New mini-zincin structures provide a minimal scaffold for members of this metallopeptidase superfamily

Christine B Trame1,2, Yuanyuan Chang3, Herbert L Axelrod2, Ruth Y Eberhardt4,5, Penelope Coggill4,5, Marco Punta4,5 and Neil D Rawlings4,5*

Abstract

Background: The Acel_2062 protein from Acidothermus cellulolyticus is a protein of unknown function. Initial sequence analysis predicted that it was a metallopeptidase from the presence of a motif conserved amongst the Asp-zincins, which are peptidases that contain a single, catalytic zinc ion ligated by the histidines and aspartic acid within the motif (HEXXHXXGXXD). The Acel_2062 protein was chosen by the Joint Center for Structural Genomics for crystal structure determination to explore novel protein sequence space and structure-based function annotation.

Results: The crystal structure confirmed that the Acel_2062 protein consisted of a single, zincin-like metallopeptidase-like domain. The Met-turn, a structural feature thought to be important for a Met-zincin because it stabilizes the active site, is absent, and its stabilizing role may have been conferred to the C-terminal Tyr113. In our crystallographic model there are two molecules in the asymmetric unit and from size-exclusion chromatography, the protein dimerizes in solution. A water molecule is present in the putative zinc-binding site in one monomer, which is replaced by one of two observed conformations of His95 in the other.

Conclusions: The Acel_2062 protein is structurally related to the zinzins. It contains the minimum structural features of a member of this protein superfamily, and can be described as a “mini-zincin”. There is a striking parallel with the structure of a mini-Glu-zincin, which represents the minimum structure of a Glu-zincin (a metallopeptidase in which the third zinc ligand is a glutamic acid). Rather than being an ancestral state, phylogenetic analysis suggests that the mini-zinzins are derived from larger proteins.

Keywords: Acel_2062, Metallopeptidase, Zincin, JCSG, Structural genomics

Background

A metallopeptidase is a proteolytic enzyme that has one or two metal ions as an integral part of its catalytic machinery located within the active site. There are many families of metallopeptidases that bind a single zinc ion required for catalysis. The zinc ion is tetrahedrally coordinated by three residues from the peptidase and a water molecule that becomes activated to be the nucleophile in the catalytic reaction. In many of these families, but not all, the residues that ligate the zinc ion (referred to here as “ligands”) are two histidines within an HEXXH motif, and a third coordinating residue that is C-terminal to this motif, which can be a glutamic acid, a histidine or an aspartic acid. Metallopeptidases with an HEXXH motif are known as zinzins. In a metallopeptidase such as thermolysin, the third zinc ligand is usually a glutamic acid, and these peptidases are known as Glu-zinzins. In a metallopeptidase such as matrix metallopeptidase 1 (MMP1), the zinc ligands are the three histidines within an HEXXHXXXGXXH motif. In MMP1 there is also an important region known as the Met-turn, in which there is a conserved methionine that structurally supports the active site. For this reason, metallopeptidases such as MMP1 are known as Met-zinzins [1]. In some zinzins the third zinc ligand may be an aspartic acid (within the motif HEXXHXXXGXXD), and there is no Met-turn; these
peptidases are known as Asp-zincins [2]. All of the zincins share structural similarities and in the MEROPS classification and database, the different families that can be recognized by sequence similarities are all included in clan MA. This clan is subdivided into three subclans. Subclan MA(E) containing the Glu-zincins, subclan MA(M) containing the Met-zincins and subclan MA(D) containing the Asp-zincins [3].

A zincin structure contains at least two subdomains: an N-terminal subdomain which includes the HEXXXH motif, and a C-terminal subdomain that includes the third zinc ligand. The active site is therefore between the two subdomains. Within the zincins, the size of the C-terminal subdomain varies enormously, from being large in the matrix metallopeptidases (peptidase family M10), to being just a single helix, which is the case for snapalysin from Streptomyces lividans (peptidase family M7), which has one of the smallest sequences and structures in the clan [4].

Very recently, a minimal structure for a Glu-zincin has been determined [5], representing the smallest known member of this subclan so far discovered. No catalytic activity could be detected. The family has been provisionally assigned the name M95, but will not appear in the MEROPS database until peptidase activity has been experimentally confirmed. Lenart et al. [6] identified a number of protein families in which an HEXXXH motif was conserved. One of these families was Pfam family PF06262 (DUF1025)[Pfam:PF06262], which includes at least 400 sequences from bacteria. Members of this family have the Asp-zincin-like motif HEXXXHXXGXXD. In this paper, we report the structure of a member of this family: the Asp-zincin-like motif HEXXXHXXGXXD. In this paper, we report the structure of a member of this family: the Asp-zincin-like motif HEXXXHXXGXXD. In this paper, we report the structure of a member of this family: the Asp-zincin-like motif HEXXXHXXGXXD.

Methods

Protein expression and purification

The American Type Culture Collection (ATCC) provided the genomic DNA used to clone Acel_2062 (ATCC Number: ATCC 43068). Protein production and crystallization of the Acel_2062 protein was carried out by standard JCSG protocols [8]. Clones were generated using the Polymerase Incomplete Primer Extension (PIPE) cloning method [9]. The gene encoding Acel_2062 (GenBank: YP_873820[GenBank:YP_873820]; UniProtKB: A0LWM4 [UniProtKB:A0LWM4]) was synthesized with codons optimized for Escherichia coli expression (Codon Devices, Cambridge, MA) and cloned into plasmid pSpeedET, which encodes an expression and purification tag followed by a tobacco etch virus (TEV) protease cleavage site (MGSDKIHHHHHHHLYFQ/G) at the amino terminus of the full-length protein. Escherichia coli GeneHogs (Invitrogen) competent cells were transformed and dispensed on selective LB-agar plates. The cloning junctions were confirmed by DNA sequencing. Expression was performed in a selenomethionine-containing medium at 37°C. Selenomethionine was incorporated via inhibition of methionine biosynthesis [10], which does not require a methionine auxotrophic strain. At the end of fermentation, lysozyme was added to the culture to a final concentration of 250 μg/ml, and the cells were harvested and frozen. After one freeze/thaw cycle the cells were homogenized in lysis buffer [50 mM HEPES, 50 mM NaCl, 10 mM imidazole, 1 mM Tris(2-carboxyethyl)phosphine-HCl (TCEP), pH 8.0] and passed through a Microfluidizer (Microfluidics). The lysate was clarified by centrifugation at 32,500 x g for 30 minutes and loaded onto a nickel-chelating resin (GE Healthcare) pre-equilibrated with lysis buffer, the resin was washed with wash buffer [50 mM HEPES, 300 mM NaCl, 40 mM imidazole, 10% (v/v) glycerol, 1 mM TCEP, pH 8.0], and the protein was eluted with elution buffer [20 mM HEPES, 300 mM imidazole, 10% (v/v) glycerol, 1 mM TCEP, pH 8.0]. The eluate was buffer exchanged with TEV buffer [20 mM HEPES, 200 mM NaCl, 40 mM imidazole, 1 mM TCEP, pH 8.0] using a PD-10 column (GE Healthcare), and incubated with 1 mg of TEV protease per 15 mg of eluted protein for 2 hours at 20°–25°C followed by overnight at 4°C. The protease-treated eluate was passed over nickel-chelating resin (GE Healthcare) pre-equilibrated with HEPES crystallization buffer [20 mM HEPES, 200 mM NaCl, 40 mM imidazole, 1 mM TCEP, pH 8.0] and the resin was washed with the same buffer. The flow-through and wash fractions were combined and concentrated to 15.6 mg/ml by centrifugal ultrafiltration (Millipore) for crystallization trials.

Protein crystallization

The Acel_2062 protein was crystallized using the nanodroplet vapor diffusion method [11] with standard JCSG crystallization protocols [12]. Sitting drops composed of 100 nl protein solution mixed with 100 nl crystallization solution in a sitting drop format were equilibrated against a 50 μl reservoir at 277 K for 72 days prior to harvest. The crystallization reagent consisted of 24% polyethylene glycol 8000, 0.167 M calcium acetate, 0.1 M MES pH 6.17. Glycerol was added to a final concentration of 20% (v/v) as a cryoprotectant. Initial screening for diffraction was carried out using the Stanford Automated Mounting System (SAM) [13] at the Stanford Synchrotron Radiation Lightsource (SSRL, Menlo Park, CA). Data were collected at 3 wavelengths corresponding to the inflection(l1), high remote(l2) and peak energy(l0) of a selenium MAD (multi-wavelength anomalous dispersion) experiment at 100 K using a MARCCD 325 detector (Rayonix) at Stanford Synchrotron Radiation Lightsource (SSRL) beamline 9_2.
Data processing was carried out using XDS [14] and the statistics are presented in Table 1. The structure was determined by the MAD method using programs SHELX [15] and autoSHARP [16], and refinement was carried out using REFMAC5 [17]. The structure was validated using the JCSG Quality Control server (http://smb.slac.stanford.edu/jcsg/QC).

### Table 1 Summary of crystal parameters, data collection, and refinement statistics for Acel_2062 from *Acidothermus cellulolyticus* 11b [PDB:3e11]

| Data collection | λ_1 MADSe | λ_2 MADSe | λ_3 MADSe |
|-----------------|-----------|-----------|-----------|
| Wavelength (Å)  | 0.97949   | 0.91837   | 0.97897   |
| Resolution range (Å) | 29.437-1.800 | 29.412-1.801 | 29.399-1.803 |
| No. of observations | 49,068 | 49,153 | 48,819 |
| No. of unique reflections | 19,523 | 19,533 | 19,440 |
| Completeness (%) | 95.3 (94.6)^a | 95.5 (93.5)^a | 94.9 (90.2)^a |
| Mean μ(I) | 8.50 (1.69)^a | 8.31 (1.67)^a | 8.08 (1.51)^a |
| Rmerge on I (%) | 6.3 (48.0)^a | 6.4 (45.7)^a | 6.8 (55.5)^a |
| Rmerge on I (%) | 9.1 (63.8)^a | 9.6 (64.7)^a | 10.5 (74.3)^a |
| Highest resolution shell (Å) | 1.86-1.80 | 1.86-1.80 | 1.86-1.80 |

### Determination of oligomeric state

The oligomeric state of the Acel_2062 protein in solution was determined using a 0.8 × 30 cm² Shodex Protein KW-803 size exclusion column (Thomson Instruments) pre-calibrated with gel filtration standards (Bio-Rad). The mobile phase consisted of 20 mM Tris pH 8.0, 150 mM NaCl, and 0.02% (w/v) sodium azide. The apparent molecular weight was calculated using the Bio-Rad Gel Filtration Standard set (#151-1901) and a linear regression of log10MW.

### Bioinformatics

To find homologues of the Acel_2062 protein, a Blastp search was conducted against the non-redundant protein sequence database at NCBI [21], using standard parameters. Structure diagrams were prepared using PyMol. Domain diagrams were taken from Pfam release 27 [22]. Secondary structure topology diagrams were generated by HERA [23] and downloaded from PDBsum website (http://www.ebi.ac.uk/pdbsum/). The alignment was prepared using MAFFT [24] and ESPript 2.2 (http://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi). PISA analysis [25] of the dimer interface was performed using the PDBe server at the European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/prot_int/). The electrostatic surface was displayed using PyMol (http://www.pymol.org/) and a Delphi [26] embedded script kindly provided by Qingping Xu. Coot [27] was used to superimpose structures from the Protein Data Bank (PDB). Molecular graphics and visualization were performed using PyMol (http://www.pymol.org/) and PymolMolviewer (http://www.pymolmolviewer.com/).

### Figure 1 Crystal structure

The crystal structure of the Acel_2062 protein from *Acidothermus cellulolyticus*, shown in magenta, reveals a single zincin-like domain. The putative zinc ligands (His95, His99 and Asp105), the putative catalytic Glu96, and Tyr113 are shown as sticks.
analyses were performed with the UCSF Chimera package [28]. The theoretical pI was calculated using the Expasy website (http://web.expasy.org/compute_pi).

Results and discussion

Structure description

The crystal structure of the Acel_2062 protein was determined to 1.8 Å resolution using the MAD phasing method. Atomic coordinates and experimental structure factors to 1.8 Å resolution [PDB:3e11] have been deposited in the Protein Data Bank (www.wwpdb.org, [29]). Data-collection, model and refinement statistics are summarized in Table 1.

The final model includes two protein molecules (residues 1–113), two acetate molecules, two (presumed to be structural) calcium ions and 193 water molecules in the asymmetric unit. The calcium ions are near the centre of the dimer interface, and may be important for dimerization. The calcium ions are coordinated by Asp18, Glu38 and via waters by Glu15. No zinc was found in the structure, either because little zinc was present during purification and crystallization, the enzyme is in latent state or the protein is not an enzyme. The Matthews coefficient ($V_M$) is 2.07 Å$^3$/Da and the estimated solvent content is 40.54%.

The Ramachandran plot produced by MolProbity [31] shows that 98.0% of the residues are in favoured regions, with no outliers.

Figure 2 Sequence alignment of the Acel_2062 protein and a selection of its homologues. The UniProt accession and the range of the peptidase domain are shown on the left. The zincin motif is boxed in red. Conserved residues shown in white text highlighted in red. Key to sequences (ordered locus name, species): A0LWM4[UniProt:A0LWM4] (Acel_2062, Acidothermus cellulolyticus), F2NA85[UniProt:F2NA85] (Corg_0144, Coriobacterium glomerans), COACH6[UniProt:COACH6] (ObaDRAFT_6101, Diplaphaera colitermitum), E6BJL2[UniProt:E6BJL2] (ES5_13183, Dietzia cincta), A5PE7[UniProt:A5PE7] (DE2_26213, Ethyrobacter sp. SD-21), A1B0F0[UniProt:A1B0F0] (Paracoccus sp. SD-21), E8L598[UniProt:E8L598] (Met49242DRAFT_2641, Methylocystis sp. ATCC 49242), Q08VI2[UniProt:Q08VI2] (COCOR_02006, Corallococcus coralloides), E2T6Y2[UniProt:E2T6Y2] (TMBG_02375, Mycobacterium tuberculosis), Q5SR8L[UniProt:Q5SR8L] (TTHA0227, Thermus thermophilus), F2NR95[UniProt:F2NR95] (Marky_0229, Oceanithermus profundus).
Glu43 and Glu106 on chain A and Glu15, Glu43 and Glu110 on chain B had poor electron density and were omitted from the model. For monomer A, the structure is composed of four helices, two 3_10 helices and three beta strands (see Figure 1). In monomer B, the N-terminus forms a fourth strand and the dimer is formed from this strand inserting into the beta sheet of monomer A.

The Acel_2062 protein was crystallized as a dimer, with the putative active sites at opposite ends of the dimer. PISA analysis of the structure indicates that the solvent-excluded surface area for the proposed dimer is ~639 Å². From size exclusion chromatography, the molecular weight of the Acel_2062 protein in solution is estimated to be 26,824, which is a ratio of 2.09 over the expected molecular weight of the monomer (12,833) indicating that the protein exists as a dimer (Additional file 1: Figure S1).

The crystal structure of Acel_2062 protein has a minimal zincin-like fold because it retains a three-stranded mixed beta-sheet (rather than the five-stranded beta-sheet of many zincins), the loops are much shorter and the overall sequence length of all members of this family (~110 aa) is significantly shorter than the average for a matrix metalloprotease (MMP)-like domain (~160 aa). The distant homology prediction program FFAS [32] recognizes this similarity, with a marginal statistical significance (Z-score of −8.9 as compared to −9.5 as the significant threshold), suggesting that DUF1025 family is distantly related to metalloproteases.

There is no signal peptide, and the Acel_2062 protein is presumably intracellular.

### Putative active site
From the presence of the HEXXHXXGXXD motif, the potential zinc ligands in the Acel_2062 protein are predicted to be His95, His99 and Asp105, and Glu96 is predicted to be a catalytic residue. In the crystal structure (PDB:3E11), the Glu96 is hydrogen-bonded to five water molecules in both monomers. Conservation of the active site residues is shown in Figure 2. Of the two active sites in both monomers A and B, the one in B is empty and the other in A is occupied by a single water molecule which is coordinated by His95 (at 2.7 Å), His99 (at 3.3 Å) and...
Asp105 (at 2.9 Å), exactly where the zinc ion would be expected to be. Delphi calculations show that the entire pocket is very acidic. There are five negatively charged residues, plus the carboxyl group from the C-terminus in that area. The surface electrostatic potential is shown in Figure 3. The molecule is negatively charged overall (with 15 Glu and 11 Asp compared to one Lys and eight Arg), with a theoretical PI calculated to be 4.29. This may be an adaptation to the acidic hot spring environment in which *Acidothermus cellulolyticus* lives.

In the Acel_2062 protein, and other members of the family, the HEXXH motif is very close to the C-terminus. Like Glu- and Asp-zincins, there is no Met-turn. The first His (His95) of the HEXXH motif exists in two conformations in monomer B. In the more stable conformation, His95 is electrostatically bound to the carboxyl group of the C-terminal residue (Tyr113). This is the situation also found in monomer A, so it is possible that the water (and presumably also the zinc) only bind when His95 and Tyr113 interact. The structure of the Acel_2062 protein was superimposed upon that of the repolysin BaP1 peptidase from the snake *Bothrops asper* [PDB:2w14] and this clearly shows that Tyr113 occupies a similar position to the methionine of the Met-turn in the BaP1 peptidase (see Figure 4). So although there is no Met-turn, Tyr113 may compensate for it.

**Sequence similarities**

Over 500 homologues of the Acel_2062 protein were found from the Blastp search. In 80 of these proteins, the HEXXHXXGXXD motif is not conserved. The third zinc ligand has been replaced in many of these homologues, often with glutamic acid, which is the third zinc ligand in Glu-zincins.

All of the homologues are from bacteria belonging to seven different phyla. Most homologues come from species in the phylum Firmicutes (294), which are Gram-positive bacteria. There are 185 homologues from species in the phylum Proteobacteria, fourteen from Chloroflexi, five from Planctomycetes, three from Verrucomicrobia, and one each from species in the phyla Caldiserica and Nitrospirae. Most species have only one homologue, but *Stigmatella aurantiaca* has two, though only one has the third zinc ligand conserved.

The different Pfam domain architectures for members of this family are shown in Figure 5. The vast majority of homologues have the simple domain architecture of the Acel_2062 protein. Seventeen homologues include tetratricopeptide repeats (TPR), which mediate protein-protein interactions and the assembly of multi-protein complexes [33]. A TPR repeat motif consists of several tandem repeats of a 34-residue sequence. Eight proteins with TPR repeats are predicted to have signal peptides and are presumably secreted. A homologue

![Figure 5](https://example.com/figure5.png)

**Figure 5 Domain architectures for the Acel_2062 protein and its homologues.** The different domain architectures for proteins containing an M94 metallopeptidase domain is shown. The UniProt identifier and source organism are given for an example of each domain architecture. Key to domains: DUF1025 is the metallopeptidase domain; TPR, tetratricopeptide repeats; TM, transmembrane region; Septum form, a domain found in proteins predicted to play a role in septum formation during cell division; S, signal peptide.
from *Mycobacterium tuberculosis* possesses an N-terminal transmembrane domain and a domain found in proteins known to be important for septum formation during spore formation [34].

**Structural similarities**

Amongst known metallopeptidases, DALI analysis shows that the Acel_2062 protein structure is most similar to that of a Met-zincin from the archaean *Methanocorpusculum labreanum* (peptidase family M54; archaealysins or archaemetzincins [PDB:3lmc]). The Acel_2062 protein structure is also similar to the carboxy-terminal domain (residues 511 to 624) of *Escherichia coli* HtpG/Hsp90 protein, which is a chaperone protein. This C-terminal domain is important for dimerization, but the mechanism of dimerization via the C-terminal helices is completely different to that of the Acel_2062 dimer. Two of the helices and the beta sheet can be superimposed, and the beta strands run in the same direction. The relationship between Hsp90 and a zincin has not previously been recognized in either the SCOP [35] or CATH [36] databases, and suggests a common evolutionary origin. These structural relationships cover the entire Acel_2062 protein sequence.

The Met-turn is important in all Met-zincins because the methionine is crucial for structurally stabilizing the active site. Even in snapalysins from *Streptomyces*, which also have short sequences, there is a conserved methionine approximately ten residues C-terminal to the third zinc ligand. In the Acel_2062 protein although there are no residues that correspond to the C-terminal 36 residues of the archaemetzincin, which includes the essential Met168, the C-terminal Tyr113 occupies a similar position to the methionine. Although it is tempting to suggest that this tyrosine performs a similar role, it should be noted that the tyrosine is present in only 75 homologues in the family and replaced by tryptophan in a further 220 homologues and by phenylalanine in a homologue from *Rhodococcus* sp. AW25M09. In a further 221 homologues, the full-length sequence falls short of Tyr113. In astacin, a tyrosine (Tyr149) has been shown to be an important residue, contributing to transition state binding, which can be replaced with much reduced efficiency by phenylalanine [37]. If proven to be a peptidase, the Acel_2062 protein and its homologues would form family M94 in MEROPS.

The structure of another putative metallopeptidase, the TTHA0227 protein from *Thermus thermophilus* ([PDB:2ejq]; unpublished), is similar. The TTHA0227 protein is also a dimer but has been crystallized with a single magnesium ion, and the structure also lacks any zinc ions. This is also an acidic protein with a theoretical pI calculated to be 4.53. The TTHA0227 protein also contains the Asp-zincin metal-binding motif, and the potential zinc coordinating residues are His95, His99 and Asp109. Chain A is shorter at the C-terminus, lacking residues 109–130, which means that the third zinc ligand is missing. The putative catalytic residue is Glu96. There is a four-residue insert preceding the essential glycine (Gly106, Additional file 2: Figure S2). In chain A, the final helix is continuous, whereas in chain B, there is a turn and there are two, opposing helices. This second helix does not superimpose with the final helix in the 3E11 structure, because it is pointing in the opposite direction so that the faces of the helices that oppose each other in 2EJQ are different from those in 3E11 (Figure 6).

There are several orthologues of the TTHA0227 protein which have the insert within the Asp-zincin motif. In the homologues from *Oceanithermus profundus* and *Meiothermus ruber* the glycine (Gly106), which in Met-zincins is important for the turn that permits the zinc ligands to face one another, is replaced by tryptophan. It had been thought that only a glycine was acceptable in this position [38], although several bacterial homologues of pappalysin, family M43, have asparagine at this position, including a homologue from *Methanosarcina acetivorans* for which the crystal structure has been solved [39], and a homologue from *Cytophaga hutchinsonii*.
Chut1718 gene product) has threonine at this position. Comparisons of the structural topologies of various Met-zincins and the mini-Glu-zincin are shown in Figure 7.

The structure of the Acel_2062 protein represents the minimum sequence known for a zincin domain. The question remains: is this a situation that has developed within this family, or is it a relic of the ancestral zincin gene? One way to answer this question is to look at the species distribution of peptidases from the various families in the clan. Table 2 shows the phyletic distribution of all families within clan MA. The number of phyla within each of the three superkingdoms (Archaea, Bacteria, Eukaryota) containing at least one homologue within each family is shown. Amongst the Glu-zincins, families M1, M3, M32 and M48 are widely distributed in phyla from all three superkingdoms (M41 is also widely distributed in bacteria and eukaryotes, but is absent from archaea). The mini-Glu-zincins, from family M95, have a much narrower distribution and are absent from eukaryotes. These observations imply that the last common universal ancestor most likely possessed a homologue from each of these families, and that the larger Glu-zincin structure is the ancestral state. The distribution of Asp- and Met-zincins is more restricted in all families, and the only family that is well represented in all domains of life is family M54, the archaelysins. There are homologues from archaea (93 species), bacteria (47 species) and eukaryotes (69 species), though these are...
| Subclan | Family | Name                                      | Archaea | Bacteria | Eukaryota |
|---------|--------|-------------------------------------------|---------|----------|-----------|
|         |        | **Total phyla**                           | 5       | 28       | 49        |
| MA(E)   | M1     | Aminopeptidase N                          | 3       | 21       | 31        |
|         | M2     | Angiotensin-converting enzyme              | -       | 6        | 11        |
|         | M3     | Thimet oligopeptidase                     | 3       | 25       | 28        |
|         | M4     | Thermolysin                               | 1       | 10       | 4         |
|         | M5     | Mycolysin                                 | -       | 2        | -         |
|         | M9     | Microbial collagenase                     | 1       | 6        | -         |
|         | M13    | Neprolysins                               | 1       | 12       | 23        |
|         | M26    | IgA metalloendopeptidase                  | 1       | 4        | -         |
|         | M27    | Tentoxylins                               | -       | 1        | -         |
|         | M30    | Hycolysin                                 | 1       | 9        | -         |
|         | M32    | Carboxypeptidase Taq                       | 5       | 16       | 7         |
|         | M34    | Anthrax lethal factor                     | -       | 1        | -         |
|         | M36    | Fungalysin                                | -       | 6        | 5         |
|         | M41    | FtsH endopeptidase                        | -       | 27       | 36        |
|         | M47    | Metallopeptidase PRSM1                    | -       | -        | 1         |
|         | M48    | Metallopeptidase STE24                    | 4       | 25       | 28        |
|         | M49    | Dipeptidyl-peptidase III                  | -       | 6        | 21        |
|         | M56    | BlaR1 peptidase                           | -       | 11       | -         |
|         | M60    | Enhancin                                  | -       | 3        | 1         |
|         | M61    | Glycyl aminopeptidase                     | 2       | 10       | 2         |
|         | M65    | YugP protein (Bacillus)                   | -       | 3        | -         |
|         | M69    | F19G6.4 protein (Caenorhabditis)          | -       | -        | 1         |
|         | M70    | Surface protein (Ehrlichia)               | -       | 1        | -         |
|         | M76    | Atp23 peptidase                           | -       | -        | 21        |
|         | M78    | ImmA peptidase                            | -       | 10       | -         |
|         | M85    | NleC peptidase                            | -       | 1        | -         |
|         | M90    | MtFA peptidase                            | -       | 9        | -         |
|         | M91    | NleD peptidase                            | -       | 1        | -         |
|         | M93    | BACCAC_01431 protein (Bacteroides)        | -       | 1        | -         |
|         | M95    | Proabylysin                               | 1       | 5        | -         |
| MA(M)   | M6     | Immune inhibitor A                        | 1       | **12**   | 3         |
|         | M7     | Snapalysin                                | -       | 1        | -         |
|         | M8     | Leishmanalysin                            | -       | 7        | 20        |
|         | M10    | Matrix metallopeptidases                  | 2       | 11       | 14        |
|         | M11    | Autolysin                                 | -       | 2        | 2         |
|         | M12    | Astacins/reprolysins                      | 1       | 8        | **26**    |
|         | M35    | Deuterolysin                              | -       | 3        | 3         |
|         | M39    | YL108W protein (Saccharomyces)            | -       | -        | 1         |
|         | M43    | Cytophagalysin                            | 1       | 6        | 14        |
|         | M54    | Archaelysin                               | 4       | **12**   | 8         |
|         | M57    | PrtB protein (Mycopoccus)                 | -       | 3        | -         |
|         | M59    | Putative zinc metalloprotease (Vibrio)    | -       | 1        | -         |
|         | M62    | Membrane metalloprotease (Euryarchaeota)  | 1       | 3        | -         |
Table 2 Phyletic distribution of experimentally confirmed and hypothetical peptidase families in MEROPS clan MA (Continued)

| M64 | IgA peptidase (Clostridium) | - | 5 | 1 |
| M66 | StCE peptidase | - | 3 | 1 |
| M68 | JHP0742 protein (Helicobacter) | 1 | 3 | - |
| M71 | PA0478 protein (Pyrobaculum) | 1 | - | - |
| M72 | Peptidyl-Asp metallopeptidase | - | 7 | 1 |
| M80 | Wss1 peptidase | - | - | 12 |
| M83 | DR2310 peptidase | - | 2 | - |
| M84 | mprBi peptidase | 1 | 1 | - |
| M94 | Acel_2062 protein (Acidothermus) | - | 9 | - |

Families are ordered by subclan and name. The number of phyla in each of the superkingdoms Archaea, Bacteria and Eukarya from which at least one homologue is known are shown. Where the number of phyla exceeds half the known phyla in the superkingdom, the number is highlighted in bold. The family in each subclan with examples from the most phyla in each superkingdom is shown as bold, italics text. A hyphen indicates no examples are known in that superkingdom. Rows in italics are families of putative peptidases that have not yet been experimentally characterized as such. The MEROPS identifier for such a family is provisional and may be subject to change.

found in less than half of the bacterial and eukaryote phyla. It is likely that an archaealysin most closely represents the ancestral Met-zincin structure. Archaealysin possesses the Met-turn [40,41], so the implication is that the Met-turn has been lost from an ancestor of family M94 and functionally replaced by a C-terminal aromatic residue (tyrosine or phenylalanine). The much narrower distribution of members of M94 supports the hypothesis that the family is a more recent development.

Conclusions
The Acel_2062 protein from *Acidothermus cellulolyticus* is a protein of unknown function, but was predicted to be a metallopeptidase from the presence of a motif (HEXXHXGXGD) conserved amongst the Asp-zincins, which contain a single, catalytic zinc ion ligated by the histidines and aspartic acid within the motif. The tertiary structure of the Acel_2062 protein was determined by the Joint Center for Structural Genomics, and confirmed the presence of a single, zinc-like metallopeptidase-like domain. In our crystallographic model there are two molecules in the asymmetric unit and from size-exclusion chromatography, the protein dimerizes in solution. A water molecule is present in the putative zinc-binding site in one monomer, which is replaced by one of two observed conformations of His95 in the other. The C-terminal Tyr113 may be important for stabilizing the putative active site. Additional experimentation would be required to prove that the Acel_2062 protein is a metallopeptidase.

Although the Acel_2062 protein is structurally related to the zincins, it contains the minimum structural features of a member of this protein superfamily, and can be described as a "mini- zincin". There is a striking parallel with the structure of a mini-Glu-zincin, which represents the minimum structure of a Glu-zincin (a metallopeptidase in which the third zinc ligand is a glutamic acid). Rather than being an ancestral state, phylogenetic analysis suggests that the mini-zincins are derived from larger proteins.

Additional files

**Additional file 1: Figure S1.** Molecular weight determination by size-exclusion chromatography. A) Elution profile for Bio-Rad Gel Filtration Standard set (#151-1901) comprising vitamin B₁₂ (1,350 Da), horse myoglobin (17 kDa), chicken ovalbumin (44 kDa), bovine gamma-globulin (158 kDa) and bovine thyroglobulin (670 kDa). B) Elution profile for the Acel_2062 protein.

**Additional file 2: Figure S2.** Crystal structure of the TTHA0227 protein from *Thermus thermophilus*. The PDB entry 2EJQ [PDB:2ejq] is a dimer, and because different elements are missing from both monomers, the structure of both monomers is shown. Chain A is shown in beige, chain B in cyan. Residues described below and in the text are labelled and shown as sticks. The monomers differ in terms of the residues that cannot be resolved. Chain A lacks the C-terminal residues Asp110–Glu130 and chain B lacks residues Pro54-Leu64 as well as the C-terminal Gly-Glu-Gly residues. Also in 2EJQ, Asp109 is too far away from the other potential zinc ligands to be a ligand itself. There are other potential zinc ligands in residues 110–130, including Glu113, Asp114, and Asp119. Only Asp119 is close enough to the imidazolium rings of the histidines to act as the third zinc ligand. Unfortunately, this Asp119 is poorly conserved, whereas Asp109 is well conserved. Because Asp109 and the final helix are close to the dimer interface, the structure here may be distorted because of the dimerization.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
CBT performed X-ray structure determination and prepared some of the figures; NDR, YC, HLA, RYE, PC and MP analysed the sequence-structure-function relationships and prepared the manuscript, tables and the other figures. All authors read and approved the final manuscript.

Acknowledgements
We are grateful to the Sanford Burnham Medical Research Institute for hosting the DUF Annotation Jamboree in June 2013 that allowed the authors to collaborate on this work. We would like to thank all the participants of this workshop for their intellectual contributions to this work: L. Aravind, Alex Bateman, Debanu Das, Robert D. Finn, Adam Godzik, William Hwang, Lukasz Jaroszewski, Alexey Murzin, Padmaja Natarajan, Daniel Rigden,
Maya Sedova, Anna Sheydina, John Wooley. We thank the members of the JCSG high-throughput structural biology pipeline for their contribution to this work. This work was supported in part by National Institutes of Health grant US4 GM094586 from the NIGMS Protein Structure Initiative to the Joint Center for Structural Genomics; intramural funds of the National Library of Medicine, USA, to LA; NIH grant R01GM014579 to AG; Howard Hughes Medical Institute to RDF; and Wellcome Trust grant WT077044/Z/05/Z for funding for open access charges. Portions of this research were carried out at the Stanford Synchrotron Radiation Lightsource, a Directorate of SLAC National Accelerator Laboratory and an Office of Science User Facility operated for the U.S. Department of Energy Office of Science by Stanford University. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Department of Energy Office of Science by Stanford University. Portions of this research were carried out at the National Institute of General Medical Sciences (including P41GM103393). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS, NCRR or NIH.

Author details
1 Joint Center for Structural Genomics, La Jolla, CA 92037, USA. 2 Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. 3 Sandford-Burnham Institute, La Jolla, CA 92037, USA. 4 Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK. 5 European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire CB10 1SD, UK.

Received: 30 July 2013 Accepted: 17 December 2013

Published: 3 January 2014

References
1. Stöcker W, Grams F, Baumann U, Reinemer P, Gomis-Rüth FX, McKay DB, 1 Joint Center for Structural Genomics, La Jolla, CA 92037, USA. 2 Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. 3 Sandford-Burnham Institute, La Jolla, CA 92037, USA. 4 Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK. 5 European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire CB10 1SD, UK.

1. Stöcker W, Grams F, Baumann U, Reinemer P, Gomis-Rüth FX, McKay DB, Bode W: The metalloenzymes: topological and sequential relations between the atacins, adamalins, serralysins, and matrixins (collagenases) define a superfamily of zinc-peptidases. Protein Sci 1995, 4:832–840.
2. Fujiwara N, Ee CE, Nakajima T, Ichishima E, Aspzinin, a family of metalloendopeptidases with a new zinc-binding motif. Identification of a new zinc-binding sites (His128, His132, and Asp164) and three catalytically crucial residues (Glu129, Asp143, and Tyr106) of deuterolysin from Aspergillus oryzae by site-directed mutagenesis. J Biol Chem 1999, 274:24105–24101.
3. Rawlings ND, Barrett AJ: Evolutionary families of metallopeptidases. Methods Enzymol 1995, 248:183–228.
4. Kurisu G, Kinosita T, Sugimoto A, Nagara A, Kai Y, Kasai N, Harada S: The Pfam protein families database. Nucleic Acids Res 2013, 41:D8-D20.
5. Punta M, Coggill PC, Eberhardt RY, Misty J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD: The Pfam protein families database, Nucleic Acids Res 2012, 40:D290–D301.
6. Hutchinson EG, Thornton JM: HERA—a program to draw schematic diagrams of protein secondary structures. Proteins 1990, 8:203–212.
7. Karch K, Standley DM: MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013, 30:772–780.
8. Krissinel E, Henrick K: Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007, 372:774–797.
9. Li L, Li C, Sarkar S, Zhang J, Witham S, Zhang Z, Wang L, Smith N, Petukh M, Alexov E: DELPHI: a comprehensive suite for DELPHI software and associated resources. BMC Biophys 2012, 6:17.
10. Wu H, Lohkamp B, Scott WG, Cowtan K: Features and development of Coot. Acta Crystallogr Sect D Biol Crystallogr 2010, 66:496–501.
11. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 2004, 25:1605–1621.
12. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Pfister A, Quesada M, Quinn GB, Ramos AG, Westbrook JD, Young J, Zardecki C, Berman HM, Bourne PE: The RCSB protein data bank: new resources for research and education. Nucleic Acids Res 2013, 41: D475–D482.
13. Matthews BW: Solvent content of protein crystals. J Mol Biol 1968, 33:491–497.
14. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Klaesemann G, Merz KM, et al: MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Sect D Biol Crystallogr 2010, 66:122–128.
15. Jaroszewski L, Li Z, Cai XH, Weber C, Godzik A: Inference of macromolecular assemblies from cryo-electron microscopy images. J Mol Biol 2007, 212:802–813.

Page 11 of 12

BMC Bioinformatics 2014, 15:1
http://www.biomedcentral.com/1471-2105/15/1

Trame et al. BMC Bioinformatics 2014, 15:1
Page 11 of 12
http://www.biomedcentral.com/1471-2105/15/1

16. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Sraimong G, Nickel HE, McMullan D, Shin T, Vincent J, Rabbi A, Birin LS, Miller MD, McPhills TM, Miller MA, Schelbe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger MA, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC: An approach to rapid protein crystallization using nanodroplets. J Appl Crystallogr 2002, 35:278–281.
17. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Sraimong G, Nickel HE, McMullan D, Shin T, Vincent J, Rabbi A, Birin LS, Miller MD, McPhills TM, Miller MA, Schelbe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger MA, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC: An approach to rapid protein crystallization using nanodroplets. J Appl Crystallogr 2002, 35:278–281.
18. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Sraimong G, Nickel HE, McMullan D, Shin T, Vincent J, Rabbi A, Birin LS, Miller MD, McPhills TM, Miller MA, Schelbe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger MA, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC: An approach to rapid protein crystallization using nanodroplets. J Appl Crystallogr 2002, 35:278–281.
19. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, Kreusch A, Sraimong G, Nickel HE, McMullan D, Shin T, Vincent J, Rabbi A, Birin LS, Miller MD, McPhills TM, Miller MA, Schelbe D, Canaves JM, Guda C, Jaroszewski L, Selby TL, Elsliger MA, Wooley J, Taylor SS, Hodgson KO, Wilson IA, Schultz PG, Stevens RC: An approach to rapid protein crystallization using nanodroplets. J Appl Crystallogr 2002, 35:278–281.
functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 2013, 41:D499–D507.

37. Yiallouros I, Grosse Berkhoff E, Stöcker W: The roles of Glu93 and Tyr149 in astacin-like zinc peptidases. FEBS Lett 2000, 484:224–228.

38. Gomis-Ruth FX: Structural aspects of the metzincin clan of metalloendopeptidases. Mol Biotechnol 2003, 24:157–202.

39. Garcia-Castellanos R, Tallant C, Marreo A, Sola M, Baumann U, Gomis-Ruth FX: Substrate specificity of a metalloprotease of the pappalysin family revealed by an inhibitor and a product complex. Arch Biochem Biophys 2007, 457:57–72.

40. Waltersperger SM, Widmer C, Baumann U: Crystal structure of archaemetzincin AmzA from Methanopyrus kandleri at 1.5 A resolution. Proteins 2010, 78:2720.

41. Graef C, Schacherl M, Waltersperger S, Baumann U: Crystal structures of archaemetzincin reveal a moldable substrate-binding site. PLoS One 2012, 7:e3863.

doi:10.1186/1471-2105-15-1
Cite this article as: Trame et al: New mini-zincin structures provide a minimal scaffold for members of this metallopeptidase superfamily. BMC Bioinformatics 2014 15:1.