and given the plethora of uncharacterized proteins in the human gut microbiome, it is clear that a combination of structural, bioinformatics and functional approaches will be needed to

address these unknown functions. For example, the *B. adolescentis* genome has been fully sequenced (Genbank: AP009256) and 31 % of the proteins (511 of 1632 encoded proteins) are annotated as “hypothetical proteins” with no known function. Our results address one such protein of unknown function, and illustrate the power of using the *THEMATICS* predictions of putative active sites, coupled to structure determination of proteins of ‘unknown’ function, from which local structural comparisons can be used to predict biochemical function.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Genomic DNA from *B. adolescentis* ATCC 15703 was obtained from the American Type Culture Collection (ATCC). This work was supported by NIH Protein Structure Initiative grants from the National Institute of General Medical Sciences (www.nigms.nih.gov); Grant Numbers: U54 GM094586 and U54 GM074898. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the NIH. Portions of this research were carried out at the Stanford Synchrotron Radiation Lightsource (SSRL), SLAC National Accelerator Laboratory, and the Advanced Light Source (ALS). The SSRL is a Directorate of SLAC National Accelerator Laboratory and an Office of Science User Facility operated for the U.S. Department of Energy Office of Science by Stanford University. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research, and by the National Institutes of Health (National Center for Research Resources, Biomedical Technology Program (P41RR001209), and the National Institute of General Medical Sciences). The ALS is supported by the Director, Office of Science, Office of Basic Energy Sciences, Materials Sciences Division of the US Department of Energy under Contract No. DE-AC03-76SF00098 at Lawrence Berkeley National Laboratory. The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health, National Institute of General Medical Sciences. Portions of this work were also supported by the National Science Foundation under grants MCB-0517292 and MCB-0843603 (awarded to MJO) and NSF Career Award MCB-0845033 (awarded to PJB), and New Faculty Award from the Camille and Henry Dreyfus Foundation (awarded to PJB). PJB is a Cottrell Scholar of the Research Corporation for Science Advancement. JK acknowledges the PA State System of Higher Education for a Faculty Professional Development Council award.

REFERENCES

1. Yuhara T, Isojima S, Tsuchiya F, Mitsuoka T. On the intestinal flora of bottle-fed infants. *Bifidobacteria Microflora*. 1983; 2:33–39.
2. Arunachalam, KD., editor. The role of bifidobacteria in nutrition, medicine, and technology. Vol. Volume 19. USA: Elsevier Science Inc.; 1999. p. 1559-1597.
3. Homma N. Bifidobacteria as a resistance factor in human beings. *Bifidobacteria Microflora*. 1988; 7:35–43.
4. Reuter G. Bifidobacteria culture as components of yogurt-like products. *Bifidobacteria Microflora*. 1990; 9:107–118.
5. Huis in't Veld HJ, Haenaar R. Probiotics and Health in man and animal. *J Chem Technol Biotechnol*. 1991; 51:562–577.
6. Ondrechen MJ, Clifton JG, Ringe D. *THEMATICS*: A simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci USA*. 2001; 98:12473–12478. [PubMed: 11606719]
7. Ko J, Murga LF, Andre P, Yang H, Ondrechen MJ, Williams RJ, Agunwamba A, Budil DE. Statistical criteria for the identification of protein active sites using Theoretical Microscopic Titration Curves. *Proteins*. 2005; 59:183–195. [PubMed: 15739204]
8. Wei Y, Ko J, Murga LF, Ondrechen MJ. Selective prediction of interaction sites in protein structures with *THEMATICS*. *BMC Bioinformatics*. 2007; 8:119–133. [PubMed: 17419878]
9. Wei Y, Ringe D, Wilson MA, Ondrechen MJ. Identification of functional subclasses in the DJ-1 superfamily proteins. *PLoS Comput Biol*. 2007; 3(e10):120–126.

10. Parasuram R, Lee JS, Yin P, Somarowthu S, Ondrechen MJ. Functional classification of protein 3D structures from predicted local interaction sites. *J Bioinform Comput Biol.* 2010; 8 Suppl 1:1–15. [PubMed: 21155016]
11. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Spraggon G, Klock HE, McMullan D, Shin T, Vincent J, Robb A, Brinen LS, Miller MD, McPhillips TM, Miller MA, Scheibe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger MA, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC. Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proc Natl Acad Sci USA.* 2002; 99:11664–11669. [PubMed: 12193646]
12. Klock HE, Koesema EJ, Knuth MW, Lesley SA. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins.* 2008; 71:982–994. [PubMed: 18004753]
13. Van Duyne GD, Standaert RF, Karplus PA, Schreiber SL, Clardy J. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J Mol Biol.* 1993; 229:105–124. [PubMed: 7678431]
14. Santarsiero BD, Yegian DT, Lee CC, Spraggon G, Gu J, Scheibe D, Uber DC, Cornell EW, Nordmeyer RA, Kolbe WF, Jin J, Jones AL, Jaklevic JM, Schultz PG, Stevens RC. An approach to rapid protein crystallization using nanodroplets. *J Appl Cryst.* 2002; Volume 35:278–281.
15. Elsliger MA, Deacon AM, Godzik A, Lesley SA, Wooley J, Wuthrich K, Wilson IA. The JCSG high-throughput structural biology pipeline. *Acta Crystallogr Sect F Struct Biol Cryst Commun.* 2010; 66(Pt 10):1137–1142.
16. Cohen AE, Ellis PJ, Miller MD, Deacon AM, Phizackerley RP. An automated system to mount cryo-cooled protein crystals on a synchrotron beamline, using compact sample cassettes and a small-scale robot. *J Appl Cryst.* 2002; 35:720–726.
17. McPhillips TM, McPhillips SE, Chiu HJ, Cohen AE, Deacon AM, Ellis PJ, Garman E, Gonzalez A, Sauter NK, Phizackerley RP, Soltis SM, Kuhn P. Blu-Ice and the distributed control system: Software for data acquisition and instrument control at macromolecular crystallography beamlines. *J Synchrotron Rad.* 2002; 9:401–406.
18. Kabsch W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J Appl Cryst.* 1993; 26:795–800.
19. Kabsch W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr.* 2010; 66:133–144. [PubMed: 20124693]
20. Kabsch W. XDS. *Acta Crystallogr D Biol Crystallogr.* 2010; 66:125–132. [PubMed: 20124692]
21. Schneider TR, Sheldrick GM. Substructure solution with SHELXD. *Acta Crystallogr D Biol Crystallogr.* 2002; 58(Pt 10 Pt 2):1772–1779. [PubMed: 12351820]
22. Vonrhein C, Blanc E, Roversi P, Bricogne G. Automated structure solution with autoSHARP. *Methods Mol Biol.* 2007; 364:215–230. [PubMed: 17172768]
23. Cohen SX, Morris RJ, Fernandez FJ, Ben Jelloul M, Kakaris M, Parthasarathy V, Lamzin VS, Kleywegt GJ, Perrakis A. Towards complete validated models in the next generation of ARP/wARP. *Acta Crystallogr D Biol Crystallogr.* 2004; 60(Pt 12 Pt 1):2222–2229. [PubMed: 15572775]
24. McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ. Likelihood-enhanced fast translation functions. *Acta Crystallogr D Biol Crystallogr.* 2005; 61:458–464. [PubMed: 15805601]
25. Storoni LC, McCoy AJ, Read RJ. Likelihood-enhanced fast rotation functions. *Acta Crystallogr D Biol Crystallogr.* 2004; 60:432–438. [PubMed: 14993666]
26. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr.* 2004; D60:2126–2132. [PubMed: 15572765]
27. Winn MD, Murshudov GN, Papiz MZ. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol.* 2003; 374:300–321. [PubMed: 14696379]
28. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr.* 1994; 50:760–763. [PubMed: 15299374]
29. Cruickshank DW. Remarks about protein structure precision. *Acta Crystallogr D Biol Crystallogr.* 1999; 55:583–601. [PubMed: 10089455]

30. Leslie AGW. Recent changes to the MOSFLM package for processing film and image plate data. Joint CCP4+ESF-EAMCB Newsletter on Protein Crystallography. 1992; (No. 26)
31. Kleywegt GJ, Jones TA. xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. Acta Crystallogr D Biol Crystallogr. 1996; 52(Pt 4):826–828. [PubMed: 15299647]
32. Yang H, Guranovic V, Dutta S, Feng Z, Berman HM, Westbrook JD. Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 2004; 60(Pt 10):1833–1839. [PubMed: 15388930]
33. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003; 50(3):437–450. [PubMed: 12557186]
34. Vriend G. WHAT IF: A molecular modeling and drug design program. J Mol Graph. 1990; 8:52–56. [PubMed: 2268628]
35. Terwilliger TC. SOLVE and RESOLVE: automated structure solution and density modification. Methods Enzymol. 2003; 374:22–37. [PubMed: 14696367]
36. Kleywegt GJ. Validation of protein models from Calpha coordinates alone. J Mol Biol. 1997; 273:371–376. [PubMed: 9344745]
37. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007; 372:774–797. [PubMed: 17681537]
38. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Res. 2005; 33:D266–D268. [PubMed: 15608193]
39. Beuning PJ, Simon SM, Godoy VG, Jarosz DF, Walker GC. Characterization of Escherichia coli translesion synthesis polymerases and their accessory factors. Methods Enzymol. 2006; 408:318–340. [PubMed: 16793378]
40. Matthews BW. Solvent content of protein crystals. J Mol Biol. 1968; 33(2):491–497. [PubMed: 5700707]
41. Davis IW, Murray LW, Richardson JS, Richardson DC. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res. 2004; 32:W615–W619. [PubMed: 15215462]
42. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure. 1997; 5:1093–1108. [PubMed: 9309224]
43. Aravind L, Koonin EV. Phosphoesterase domains associated with DNA polymerases of diverse origins. Nucleic Acids Res. 1998; 26:3746–3752. [PubMed: 9685491]
44. Hosfield D, Guan Y, Haas B, Cunningham R, Tainer J. Structure of the DNA repair enzyme endonuclease IV and its DNA complex: double-nucleotide flipping at abasic sites and three-metal-ion catalysis. Cell. 1999; 98:397–408. [PubMed: 10458614]
45. Kim EE, Wyckoff W. Reaction mechanism of alkaline phosphatase based on crystal structures. Two-metal ion catalysis. J Mol Biol. 1991; 218:449–464. [PubMed: 2010919]
46. Omi R, Goto M, Miyahara I, Manzoku M, Ebihara A, Hirotsu K. Crystal structure of monofunctional histidinol phosphate phosphatase from Thermus thermophilus HB8. Biochemistry. 2007; 46:12618–12627. [PubMed: 17929834]
47. Hough E, Hansen LK, Birknes B, Jynge K, Hansen S, Hordvik A, Little C, Dodson E, Derewenda Z. High-resolution (1.5 Å) crystal structure of phospholipase C from Bacillus cereus. Nature. 1989; 338:357–360. [PubMed: 2493587]
48. Aaron JA, Christianson DW. Trinuclear metal clusters in catalysis by terpenoid synthases. Pure Appl Chem. 2010; 82:1585–1597. [PubMed: 21562622]
49. Marin M, Pieper DH. Novel metal-binding site of Pseudomonas reinekei MT1 trans-dienelactone hydrolase. Biochem Biophys Res Commun. 2009; 390:1345–1348. [PubMed: 19895788]
50. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. Nucleic Acids Res. 2004; 32:W549–W554. [PubMed: 15215447]

51. Vincent F, Yates D, Garman E, Davies G, Brannigan J. The three-dimensional structure of the N-acetylglucosamine-6-phosphate deacetylase, NagA, from *Bacillus subtilis*: a member of the urease superfamily. *J Biol Chem*. 2004; 279:2809–2816. [PubMed: 14557261]
52. Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W. WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res*. 2003; 31:3795–3798. [PubMed: 12824421]
53. Friedberg I, Harder T, Godzik A. JAFa: a protein function annotation meta-server. *Nucleic Acids Res*. 2006; 34:W379–W381. [PubMed: 16845030]
54. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001; 17:847–848. [PubMed: 11590104]
55. Glanville JG, Kirshner D, Krishnamurthy N, Sjolander K. Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis. *Nucleic Acids Res*. 2007; 35:W27–W32. [PubMed: 17488835]
56. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*. 2004; 60:2256–2268. [PubMed: 15572779]
57. Holm L, Kaariainen S, Wilton C, Plewczynski D. Using Dali for structural comparison of proteins. *Curr Protoc Bioinformatics*. 2006; Chap 5(Unit 5 5)
58. Bailey S, Wing RA, Steitz TA. The structure of *T. aquaticus* DNA polymerase III is distinct from eukaryotic replicative DNA polymerases. *Cell*. 2006; 126:893–904. [PubMed: 16959569]
59. Teplyakov A, Obmolova G, Khil PP, Howard AJ, Camerini-Otero RD, Gilliland GL. Crystal structure of the *Escherichia coli* YcdX protein reveals a trinuclear zinc active site. *Proteins*. 2003; 51:315–318. [PubMed: 12661000]
60. Kawano S, Nakashima T, Kakuta Y, Tanaka I, Kimura M. Crystal structure of protein Ph1481p in complex with protein Ph1877p of archaeal RNase P from *Pyrococcus horikoshii* OT3: implication of dimer formation of the holoenzyme. *J Mol Biol*. 2006; 357:583–591. [PubMed: 16430919]
61. Jozic D, Kaiser JT, Huber R, Bode W, Maskos K. X-ray structure of isoaspartyl dipeptidase from *E.coli*: a dinuclear zinc peptidase evolved from amidohydrolases. *J Mol Biol*. 2003; 332:243–256. [PubMed: 12946361]
62. Hall RS, Xiang DF, Xu C, Raushel FM. N-Acetyl-D-glucosamine-6-phosphate deacetylase: substrate activation via a single divalent metal ion. *Biochemistry*. 2007; 46:7942–7952. [PubMed: 17567047]
63. Binkowski TA, Naghibzadeh S, Liang J. CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Res*. 2003; 31:3352–3355. [PubMed: 12824325]
64. Tong W, Wei Y, Murga LF, Ondrechen MJ, Williams RJ. Partial Order Optimum Likelihood (POOL): Maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. *PLoS Comput Biol*. 2009; 5:e1000266.
65. Somarowthu S, Yang H, Hildebrand DGC, Ondrechen MJ. High-performance prediction of functional residues in proteins with machine learning and computed input features. *Biopolymers*. 2011 Jan 20. [Epub ahead of print].
66. Sankaraman S, Sjolander K. INTREPID: INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics*. 2008; 24:2445–2452. [PubMed: 18776193]
67. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol*. 2009; 5:e1000585.
68. Gold MG, Smith FD, Scott JD, Barford D. AKAP18 contains a phosphoesterase domain that binds AMP. *J Mol Biol*. 2008; 375:1329–1343. [PubMed: 18082768]

Figure 1.

Crystal structure of YP_910028.1 from *B. adolescentis*. (A) Stereo ribbon diagram of YP_910028.1 monomer color-coded by domain, PHP domain (residues 8–104 and 179–293, green) and an insert domain (residues 105–178, cyan). Helices (H1–H12) and β -strands (β 1– β 8) are indicated. Zn (blue grey) and Fe (orange) ions are shown as spheres, and phosphate ions are shown as red sticks. (B) Diagram showing the secondary structure elements of YP_910028.1 superimposed on its amino-acid sequence. The α -helices (H1–4, H6–10 and H12,), 3_{10} -helices (H5 & H11) and β -strands (β 1– β 8) are indicated. β -turns, and γ -turns are designated by their respective Greek letters (β , γ). The β -hairpin is depicted as a red loop and disordered residues as dashes.

Figure 2.

Superposition of YP_910028.1 (green) with Dali hits (from top, left to right): DNA polymerase III alpha subunit (DNA pol, blue, 2.9 Å RMSD, 195 C α atoms aligned), TM0559 (red, 2.8 Å RMSD, 172 C α atoms aligned), YcdX (light pink, 3.1 Å RMSD, 165 C α atoms aligned), histidinol phosphate phosphatase (HPP, yellow, 3.8 Å RMSD, 166 C α atoms aligned), ribonuclease P protein component 3 (RNase P, magenta, 3.9 Å RMSD, 145 C α atoms aligned), isoaspartyl dipeptidase (IADP, grey, 3.2 Å RMSD, 160 C α atoms aligned) and N-acetylglycosamine-6-phosphate deacetylase (NADP, orange, 3.7 Å RMSD, 156 C α atoms aligned).

Figure 3.

Active site and nucleotide binding of YP_910028.1: **(A)** Stereo image of the active site region of YP_910028.1. The metal ions, Zn-301 (blue grey) and Fe-302 & Fe-303 (orange) are shown as spheres and PO_4 (red) as sticks. Residues involved in the metal coordination are shown as sticks (carbon atoms are green, phosphate orange, oxygen red and nitrogen blue). The backbone is shown in the background as a green ribbon. **(B)** Surface representation of the structure of YP_910028.1 showing the PHP domain (green) and the extra insertion domain (residues 105–178, cyan). The region around the THEMATICs-predicted active site is shown in magenta. The side chains of the partially buried residues are represented as sticks under the surface. **(C)** Thermofluor melting curve of YP_910028.1 in the presence of no ligand (dark blue circle, $\Delta T_m = 0.0^\circ\text{C}$), ATP (green square, $\Delta T_m = 8.0^\circ\text{C}$), ADP (sky blue circle, $\Delta T_m = 11.0^\circ\text{C}$), AMP (yellow triangle, $\Delta T_m = 5.5^\circ\text{C}$), Adenosine (red square, $\Delta T_m = 0.0^\circ\text{C}$) and FAD (orange circle, $\Delta T_m = 3.0^\circ\text{C}$). **(D)** Superposition of Apo-YP_910028.1 (green) and AMP-YP_910028.1 (gray; 0.3 Å RMSD, 286 Ca atoms aligned). AMP shown as sticks (carbon atoms are yellow, phosphate orange, oxygen red and nitrogen blue). Residues interacting with AMP shown as sticks (carbon atoms are magenta, phosphate orange, oxygen red and nitrogen blue). Hydrogen bonds are shown as black dashed lines. Omit $2F_o - F_c$ electron density map around the AMP is shown as a green mesh contoured at 1σ .

(A)

(B)

Figure 4.

THEMATICS prediction of the active site of YP_910028.1 and other structurally related proteins (A) The THEMATICS-positive cluster of the query protein YP_910028.1 (green) is superimposed on that of the aligned DNA polymerase III alpha subunit (DNA pol; blue). The backbone of the query protein is rendered as a green ribbon. The predicted active site residues for YP_910028.1 are shown in green and those for DNA pol are shown in blue. The two non-conserved residues are highlighted by dashed circles. (B) The THEMATICS-positive cluster of YP_910028.1 superimposed on that of the aligned HPP active site. Predicted active residues for the query protein are shown in green, with the backbone rendered as a green ribbon. The predicted residues for HPP are shown in orange. The only

position that is not a perfect match by residue type is that of D189 and E230; these have chemically equivalent carboxylate groups on their side chains and are highlighted by a dashed circle.

Figure 5.

Para-nitrophenylphosphate (pNPP) phosphatase activity of YP_910028.1. Increasing amounts of YP_910028.1 (■) and three concurrently purified negative controls: NP_976866.1 (●), YP_508449.1 (▲) and YP_754858 (○) were assayed in the presence of pNPP. Note that the data points for the negative controls coincide and are not all visible.

Figure 6.

Polymerase and DNA proofreading activity assays for YP_910028.1 **(A)** Primer extension assay of DNA polymerase activity. YP_910028.1 was assayed at three different concentrations, as indicated above the gel. Assays were conducted under high salt conditions at room temperature (H), low salt conditions at room temperature (L), or high salt conditions at 37°C (labeled 37°C). Klenow was included as a positive control. Labeled (32 P) primer and template are included as gel standards. Time points were taken at 0, 15, 30, 50, and 80 or 82 minutes. **(B)** Hydrolysis activity on matched versus mismatched DNA ends. The letter above each set of time points indicates the nucleotide present at the end of the primer (X=dA, dC, dG, dT, where X=dC is the correctly matched primer end). Enzyme

concentrations are indicated at the top of the gel. Time points were taken at 0, 15, 30, and 60 minutes.

Table I

Summary of crystal parameters, data collection, and refinement statistics for YP_910028.1

	Apo-YP_910028.1	AMP-YP_910028.1
PDB ID	3e0f	3o0f
Space group	P 6 ₅	P 6 ₅
Unit cell parameters	a=b=102.50Å, c=54.17Å	a=b=103.34Å, c=54.74Å
Data collection	λ1 SADSe	Single Wavelength (MR)
Wavelength(Å)	0.97845	0.9765
Resolution range(Å)	29.6-2.40	44.8-1.94
Number of observations	136,119	268,826
Number of unique reflections	12,857	24,784
Completeness (%)	99.9 (99.8)	99.6 (97.7)
Mean I/σ(I)	15.2 (3.9)	12.5 (2.6)
R _{merge} on I (%)	12.8 (56.7)	16.7 (79.9)
Highest resolution shell (Å)	2.46-2.40	2.01-1.94
Model and refinement statistics		
Resolution range (Å)	29.6-2.40	44.8-1.94
No. of reflections (total) [#]	12,839	24,761
No. of reflections (test)	626	1,241
Completeness (% total)	99.9	99.6
R _{cryst}	0.149	0.160
R _{free}	0.207	0.192
Cutoff criteria	F >0	F >0
Stereochemical parameters		
Restraints (RMSD observed)		
Bond angle (°)	1.47	1.61
Bond length (Å)	0.014	0.016
Average isotropic B-value (Å ²) [*]	30.6	23.8
Wilson B-value (Å ²)	36.5	19.0
ESU based on R _{free} (Å)	0.22	0.12
Number of protein residues / atoms	286/2208	287/2214
Number of water molecules	103	183
Number of ions	6	4
Number of ligands	0	3

Highest resolution shell is shown in parentheses.

ESU = Estimated overall coordinate error ²⁹ $R_{\text{merge}} = \sum_i |F_{\text{obs}} - \langle I(hkl) \rangle| / \sum_i I_i(hkl)$ $R_{\text{cryst}} = \sum_i |F_{\text{obs}} - F_{\text{calc}}| / \sum_i F_{\text{obs}}$ where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively. R_{free} = as for R_{cryst} , but for 4.9% (Apo-YP_910028.1) and 5.0% (AMP-YP_910028.1) of the total reflections chosen at random and omitted from refinement.

[#]Typically, the number of unique reflections used in refinement is slightly lesser than the total number that were integrated and scaled. Reflections are excluded due to systematic absences, negative intensities and rounding errors in the resolution limits and cell parameters.

^{*}This value represents the total B that includes TLS and residual B components.

Table II

The top non-redundant 3D structural matches for YP_910028.1

Name	PDB ID	Z score	RMSD (Å)	No. of Cu atoms aligned	Sequence ID	Ligands found in the active site region of YP_910028.1
YP_910028.1 (Query structure)	3e0f	54.5	0.0	286	100%	Zn-301, Fe-302, Fe-303 PO4-304, PO4-305
DNA Polymerase III alpha subunit (DNA pol)	2hpi	17.0	2.9	195	17%	Zn-301, Zn-401
TM0559 – hypothetical protein	2anu	15.5	2.8	172	22%	Zn-405, Zn-505, Zn-605, Zn-705 Cl-1001, Cl-1002
YcdX – hypothetical protein	1m68	12.0	3.2	167	22%	Zn-300, Zn-301, Zn-302
Histidinol phosphate phosphatase (HPP)	2yz5	10.6	3.8	166	21%	Zn-501, Fe-502, Fe-503 PO4-2001
Ribonuclease P protein component 3(RNase P)	2czv	9.2	3.9	145	8%	ACT-2001
Isoaspartyl dipeptidase (IADP)	1po9	8.8	3.2	160	16%	Zn-601, Zn-602
N-Acetylglucosamine-6-phosphate deacetylase (NAPD)	2p50	8.5	3.7	156	18%	Zn-601

Table III

THEMATICS-predicted residues for YP_910028.1

Protein	PDB ID	Structurally aligned residues – <i>THEMATICS</i> predictions in bold									
		A	B	C	D	E	F	G	H	I	J
YP910028.1	3e0f	H17	H19	H49	E74	T76	H85	H202	E230	D260	H262
Histidinol phosphate phosphatase (HPP)	2yz5	H5	H7	H38	E80	D82	H108	H154	D189	D224	H226
DNA Polymerase III alpha subunit (DNA pol)	2hpi	H11	H13	H47	Y71	Y74	L96	C145	E180	D212	H214
Isoaspartyl dipeptidase (IADP)	1po9	H68	H70	G105	K162	A164	gap	H201	T229	D285	N287
N-Acetylglucosamine-6-phosphate deacetylase (NAPD)	2p50	Q57	N61	I100	V169	L171	gap	H195	G244	D273	T275
Ribonuclease P protein component 3(RNase P)	2czv	D13	R15	K35	L58	S60	gap	S102	G128	S168	Q170
TM0559	2anu	H13	H15	H45	E97	T99	H106	H140	E170	D193	H195
YcdX	1m68	H7	H9	H40	E73	N75	H101	H131	E156	D192	H194

The residues in **boldface** are predicted by *THEMATICS* to be functionally important. Each of the columns A–J represents a structurally aligned position; thus residues in the same column are superimposed in the 3D structural alignment. The query protein YP_910028.1 is listed first; then the 3D structure matches that have been characterized previously are listed in order from the best to the worst local, active site structural match. The last two proteins are uncharacterized.

Table IV

Top 8% residues predicted by POOL for the apo structure of YP_910028.1

Rank	Residue	Contact
1	Glu74	Metal Binding(Fe)
2	His49	Metal Binding(Zn)
3	His85	Metal Binding(Fe)
4	Asp260	Metal Binding (Fe)
5	His19	Metal Binding (Fe)
6	His17	Metal Binding (Fe)
7	His262	Metal Binding (Zn)
8	His202	Metal Binding (Fe)
9	Asp24	Metal Binding (Zn)
10	Lys266	AMP
11	Arg208	(His85)
12	Gln207	--
13	Arg145	AMP
14	Asp205	(His202)
15	Phe165	AMP
16	Pro146	AMP
17	Ser83	(His85)
18	Thr76	AMP
19	Tyr176	AMP
20	Arg112	AMP
21	Arg109	AMP
22	His233	(His17, His202, D260)
23	Asn209	--