WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

The X-ray crystallographic structure and activity analysis of a *Pseudomonas*-specific subfamily of the HAD enzyme superfamily evidences a novel biochemical function

Ezra Peisach,[1] Liangbing Wang,[2] A. Maxwell Burroughs,[3,4] L. Aravind,[3] Debra Dunaway-Mariano,[2]* and Karen N. Allen[1,4]*

[1] Department of Physiology and Biophysics, Boston University School of Medicine, 715 Albany Street, Boston, Massachusetts 02118-2394

[2] Department of Chemistry, University of New Mexico, Albuquerque, New Mexico 87131

[3] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

[4] Bioinformatics Program, Boston University, Boston, Massachusetts 02215

ABSTRACT

The haloacid dehalogenase (HAD) superfamily is a large family of proteins dominated by phosphotransferases. Thirty-three sequence families within the HAD superfamily (HADSF) have been identified to assist in function assignment. One such family includes the enzyme phosphoacetaldehyde hydrolase (phosphonatase). Phosphonatase possesses the conserved Rossmanniod core domain and a C1-type cap domain. Other members of this family do not possess a cap domain and because the cap domain of phosphonatase plays an important role in active site desolvation and catalysis, the function of the capless family members must be unique. A representative of the capless subfamily, PSPTO_2114, from the plant pathogen Pseudomonas syringae, was targeted for catalytic activity and structure analyses. The X-ray structure of PSPTO_2114 reveals a capless homodimer that conserves some but not all of the intersubunit contacts contributed by the core domains of the phosphonatase homodimer. The region of the PSPTO_2114 that corresponds to the catalytic scaffold of phosphonatase (and other HAD phosphotransfereases) positions amino acid residues that are ill suited for Mg^{+2} cofactor binding and mediation of phosphoryl group transfer between donor and acceptor substrates. The absence of phosphotransferase activity in PSPTO_2114 was confirmed by kinetic assays. To explore PSPTO_2114 function, the conservation of sequence motifs extending outside of the HADSF catalytic scaffold was examined. The stringently conserved residues among PSPTO_2114 homologs were mapped onto the PSPTO_2114 three-dimensional structure to identify a surface region unique to the family members that do not possess a cap domain. The hypothesis that this region is used in protein–protein recognition is explored to define, for the first time, HADSF proteins which have acquired a function other than that of a catalyst.

Proteins 2008; 70:197–207.

Key words: haloacid dehalogenase; Pseudomonas; protein: protein interactions; functional genomics; substrate screening; HNOB domain; phosphonatase.

INTRODUCTION

The HAD (haloacid dehalogenase) enzyme superfamily includes phosphatases, ATPases, phosphomutases, phosphonatases, and dehalogenases.[1–3] The phosphatases are especially ubiquitous, having evolved substrate recognition elements which allow the conserved catalytic scaffold (a Rossmannoid fold, hereafter referred to as the core domain) to act on a wide range of metabolites. Most members of the HADSF contain inserts, termed cap domains, placed in the middle of the hairpin motif at the end of β-strand 1 (class C1) or the linker immediately downstream of β-strand 3 (class C2) of the core domain.[4] The

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Coordinates have been deposited in the protein data bank with accession code 2ODA.

Grant sponsor: NIH; Grant number: GM61099; Grant sponsors: Office of Biological and Environmental Research and of Basic Energy Sciences, U.S. Department of Energy; National Center for Research Resources, National Institutes of Health; National Library of Medicine (Intramural Research Program), National Institutes of Health.

*Correspondence to: Karen Allen, Department of Physiology and Biophysics–R702, 715 Albany St., Boston, MA 02118. E-mail: drkallen@bu.edu or Debra Dunaway-Mariano, Department of Chemistry, University of New Mexico, Albuquerque, NM 87131. E-mail: dd39@unm.edu.

Received 21 January 2007; Revised 26 March 2007; Accepted 29 March 2007
Published online 24 July 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21583

cap domains have been independently inserted at one of these two locations on multiple occasions during the course of evolution. The cap domains of the C1 and C2 HADSF members function to desolvate the active site during catalytic turnover, and to contribute catalytic and/or substrate binding residues to the active site housed in the core domain.[5–7] A third class of the HADSF, class C0, lacks a cap domain insertion.[4,8]

Phylogenetic analysis has been carried out to divide the HADSF into 33 families.[4] Although diversification of chemical function and substrate recognition is coupled with cap domain variation, the division of the families is not. Function assignment to the HAD families has been undertaken by numerous laboratories. Our approach typically combines bioinformatics, X-ray crystallographic analysis, and substrate-activity screening.[9,10] Computational methods are employed to identify novel proteins containing sequence motifs that may support a unique function, after which targeted X-ray crystallographic analysis is carried out to define the three-dimensional structure. The electrostatic and steric features of the active site thus defined are used along with information derived from genomic analysis to identify potential substrates. In the present work, we applied this approach to function identification of the targeted protein PSPTO_2114 (accession Q884H9; gi: 28869317) from the plant pathogen *Pseudomonas syringae* pv. *tomato*. PSPTO_2114 is a member of the recently described PA2803 subfamily found in Pseudomonads.[4] The PA2803 subfamily belongs to a larger family of the HADSF that includes the well characterized enzyme phosphonoacetaldehyde hydrolase (phosphonatase), and is thus named the "phosphonatase family." A predominant structural feature of the PA2803 subfamily in general and PSPTO_2114 in particular, which sets it apart from the other members of the phosphonatase family is the absence or abbreviation of the cap domains insert within its amino acid sequence. Phosphonatase, a C1 HADSF member, utilizes the cap domain to desolvate the catalytic site as well as to position a Lys residue to activate the phosphonoacetaldehyde via Schiff base formation (Fig. 1). Thus, the absence or abbreviation of the cap domain is a clear indication of divergence in function. Furthermore, PSPTO_2114 and the other PA2803 members do not conserve the key core domain active-site residues, located within motifs 1–4, which are required for Mg^{+2} cofactor binding, substrate binding, and catalysis (Fig. 1). We are intrigued by this subfamily of proteins, because the primary structures argue against a catalytic function, and therefore constituted the first example of the adaptation of the HADSF fold for a noncatalytic role.

Herein, we report the bioinformatic identification, cloning, and expression of the *P. syringae* pv. *tomato* PSPTO_2114 gene in *E. coli* and the method used to purify the protein product to homogeneity. The X-ray crystal structure determination of the protein to 1.9 Å resolution reveals a homodimeric structure composed of the conserved Rossmannoid fold core domain. The active site is devoid of the HADSF core catalytic residues and the metal-ion cofactor that typify the phosphotransferase branch of the HADSF. We hypothesize that the elongated surface conserved within the PA2803 subfamily is used in protein–protein interaction. Although the identity of the putative protein partner is unknown, conserved gene context provides insight into the biochemical process in which PA2803 subfamily members might participate.

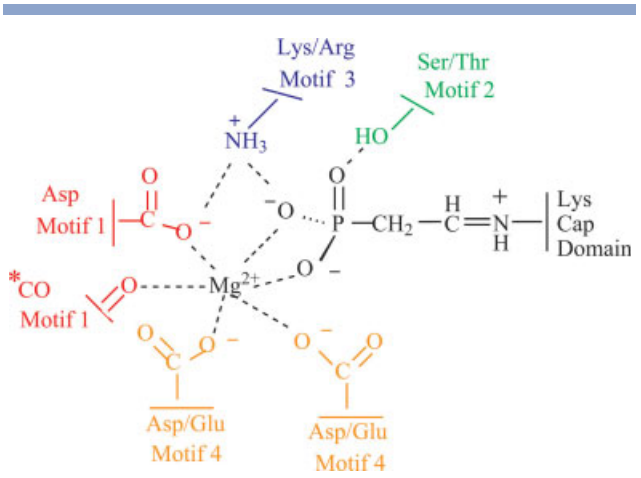

Figure 1

*Schematic depiction of residues that form the phosphonatase catalytic site (Motifs 1–4, plus the Schiff base Lys from the cap domain). The Schiff base of lysine is shown with the Asp nucleophile aligned for attack. The Ala from Motif 1 that is conserved in the phosphonatase subfamily is denoted with a *.*

MATERIALS AND METHODS

Except where indicated, all chemicals were obtained from Sigma-Aldrich. Primers, T4 DNA ligase, and restriction enzymes were from Invitrogen. *Pfu* polymerase and the pET23b vector kit were from Stratagene. Host cells were purchased from Stratagene. Genomic DNA from *Pseudomonas syringae* pv. *tomato* (ATCC BAA-871) was purchased from ATCC.

Target selection

The PA2803 subfamily was identified through BLASTP searches[11] performed on the nonredundant (NR) database of protein sequences using known phosphonatase sequences as queries. For example, a search initiated with the *Bacillus cereus* phosphonatase (gi: 18254515) recovers a PA2803 member from *Pseudomonas syringae* (*e*-value = 0.009; gi: 28852558). Surprisingly, the catalytic motifs typical of the HADSF, which are essential to phosphonatase activity, were disrupted. However, reverse searches were performed with the *P. syringae* sequence, and the re-

DOI 10.1002/prot

covery of the *B. cereus* phosphonatase sequence at a significant expectation value (*e*-value = $2e^{-04}$) supported the existence of a genuine evolutionary relationship between the sequences. To further probe this relationship, a multiple alignment of the PA2803 subfamily was constructed using the T-COFFEE program[12] and the secondary structure of the proteins in the subfamily was predicted by feeding the alignment into the JPRED2 program.[13] The JPRED2 output revealed an alternating α/β pattern, congruous with the Rossmannnoid fold found in other members of the HADSF, confirming the presence of the HAD fold in the PA2803-like proteins.

Cloning, expression, and purification

The DNA encoding the gene PSPTO_2114 from *Pseudomonas syringae* pv. *tomato* was amplified by PCR using the genomic DNA from *Pseudomonas syringae* pv. *tomato* (ATCC BAA-871), and *Pfu* DNA polymerase. Oligonucleotide primers (5′-GCTCG CGCCGCGCCATATGCCTTTACCAACC) and (5′-AGTGACTCAAGACGGATCCAGTAAGCTCGC) containing restriction endonuclease cleavage sites *Nde*I and *Bam*HI (underlined) were used in the PCR reaction.

The pET-23b vector, cut by restriction enzymes *Nde*I and *Bam*HI, was ligated with the PCR product that had been isolated and digested with the same restriction enzymes. The ligation product was used to transform *Escherichia coli* BL21(DE3) competent cells (Stratagene) and the plasmid DNA produced was purified using the Qiaprep Spin Miniprep Kit. The gene sequence was confirmed by DNA sequencing carried out by the Center for Genetics in Medicine at the Biochemistry and Molecular Biology Department, University of New Mexico. Transformed cells (10 L) were grown at 25°C with agitation at 200 rpm in Luria broth containing 50 μg/mL ampicillin to an OD_{600} of 0.6–1.0 and induced for 4 h at 25°C at a final concentration of 0.4 m*M* isopropyl α-D-thiogalactopyranoside (IPTG). The cells were harvested by centrifugation (7855 g for 15 min at 4°C) to yield 3 g/L of culture medium. The cell pellet was suspended in 10 mL of ice-cold buffer A/1g wet cell consisting of 50 m*M* K^+-Hepes (pH 7.0 at 25°C), 5 m*M* Mg^{2+} and 1 m*M* DTT. The cell suspension was passed through a French press at 1200 PSIG before centrifugation at 48,384 g for 30 min at 4°C. The supernatant was loaded onto a 40 cm × 5 cm DEAE Sepharose column, which was eluted with a 2 L linear gradient of KCl (0–0.5 M) in buffer A. The column fractions were analyzed by SDS-PAGE.

The desired protein fractions from DEAE were combined, adjusted to 15% $(NH_4)_2SO_4$ (W/V), and loaded onto an 18 cm × 3 cm butyl-Sepharose column pre-equilibrated with buffer A containing 15% $(NH_4)_2SO_4$. The column was eluted with a 0.5 L linear gradient of $(NH_4)_2SO_4$ (15–0%) in buffer A. The desired protein fractions shown to be homogeneous by SDS-PAGE were combined, dialyzed against buffer A, and concentrated at 4°C using a 10K Amicon Ultra Centrifugal filter (Millipore) then stored at −80°C. The protein concentration was determined by the Bradford method[14] and from the protein absorbance at 280 nm (ε = 76335 M^{-1} cm^{-1}). The yield of PSPTO_2114 was 10 mg/g wet cell. Selenomethionione (SeMet)-substituted PSPTO_2114 was overexpressed in a methionine auxotroph of *E. coli* BL21(DE3), grown in a defined media containing 40 mg/L SeMet and purified using the protocol described for the native protein.

Molecular mass determination

The theoretical subunit molecular mass of recombinant PSPTO_2114 was calculated from the amino acid composition (derived from the gene sequence) using the EXPASY Molecular Biology Server program Compute pI/MW.[15] The subunit size was determined by SDS-PAGE analysis with molecular weight standards from Invitrogen, and the subunit mass was determined by MS-ES mass spectrometry (Mass Spectrometry Lab, University of New Mexico). The molecular size of native recombinant PSPTO_2114 was determined using gravity-flow gel-filtration techniques. PSPTO_2114 was subjected to chromatography at 4°C on a calibrated (Pharmacia Gel Filtration Calibration Kit) 1.5 × 180 cm Sephacryl S-200 column (Pharmacia) using buffer A at a flow rate of 1 mL/min, as eluant. The molecular weight was determined from the elution volume using a plot of log(molecular weight) of a standard protein *vs.* elution volume as reference.

Protein crystallization

Crystals were obtained via the vapor-diffusion/hanging drop method at 4°C using a precipitant solution consisting of 15–20% PEG 3350, 150–200 m*M* sodium formate, 100 m*M* Hepes (pH 7.3 at ambient temperature) and 5 m*M* $MgCl_2$. Freshly thawed protein was diluted to 12 mg/mL in 2 m*M* DTT and 5 m*M* $MgCl_2$ combined in equal volumes (1.5 μL) of protein and precipitant and suspended over 0.5 mL of precipitant. It was critical to the formation of single crystals to form an elongated drop, wherein crystals nucleated in the narrow portion of the drop. Failure to do so resulted in many nucleation events uniformly along the entire circumference of the drop. Crystals are temperature sensitive and grow as rods 0.1 × 0.1 × 0.5 mm in size at 4°C. Single crystals of SeMet protein were transferred to Paratone-N at 4°C and frozen in liquid nitrogen. X-ray diffraction data were collected at beamline X12C at the National Synchrotron Light Source (NSLS), Brookhaven National Laboratory on an ADSC Q210 detector and processed with the HKL2000[16] program suite. The crystal structure was solved by the multiwavelength anomalous dispersion (MAD) method using the Selenium edge (Table I). The programs SOLVE and RESOLVE (version 2.11)[17] were used to locate the eight selenium sites, automati-

Table I
Summary of Data Collection and Refinement Statistics

Datasets	Remote	Peak	Edge
Wavelength (Å)	0.9500	0.9791	0.9798
Space group		$P2_12_12_1$	
Unit-cell dimensions (Å, degree)		$a = 48.61$, $b = 74.43$, $c = 106.35$, $\alpha = \beta = \gamma = 90$	
Resolution range (outer shell) (Å)		50–1.9 (1.97–1.9)	
Total reflections	683,234	654,397	653,527
Unique reflections	31,224	30,874	30,826
$I/\sigma(I)$ (outer shell)	43.1 (8.1)	41.2 (7.3)	39.6 (6.9)
R_{sym} (outer shell) (%)[a]	5.9 (18.0)	6.5 (18.3)	5.7 (18.6)
Completeness (outer shell) (%)	99.0 (90.4)	98.0 (83.8)	98.2 (84.9)
Figure of merit (RESOLVE 2.1 Å)		0.85	
f′/f″ (e^-)	−2.4/4.1	−5.8/5.5	−8.5/3.0
R_{work}/R_{free}[b]		0.17/0.22	
Monomers per asymmetric unit		2	
Atoms per asymmetric unit		3445	
Average B ($Å^2$)			
Protein		12.5	
Water		20.8	
Hepes and Mg^{+2}		15.1	
RMSD bonds (Å)/angles (degree)		0.013/1.423	
Improper (degree)		0.005	
PDB ID		2ODA	

[a] $R_{sym} = \Sigma_i |I_i - \langle I \rangle| / \Sigma \langle I \rangle$, where I_i is an individual intensity measurement and $\langle I \rangle$ is the average intensity for all measurements of the reflection i.
[b] $R_{work}/R_{free} = \Sigma||F_o| - |F_c||/\Sigma|F_o|$, where F_o and F_c are the observed and calculated structure factors, respectively. R_{free} is calculated for 10% of reflections randomly chosen and excluded from refinement.

cally trace >85% of the protein backbone, and correctly identify the two monomers in the asymmetric unit. Refinement was carried out in CNS[18] and REFMAC5,[19] and model rebuilding in COOT.[20] The final model, refined to 1.9 Å resolution ($R_{work} = 0.17$, $R_{free} = 0.22$), consists of a dimer of 392 amino acid residues, 536 water molecules, 2 Hepes molecules, and 1 Mg^{2+} (found on the surface at the dimer interface) (see Table I for final model statistics). The first four residues (1–4) of chain A and the first residue (1) of chain B within the dimer were not visible and were excluded from the final model.

Substrate-activity screens

The rate of *p*-nitrophenyl phosphate (PNPP) hydrolysis was determined by monitoring the increase in absorbance at 410 nm ($\Delta\varepsilon = 18.4\ mM^{-1}cm^{-1}$) at 25°C. The 0.5 mL assay mixtures contained 20 μ*M* PSPTO_2114, 50 m*M* Hepes (pH = 7.0), 5 m*M* $MgCl_2$, and 1 m*M* PNPP. Hydrolyses of all other phosphate esters were monitored using the Enzcheck phosphate assay kit (Invitrogen) to detect total phosphate release determined by monitoring the increase in absorbance at 360 nm at 25°C. The 1 mL assay mixture initially contained 50 m*M* Tris-HCl buffer (pH = 7.5), 1 m*M* $MgCl_2$, 1 m*M* substrate 20 μ*M* PSPTO_2114, 0.2m*M* MESG (2-amino-6-mercapto-7-methylpurine ribonucleoside) and 1U purine nucleoside phosphorylase. In parallel, the background level of phosphate release was measured using a control reaction, which excluded the PSPTO_2114.

Phosphonatase and phosphoglucomutase activities were tested using the reaction conditions listed above in conjunction with the published assay methods.[3,21]

RESULTS AND DISCUSSION

Identification of a lineage-specific phosphonatase variant

The PA2803 subfamily of the HADSF was identified through sequence homology searches performed on the nonredundant (NR) database of protein sequences using known phosphonatase sequences as queries. Multiple-sequence alignment and secondary-structure predictions revealed the loss or abbreviation of the helical "cap" domain characteristic of the phosphonatase proteins. These searches also showed that the subfamily was restricted to various *Pseudomonas* strains (represented in Fig. 2), suggesting that the protein arose through a lineage-specific duplication of the phosphonatase protein; a scenario supported by the presence of both phosphonatase and the variant with no cap in most *Pseudomonas* strains. In fact, of the strains with completely sequenced genomes carrying a PA2803-like protein, only *syringae* pv. *tomato* str. DC3000 and *syringae* pv. *syringae* B728a lack phosphonatase, indicating that a lineage-specific loss of the phosphonatase gene may have occurred in *syringae*. Cap-do-

DOI 10.1002/prot

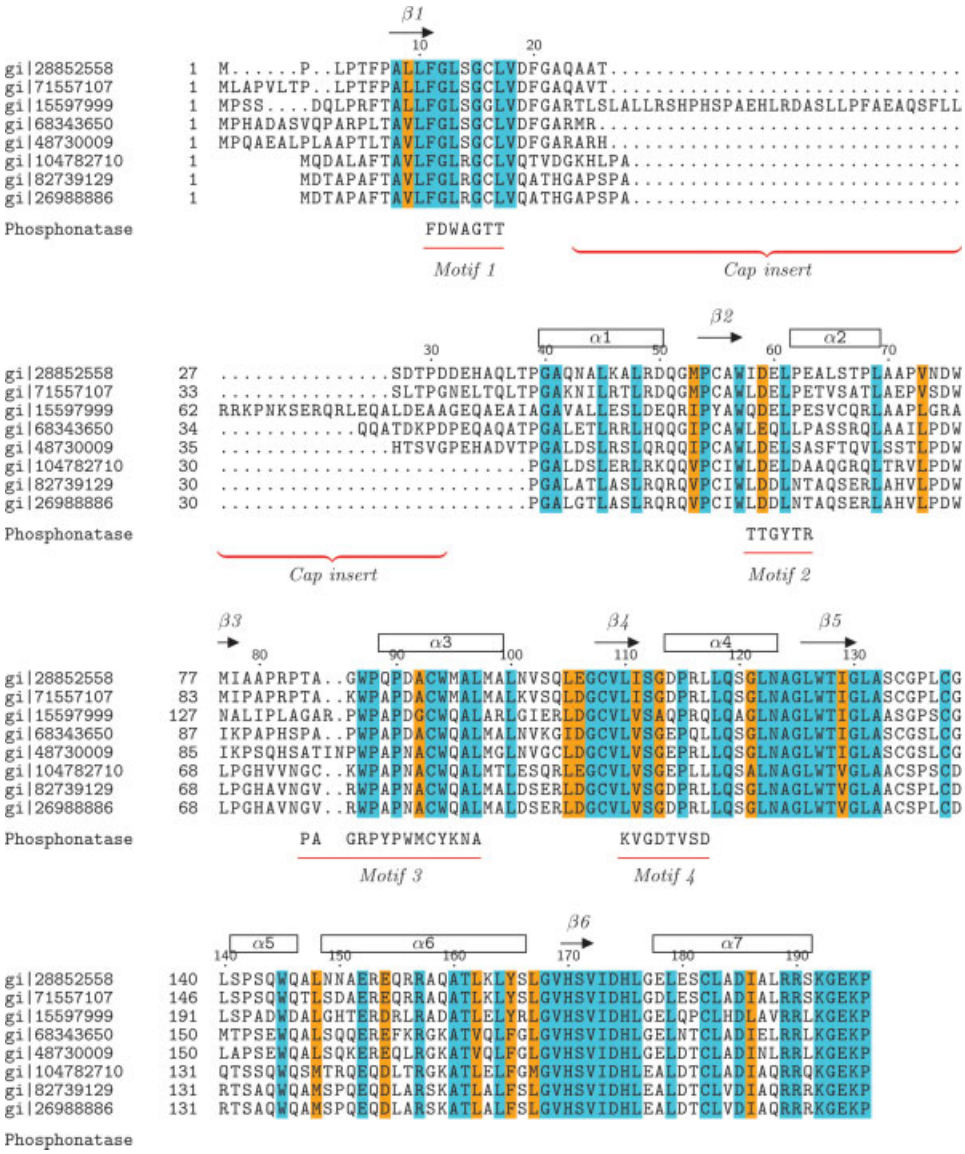

Figure 2

Sequence alignment of members of the PA2803 subfamily. Secondary structure assignments (noted along the top) are from the current structure: gi28852558 (PSPTO_2114) P. syringae *pv.* tomato *str. DC3000; gi71557107* P. syringae *pv.* phaseolicola *1448A; gi15597999 PA2803 from* P. aeruginosa *PAO1; gi68343650* P. fluorescens *Pf-5; gi48730009* P. fluorescens *PfO-1; gi104782710 PSEEN3700 from* P. entomophila *L48; gi82739129* P. putida *F1; gi26988886* P. putida *KT2440. Sequence Motifs 1–4 used to identify HADSF members are underlined in red and labeled, and the corresponding motifs from* B. cereus *phosphonatase included for comparison (as well as the location of the insertion point of the phosphonatase cap (note that in phosphonatase the cap insert would be >60 amino acids in length, thus any insert in this region in the PA2803 subfamily would be too small to form a typical cap).*

main loss is consistent with the lineage-specific duplication scenario, as duplication is often accompanied by functional divergence.[22] Sequence motifs 1–4, which comprise the catalytic framework of not just the phosphonatase family but the entire HADSF, are not evident in the PA2803 subfamily (Fig. 2). These findings were verified by the X-ray crystal structure of the PA2803 subfamily member *P. syringae* PSPTO_2114 (*vide infra*).

Structure determination of recombinant *P. syringae* PSPTO_2114

Purification

Recombinant *P. syringae* PSPTO_2114 was purified to homogeneity using a column chromatography-based protocol with an overall yield of 10 mg/g wet cells. The theoretical molecular weight of PSPTO_2114 minus the N-

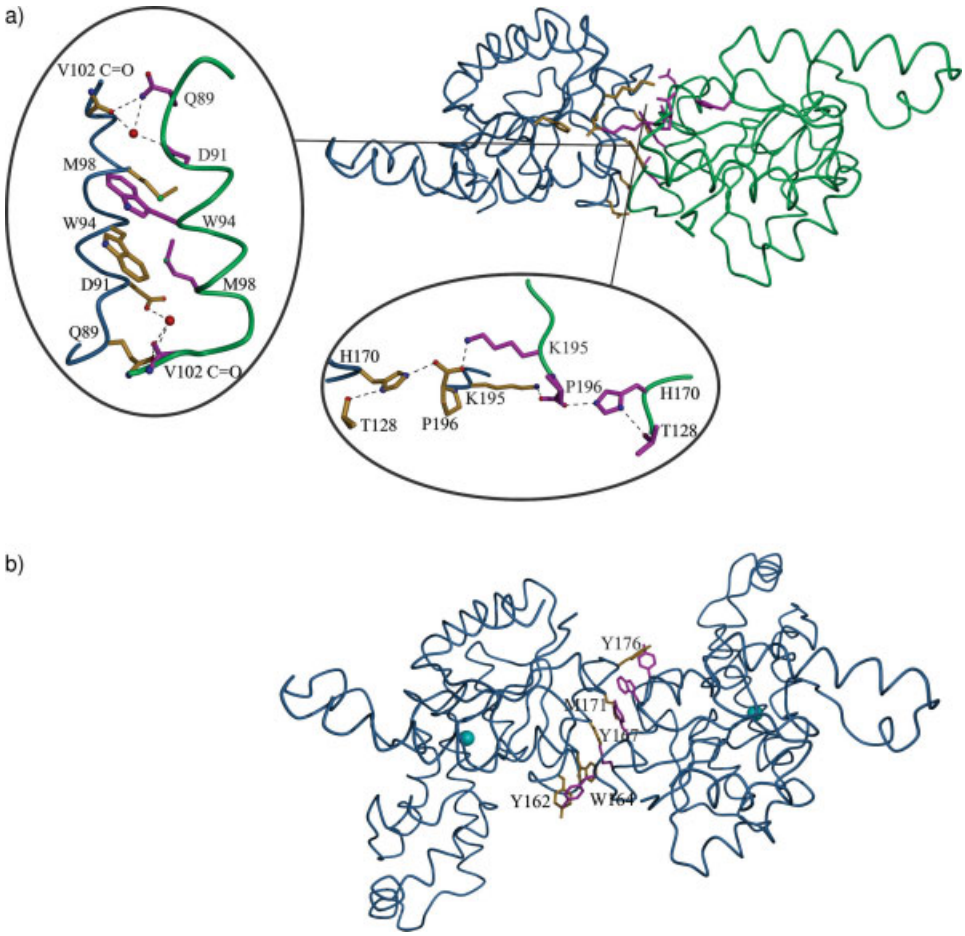

Figure 3

***A**: Structure of dimeric PSPTO_2114 with the residues lining the dimer interface in orange (monomer A) and magenta (monomer B). The magnified regions show the dimer interface and the Lys–Lys stacking interactions making up the oligomeric contacts. **B**: Structure of dimeric* B. cereus *phosphonatase shown in the same relative orientation as PSPTO_2114 in panel A with the residues lining the dimer interface in orange (monomer A) and magenta (monomer B).*

terminal Met as calculated from the amino acid sequence is 20,704 Da. This value agrees with the experimental molecular weight of 20,702 Da measured by mass spectrometry, indicating that the N-terminal Met is removed by posttranslational modification. The subunit size of PSPTO_2114 estimated by SDS-PAGE analysis is ~21 kDa, whereas the native protein size determined by gel-filtration chromatography is ~47 kDa. These results are consistent with a homodimeric quaternary structure, which is also observed for the crystalline enzyme by X-ray crystallography.

Overall fold

The structure of PSPTO_2114 was determined to 1.9 Å using the MAD phasing method (Table I). The native structure of this protein reveals an elliptical dimeric protein with overall dimensions of 44 Å × 45 Å × 83 Å [Fig. 3(A)]. The monomer is a modified Rossmannoid fold consisting of a six-stranded parallel β-sheet surrounded by seven α-helices, characteristic of the core domain of HADSF members. As predicted, the cap domain present in the closely related phosphonatase [Fig. 3(B)] is extremely abbreviated. A multiple sequence alignment of the PA2803 subfamily (Fig. 2) shows that the region corresponding to the cap domain has become truncated to varying degrees in different members of the subfamily. Thus, the original function of the phosphonatase cap is not retained in the PA2803.

A DALI[23] search using the PSPTO_2114 coordinates identifies *Bacillus cereus* phosphonatase (class C1) as the most similar structure ($Z = 21.3$) followed by *Lactococcus lactis* β-phosphoglucomutase (class C1, $Z = 14.0$). The RMSD between a single monomer of phosphonatase (pdb accession 1FEZ) and PSPTO_2114 is 1.62 Å for 166 Cα atoms.

DOI 10.1002/prot

Dimer interface

Examination of the PSPTO_2114 subunit interface reveals an arrangement similar to that of phosphonatase, with symmetrical packing of the two α3 helices and favorable amino acid side chain interactions across the interface [Fig. 3(A,B)]. The buried surface area in PSPTO_2114 is 5,085 Å^2 compared to that of phosphonatase at 6,827 Å^2, confirming the similarity in the size of the dimer interface. The dimerization mode common to phosphonatase and PSPTO_2114 has not been observed in other HADSF members and thus, appears to be a specific trait of the phosphonatase family.

Examination of the interacting residues at the respective phosphonatase and PSPTO_2114 dimer interfaces reveals both sites of conservation and sites of divergence. In *B. cereus* phosphonatase, conserved residues Met171, Trp164, Tyr162, Tyr167, and Tyr176 pair across the interface and much of the driving force for subunit association is based on desolvation of the hydrophobic side chains. The phosphonatase Met171 and Tyr167 positions correspond to Met98 and Trp94 in PSPTO_2114. The residues are conserved within the PA2083 subfamily, and are thus a shared dimerization motif across the phosphonatase family. In contrast, the *B. cereus* phosphonatase Trp164 is replaced by Asp91 in PSPTO_2114, and a water molecule serves to bridge the two Asp91 residues across the dimer interface. Thus, while the amino acid position for interaction is retained, the nature of the interaction has been altered. The Tyr162 of phosphonatase is replaced with Gln89 in PSPTO_2114 [Fig. 3(A)] and with Ala in other PA2083 subfamily members. Although the Gln89 in PSPTO_2114 appears to interact with Asn101 (3.5 Å) and Val102 of the opposing subunit, the interaction distances are large (3.5 and 3.3 Å). It is unlikely that this particular position is important to dimerization in the PA2083 subfamily, and thus it constitutes divergence. The *B. cereus* phosphonatase Tyr176 does not have a spatial counterpart in PSPTO_2114, also indicative of loss of function and divergence.

In addition to the dimer described above, the PSPTO_2114 dimer is stabilized by contributions from paired Lys195 residues [inset, Fig. 3(A)]. The hydrocarbon side-chains of the Lys195-Lys195 pair are stacked and desolvated while each Nε-ammonium group engages in hydrogen-bond formation with the carbonyl oxygen of Pro196 on the opposing protomer. The conservation of the C-terminus sequence motif GEKP among PA2803 subfamily members suggests that this hydrogen-bond network serves a common function. The glycine caps the C-terminal helix whereas the Lys and Pro extend across the dimer interface. The C-terminal Pro196 is positioned in each protomer by hydrogen-bond interactions with NE2 of His170, which in turn is oriented through a hydrogen bond from ND1 to the carboxylate side chain of Thr128 [Fig. 3(A)]. Both His170 and Thr128 residues are stringently conserved within the PA2803 subfamily. Among the *Pseudomonas* phosphonatase sequences, the C-terminal sequence is GEMPXXX. This motif is equivalent to the PSPTO_2114 sequence GEKP, however the bridging Lys–Pro interaction is a unique feature of the PA2803 subfamily, and absent in phosphonatase.

HADSF active site

The four conserved motifs (1–4) comprising the active site of the HADSF are highlighted in the representation of the active-site structure of phosphonatase [Fig. 4(A,C)]. A comparison of the active sites of phosphonatase and PSPTO_2114 shows remarkable differences [Fig. 4(C,D)]. In PSPTO_2114, the nucleophile Asp12 of phosphonatase has been replaced by the stringently conserved Gly12. Without the nucleophilic Asp, it is unlikely that PSPTO_2114 retains the phosphoryl-transfer activity of the HADSF. This conclusion is supported by the screens for phosphonatase, phosphatase, and phosphomutase activity wherein no activity above background was observed for any substrate tested. The panel of substrates included phosphonoacetaldehyde and glucose 1-phosphate, the substrates for the closely related HADSF members (*vide supra*) phosphonatase and β-phosphoglucomutase, and substrates of known HADSF phosphatases viz. phosphoserine, phosphoglycolate, fructose 6-phosphate, *N*-acetyl glucoseamine 6-phosphate, and *p*-nitrophenyl phosphate.

Examination of the other catalytic HAD motifs reveals additional replacements of catalytic and cofactor-binding residues. Specifically, the hydrogen-bond donor of loop 2, Thr126 of phosphonatase, is replaced by Asp59, which could assume the same phosphate binding role only if it were protonated. The Loop 3 Arg160 of phosphonatase, which provides electrostatic stabilization of the phosphoryl group and orients the Asp nucleophile, is replaced with Trp87. The loop 4 of phosphonatase positions two Asp residues for Mg^{+2} binding, consistent with the minimal requirement of two carboxylate residues for the HADSF phosphotransferases.[24] PSPTO_2114 positions only one such residue, Asp114, and in this manner is similar to the HADSF dehalogenases, which do not possess a metal cofactor binding site.[24] PSPTO_2114 was crystallized from a solution containing 5 m*M* Mg^{2+}, yet the X-ray crystal structure evidences no Mg^{2+} bound to the active site cleft.

The lack of catalytic and metal ion cofactor-binding residues indicates that PSPTO_2114 does not catalyze phosphoryl transfer. The lack of the nucleophilic Asp also nullifies the possibility that PSPTO_2114 functions in carbon group transfer as exemplified by the dehalogenases of the HADSF. In fact, examination of the conservation of the residues overlapping the canonical phosphatase active site (Fig. 2) shows that only Trp89 and Gly12 are absolutely conserved among PSPTO_2114 orthologs from different Pseudomonads. The lack of

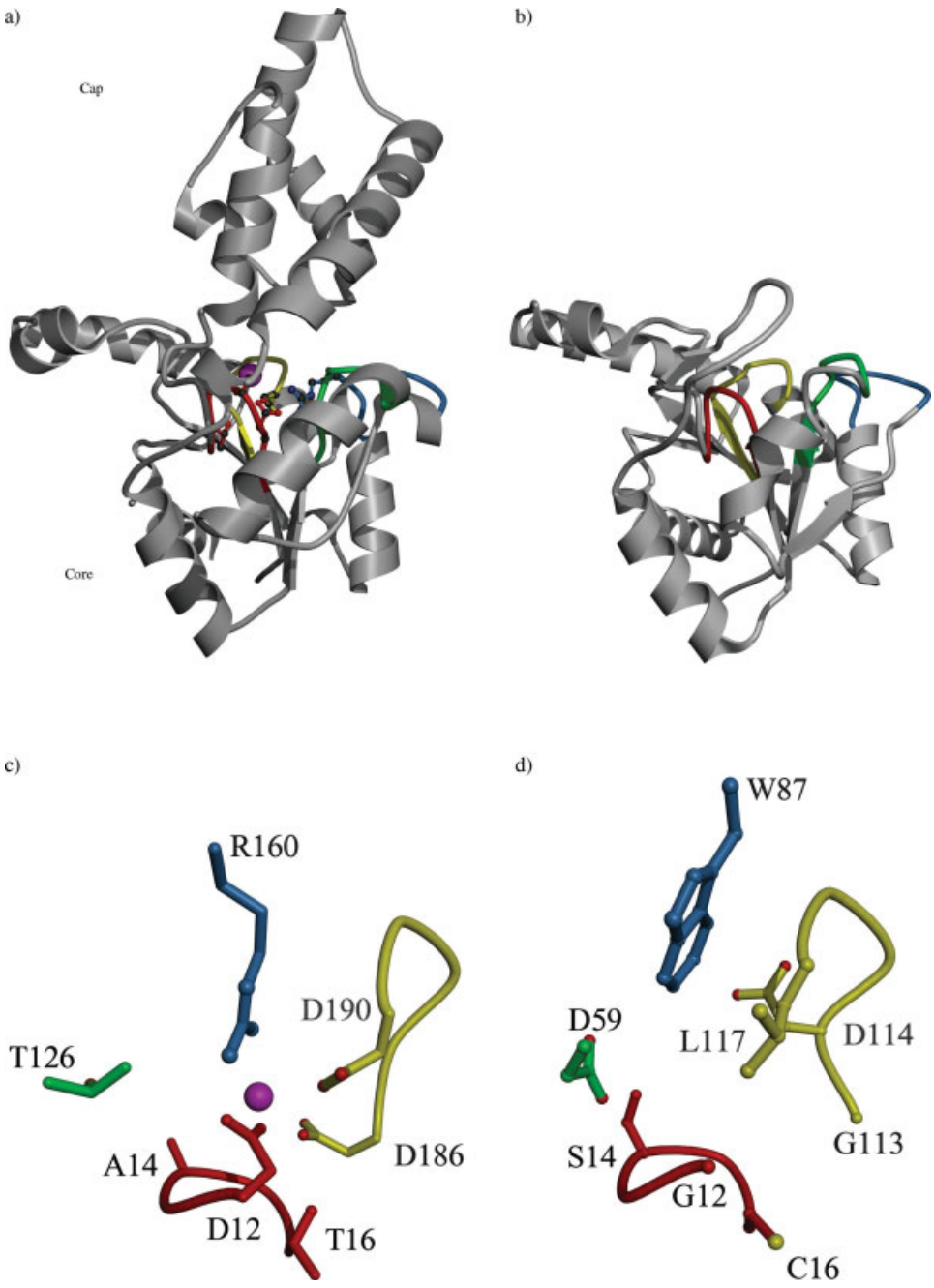

Figure 4

*Comparison of phosphonatase and PSPTO_2114. **A**: Ribbon diagram of one subunit of phosphonatase (pdb accession code 1FEZ) and (**B**) PSPTO_2114. The active site in phosphonatase is indicated by the cofactor Mg^{2+} (magenta sphere) with motifs 1–4 colored red, green, blue, and yellow, respectively (as in Fig. 1). The corresponding ribbon diagram of PSPTO_2114 uses the same color scheme. There is no domain in PSPTO_2114 corresponding to the mobile cap domain of phosphonatase. **C**: Close up of the active site of phosphonatase with coloring as in panel A (yellow backbone) and (**D**) the corresponding structure of PSPTO_2114 (cyan backbone).*

amino acid conservation argues against use of this pocket as an enzymatic active site. A search carried out with ProFunc[25,26] failed to identify a plausible catalytic site in PSPTO_2114, and, notably, the vestigial phosphonatase/HADSF active site was not found.

Conservation of binding surfaces

If PSPTO_2114 is not an enzyme, what function does it provide to the *Pseudomonad*? Mean distance tests performed using the MEGA program[27] established the rate of sequence divergence within the PA2803 subfamily as

DOI 10.1002/prot

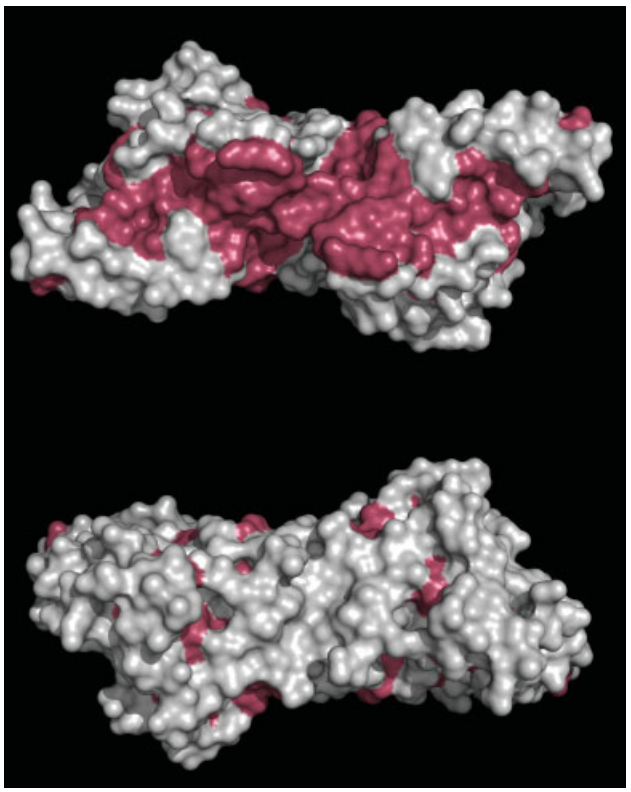

Figure 5

Surface representation of PSPTO_2114 in which absolutely conserved residues identified in Figure 2 are mapped onto the structure in dark pink. The top and bottom views are rotated 180 along the x *axis.*

significantly higher than the set of all classical phosphonatase sequences found in *Pseudomonas*, a finding consistent with other catalytically inactive proteins which have assumed a secondary binding function.[28] Evolution of a binding function can easily be envisaged if the substrate-binding residues are separate in identity from catalytic residues. Selective replacement of the catalytic residues would produce a protein that could selectively bind but not transform the physiological substrate. Alternatively, if the protein scaffold is suited for presenting an extensive binding surface, the protein may be recruited to bind to a macromolecule (either protein or DNA/RNA). The patterns of conservation of residues in the PA2803 subfamily should allow the identification of binding surfaces critical to function. PSPTO_2114 is conserved in at least eight different genomes in *Pseudomonas* with 50–92% identity between sequences (Fig. 2). By mapping the stringently conserved residues onto the PSPTO_2114 three-dimensional structure (Fig. 5) we identified one contiguous region of the protein dimer surface that is conserved whereas all other surface regions are not conserved. Calculation of the electrostatic potential surface shows a normal distribution of charges within the conserved patch. The total surface area of the conserved patch is 2,432 Å^2 which is considerably larger than the average 800 Å^2 buried in a single patch protein–protein interface.[29] Thus, it may be that the entire conserved patch is not involved in forming an interface or there may be two independent binding sites of half the area.

Because a bias is created by the multiple-sequence alignment of PSPTO_2114 with sequences from other *Pseudomonads* the validity of the analysis was tested by repeating the mapping exercise with an alignment of the four phosphonatase sequences derived from *Pseudomona*ds with the X-ray crystal structure of the *B. cereus* phosphonatase (PDB ID 1SWW). For comparison, a member of a different fold family, serralysin (an alkaline protease), from six *Pseudomonads* was similarly mapped onto the structure of the *P. aeruginosa* serralysin (PDB ID 1AKL). It is expected that the evolutionary drift among the *Pseudomonad* orthologs is similar and that the protein surface residues that do not perform specialized roles in function, folding or stability will diverge accordingly. The pattern of residue conservation on the surfaces of phosphonatase and serralysin (Supplementary material Figs. S1 and S2) shows extended regions of well-conserved clusters distributed on all sides of the molecule (typical surface area is 1,200–2,000 Å^2) in stark contrast to the pattern of residue conservation on the PSPTO_2114 surface. Thus, we conclude that the large conserved surface region of the PSPTO_2114 orthologs identifies a site of function, as there is no apparent structural role that these residues might fill.

Functional assessment derived from gene context

As revealed by computational analyses, it appears that the PA2803 subfamily has undergone a secondary lineage-specific loss of the C1 cap domain. The former cap region as well as the vestigial active site is on the face of the PSPTO_2114 dimer that is not conserved. This result can be contrasted to the results for the HAD member MDP-1, which has been identified as a protein sugar phosphatase,[30] where the conserved surface encircles the active site, acting as a docking surface for the protein substrate.[31] Because the electrostatic potential of the conserved surface is not positive, there is no evidence that the surface would be involved in DNA or RNA binding, as is typically mediated by conserved Arg/Lys residues.[32] Together, the sequence and structural evidence point to a non-catalytic role for PSPTO_2114 that involves the binding of a polar macromolecule.

In an attempt to further elucidate functional information for this subfamily of proteins we surveyed their gene neighborhoods. The products of genes co-occurring in the same neighborhood in multiple, sufficiently evolutionary diverse genomes tend to interact physically and

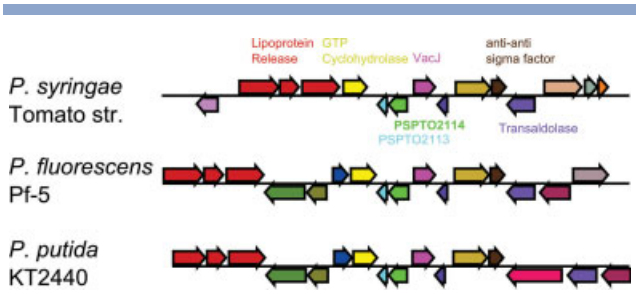

Figure 6

Gene context of PSPTO_2114 in Pseudomonads. Individual genes with similar function are identified.

functionally.[33–35] The PA2803 subfamily associates with several neighboring genes conserved across the Pseudomonads that encode proteins relating to the lipoprotein release pathway and the anti-anti sigma-stimulating stress response pathway (Fig. 6). However, given the close evolutionary distance of the genomes being investigated and the observation that many of these genes are transcribed in the opposite direction relative to the PSPTO_2114-like proteins, it seems unlikely these proteins form meaningful interactions. Nevertheless, a small, uncharacterized gene typically 90 residues in length, and tightly coupled immediately downstream (within ~100 base pairs and transcribed in the same direction) of the PSPTO_2114-like proteins was identified (PSPTO_2113). These two genes appear to form a small "gene island," set aside from other genes in the neighborhood, likely comprising a coregulated unit which in some cases is combined with a third member, a gene coding for a protein with an HD phosphohydrolase fold.[36] Secondary structure predictions of the protein product of the previously uncharacterized PSPTO_2113 revealed a tetra-α-helical fold which positions a stringently conserved histidine residue between the third and fourth helices (supplementary material Fig. S3), reminiscent of certain heme-binding proteins of the Heme NO Binding (HNOB) domain.[37] This newly characterized protein is found in other bacteria as a solo domain; suggesting that instead of acting as a sensor for the activation of another domain (as is the case of the heme-binding HNOB domains which sense nitric oxide for soluble guanyl cyclases), these proteins may be involved in the simple transporting/transferring of heme or other prosthetic groups. The PA2803 subfamily of proteins could potentially bind to proteins or ligands, which would interact with and possibly receive or transfer a heme protein to/from this tetra-helical family of proteins. However, this putative protein–protein binding function cannot currently be simulated because of the absence of a homologous structure on which to base a model of the tetra-helical PSPTO_2113. Therefore, creation of a docked complex of this protein with PSPTO_2114 is impossible. The nature of the hypothesized binding function must await future biochemical characterization of these proteins.

ACKNOWLEDGMENTS

Data were collected at beamline X12C of the National Synchrotron Light Source.

REFERENCES

1. Koonin EV, Tatusov RL. Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. J Mol Biol 1994;244:125–132.
2. Collet JF, van Schaftingen E, Stroobant V. A new family of phosphotransferases related to P-type ATPases. Trends Biochem Sci 1998;23:284.
3. Baker AS, Ciocci MJ, Metcalf WW, Kim J, Babbitt PC, Wanner BL, Martin BM, Dunaway-Mariano D. Insights into the mechanism of catalysis by the P-C bond-cleaving enzyme phosphonoacetaldehyde hydrolase derived from gene sequence analysis and mutagenesis. Biochemistry 1998;37:9305–9315.
4. Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L. Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. J Mol Biol 2006;361:1003–1034.
5. Morais MC, Zhang W, Baker AS, Zhang G, Dunaway-Mariano D, Allen KN. The crystal structure of *Bacillus cereus* phosphonoacetaldehyde hydrolase: insight into catalysis of phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily. Biochemistry 2000;39:10385–10396.
6. Li YF, Hata Y, Fujii T, Hisano T, Nishihara M, Kurihara T, Esaki N. Crystal structures of reaction intermediates of L-2-haloacid dehalogenase and implications for the reaction mechanism. J Biol Chem 1998;273:15035–15044.
7. Shin DH, Roberts A, Jancarik J, Yokota H, Kim R, Wemmer DE, Kim SH. Crystal structure of a phosphatase with a unique substrate binding domain from *Thermotoga maritima*. Protein Sci 2003; 12:1464–1472.
8. Selengut JD. MDP-1 is a new and distinct member of the haloacid dehalogenase family of aspartate-dependent phosphohydrolases. Biochemistry 2001;40:12704–12711.
9. Lu Z, Dunaway-Mariano D, Allen KN. HAD superfamily phosphotransferase substrate diversification: structure and function analysis of HAD subclass IIB sugar phosphatase BT4131. Biochemistry 2005;44:8684–8696.
10. Tremblay LW, Dunaway-Mariano D, Allen KN. Structure and activity analyses of *Escherichia coli* K-12 NagD provide insight into the evolution of biochemical function in the haloalkanoic acid dehalogenase superfamily. Biochemistry 2006;45:1183–1193.
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
12. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000; 302:205–217.
13. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
14. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 1976;72:248–254.

DOI 10.1002/prot

15. Appel RD, Bairoch A, Hochstrasser DF. A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. Trends Biochem Sci 1994;19:258–260.
16. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
17. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr 1999;55 (Part 4):849–861.
18. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54 (Part 5):905–921.
19. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 1997;53 (Part 3):240–255.
20. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004;60 (Part 12, Part 1):2126–2132.
21. Zhang G, Dai J, Wang L, Dunaway-Mariano D, Tremblay LW, Allen KN. Catalytic cycling in β-phosphoglucomutase: a kinetic and structural analysis. Biochemistry 2005;44:9404–9416.
22. Aravind L, Iyer LM, Koonin EV. Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr Opin Struct Biol 2006;16:409–419.
23. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234.
24. Zhang G, Morais MC, Dai J, Zhang W, Dunaway-Mariano D, Allen KN. Investigation of metal ion binding in phosphonoacetaldehyde hydrolase identifies sequence markers for metal-activated enzymes of the HAD enzyme superfamily. Biochemistry 2004;43: 4990–4997.
25. Laskowski RA, Watson JD, Thornton JM. From protein structure to biochemical function? J Struct Funct Genomics 2003;4:167–177.
26. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 2005;33 (Web Server issue):W89–W93.
27. Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 2004;5:150–163.
28. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci 2006;15:2120–2128.
29. Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins 2002;47:334–343.
30. Fortpied J, Maliekal P, Vertommen D, Van Schaftingen E. Magnesium-dependent phosphatase-1 is a protein-fructosamine-6-phosphatase potentially involved in glycation repair. J Biol Chem 2006; 281:18378–18385.
31. Peisach E, Selengut JD, Dunaway-Mariano D, Allen KN. X-ray crystal structure of the hypothetical phosphotyrosine phosphatase MDP-1 of the haloacid dehalogenase superfamily. Biochemistry 2004;43:12770–12779.
32. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 2003;31:7189–7198.
33. Aravind L. Guilt by association: contextual information in genome analysis. Genome Res 2000;10:1074–1077.
34. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol 2000; 18:609–613.
35. Huynen M, Snel B, Lathe W, Bork P. Exploitation of gene context. Curr Opin Struct Biol 2000;10:366–370.
36. Aravind L, Koonin EV. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. Trends Biochem Sci 1998; 23:469–472.
37. Iyer LM, Anantharaman V, Aravind L. Ancient conserved domains shared by animal soluble guanylyl cyclases and bacterial signaling proteins. BMC Genomics 2003;4:5.