Functional Identification and Structure Determination of Two Novel Prolidases from cog1228 in the Amidohydrolase Superfamily

ARTICLE *in* BIOCHEMISTRY · AUGUST 2010

Impact Factor: 3.02 · DOI: 10.1021/bi100897u · Source: PubMed

10 AUTHORS, INCLUDING:

Chengfu Xu
Guilin University of Aerospace Technology
23 PUBLICATIONS **288** CITATIONS
SEE PROFILE

Elena V Fedorov
Albert Einstein College of Medicine
62 PUBLICATIONS **2,151** CITATIONS
SEE PROFILE

J. Michael Sauder
Eli Lilly
63 PUBLICATIONS **2,389** CITATIONS
SEE PROFILE

Frank Raushel
Texas A&M University
312 PUBLICATIONS **9,420** CITATIONS
SEE PROFILE

Published in final edited form as:
Biochemistry. 2010 August 10; 49(31): 6791–6803. doi:10.1021/bi100897u.

Functional Identification and Structure Determination of Two Novel Prolidases from cog1228 in the Amidohydrolase Superfamily

Dao Feng Xiang[‡], **Yury Patskovsky**[Ψ], **Chengfu Xu**[‡], **Alexander A. Fedorov**[Ψ], **Elena V. Fedorov**[Ψ], **Abby A. Sisco**[‡], **J. Michael Sauder**[§], **Stephen K. Burley**[§], **Steven C. Almo**[Ψ,*], and **Frank M. Raushel**[‡,*]

[‡]Department of Chemistry, P.O. Box 30012, Texas A&M University, College Station, Texas 77842-3012

[Ψ]Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461

[§]Lilly Biotechnology Center, Eli Lilly and Company, 10300 Campus Point Dr., San Diego, California 92121

Abstract

Two uncharacterized enzymes from the amidohydrolase superfamily belonging to cog1228 were cloned, expressed and purified to homogeneity. The two proteins, Sgx9260c (gi|44242006) and Sgx9260b (gi|44479596), were derived from environmental DNA samples originating from the Sargasso Sea. The catalytic function and substrate profiles for Sgx9260c and Sgx9260b were determined using a comprehensive library of dipeptides and *N*-acyl derivative of L-amino acids. Sgx9260c catalyzes the hydrolysis of Gly-L-Pro, L-Ala-L-Pro and *N*-acyl derivatives of L-Pro. The best substrate identified to date is *N*-acetyl-L-Pro with a value of k_{cat}/K_m of 3×10^5 M^{-1} s^{-1}. Sgx9260b catalyzes the hydrolysis of L-hydrophobic L-Pro dipeptides and *N*-acyl derivatives of L-Pro. The best substrate identified to date is *N*-propionyl-L-Pro with a value of k_{cat}/K_m of 1×10^5 M^{-1} s^{-1}. Three dimensional structures of both proteins were determined by X-ray diffraction methods (PDB codes: 3MKV and 3FEQ). These proteins fold as distorted $(\beta/\alpha)_8$-barrels with two divalent cations in the active site. The structure of Sgx9260c was also determined as a complex with the N-methyl phosphonate derivative of L-Pro (PDB code: 3N2C). In this structure the phosphonate moiety bridges the binuclear metal center and one oxygen atom interacts with His-140. The α-carboxylate of the inhibitor interacts with Tyr-231. The proline side chain occupies a small substrate binding cavity formed by residues contributed from the loop that follows β-strand 7 within the $(\beta/\alpha)_8$-barrel. A total of 38 other proteins from cog1228 are predicted to have the same substrate profile based on conservation of the substrate binding residues. The structure of an evolutionarily related protein, Cc2672 from *Caulobacter crecentus*, was determined as a complex with the *N*-methyl phosphonate derivative of L-arginine (PDB code: 3MTW).

New genes are being sequenced at an exponential rate (1). Regrettably, functional annotations of the enzymes encoded by these genes have not kept pace with the rate at

[*]To whom correspondence may be addressed: (FMR) telephone: (979) 845-3373; fax: (979)-845-9452; raushel@tamu.edu. [*](SCA) telephone: (718) 430-2746; fax: (718)-430-8565; almo@aecom.yu.edu.

SUPPORTING INFORMATION AVAILABLE. Three supplementary figures showing Fo-Fc omit electron density maps for the active site of Sgx9260c complexed with *N*-methylphosphonate of L-proline (**Figure S1**), the active site of Sgx9260b complexed with sulfate (**Figure S2**) and the active site of Cc2672 complexed with *N*-methylphosphonate of L-arginine (**Figure S3**). This material is available free of charge via the Internet at http://pubs.acs.org.

which new genes have been sequenced, since it is far easier to sequence DNA than to experimentally determine the catalytic function of a given enzyme. This problem is compounded by over reliance on automated functional assignments based upon moderate to weak sequence similarities to proteins of known function. The significant number of proteins and enzymes of unknown, uncertain, or incorrect functions suggests that a substantial fraction of the biochemical landscape remains to be discovered. This wealth of new information offers an unprecedented opportunity to discover new catalytic transformations and the evolutionary relationships that define contemporary metabolism. Realization of this promise depends on a better understanding of the structural and chemical determinants of substrate binding and catalysis.

Our approach to this problem is directed at the elucidation of the structural determinants of the reaction profiles for members of the amidohydrolase superfamily (AHS). This superfamily was first recognized by Sander and Holm based upon the structural similarities among adenosine deaminase, urease, and phosphotriesterase (2–3). All of the AHS members fold as distorted $(\beta/\alpha)_8$-barrels and contain either a mononuclear or binuclear metal center within their active sites. For most of the AHS, the metal center functions to activate water for nucleophilic attack and/or enhance the electrophilic character of the target substrate. Reactions catalyzed by this superfamily include hydrolysis of ester (4–5) and amide bonds (6–10), hydration of double bonds (11), aldose/ketose isomerizations (12–14) and decarboxylations (15–17). Thus far, over 12,000 unique proteins sequences have been identified as members of the AHS from the first 1000 completely sequenced bacterial genomes (18). These sequences have been organized into 24 separate clusters of orthologous groups (COG) by the NCBI.

One of the largest of these clusters of orthologous groups in the AHS is cog1228 with approximately 1000 unique sequences. Shown in Figure 1 is a Cytoscape network analysis representation of this COG at a BLAST E-value cutoff of 10^{-80} (19). Enzymes in cog1228 have been reported to catalyze the hydrolysis of imidazolone propionate (HutI), organophosphate triesters, and dipeptides. We have previously determined three-dimensional structures of HutI from *Agrobacterium tumefaciens* (PDB code: 2PUZ) and from an organism isolated as part of the Global Ocean Sampling Project from the Sargasso Sea (PDB code: 2Q09) (20–21). These two proteins contain a single divalent cation in the active site and are localized in subcluster **1** in Figure 1. In addition, the structure of HutI from *Bacillus subtilis* (PDB code: 2G3F) has been determined (22) and this protein is present within subcluster **2** of the same figure. More recently, we have determined the substrate profiles (7–8) and X-ray structures (PDB codes: 3BE7, 3DUG, and 2QS8) for a group of peptidases that catalyze the hydrolysis of L-Xaa-L-Arg/Lys dipeptides (subcluster **4**) and another group that is specific for the hydrolysis of L-Xaa-L-hydrophobic dipeptides (subcluster **5**).

In this investigation we have examined substrate profiles for the group of enzymes found within subcluster **7** in Figure 1. Two of these proteins catalyze the hydrolysis of L-Xaa-L-Pro dipeptides and *N*-acyl derivatives of L-proline. X-ray structures of these two proteins have been determined, one of which is a complex with the *N*-methyl phosphonate derivative of L-Pro, a potent mimic of the tetrahedral intermediate that would be formed during the hydrolysis of *N*-acetyl-L-proline. Our X-ray structures of the L-Xaa-L-Arg/Lys, L-Xaa-L-hydrophobic, and L-Xaa-L-Pro dipeptidases have been utilized to formulate a model of the structural determinants for substrate recognition within cog1228.

Materials and Methods

Materials

LB broth was purchased from TPI Research Products International Corp. Chromatographic columns and resins were purchased from GE Healthcare. ICP standards were obtained from Inorganic Ventures Inc. *N*-acetyl-L-proline and Gly-L-proline were from Sigma. The dipeptide L-Pro-L-Pro was purchased from BACHEM. The tripeptide L-Gly-L-Gly-L-Pro was purchased from Aroz Technologies LLC. *N*-propionyl-L-proline was purchased from ChemBridge. Resins and protected amino acids used for solid phase peptide synthesis were purchased from Calbiochem. All other buffers, purification reagents, and chemicals used in this investigation were obtained from Sigma, unless otherwise stated. The *N*-acetyl-L-Xaa and L-Zaa-L-Xaa dipeptide libraries were synthesized as reported previously (7,10). In general, for the dipeptide libraries utilized in this investigation there was a fixed amino acid at the N-terminus (L-Zaa) and a mixture of amino acids at the C-terminus (L-Xaa)

Synthesis of L-Xaa-L-Pro Library

Preloaded *N*-Fmoc-L-Pro Wang resin (0.19 mmol) with DMF (3 mL) was shaken in a syringe for 30 minutes. After removal of the DMF, 3 mL of 20% piperidine in DMF were added and shaken for 30 minutes. The beads were washed with DMF (4 × 5 mL) and a mixture of 19 *N*-Fmoc protected amino acids was added. The amino acids included 0.01 mmol each of the following amino acids: *N*-Fmoc-L-Ala-OH, *N*-Fmoc-L-Arg(Mtr)-OH, *N*-Fmoc-L-Asn(Trt)-OH, *N*-Fmoc-L-Asp(OtBu)-OH, *N*-Fmoc-L-Glu(OtBu)-OH, *N*-Fmoc-L-Gln(Trt)-OH, *N*-Fmoc-L-Gly-OH, *N*-Fmoc-L-His(Trt)-OH, *N*-Fmoc-L-Ile-OH, *N*-Fmoc-L-Leu-OH, *N*-Fmoc-L-Lys(Boc)-OH, *N*-Fmoc-L-Met-OH, *N*-Fmoc-L-Phe-OH, *N*-Fmoc-L-Pro-OH, *N*-Fmoc-L-Ser(Trt)-OH, *N*-Fmoc-L-Thr(Trt)-OH, *N*-Fmoc-L-Trp(Boc)-OH, *N*-Fmoc-L-Tyr(tBu)-OH, and *N*-Fmoc-L-Val-OH. To this mixture was added $HOBt{\cdot}H_2O$ (35.2 mg, 0.23 mmol), DIC (28.7 mg, 0.23 mmol) and DMAP (1.0 mg) in DMF (3 mL) and then shaken for two days. The reagents were removed and the beads washed with DMF (4 × 5 mL), dichloromethane (4 × 5 mL), methanol (4 × 5 mL) and dried for several hours. To the dried beads was added cocktail R (2.0 mL of TFA/thioanisole/EDT/anisole (v/v, 90/5/3/2)) and shaken for 3 hours. To obtain the *N*-Fmoc-L-Xaa-L-Pro amino acid library, the cocktail was removed under reduced pressure and dried overnight at 50 °C. The *N*-Fmoc-L-Xaa-L-Pro library was subsequently stirred with 20% piperidine in DMF (5 mL) for 30 minutes. The L-Xaa-L-Pro dipeptide library was washed with $EtOAc/Et_2O$ (v/v, 1/5) and then dried overnight.

Synthesis of N-Butyryl-L-Proline

The synthesis of *N*-butyryl-L-proline (**3**) was initiated with the preparation of the methyl ester of *N*-butyryl-L-proline (**2**) according to the procedure outlined in Scheme 1. In a 100 mL flask was added methyl L-proline hydrochloride (**1**) (1.50 g, 9.0 mmol), triethylamine (0.91 g, 9.0 mmol), and chloroform (20 mL). The mixture was stirred for 5 minutes, and then DCC (2.22 g, 10.8 mmol), butyric acid (0.95 g, 10.8 mmol), and DMAP (1.0 mg) were added. After the mixture stirred for 3 hours at room temperature, another 50 mL of chloroform were added and then washed with aqueous HCl (2 × 10 mL), brine (2 × 20 mL) and dried. After removal of the solvent, the compound was purified by silica gel chromatography after elution with ethyl acetate/chloroform. An oily product was obtained with a yield of 85% (1.52 g).

The NMR data for compound **2** are as follows: ^{1}H NMR (300 MHz, $CDCl_3$): 4.41 (0.8H, dd, J = 3.3 7.2 Hz, C$\underline{H}$), 4.36 (0.2H, dd, J = 3.3, 7.2Hz, C$\underline{H}$), 3.68 (0.6H, s, $CO_2C\underline{H}_3$), 3.64 (2.4H, s, $CO_2C\underline{H}_3$), 3.62-3.41 (2H, m, $NC\underline{H}_2$), 2.40-182 (6H, m, $CH_3CH_2C\underline{H}_2CO$,$CHC\underline{H}_2C\underline{H}_2$), 1.68-1.52 (2H, m, $CH_3C\underline{H}_2CH_2CO$), 0.89 (2.4H, t, J = 7.5

Hz, C<u>H</u>$_3$CH$_2$CH$_2$CO), 0.85 (0.6H, t, J = 7.5 Hz, C<u>H</u>$_3$CH$_2$CH$_2$CO). ^{13}C NMR (75.4 MHz, $CDCl_3$): 17.95, 172.81, 171.89, 171.83, 59.30, 58.49, 52.45, 52.03, 46.94, 46.19, 36.24, 36.14, 31.42, 29.17, 24.77, 22.52, 18.23, 18.00, 13.85, and 13.79 ppm.

For the preparation of compound **3**, methyl *N*-butyl-L-proline (**2**) (1.5 g, 7.5 mmol), water (10 mL) THF (5 mL) were added to a 100 mL flask and then a mixture of sodium hydroxide (0.45 g, 11.3 mmol) and water (10 mL) was added drop wise. After the mixture stirred at room temperature for 1 hour, the THF was removed and the water phase neutralized to a pH value below 1.0. The final compound (**3**) was extracted with chloroform and dried. The NMR data were as follows: ^{1}H NMR (300 MHz, $CDCl_3$): 11.36 (1H, s, CO$_2$<u>H</u>), 4.50 (0.8H, dd, J = 3.3, 7.2 Hz, C<u>H</u>), 4.37 (0.2H, dd, J = 3.3, 7.2 Hz, C<u>H</u>), 3.60-3.44 (2H, m, NC<u>H</u>$_2$), 2.29-1.95 (6H, m, CH$_3$CH$_2$C<u>H</u>$_2$CO,CHC<u>H</u>$_2$C<u>H</u>$_2$), 1.66-1.58 (2H, m, CH$_3$C<u>H</u>$_2$CH$_2$CO), 0.94 (2.4H, t, J = 7.2 Hz, C<u>H</u>$_3$CH$_2$CH$_2$CO), 0.87 (0.6H, t, J = 7.2Hz, C<u>H</u>$_3$CH$_2$CH$_2$CO). ^{13}C NMR (75.4 MHz, $CDCl_3$): 172.95, 172.81, 171.89, 171.83, 59.30, 58.49, 52.45, 52.03, 46.94, 46.19, 36.24, 36.14, 31.42, 29.17, 24.77, 22.52, 18.23, 18.00, 13.85, 13.79 ppm.

Synthesis of N-Methylphosphonate-L-Proline

N-methylphosphonate-L-proline (**5**) was prepared from L-proline according to the procedure outlined in Scheme 2. In a 100 mL flask was added 1.0 g of methyl L-proline hydrochloride (**1**), triethylamine (3.0 mL), and chloroform (20 mL). The mixture was cooled with ice-water and added drop wise to a mixture of methyl methylchlorophosphonate (*34*) (0.93 g, 7.2 mmol) in chloroform (10 mL). After the mixture was stirred for 3 hours at room temperature, 50 mL of chloroform were added and then washed with aqueous HCl (2 × 10 mL), brine (2 × 20 mL) and dried. After removal of the solvent, the residue was purified by silica gel chromatography after elution with ethyl acetate/methanol (15/1). An oily product (**4**) was obtained in a yield of 75% (1.0 g).

The analytic data for compound **4** are as follows: ^{1}H NMR (300 MHz, $CDCl_3$): 4.08 (0.5H, dt, J = 6.9, 9.6 Hz, C<u>H</u>), 3.95 (0.5H, dt, J = 6.9, 9.6 Hz, C<u>H</u>), 3.42 (3H, s, CO$_2$C<u>H</u>$_3$), 3.41 (1.5H, d, J = 11.1 Hz, POC<u>H</u>$_3$), 3.28 (1.5H, d, J=11.1 Hz, POC<u>H</u>$_3$), 3.00-2.85 (2H, m, NC<u>H</u>$_2$), 2.00-1.60 (4H, m, CHC<u>H</u>$_2$C<u>H</u>$_2$), 1.18 (3H, dd, J = 16.8, 34.2 Hz, PC<u>H</u>$_3$). ^{3}P NMR (121.4 MHz, $CDCl_3$): 33.43, 32.23ppm. ^{13}C NMR (75.4 MHz, $CDCl_3$): 174.15, 174.07, 59.58, 59.52, 58.87, 58.80, 51.50, 51.44, 50.08, 49.90, 49.27, 49.19, 45.86, 45.79, 45.69, 45.62, 30.89, 30.80, 30.55, 30.45, 24.83, 24.75, 24.72, 24.66, 11.80, 11.68, 10.02, and 9.90 ppm. MS (ESI positive mode): found; 222.10 $(M+H)^+$, 244.08 $(M+Na)^+$ and $C_8H_{16}NO_4P$ required: 221.08.

To make **5** compound **4** (1.0 g, 4.5 mmol), water (6 mL) and THF (4 mL) were added together in a 50 mL flask, and then a mixture of sodium hydroxide (0.54g, 13.6 mmol) in water (10 mL) was added drop wise. After stirring at room temperature overnight, the solvent was removed and the solid residues washed with a mixture of methanol and ethyl acetate. The yield of **5** was 0.69 g (65%).

The analytical data for compound **5** are as follows: ^{1}H NMR (300 MHz, D_2O): 3.60 (1H, dt, J = 2.4, 6.9 Hz, C<u>H</u>), 3.05-2.95 (2H, m, NC<u>H</u>$_2$), 2.03-1.89 (1H, m, one of CHC<u>H</u>$_2$CH$_2$), 1.71-1.55 (3H, m, CHC<u>H</u>$_2$C<u>H</u>$_2$), 1.03 (3H, d, J = 15.6 Hz, PC<u>H</u>$_3$). ^{31}P NMR (121.4 MHz, D_2O): 28.67 ppm. ^{13}C NMR (75.4 MHz, D_2O): 184.53, 184.50, 62.06, 62.00, 47.28, 47.24, 32.55, 32.48, 25.38, 25.30, 12.99, 11.35 ppm. MS (ESI negative mode): found: 192.09 $(M-2Na+H)^-$ and $C_6H_{10}NNa_2O_4P$ required: 237.01.

Preparation of N-Methylphosphonate-L-Arginine

Compound **8** was prepared according to the procedure outlined in Scheme 3. In a 100 mL flask was added methyl L-(di-Cbz)-arginine hydrochloride (**6**) (1.00 g, 2.0 mmol),

triethylamine (1.0 mL), chloroform (20 mL) and the mixture cooled in an ice bath. The solution was added drop wise to a mixture of methyl methylchlorophosphonate (*34*) (0.38 g, 3.0 mmol) in chloroform (10 mL). After stirring at room temperature for 3 hours another 50 mL of chloroform were added. This solution was washed with aqueous HCL (2 × 10 mL), brine (2 × 20 mL) and then dried. The residue was purified by chromatography on silica gel with a mixture of ethyl acetate/methanol (20/1). The oily product (**7**) was obtained in a yield of 71% (0.71 g).

The analytical data for **7** are as follows: ^{1}H NMR (300 MHz, $CDCl_3$): 9.50 (1H, s, CbzN$\underline{H}$), 9.30 (1H, s, CbzN$\underline{H}$), 7.43-7.29 (10H, m, $2C_6\underline{H}_5CH_2$), 5.26 (2H, s, $C_6H_5C\underline{H}_2$), 5.15 (2H, s, $C_6H_5C\underline{H}_2$), 4.06-3.95 (3H, m, N$\underline{H}$, NC$\underline{H}_2$),3.66 (3H, s, $CO_2C\underline{H}_3$), 3.55 (3H, dd, J = 6.9, 12.0 Hz, POC$\underline{H}_3$), 3.35(1H, dt, J = 6.9, 15.3 Hz, C$\underline{H}$), 1.80-1.60 (4H, m, CHC$\underline{H}_2$C$\underline{H}_2$), 1.39 (3H, dd, J = 13.5, 16.8 Hz, PC$\underline{H}_3$). ^{31}P NMR (121.4 MHz, $CDCl_3$): 33.13, 32.41 ppm. ^{13}C NMR (75.4 MHz, $CDCl_3$): 174.19, 174.05, 174.00, 163.89, 160.58, 155.83, 136.91, 134.67, 128.93, 128.89, 128.48, 128.36, 128.14, 128.11, 127.97, 69.00, 66.99, 53.56, 53.41, 52.31, 50.52, 50.43, 50.33, 44.26, 31.15, 31.09, 24.58, 24.55, 14.35, 14.30, 12.60, and 12.53 ppm.

In a 50 mL flask was added compound **7** (0.7 g, 1.4 mmol), water (6 mL) and methanol (4 mL). This mixture was added drop wise to a solution of sodium hydroxide (0.17 g, 4.2 mmol) and water (8 mL). After stirring at room temperature overnight ethyl acetate (5 mL) was added to quenched the reaction. After removal of the solvent the residue was dissolved in water (15 mL) and 30 mg of palladium on carbon was added and stirred under a hydrogen atmosphere overnight. A solid residue was obtained that contained 2.8 equivalents of acetate after removal of the catalyst and solvent.

The analyticaL data for **8** are as follows: ^{1}H NMR (300 MHz, D_2O): 3.35-3.28 (1H, m, C$\underline{H}$), 3.05 (2H, t, J = 6.9 Hz, NC$\underline{H}_2$), 1.79 (8.5H, s, $C\underline{H}_3O_2Na$), 1.50-1.42 (4H, m, CHC$\underline{H}_2$C$\underline{H}_2$), 1.05 (3H, d, J = 15.6 Hz, PC$\underline{H}_3$). ^{31}P NMR (121.4 MHz, D_2O): 27.43 ppm. ^{13}C NMR (75.4 MHz, D_2O): 184182.37, 181.67, 161.84, 156.81, 56.34, 40.97, 32.26, 32.18, 24.42, 23.46, 15.27, and 13.66 ppm. MS (ESI negative mode); found: 251.09 $(M-2Na+H)^-$ and $C_7H_{15}N_4Na_2O_4P$ required: 296.06.

Cloning, Expression, and Purification of Sgx9260b and Sgx9260c

The gene sequences of Sgx9260b (gi|44479596) and Sgx9260c (gi|44242006) were obtained from the Sargasso Sea as part of the Global Ocean Sampling Project. The target selection process is described in Pieper *et. al* (23). Codon-optimized synthetic DNA (Codon Devices, Inc.) was used as a template and cloned into a custom TOPO-isomerase vector, pSGX3(BC), supplied by Invitrogen. Forward and reverse primers for Sgx9260b were ACTACGTTTCTGTTCCGTAATGG and CATGACCCTCCAGTTCATTCAC, and for Sgx9260c were ACTATTACGGTGCTGCAGGG and CACGCCCACCAGCCTGTCTTTTAACG, respectively). The clones encode Met-Ser-Leu (N-terminus) followed by the PCR product and finally Glu-Gly-His_6 (C-terminus). Miniprep DNA was transformed into BL21(DE3)-Codon+RIL expressions cells (Stratagene), expressed, and made into a 30% glycerol stock for large scale fermentation.

Expression clones of Sgx9260b and Sgx9260c were cultured using a ZYP-5052 auto-induction media. Overnight cultures were prepared from glycerol stocks that were derived from fresh transformants in LB media. The next morning, overnight cultures were used to inoculate ZYP-5052 auto-induction media. A 5–10% inoculum was achieved by adding 12–25 mL of overnight culture to 250 mL ZYP-5052 media containing 90 μg/mL kanamycin and 30 μg/mL chloramphenicol in a 2 L baffled shake flask. Cultures were grown at 37 °C for 5 hours at 225 rpm agitation. The temperature was then switched to 22 °C and grown overnight for 18 hours to allow for induction of the intended protein. Selenomethionine-

labeled Sgx9260c was produced for crystallographic structure determination using High Yield SeMet media (Orion Enterprises, Inc, Northbrook, IL). Cells were harvested using standard centrifugation for 10 minutes at 6000 rpm and frozen at −80 °C.

Cells were lysed in 20 mM TRIS pH 8.0, 0.5 M NaCl, 25 mM imidazole, and 0.1% Tween 20 by sonication. The cellular debris was removed by centrifugation for 30 minutes (39,800 × g). The supernatant was collected and incubated with 10 mL of a 50% slurry of Ni-NTA Agarose (Qiagen) for 30 minutes with gentle stirring. The sample was then poured into a drip column and washed with 50 mL of wash buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 10% glycerol, and 25 mM imidazole) to remove unbound proteins. The protein of interest was eluted using 25 mL of elution buffer (wash buffer with 500 mM imidazole). Fractions containing the protein were pooled and further purified by gel filtration chromatography on a GE Healthcare HiLoad 16/60 Superdex 200 prep grade column pre-equilibrated with gel filtration buffer (10 mM HEPES, pH 7.5, 150 mM NaCl, 10% glycerol and 5 mM DTT). Fractions containing the protein of interest were combined and concentrated to 10 mg/mL by centrifugation in an Amicon Ultra-15 10,000 Da MWCO centrifugal filter unit. Electrospray mass spectrometry was used to obtain an accurate mass of the purified protein to confirm its identity. The expression plasmids are available through the PSI Material Repository (NYSGXRC clone ID 9260b1BCt8p1 and 9260c1BCt10p1), and the DNA sequences and experimental details are available in the Protein Expression Purification and Crystallization Database (PepcDB) as TargetID “NYSGXRC-9260c”.

Purification of Sgx9260b and Sgx9260c for Kinetic Analysis

For purification of Sgx9260c, *E. coli* BL21(DE3) cells harboring the plasmid containing the gene for Sgx9260c were grown in LB media with 50 μg/mL kanamycin at 30°C. When the optical density of the cells reached an OD_{600} of ~0.6, 1.0 mM zinc acetate was added to the culture and expression of Sgx9260c was then induced with 0.5 mM isopropyl β-thiogalactoside (IPTG). The cells were allowed to grow at 16–19 °C for an additional 18 hours after IPTG induction and then harvested by centrifugation (6000 rpm for 10 minutes). The cell lysate from 15 g of cells was obtained by centrifugation after sonication in 75 mL of binding buffer (20 mM HEPES, pH 7.9, 0.5 M NaCl, and 5 mM imidazole) containing 100 μg/mL phenylmethanesulfonyl fluoride (PMSF) at 0 °C. Sgx9260c was expressed in *E. coli* BL21 (DE3) cells but the enzyme was found predominantly in the insoluble pellet after centrifugation. The clarified cell extract containing a small amount of the target protein was applied to a 24 mL column of chelating Sepharose Fast Flow resin (Amersham Biosciences) charged with Ni^{2+} and pre-equilibrated with binding buffer. The column was washed thoroughly with binding buffer until the absorbance of the flow-through at 280 nm did not change. The column was then washed with 100 mL of wash buffer (20 mM HEPES, pH 7.9, 0.5 M NaCl, and 60 mM imidazole). The His-tagged Sgx9260c was eluted with 150 mL of elution buffer (20 mM HEPES, pH 7.9, 0.25 M NaCl, and 0.5 M imidazole). The protein solution was concentrated and then applied to HiLoad 16/60 Superdex 200 gel filtration column for further purification. The protein that was eluted from the gel filtration column was more than 95% pure as judged by SDS-PAGE analysis. Approximately 1–2 mg of protein were obtained from 15 g of cell paste.

Sgx9260b was expressed in *E. coli* BL21 (DE3) cells and the enzyme was found predominantly in the insoluble pellet after centrifugation of the lysed cell mixture. Sgx9260b was purified using the same purification procedures as for Sgx9260c. About 1 mg of pure protein was obtained from 15 g of cells. Cc2672 was expressed and purified to homogeneity as previously described (21).

Apparent Molecular Weight Determination

A HiLoad 16/60 Superdex 200 gel filtration column was calibrated with standard molecular weight markers (MW-GF-1000 from Sigma) in 50 mM HEPES, pH 7.5. Sgx9260c or Sgx9260c were applied to the column and eluted with the calibration buffer for the protein markers at a flow rate of 1.0 mL per minute. The elution volume for the protein was used to calculate the apparent molecular weight of the protein.

Metal Analysis

The metal content of the purified proteins was determined with a Perkin-Elmer Analyst 700 atomic absorption spectrometer and by inductively coupled plasma emission-mass spectrometry (ICP-MS). The purified Sgx9260c contained an average of 0.7 equivalents of Zn^{2+} and the Zn^{2+} content was increased to 1.1 equivalents/subunit after incubating the as-purified protein with 2 equivalents of Zn^{2+} and 10 equivalents of $NaHCO_3$ at pH 7.5 for 24 hours. The loosely bound Zn^{2+} was removed by passage through a PD-10 desalting column. Sgx9260c with an average of 1.1 equivalents of Zn per subunit was used for all kinetic assays. Sgx9260b contained an average of 1.2 equivalents of Zn^{2+} and Cc2672 contained 2 equivalents of Zn^{2+}.

Assay Methods for Peptidase Activity

A modified colorimetric Cd-ninhydrin assay method was used to screen the catalytic activity of Sgx9260c and Sgx9260b using various dipeptide libraries (*35*). The same method was used to measure the hydrolysis of tripeptides and the hydrolysis of *N*-formyl-, *N*-acetyl-, *N*-propionyl-, and *N*-butyryl-L-proline. The absorbance at 440 nm of the yellow compound formed at 80 °C between the Cd-ninhydrin reagent and free L-proline released from the enzymatic reaction was recorded with a SPECTRAmax plate reader from Molecular Devices, while the formation of the red compound at 80 °C between the Cd-ninhydrin reagent and the other free amino acids released from the catalytic reaction was monitored at 507 nm as described previously (7–8). A standard curve for the colored compound formed between each amino acid and the Cd-ninhydrin reagent was measured under the identical conditions as that for the enzymatic activity assay. Quantitative analysis of the librated free amino acids was conducted by the Protein Chemistry Laboratory at Texas A&M University.

Screening of Dipeptide Libraries

The entire set of L-Zaa-L-Xaa dipeptide libraries (361 compounds in all) was used to examine the catalytic activity of Sgx9260c using the Cd-ninhydrin assay as described previously (7–8). Each dipeptide library consisted of a mixture of dipeptides with a fixed L-amino acid at the N-terminus (L-Zaa) but 19 different amino acids at the C-terminus (L-Xaa). L-Cysteine was not included in any of the libraries. Screening of the dipeptide libraries was initiated by mixing a fixed concentration of the dipeptide library (each dipeptide at ~0.1 mM) and various amounts of enzyme over a concentration range of 10–1000 nM. All reactions were conducted in 50 mM HEPES, pH 8.0, in 96-well microplates. The reaction mixture was incubated at 30 °C for a fixed time period and then the reaction was quenched by adding the Cd-ninhydrin reagent. The quenched reaction mixture was incubated for 5 minutes at 80 °C for color development. The absorbance of the red compound thus formed was measured at 507 nm. Control reactions without enzyme or without the dipeptide library were conducted simultaneously. The relative rates were determined by fitting the data to equation 1 where *y* is the change in absorbance at 507 nm, x is the concentration of enzyme, and *k* is the relative rate constant. In these enzyme-course measurements, the enzyme concentration was varied but the reaction time was fixed.

$$y=a\left(1-e^{-kx}\right) \quad (1)$$

Kinetic Determinations

Selected dipeptides were used as substrates for measurement of the kinetic parameters for Sgx9260c and Sgx9260b. All of the kinetic reactions were conducted in 50 mM HEPES, pH 8.0. The time courses and initial reaction rates were obtained by monitoring product formation as a function of time using the Cd-ninhydrin reagent. The kinetic constants, k_{cat}, K_m, and k_{cat}/K_m were determined by fitting the initial rate data to equation 2 where v is the initial velocity, k_{cat} is the turnover number, E_t is the enzyme concentration, A is the substrate concentration, and K_m is the Michaelis constant.

$$v/E_t=k_{cat}A/(K_m+A) \quad (2)$$

Inhibition Study

The inhibitory properties of the *N*-methyl phosphonate derivative of L-proline (**5**) on the catalytic activity of Sgx9260c were determined using *N*-acetyl-L-proline as the substrate. The Cd-ninhydrin reagent was used to measure product formation as described previously. All enzymatic reactions were conducted in 50 mM HEPES, pH 8. The concentrations of *N*-acetyl-L-proline and inhibitor **5** were varied in the range of 0.1–10 mM and 0.1–100 μM, respectively. The initial rates for substrate hydrolysis in the presence of the inhibitor at different concentrations of the substrate were fit to equation 3 for competitive inhibition where v is the initial velocity, A is the substrate concentration, I is the inhibitor concentration, k_{cat} is the turnover number, E_t is the enzyme concentration, K_m is the Michaelis constant and K_i is the competitive inhibition constant. .

$$v/E_t=k_{cat}A/(K_m(1+I/K_i)+A) \quad (3)$$

Crystallization, X-ray Data Collection, Structure Determination and Refinement for Sgx9260c and Sgx9260b

Diffraction quality crystals of protein Sgx9260b were obtained by sitting drop vapor diffusion method. In brief, 1 μL of protein solution (38 mg/mL, in the presence of 10% glycerol) was mixed with 1 μL of mother liquor composed of 100 mM ammonium sulfate and 12–20% of PEG8000. Plate-shaped crystals that appeared after 6–8 days were briefly soaked in mother liquor supplemented with 20% ethylene glycol as a cryoprotectant before flash-cooling in liquid nitrogen. Crystals of protein Sgx9260c were grown as described above by mixing 1 μL of protein solution (10 mg/mL) with an equal volume of a reservoir solution (100 mM MES, pH 6.0, 200 mM sodium acetate and 6–12% PEG8000). Plate-shaped crystals were briefly soaked into mother liquor supplemented with 25% DMSO as a cryoprotectant and then flashed-cooled in liquid nitrogen. To obtain the crystalline complex between Sgx9260c and *N*-methylphosphonate-L-proline, crystals of unliganded protein were grown as described above for unliganded Sgx9260c and soaked in a reservoir solution supplemented with 20% DMSO and 5 mM of the ligand. After 15–20 minutes of incubation, protein crystals were flash-cooled in liquid nitrogen.

X-ray diffraction data for Sgx9260b and Sgx9260c were collected from previously frozen crystals at the Advanced Photon Source beamline 31ID (Argonne National Laboratory, Argonne, IL) under standard cryogenic conditions. X-ray diffraction data for the liganded Sgx9260c structure were collected at the National Synchrotron Light Source (Brookhaven

National Laboratory, NY) beamline X29A. All diffraction data were indexed, integrated and scaled with the HKL2000 software package. Data collection and refinement statistics are provided in Table 1.

The structure of Sgx9260b was determined by the single-wavelength anomalous difference (SAD) method using diffraction data collected at the selenium absorption edge (λ = 0.979 Å) with the SHELX (CCP4 program suit, version 6.1) and SOLVE/RESOLVE (https://solve.lanl.gov/) software packages. The model was further refined by REFMAC 5.3 (CCP4 package suit) as described below.

The structure of Sgx9260c was subsequently determined via molecular replacement (MR) with MOLREP (CCP4 program suit) using the coordinates of Sgx9260b (PDB code: 3MKV) as a search model. The structure of the Sgx9260c bound to *N*-methylphosphonate-L-proline was determined by MR using the coordinates for substrate-free Sgx9260c (PDB code: 3FEQ). Both types of Sgx9260c crystals, liganded and substrate-free, were isomorphous and exhibited diffraction consistent with the triclinic space group P1. Each asymmetric unit (or the unit cell in this case) contains 16 protein monomers arranged as two virtually identical homo-octamers with 422 point symmetry, allowing for the application of non-crystallographic symmetry restraints (NCS) throughout the early stages of refinement. The atomic model for Sgx9260b corresponds to a homo-octamer of the same topology as for Sgx9260c. All atomic models were refined using REFMAC 5.3 (CCP4 package suit) and then adjusted interactively using the program COOT (24). The Sgx9260b structure (PDB code: 3MKV) was refined to a resolution limit of 2.4 Å, with R_{work} and R_{free} of 17.0% and 23.0%, respectively. The Sgx9260c substrate-free structure (PDB: 3FEQ) was refined to a resolution limit of 2.61 Å with R_{work} and R_{free} of 23.8% and 28.1%, respectively. The Sgx9260c structure with the bound *N*-methylphosphonate-L-proline (PDB code: 3N2C) was refined to a resolution limit of 2.8 Å with R_{work} and R_{free} of 22.0% and 27.3%, respectively. Final model stereochemistry was evaluated using PROCHECK (25). Figures were prepared with PYMOL (26).

Crystallization and Data Collection of Cc2672

Crystals of Cc2672 from *Caulobacter crecentus* bound to *N*-methyl phosphonate-L-arginine were grown via hanging drop vapor diffusion at room temperature versus 2.0 M ammonium sulfate, 0.1 M Bis-Tris (pH 5.5), and 1.0 mM $ZnCl_2$. The protein solution contained Cc2672 (10.1 mg/mL) in 20 mM Hepes (pH 8.0), 1.0 mM $ZnCl_2$, and 40 mM N-methyl-phosphonate-L-arginine. Crystals appeared in 8–9 days and exhibited diffraction consistent with the space group I432, with one protomer per asymmetric unit (Table 1). Prior to data collection, the crystals were cryoprotected via addition of glycerol to a final concentration of 20% (v/v). After incubation for ~10 seconds, crystals were flash-cooled in a refrigerated nitrogen gas stream. Diffraction data were collected at the NSLS X4A beamline (Brookhaven National Laboratory). Diffraction intensities were integrated and scaled with programs DENZO and SCALEPACK (27). Data collection statistics are provided in Table 1.

Structure Determination and Model Refinement for Cc2672

The structure of Cc2672 was determined by molecular replacement with the fully automated molecular replacement pipeline BALBES (28). The protein part of the Zn-dependent arginine carboxypeptidase (PDB code: 3BE7) was used as the search model yielding a partially refined structure of the Cc2672-inhibitor complex. Subsequent iterative cycles of refinement were performed including: model rebuilding with COOT (29), refinement with PHENIX (30), and automatic model rebuilding with ARP (31). Final crystallographic refinement statistics for the Cc2672 crystal complex are provided in Table 1. The atomic

model of Cc2672 bound to *N*-methyl-phosphonate-L-arginine was refined to a resolution of 1.7 Å with R_{cryst} of 18% and R_{free} of 20.3%, respectively. The final refined structure consists of residues 27–429, two well-defined Zn^{2+} ions per protomer, and one well-defined ligand, *N*-methyl-phosphonate-L-arginine inhibitor in the active site.

RESULTS

Purification of Sgx9260c and Sgx9260b

The gene for Sgx9260c was synthesized and codon optimized based on an environmental DNA sequence originally isolated from the Sargasso Sea. The predicted molecular weight of a single subunit of Sgx9260c is 44.6 kDa and the apparent molecular weight of Sgx9260c in solution was found to be 360 ± 25 kDa based on the elution volume from a calibrated gel filtration column (data not shown). These results are consistent with an oligomeric state as a homo-octamer in solution. The gene for Sgx9260b was also synthesized and the protein expressed and purified using the same procedures used to isolate Sgx9260c. The predicted molecular weight of a single subunit of Sgx9260b is 45.3 kDa and the apparent molecular weight determined from gel filtration measurements was 367 ± 26 kDa, again consistent with a homo-octamer.

Screening of Dipeptide Libraries

Nineteen L-Zaa-L-Xaa dipeptide libraries were used to screen for dipeptidase activity catalyzed by Sgx9260c. The only libraries showing measureable turnover were Gly-L-Xaa and L-Ala-L-Xaa. Significantly lower turnover rates were found for the L-Leu-L-Xaa library. To determine the C-terminal amino acid specificity of Sgx9260c for the hydrolysis of dipeptides, quantitative amino acid analysis was employed to identify the specific amino acids liberated after incubation of the enzyme with dipeptide libraries. Sgx9260c was assayed with the L-Ala-L-Xaa and L-Gly-L-Xaa dipeptide libraries. With the L-Ala-L-Xaa library, L-proline was the predominant amino acid detected. Glycine, L-methionine, L-serine, L-valine, L-phenylanaline and L-leucine were also detected using this library, but the relative reaction rates were less than 4% of that for L-proline. With the Gly-L-Xaa library, L-proline was again the dominant C-terminal amino acid detected. Other amino acids detected included L-alanine, L-methionine, L-serine, L-valine, L-phenylalanine and L-leucine but the relative rates of hydrolysis were less than 3% of the rate for the hydrolysis of Gly-L-Pro. These results document that Sgx9260c preferentially hydrolyzes dipeptides with L-proline at the C-terminus. The same methodology was used to determine the C-terminal specificity of Sgx9260b with the L-Ala-L-Xaa and Gly-L-Xaa dipeptide libraries. With both libraries the predominant amino acid detected was L-proline. None of the other amino acids were detected at a rate greater than 3% of the rate observed for the appearance of L-proline (except for the fixed amino acid at the N-terminus).

N-terminal Specificity of Sgx9260c and Sgx9260b

To obtain a more quantitative determination of the N-terminal specificity of these two proteins toward dipeptide substrates, an L-Xaa-L-Pro dipeptide library was synthesized and assayed. The free amino acid products detected after incubation of Sgx9260c with the L-Xaa-L-Pro library were proline, glycine, alanine, serine, threonine, tyrosine, phenylalanine and leucine. The relative reaction rates are presented in Figure 2a. Sgx9260c preferentially hydrolyzes dipeptides with small N-terminal residues, such as glycine and alanine. The N-terminal specificity of Sgx9260b was also examined in the same way. This enzyme was significantly more promiscuous for the N-terminal amino acids as shown in Figure 2b. Favored N-terminal amino acids include tyrosine, tryptophan, phenylalanine, and leucine.

Hydrolysis of N-Acy-L-Amino Acids, Tripeptides, and Gly-Sarcosine

Sgx9260c was assayed with a library of individual *N*-acetyl-L-amino acids. The predominant compound from this library that functioned as a substrate for Sgx9260c was *N*-acetyl-L-proline but lower rates of hydrolysis were detected for *N*-acetyl-L-alanine, *N*-acetyl-L-methionine and *N*-acetyl-L-valine. Sgx9260c was also found to hydrolyze *N*-acetyl hydroxy-L-proline, *N*-formyl-L-proline, *N*-propionyl-L-proline, *N*-butyryl-L-proline and Gly-sarcosine. Sgx9260b hydrolyzed the same set of *N*-acetyl-L-amino acids as Sgx9260c and also hydrolyzed *N*-acetyl hydroxy-L-proline, *N*-formyl-L-proline, *N*-propionyl-L-proline, *N*-butyryl-L-proline and Gly-sarcosine.

Kinetic Constants for Sgx9260c and Sgx9260b

The kinetic constants for Sgx9260c were determined using the following substrates: Gly-L-Pro, L-Ala-L-Pro, *N*-acetyl-L-proline, *N*-acetyl hydroxy-L-proline, *N*-formyl-L-proline, *N*-propionyl-L-proline, *N*-butyryl-L-proline and Glysarcosine. The kinetic data were fit to equation 2, and the values of k_{cat}, K_m and k_{cat}/K_m for the hydrolysis of the selected substrates are listed in Table 1. Sgx9260c was found to have the highest turnover number for the hydrolysis of *N*-acetyl-L-proline. The values of k_{cat}/K_m for the hydrolysis of the two dipeptides tested (Gly-L-Pro and L-Ala-L-Pro) are about two orders of magnitude smaller than obtained for *N*-acetyl-L-Pro. Sgx9260c was also assayed with the dipeptide L-Pro-L-Pro and the tripeptide Gly-Gly-L-Pro, but no activity could be detected with either compound. Kinetic parameters for Sgx9260b were measured using a variety of L-Xaa-L-Pro dipeptides as substrates. Gly-L-Pro was found to have the highest turnover number and *N*-propionyl-L-Pro had the best value of k_{cat}/K_m. The kinetic constants are presented in Table 2.

Inhibition by N-Methyl Phosphonate L-Proline

The kinetic behavior of Sgx9260c was determined in the presence of the *N*-methyl phosphonate derivative of L-proline (compound **5**) using *N*-acetyl-L-proline as substrate. This compound was found to be a competitive inhibitor with an inhibition constant, K_i, of 7.5 ± 1.0 μM from a fit of the data to equation 3 (data not shown). The *N*-methyl phosphonate derivative of L-arginine (**8**) was tested as a competitive inhibitor of Cc2672. This compound inhibits this enzyme with a K_i value of 92 ± 10 nM when L-Ala-L-Arg was used as the substrate.

Three-Dimensional Structures of Sgx9260c and Sgx9260b

The crystal structures of Sgx9260c and Sgx9260b were determined to resolutions of 2.6 Å and 2.4 Å, respectively (PDB codes: 3FEQ for Sgx9260c and 3MKV for Sgx9260b). The ribbon diagrams for Sgx9260c and Sgx9260b are presented in Figures 3a and 3b, respectively. Each protein subunit consists of two domains; a small predominantly β-barrel domain and a larger catalytic domain that is composed of a distorted $(\beta/\alpha)_8$-barrel. The structures for Sgx9260b and Sgx9260c are topologically identical (Figure 3). Moreover, although the overall amino acid sequence identity between the two proteins is 57%, active site residues are 100% identical, which explains their overlapping catalytic specificities. The catalytic site of each monomer is buried within the large $(\beta/\alpha)_8$-barrel domain. The zinc-binding site is composed of six conserved residues that include an aspartate, a carboxylated lysine, and four histidine residues. In Sgx9260c, His-63, His-65, His-229, His-249, Asp-321 and Lys-188 comprise the metal-binding site. His-63, His-65 and Asp-321 are coordinated to Zn_α (Figure 4a), whereas His-229 and His-249 are coordinated to Zn_β (Figure 4a).

In all members of the amidohydrolase superfamily with binuclear metal centers, the lysine residue is carboxylated and bridges the two metal ions (23). However, in the substrate-free Sgx9260c structure, Lys-188 is not carboxylated, which is consistent with our observation

that the metal-binding site of Sgx9260c is only partially occupied by Zn^{2+}, and partially filled with water. The final refined atomic model for Sgx9260c shows occupancy values for Znatoms of 0.5. In contrast, the structure of Sgx9260b (PDB code: 3MKV) exhibits a fully occupied Zn-binding site with Lys-191 being completely carboxylated and coordinated to the adjacent Zn atoms. These differing observations reflect the prevailing view that only in the presence of Zn^{2+} is the active site lysine carboxylated. In the structure of Sgx9260c bound with inhibitor **5** (PDB code: 3N2C), the zinc-binding sites are fully occupied and Lys-188 is carboxylated.

Structure of Sgx9260c-Inhibitor Complex

The three-dimensional structure of Sgx9260c bound to the *N*-methylphosphonate derivative of L-proline was determined at a resolution limit of 2.8 Å (PDB code: 3N2C). Binding of this inhibitor in the active site caused no gross structural rearrangements versus the inhibitor-free structure of Sgx9260c (PDB code: 3FEQ). The two zinc atoms of the inhibitor-bound structure are bridged by a fully carboxylated Lys-188. The *pro*R oxygen of the phosphonate moiety of compound **5** bridges with the two metals at distances of 2.1 and 2.0 Å and the two zinc ions are separated by 3.4 Å. The structure of the active site with the bound inhibitor is presented in Figure 5a.

Structure of Cc2672-Inhibitor Complex

The three-dimensional structure of Cc2672 from *Caulobacter crecentus* was determined in the presence of the *N*-methyl phosphonate derivative of L-arginine (compound **9**) at a resolution limit of 1.7 Å (PDB code: 3MTW). The active site is formed by the six conserved residues that coordinate the two divalent cations (His-90, His92, His-252, His-272, Asp344, Lys-211). The orientation of the inhibitor with respect to the two divalent cations is illustrated in Figure 5b. In this structure the phosphonate moiety bridges the two divalent cations. The *pro*R oxygen is 1.9 Å from Zn_{α} while the *pro*S oxygen is 1.9 Å from Zn_{β}. The distance between the two metal ions is 4.0 Å.

DISCUSSION

Substrate Profiles

Two previously uncharacterized enzymes from the amidohydrolase superfamily were purified to homogeneity and shown to catalyze the hydrolysis of L-Xaa-L-Pro dipeptides and *N*-acyl-L-Pro derivatives. Sgx9260c catalyzes the hydrolysis of *N*-acetyl-L-Pro with a k_{cat}/K_m that exceeds $10^5 M^{-1}s^{-1}$. This enzyme also catalyzes the hydrolysis of Gly-L-Pro and L-Ala-L-Pro but the values of k_{cat}/K_m are about two orders of magnitude lower than for *N*-acetyl-L-Pro. In contrast, the best substrate for Sgx9260b is *N*-propionyl-L-Pro with a k_{cat}/K_m of $\sim 10^5 M^{-1}s^{-1}$. This enzyme will also catalyze the hydrolysis of L-Xaa-L-Pro dipeptides with values of k_{cat}/K_m exceeding $10^4 M^{-1}s^{-1}$. Sgx9260b hydrolyzes a broader set of dipeptides than does Sgx9260c. In addition to alanine and glycine this enzyme will accept most of the hydrophobic amino acids at the N-terminus. The metabolic role of these enzymes has not been firmly established but it is likely that they are involved in protein and amino acid salvage pathways.

Structure Elucidation

Sgx9260c and Sgx9260b were crystallized and their structures determined at 2.4 and 2.6 Å resolution, respectively. Both enzymes exhibit active sites typical for Type I binuclear metal centers in the amidohydrolase superfamily (3). In these structures, the two divalent cations are ligated by one aspartate and four histidine residues, and typically bridged by a carboxylated lysine and a hydroxide ion. However, in the Sgx9260c inhibitor-free structure,

the active site lysine residue is not fully carboxylated, presumably because of partial zinc-binding site occupancies. The structure of Sgx9260c was also determined in the presence of the potent inhibitor that mimics the tetrahedral intermediate predicted to form during the hydrolysis of *N*-acetyl-L-Pro by this enzyme. In this structure Lys-188 is carboxylated and the phosphonate moiety displaces the bridging hydroxide and connects the two metal ions. The orientation of the inhibitor in the active site is consistent with a mechanism for nucleophilic attack of the bridging hydroxide on the *re*-face of the amide bond of the substrate. This is the same stereochemical face of the amide bond that is recognized by all of the other structurally characterized members of the AHS (3,5–8). In addition, the closest residues in the active site of Sgx9260c and Sgx9260b that can facilitate a proton transfer to the leaving group nitrogen of proline are Asp-321 and Asp-323, respectively. These two residues occur at the C-terminal end of β-strand 8 and are also coordinated to Zn_{α}.

Substrate Recognition Determinants

The structure of the *N*-methyl phosphonate derivative of L-Pro bound to Sgx9620c has revealed how the proline leaving group is recognized by the active site. In this structure, the original carbonyl oxygen is represented by the *pro*S oxygen and the attacking hydroxide is represented by the *pro*R oxygen. During nucleophilic attack the carbonyl group is polarized by Zn_{β} and the conserved His-140, found in the loop that follows β-strand 3. The free carboxylate of the substrate (proline) is hydrogen bonded to the hydroxyl group of the conserved Tyr-231in the loop that follows β-strand 5. This carboxylate also makes hydrogen bond interactions with the backbone amide nitrogen atoms of Ala-196 and Val-195, located in the loop that follows β-strand 4. The binding site for the proline side chain is relatively compact, and is formed by Thr-271, Tyr-273, Leu-278, and Val-295, all of which reside in the loop following β-strand 7. These binding interactions are highlighted in Figure 6a.

The recognition elements for the proline substituent in Sgx9620c can be compared with those for arginine in a related enzyme in the same superfamily, Cc2672. This enzyme hydrolyzes dipeptides terminating in L-arginine and the structural determinants responsible for recognition of the *N*-methyl phosphonate derivative of L-arginine (**8**) are shown in Figure 6b. The histidine that polarizes the carbonyl group (in addition to Zn_{β}) during nucleophilic attack is also present in the active site of this enzyme (His-171). The α-carboxylate of the arginine moiety is hydrogen bonded to His-254 and the backbone amide groups of Phe-219 and Val-218. These interactions are nearly the same as those observed for Sgx9620c, except that the tyrosine in Sgx9620c is replaced with a conserved histidine residue. This function is also served by a histidine residue in the group of enzymes that is specific for the hydrolysis of L-Xaa-L-hydrophobic dipeptides (subcluster **5** in Figure 1). In Cc2672, the guanidino side chain is recognized by hydrogen bonding interactions with Asp-294 from the C-terminal end of β-strand 7 and Asp-318 in the loop that follows β-strand 7. In the subcluster of sequences specific for the hydrolysis of L-Xaa-L-hydrophobic dipeptides, these residues are replaced by threonine and alanine, respectively. Thus, the identifies of residues occurring in the loop that follows β-strand 7 determine the substrate specificity for the hydrolysis of dipeptides or *N*-acylated amino acid derivatives. A sequence alignment of the structurally and functionally characterized enzymes from subclusters **4** (Sgx9355e, Cc0300), **5** (Sgx9359b, Cc2672, and Cc3125), and **7** (Sgx9260c and Sgx9260b) from cog1228 is provided in Figure 7.

Prediction of Catalytic Function

Now that the structural determinants for the recognition of substrates for two proteins within subcluster **7** have been identified, this information can be used to predict the specificity for the remainder of the proteins within this subcluster. Table 3 compares those residues critical for substrate recognition within this subcluster. All of these subcluster members possess the tyrosine residue at the C-terminal end of β-strand 5 for interaction with the α-carboxylate of

the protein substrate. All of these proteins also contain the conserved threonine at the end of β-strand 7 and all of these proteins contain either a valine or isoleucine in the loop that follow β-strand 7. Therefore, it is highly probable that all members of subcluster **7** catalyze the hydrolysis of *N*-substituted L-Pro derivatives.

Functional Annotation of cog1228

Subclusters **1** and **2** in Figure 1 contain the enzymes known to catalyze the hydrolysis of imidazolone propionate (HutI). At a BLAST E-value of 10^{-70} subcluster **16** merges with subcluster **2**. A sequence alignment of subclusters **2** and **16** indicates that the conserved residues responsible for substrate binding are retained, and thus the enzymes contained within subcluster **16** are predicted to catalyze the same chemical reaction as those found in subclusters **1** and **2**. Subcluster **4** contains the dipeptidases (such as Cc2672) that recognize arginine and/or lysine at the C-terminus. At a BLAST E-value of 10^{-70}, subcluster **15** merges with subcluster **4**. However, this group of proteins does not share the canonical HxH motif at the end of β-strand-1, but does exhibit an ExN motif. This subcluster also lacks the HxH or HxY motif at the end of β-strand 5. Finally, the conserved aspartate at the end of β-strand 7, required for the recognition of the guanidino group of arginine, is replaced with a threonine. Therefore, we think that it is highly unlikely that subcluster **15** catalyzes the hydrolysis of arginine containing dipeptides.

Subcluster **6** merges with subcluster **7** at an E-value cutoff of 10^{-70}. The recognition elements for the hydrolysis of dipeptide substrates are retained in subcluster **6**. The primary difference between these two subclusters is the identity of the residue that immediately follows β-strand 7, which appears to determine C-terminal amino acid specificity for dipeptide substrates. In subcluster **7** this residue is threonine, whereas this residue is predominantly a glutamine in subcluster **6**. The effect of this substitution on the substrate specificity of subcluster **6** is uncertain. Functional annotations of those subclusters (**6, 11, 15**, and **16**), most closely aligned to with the experimentally annotated subclusters (**1, 2, 4, 5**, and **7**) are currently being addressed by a combination of computational docking and compound library screening.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the NYSGXRC protein production team for providing purified samples of Sgx9260b and Sgx9260c.

This work was supported in part by the NIH (GM 071790 and GM 074945) and the Hackerman Advanced Research Program (010366-0034-2007). X-ray coordinates and structure factors for Sgx9260b, Sgx9260c and Cc2672 have been deposited in the Protein Data Bank (PDB accession codes: 3MKV, 3FEQ, 3N2C, and 3MTW).

Abbreviations

AHS	amidohydrolase superfamily
COG	clusters of orthologous groups

REFERENCES

1. Wu CH, Aweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumander R, O'Donova C, Redaschi N, Suzek B. The

universal protein resource (Uniprot): An expanding universe of protein information. Nucleic Acids Res. 2006; 34:D187–D191. [PubMed: 16381842]

2. Holm L, Sander C. An evolutionary treasure: unification of a broad set of amidohydrolase related to urease. Proteins. 1997; 28:72–82. [PubMed: 9144792]
3. Seibert CM, Raushel FM. Structural and catalytic diversity within the amidohydrolase superfamily. Biochemistry. 2005; 44:6383–6391. [PubMed: 15850372]
4. Afriat L, Roodveldt C, Manco G, Tawfik DS. The latent promiscuity of newly identified microbial lactonase is linked to a recently diverged phosphotriesterase. Biochemistry. 2006; 45:13677–13684. [PubMed: 17105187]
5. Xiang DF, Kolb P, Fedorov AA, Meier MM, Fedorov LV, Nguyen TT, Sterner R, Almo SC, Shoichet BK, Raushel FM. Functional annotation and three-dimensional structure of Dr0930 from *Deinococcus radiodurans*, a close relative of phosphotriesterase in the amidohydrolase superfamily. Biochemistry. 2009; 48:2237–2247. [PubMed: 19159332]
6. Marti-Arbona R, Fresquet V, Thoden JB, Davis ML, Holden HM, Raushel FM. Mechanism of the reaction catalyzed by iso-aspartyl dipeptidase from *Escherichia coli*. Biochemistry. 2005; 44:7115–7124. [PubMed: 15882050]
7. Xiang DF, Patskovsky Y, Xu C, Meyer AJ, Sauder JM, Burley SK, Almo SC, Raushel FM. Functional identification of incorrectly annotated prolidases from the amidohydrolase superfamily enzymes. Biochemistry. 2009; 48:3730–3742. [PubMed: 19281183]
8. Xiang DF, Kumaran D, Brown AC, Sauder JM, Burley SK, Swaminathan S, Raushel FM. Functional annotation of two new carboxypeptidases from the amidohydrolase superfamily of enzymes. Biochemistry. 2009; 48:4567–4576. [PubMed: 19358546]
9. Cummings JA, Nguyen TT, Fedorov AA, Kolb P, Xu C, Fedorov EV, Shoichet BK, Barondeau DP, Almo SC, Raushel FM. Structure, mechanism, and substrate profile for Sco3058: the closest bacterial homologue to human renal dipeptidase. Biochemistry. 2010; 49:611–622. [PubMed: 20000809]
10. Cummings JA, Fedorov AA, Xu C, Brown S, Fedorov E, Babbitt PC, Almo SC, Raushel FM. Annotating enzymes of uncertain function: the deacylation of D-amino acids by members of the amidohydrolase superfamily. Biochemistry. 2009; 48:6469–6481. [PubMed: 19518059]
11. Hara H, Masai E, Katayama Y, Fukuda M. The 4-oxalomesaconate hydratase gene, involved in the protocatechuate 4,5-cleavage pathway, is essential to vanillate and syringate degradation in *Sphingomonas paucimolbillis* SYK-6 hydration reaction. J. Bacteriol. 2000; 182:6950–6957. [PubMed: 11092855]
12. Williams L, Nguyen T, Li Y, Porter T, Raushel FM. Uronate isomerase: a nonhydrolytic member of the amidohydrolase superfamily with an ambivalent requirement for a divalent metal ion. Biochemistry. 2006; 45:7453–7462. [PubMed: 16768441]
13. Nguyen TT, Brown T, Fedorov AA, Fedorov EV, Babbitt PC, Almo SC, Raushel FM. At the periphery of the amidohydrolase superfamily: Bh0493 from *Bacillus halodurans* catalyzes the isomerization of D-galacturonate to D-tagaturonate. Biochemistry. 2008; 47:1194–1206. [PubMed: 18171028]
14. Nguyen TT, Fedorov AA, Williams L, Fedorov EV, Li Y, Xu C, Almo SC, Raushel FM. The mechanism of the reaction catalyzed by uronate isomerase illustrates how an isomerase may have evolved from a hydrolase within the amidohydrolase superfamily. Biochemistry. 2009; 48:8879–8890. [PubMed: 19678710]
15. Li T, Iwaki H, Fu R, Hasegawa Y, Zhang H, Liu A. α-Amino-β-carboxymuconic-ε-semialdehyde decarboxylase (ACMSD) is a new member of the amidohydrolase superfamily. Biochemistry. 2006; 45:6628–6634. [PubMed: 16716073]
16. Martynowski D, Eyobo YT, Yang K, Liu A, Zhang H. Crystal structure of alpha-amino-beta-carboxymuconate-epsilon-semialdehyde decarboxylase: insight into the active site and catalytic mechanism of a novel decarboxylation reaction. Biochemistry. 2006; 45:10412–10421. [PubMed: 16939194]
17. Liu A, Zhang H. Transition metal-catalyzed monoxide decarboxylation reactions. Biochemistry. 2006; 45:10407–10411. [PubMed: 16939193]

18. Pegg SC-H, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry. 2006; 45:2545–2555. [PubMed: 16489747]
19. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS ONE. 2009; 4:e4345. [PubMed: 19190775]
20. Tyagi R, Kumaran D, Burley SK, Swaminathan S. X-ray structure of imidazolone propionase from *Agrobacterium tumefaciens* at 1.87Å resolution. Proteins. 2007; 69:652–658. [PubMed: 17640072]
21. Tyagi R, Eswaramoorthy S, Burley SK, Raushel FM, Swaminathan S. A common catalytic mechanism for proteins of the HutI family. Biochemistry. 2008; 47:5608–5615. [PubMed: 18442260]
22. Yu Y, Liang YH, Brostromer E, Quan JM, Panjikar S, Dong YH, Su XD. A catalytic mechanism revealed by the crystal structures of the imidazolone propionase from *Bacillus Subtilis*. J. Biol. Chem. 2006; 281:36929–36936. [PubMed: 16990261]
23. Pieper L, Chiang R, Seffernick JJ, Brown SD, Glasner ME, Kelly L, Eswar N, Sauder JM, Bonanno JB, Swaminathan S, Burley SK, Zheng X, Chance MR, Almo SC, Gerlt JA, Raushel FM, Jacobson MP, Babbitt PC, Sali A. Target selection and annotation for the structural genomics of the amidohydrolase and enolase superfamilies. J. Struct. Funct. Genomics. 2009; 10:107–125. [PubMed: 19219566]
24. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of *Coot*. Acta Cryst. 2010; D66:486–501.
25. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK- a program to check the stereochemical quality of protein structures. J. App. Cryst. 1993; 26:283–291.
26. Delano, WL. The PyMOL molecular graphics system. Delano Scientific LLC; San Carlo, CA: 2002. http://www.pymol.org
27. Otwinowski, Z.; Minor, W. Processing of X-ray diffraction data collected in oscillation mode. In: Carter, CWJ.; Sweet, RM.; Abelson, JN.; Simon, MI., editors. Methods in Enzymology. Academic Pres; New York: 1997. p. 307-326.
28. Long F, Vagin A, Young P, Murshudov GN. BALBES: a molecular replacement pipeline. Acta Cryst. 2008; D 64:125–132.
29. Emsley P, Cowtan L. Coot: model-building tools for molecular graphics. Acta Cyst. 2004; D60:2126–2132.
30. Adams PD, Afonine, P. V. Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson JS, Terwilliger TC, Zwart PH. PHENIX: a comprehensive python-based system for macromolecular structure solution. Acta Cryst. 2010; D66:213–212.
31. Lamzin VS, Wilson KS. Automated refinement of protein models. Acta Cryst. 1993; D49:129–147.

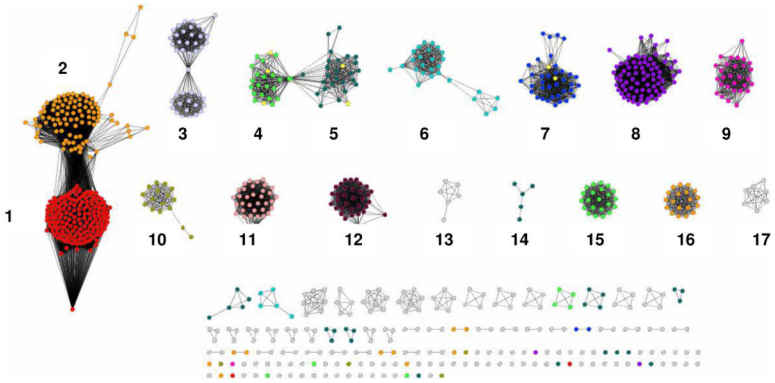


Figure 1.

Cytoscape network (www.cytoscape.org) representation of the sequence relationships in cog1228 from the amidohydrolase superfamily. Each node in the network represents a single sequence and each edge (depicted as lines) represents the pairwise connection between two sequences with the most significant BLAST E-value (better than 1×10^{-80}). Lengths of edges are not meaningful except that sequences in tightly clustered groups are relatively more similar to each other than sequences with few connections. In subcluster **4,** the yellow-colored nodes catalyze the hydrolysis of L-Xaa-L-Arg/Lys dipeptides. In subcluster **5,** the yellow colored nodes catalyze the hydrolysis of L-Xaa-L-hydrophobic dipeptides. Subclusters **1** and **2** catalyze the hydrolysis of imidazolone propionate (HutI). Additional information is presented in the text.

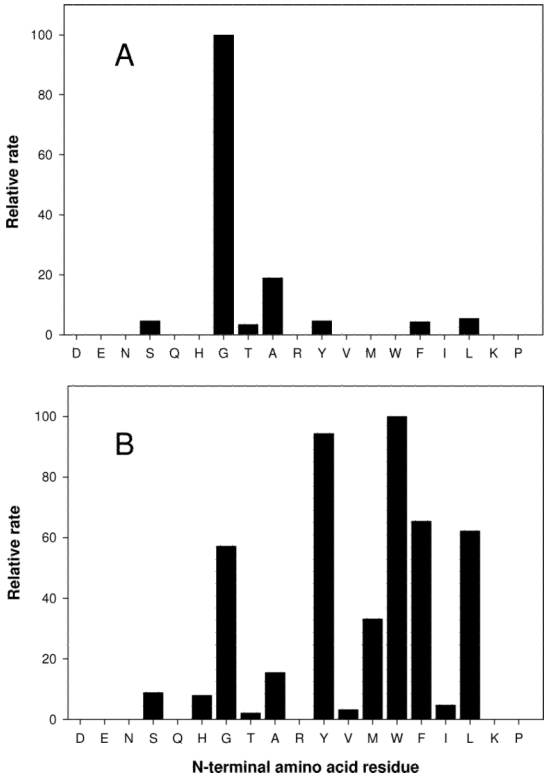


Figure 2.
Relative reaction rates for hydrolysis of L-Xaa-L-Pro dipeptides by Sgx9260C (A) and for Sgx9260b (B). Additional details are found in the text.

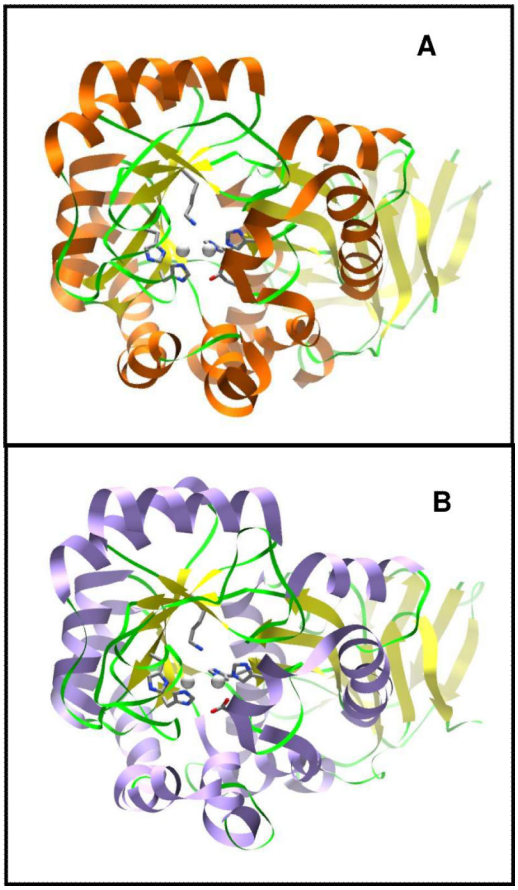


Figure 3.
Ribbon representations for the three dimensional structures of Sgx9260c (A) and Sgx9260b (B).

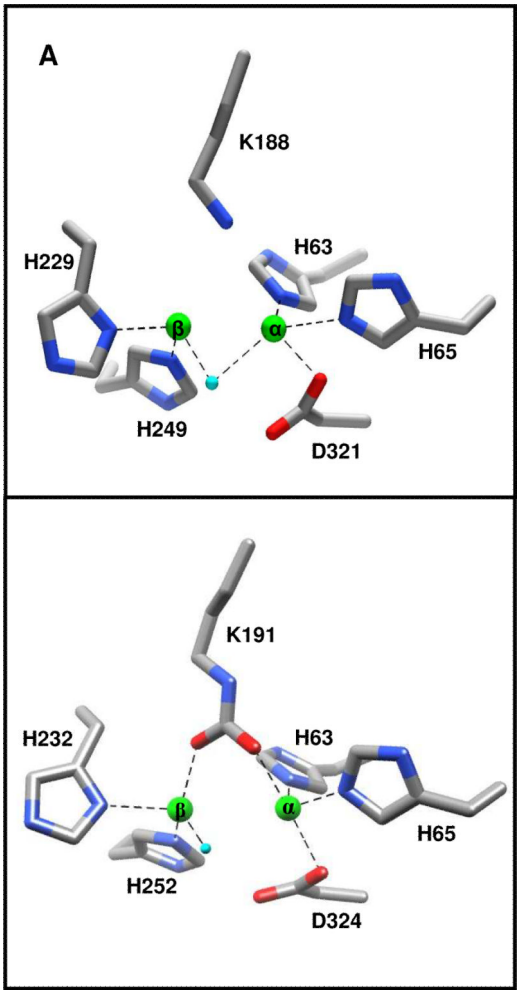


Figure 4.
Metal coordination scheme for Sgx9260c (A) and Sgx9260b (B).

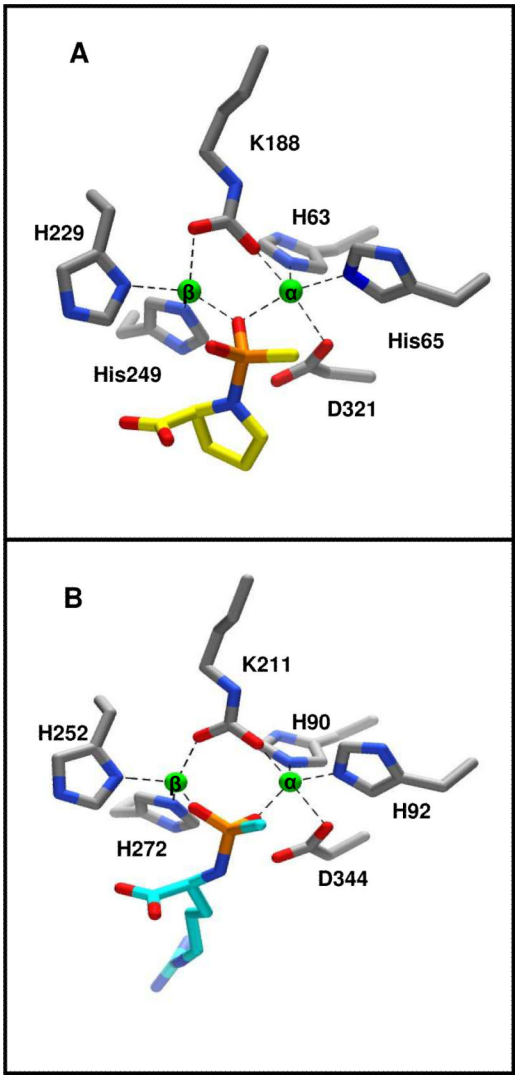


Figure 5.
Structure of the active site of Sgx9260c bound to the *N*-methyl phosphonate derivative of L-proline (A) and the active site of Cc2672 bound to the *N*-methyl phosphonate derivative of L-arginine (B).

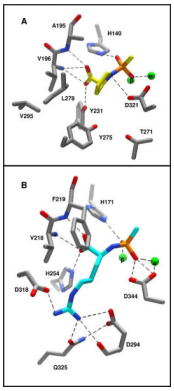


Figure 6.
Structures of the active sites of Sgx9260c (A)and Cc2672 (B) showing interactions with an *N*-methyl phosphonate mimic of the catalytic tetrahedral intermediate.

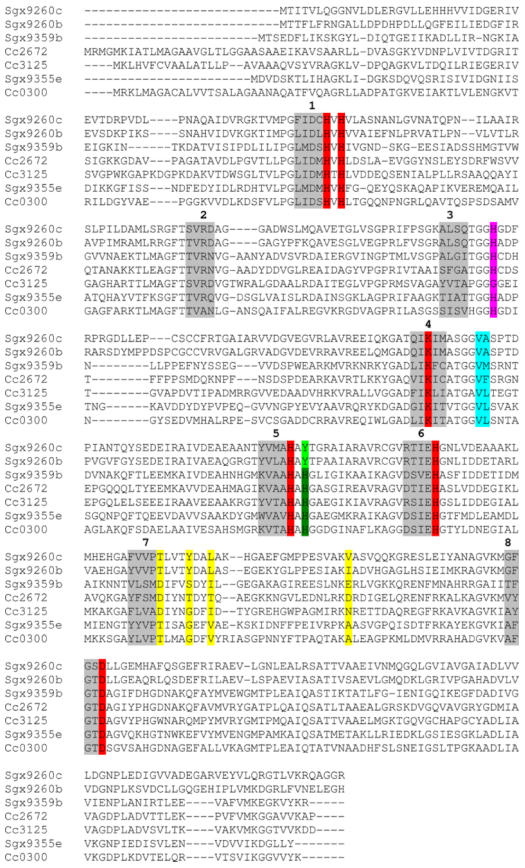
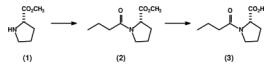
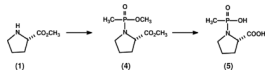


Figure 7.

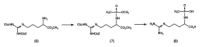
Amino acid sequence comparisons for Sgx9260c and Sgx9260b (from subcluster **7**) and two other groups of peptidases from cog1228 (subclusters **4** and **5**) which have also been experimentally characterized. These proteins include the following: Sgx9260c (gi|44242006), Sgx9260b (gi|44479596), Sgx9359b (gi|44368820), Cc2672 (gi|16126907), Cc3125 (gi|16127355), Sgx9355e (gi|44371129), and Cc0300 (gi|16124555). In this figure the residues that coordinate either of the two divalent cations in the active site are highlighted in red. The conserved histidine residue that polarizes the carbonyl group to be hydrolyzed is highlighted in magenta. The two residues whose backbone amide groups hydrogen bond with the α-carboxylate of substrates are highlighted in blue. The tyrosine or histidine residues that form hydrogen bonds with the substrate α-carboxylate of the substrate are highlighted in green. The residues from the loop that follows β-strand **7** that interact with the side chain of the C-terminal amino acid in dipeptide substrates are highlighted in yellow. The 8 β-strands from the $(\beta/\alpha)_8$-barrel core are highlighted in gray.



Scheme 1.



Scheme 2.



Scheme 3.

Table 1

Data collection and refinement statistics

PDB id	**3MKV**	**3FEQ**	**3N2C**	**3MTW**
Space group	P 1 21 1	P 1	P 1	I432
Unit-cell lengths (Å)	a = 92.0 b = 198.6 c = 104.39	a = 113.2 b = 108.0 c = 171.1	a = 113.2 b = 108.0 c = 170.8	a = 200.3 b = 200.3 c = 200.3
Unit-cell angles(°)	alpha = 90.0 beta = 107.8 gamma = 90.0	alpha = 81.8 beta = 80.4 gamma = 74.4	alpha = 81.3 beta = 80.5 gamma = 73.8	alpha = 90.0 beta = 90.0 gamma = 90.0
Diffraction protocol	single wavelength	single wavelength	single wavelength	single wavelength
Wavelength used (Å)	0.9793	0.9790	0.9790	0.9791
Resolution range(Å)	2.400 –40.000	2.61 –50.0	2.80 –50.000	1.7 – 35.0
Unique reflections (N)	140169	227280	186317	73648
Redundancy	2.5	3.1	1.9	6.1
Completeness (%)	92.6 (85.2)	98.0 (97.1)	98.3 (97.2)	98.6
*R*merge	0.090 (0.88)	0.206 (0.76)	0.180 (null)	0.067
<*I*/*sigmaI*>	5.6(0.8)	3.2(1.0)	3.5(0.5)	14.9
<FOM> (overall)	0.96	0.92	0.94	0.86
*R*factor	0.17	0.24	0.22	0.18
*R*free	0.23	0.28	0.27	0.203
R.m.s. bonds (Å)	0.006	0.009	0.007	0.006
R.m.s. angles (°)	0.99	1.080	1.14	1.06
Average B factor (Å^2)	53.1	58.12	85.98	21.05

Table 2

Kinetic Constants of Sgx9260c and Sgx9260b[a]

enzyme	substrate	k_{cat} (s^{-1})	K_m (mM)	k_{cat} / K_m
Sgx9260c	*N*-acetyl-L-proline	227 ± 9	0.8 ± 0.1	(2.8 ± 0.4) × 10^5
	N-propionyl-L-proline	141 ± 6	1.9 ± 0.2	(7.4 ± 0.8) × 10^4
	N-butyryl-L-proline	146 ± 6	2.6 ± 0.3	(5.6 ± 0.6) × 10^4
	N-acetyl-hydroxy-proline	49 ± 3	4.3 ± 0.6	(1.1 ± 0.2) × 10^4
	N-formyl-L-proline	16 ± 1	1.9 ± 0.3	(8.4 ± 1.3) × 10^3
	N-acetyl-L-alanine			(1.9 ± 0.2) × 10^3
	N-acetyl-L-methionine			(7.1 ± 0.3) × 10^2
	Gly-L-Pro	18 ± 2	6 ± 1	(3.1 ± 0.6) × 10^3
	L-Ala-L-Pro			(1.4 ± 0.1) × 10^3
	Gly-sarcosine			(2.9 ± 0.1) × 10^2
Sgx9260b	*N*-propionyl-L-proline	130 ± 5	1.2 ± 0.1	(1.1 ± 0.1) × 10^5
	N-acetyl-L-proline	182 ± 6	2.5 ± 0.2	(7.3 ± 0.6) × 10^4
	N-butyryl-L-proline	70 ± 2	0.9 ± 0.1	(7.8 ± 0.9) × 10^4
	N-formyl-L-proline	74 ± 4	3.1 ± 0.3	(2.4 ± 0.3) × 10^4
	N-acetyl-hydroxy-L-proline			(5.4 ± 0.5) × 10^3
	L-Trp-L-Pro	138 ± 7	2.7 ± 0.3	(5.1 ± 0.6) × 10^4
	L-Pro-L-Pro	98 ± 5	4.7 ± 0.5	(2.1 ± 0.2) × 10^4
	Gly-L-Pro	332 ± 20	7.1 ± 0.7	(4.7 ± 0.5) × 10^4
	L-Leu-L-Pro	304 ± 16	11 ± 1	(2.9 ± 0.3) × 10^4
	L-Ala-L-Pro			(5.4 ± 0.1) × 10^3
	Gly-sarcosine	0.8		

[a]The kinetic constants were obtained in 50 mM HEPES, pH 8.0, from fits of the data to equation 2.

Table 3

Proteins Predicted to Share Substrate Specificity with Sgx9260c and Sgx9260b

gi number	Organism	Locus tag	Identity of residues equivalent to those in Sgx9260c (Y231, T271, Y275, L278, V295)
218681969	Environmental sequence		HAYTG…PTLVTYDAL…KVASVQ
116691839	*B. cenocepacia* HI2424	Bcen2424_3741	HAYTG…PTLVTYDAL…KVASVQ
78063201	*B. sp. 383*	Bcep18194_B2354	HAYTG…PTLVTYDAL…KVASVQ
170736163	*B. cenocepacia* MC0-3	Bcenmc03_3782	HAYTG…PTLVTYDAL…KVASVQ
254248918	*B. cenocepacia* PC184	BCPG_03772	HAYTG…PTLVTYDAL…KVASVQ
206562642	*B. cenocepacia* J2315	BCAM0782	HAYTG…PTLVTYDAL…KVASVQ
221197215	*B. multivorans* CGD2M	BURMUCGD2M_5614	HAYTG…PTLVTYDAL…KVASVQ
161521530	*B. multivorans* ATCC 17616	Bmul_4994	HAYTG…PTLVTYDAL…KVASVQ
221210923	*B. multivorans* CGD1	BURMUCGD1_5177	HAYTG…PTLVTYDAL…KVASVQ
107026964	*B. cenocepacia* AU 1054	Bcen_4622	HAYTG…PTLVTYDAL…KVASVQ
167840518	*B. thailandensis* MSMB43	Bpse38_010100027834	HAYTA…PTLVTYDAL…KVASVR
91782797	*B. xenovorans* LB400	Bxe_A3028	HAYTG…PTLVTYDAL…KVETVR
73540593	*R. eutropha* JMP134	Reut_A0890	HAYTG…PTLVTYDAL…KIETVR
116695448	*R. eutropha* H16	H16_B1507	HAYTG…PTLVTYDAL…KIETVR
66045765	*P. syringae pv. syringae* B728a	Psyr_2529	HAYTA…PTQVTYEML…KIEDVR
28869987	*P. syringae pv. tomato str.* DC3000	PSPTO_2801	HAYTA…PTQVTYEML…KIEDVR
213966974	*P. syringae pv. tomato* T1	PSPTOT1_2352	HAYTA…PTQVTYEML…KIEDVR
257487754	*P. syringae pv. tabaci* ATCC 11528	PsyrptA_020100031063	HAYTA…PTQVTYEML…KIEDVR
71733642	*P. syringae pv. phaseolicola* 1448A	PSPPH_2686	HAYTA…PTQVTYEML…KIEDVR
146341656	*B. sp.* ORS278	BRADO4760	HAYTA…PTLVTYEAL…KIADVR
187924074	*B. phytofirmans* PsJN	Bphyt_2089	HAYTP…PTLVTYDAL…KVADVH
158430682	*unidentified*		HAYTP…PTLVTYDAL…KIADVH
27380605	*B. japonicum* USDA 110	bll5494	HAYTA…PTLVTYEAL…KIADVR
220921543	*M. nodulans* ORS 2060	Mnod_1549	HLYTD…PTLVTYEAL…KIDDVR
134100988	*S. erythraea* NRRL 2338	SACE_4455	HAYTA…PTLVTYWAL…KVDEVL
158424623	*A. caulinodans* ORS 571	AZC_2999	HLYTD…PTLVTYDKL…KIDNVR
220923856	*M. nodulans* ORS 2060	Mnod_3965	HLYTD…PTLVTFDKL…KVKVVQ
170741991	*M. sp.* 4–46	M446_3838	HLYTD…PTLVTFDKL…KVKVVQ
256376875	*A. mirum* DSM 43827	Amir_2756	HAYTA…PTLVTYWAL…KVDDVL
119383788	*P. denitrificans* PD1222	Pden_1037	HAYTN…PTLSAYTTM…KIRPVL
89067674	*O. granulosus* HTCC2516	OG2516_00694	HLYAD…PTNITYDLL…KVADVR
258653004	*N. multipartita* DSM 44233	Namu_2826	HAYTA…PTLVTYWAL…KVDDVL
62424497	*B. linens* BL2	BlinB01001910	HAYTA…PTLVTYERL…KVDDVL
111025581	*R.jostii R*HA1	RHA1_ro08799	HAYTP…PTLVTYHAL…KVSEVL
159473555	*C. reinhardtii*	CHLREDRAFT_49061	HAYTP…PTLVTYQEL…KVGNAL
83594741	*R.rubrum* ATCC 11170	Rru_A3412	HAYTG…PTLATYHAL…KVYTVL
226457772	*M. pusilla* CCMP1545	MICPUCDRAFT_68936	HAYTD…PTLVTYERL…KVADLV
255085764	*M. sp.* RCC299	MICPUN_75663	HAYTD…PTLITYDRL…KVGNLV