

Research Paper

***Schistosoma mansoni* cathepsin D1: biochemical and biophysical characterization of the recombinant enzyme expressed in HEK293T cells**

Araujo-Montoya, BO¹; Senger, MR¹; Gomes, BF¹; Harris, G², Owens, RJ^{2,3}; Silva-Jr, FP^{1,*}

¹ *Laboratório de Bioquímica Experimental e Computacional de Fármacos, LaBECFar, Instituto
Oswaldo Cruz, Fundação Oswaldo Cruz. Avenida Brasil 4365, CEP 21040-360, Rio de Janeiro, RJ,
Brazil.*

²*Research Complex at Harwell, R92 Rutherford Appleton Laboratory, Didcot, OX11 0FA, UK*

³*The Division of Structural Biology, Henry Wellcome Building for Genomic Medicine
Roosevelt Drive, Oxford, OX3 7BN, UK*

*Corresponding author:

F.P.S.-J.: phone, + 55 21 3865 8248; fax, +55 21 25903495; E-mail, floriano@ioc.fiocruz.br.

Abstract

Schistosomes express a variety of aspartyl proteases (APs) with distinct roles in the helminth pathophysiology, among which degradation of host haemoglobin is key, since it is the main amino acid source for these parasites. A cathepsin D-like AP from *Schistosoma mansoni* (SmCD1) has been used as a model enzyme for vaccine and drug development studies in schistosomes and yet a reliable expression system for readily producing the recombinant enzyme in high yield has not been reported. To contribute to further advancing the knowledge about this valuable antischistosomal target, we developed a transient expression system in HEK 293T mammalian cells and performed a biochemical and biophysical characterization of the recombinant enzyme (rSmCD1). It was possible to express a recombinant C-terminal truncated form of SmCD1 (rSmCD1 Δ CT) and purify it with high yield (16 mg/L) from the culture supernatant. When analysed by Size-Exclusion Chromatography and multi-angle laser light scattering, rSmCD1 Δ CT behaved as a dimer at neutral pH, which is unusual for cathepsins D, turning into a monomer after acidification of the medium. Through analytical ultracentrifugation, the dimer was confirmed for free rSmCD1 Δ CT in solution as well as stabilization of the monomer during interaction with pepstatin. The mammalian cell expression system used here was able to produce rSmCD1 Δ CT with high yields allowing for the first time the characterization of important kinetic parameters as well as initial description of its biophysical properties.

Key words

Cathepsin D, schistosome, HEK 293T cells, analytical ultracentrifugation, SEC-MALS, dimer.

Abbreviations

rSmCD1: recombinant *Schistosoma mansoni* cathepsin D1, rSmCD1 Δ CT: recombinant C-terminal truncated *Schistosoma mansoni* cathepsin D1, SDS-PAGE: Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis, IMAC: Immobilized-Metal Affinity Chromatography, SEC: Size-Exclusion Chromatography, MALS: Multi-angle laser light scattering, AUC: analytical ultracentrifugation.

1. Introduction

The coordinated activity of both cysteinyl and aspartyl proteases from blood-feeding helminths, such as *Schistosoma mansoni*, degrades haemoglobin, the main amino acid source for these parasites [1, 2]. Aspartyl proteases are acidic enzymes that are found in most organisms, from humans to viruses, and can be inhibited by pepstatin, which is used as a diagnostic test for aspartyl proteases activity [3]. Most of aspartyl proteases of family A1 (MEROPS-Peptidase Database Cathepsin D A01.009), which includes pepsin and chymosin, are expressed as zymogens possessing a pro-peptide of up to 50 amino acids long that is cleaved upon activation at acidic pH [4]. All aspartyl proteases have characteristic sequences in the regions of the two catalytic aspartyl residues: (hydrophobic-F/I/L-D-T-G-S) in the N-terminal domain, and a corresponding (hydrophobic-D-T/S-G-S/T) in the C-terminal domain [3]. These enzymes can accommodate up to eight amino acids in the binding cleft of the active site and prefer hydrophobic amino acids at both sides of the cleaved bond.

Schistosomes express a specific member of the A1 family, similar to vertebrate cathepsin D, which plays a major role in the digestion of haemoglobin [5, 6]. This cathepsin D-like enzyme has proved to be essential for the metabolism of the parasite, especially during the larval stage [7]. Cathepsin D from *S. mansoni* (SmCD1) possesses 84% identity to the orthologue from

1 *Schistosoma japonicum* and around 55% similarity to homologues from vertebrates and other
2 invertebrates. The mature enzyme is predicted to be 377 amino acids long with a molecular
3 mass of 41.2 kDa. A striking feature of SmCD1 is the presence of a signature sequence for
4 trematode cathepsin D, which consists of a C-terminal extension of 43 amino acid residues that
5 has similarity to the C-terminal extensions found in cathepsins D from other liver trematodes
6 (*Clonorchis sinensis*, *Fasciola hepatica*, *F. gigantica* and *Opisthorchis viverrini*) and is absent
7 from any other cathepsins. The exon structure of the SmCD1 gene is similar to human cathepsin
8 D; the C-terminal extension is coded in the last exon of the SmCD1 gene [8]. Furthermore, our
9 group has deposited the full-length ORFs of two additional isoforms (SmCD2 and 3) that could
10 be cloned from adult *S. mansoni* parasites. A study detecting enrichment of the SmCD2
11 transcript in the female gastrodermis compared to the whole female body [9] has confirmed the
12 importance of this family of proteases in parasite's biology. Nonetheless, despite these few
13 findings on the biological role of the parasite's enzymes carried out so far, structural and
14 functional characterization at the molecular level still is scarce.

15 Haemoglobinolytic activity from an aspartyl protease was first demonstrated in *S. mansoni*
16 extracts by Cesari and collaborators in 1998 [10]. Subsequently, recombinant truncated forms
17 of this enzyme from both *S. mansoni* and *S. japonicum* were expressed in insect cells and shown
18 to proteolyse haemoglobin *in vitro*, mapping its cleavage sites and demonstrating by immune-
19 staining that SmCD1 is localised to the gastrodermis of the parasite [11]. Silva-Jr and
20 collaborators in 2002 noted that both *S. mansoni* and *S. japonicum* SmCD1 homologues show
21 a preference for cleaving substrates with Pro at P1' site, a feature previously only attributed to
22 retroviral proteases [12]. This observation suggested the idea that the HIV-1 protease inhibitors
23 could be exploited in the design of inhibitors selective for SmCD1 over mammalian enzymes
24 [12]. This hypothesis was tested in 2006, when Delcroix *et al.* observed the nearly complete
25 disruption of haemoglobin degradation in the parasite's gastrodermis after incubation with the

commercially available HIV-1 protease inhibitor Lopinavir [2]. Subsequently, RNAi studies confirmed the importance of the enzyme in the immature stages, impairing haemoglobin degradation and consequently, larval maturation [13].

In this context, SmCD1 proved to have great potential as a therapeutic target for anti-schistosome treatment. Hence, we have developed an expression system for readily producing a recombinant C-terminal truncated form of the enzyme in high yield using mammalian cells and report the biochemical and biophysical characterisation of the purified enzyme.

2. Materials and Methods

2.1 Biological material

Adult worms were obtained by perfusion from infected Balb/c mice. Total RNA was extracted from homogenized adult worms using TRIzol reagent (Life Technologies) according to manufacturer instructions. cDNA was synthesized from total RNA using SuperScriptTM III First-Strand Synthesis System (Life Technologies).

2.2 SmCD1 cloning and expression

SmCD1 was expressed as a recombinant truncated proenzyme form, lacking the trematode-conserved C-terminal 43 residues (aa 13-385, rProSmCD1 Δ CT). Amplification was done using pFastBac1-proSmCD1 or cDNA as a template for either proenzyme or preproenzyme (acc. U60995.1) and inserted into the expression vectors pOPING, H and E (supplementary figure 1) by Infusion cloning [14]. Transient expression of rProSmCD1 Δ CT was carried out in HEK 293T cells. Briefly, $1.5\text{--}2 \times 10^6$ cells/mL were transfected (GeneJuice, Life Technologies) with $\sim 1 \mu\text{g}$ of each construct and after 72 h, media were collected and analysed by western blotting with a monoclonal anti-polyhistidine antibody (Sigma). Large-scale production of rProSmCD1 Δ CT was carried out by transfecting HEK 293T cells grown in either T-175 flasks (20 x 50 ml culture volume), or roller bottles (4 x 250 mL culture volume) with 1 mg/L culture

volume of the positive-expression construct (from previous screening) in a solution of 1 mg/mL PEI (Polyethyleneimine, Sigma) [15]. Media were harvested after 3-5 days.

2.3 Protein purification

Media collected from the large-scale expression in HEK 293T cells were used as the source material to purify the recombinant protein produced and secreted by the mammalian cells. A 5 mL HisTrap FF column *in tandem* with a HiLoad 16/60 Superdex 200 pg column connected to an Äkta Xpress system was used to carry out protein purification. Affinity chromatography wash/sample buffer (Tris 50 mM, NaCl 500 mM, Imidazole 30 mM, pH 7.5) was allowed to flow at a speed of 5 mL/min during the IMAC, as well as the elution buffer (Tris 50 mM, NaCl 500 mM, Imidazole 500 mM, pH 7.5). Gel filtration buffer (Tris 20 mM, NaCl 300 mM, pH 7.5) was pumped at a speed of 1.2 mL/min for the SEC. The elution peak collected during IMAC was injected into the SEC column and the fractions collected every 2 mL. These fractions were analysed by 4-10% SDS-PAGE for purity and quantity. Fractions containing the protein of interest were pooled and concentrated for subsequent biophysical and biochemical experiments. Protein quantification was made using UV absorbance and theoretical molar extinction coefficients for both forms applying the Lambert-Beer law.

2.4 Activity assays

For haemoglobinolytic activity experiment, activity buffer 2X was prepared (200 mM sodium acetate-HCl, 200 mM NaCl, pH 3.8). 50 µL of this buffer was mixed with 20 µg rProSmCD1ΔCT and 100 µg haemoglobin (Sigma), plus distilled water, to make a final volume of 100 µL. This mixture was incubated at 37 °C for a period of 4-16 h and then analysed by SDS-PAGE. Negative (no enzyme added) control used distilled water instead of protein.

For the kinetic experiments, enzyme (25 ng) was added to the reaction mixture containing 100 mM sodium acetate-HCl + 100 mM NaCl at pH 3.5, in a final volume of 100 μ L. The reaction was started with the addition of 2 μ M of a FRET substrate (7-methoxycoumarin-4-acetyl-GKPILFFRLK(DNP)-D-R-amide, Sigma) and activity recorded during 5 min at 37 °C. For pH variation experiments in the range of 2 to 4.5, the buffer 100 mM sodium acetate-HCl + 100 mM NaCl was used while for pH 7.4 a 100 mM Tris + 100 mM NaCl buffer was used. To calculate the kinetic parameters of the enzymes, Michaelis-Menten constant (K_m) and maximum velocity (V_{max}), eight points of substrate concentration were used. To determine the inhibitor concentration at which enzyme activity is reduced by 50% (IC_{50}), serial 10-fold dilutions of pepstatin (catalogue No. P5318), ranging from 10,000 to 0.0003 nM), were tested by pre-incubating 10 min with rProSmCD1 Δ CT prior to addition of substrate. Experiments were read in a FlexStation III system (Molecular Devices) set for reading fluorescence, with an excitation wavelength of 310 nm and emission at 420 nm. The collected data were plotted as Relative Fluorescence Units (RFU). Initial velocity was obtained by the slope, calculated by linear regression, of the 0-180 seconds of the linear part of the progress curve giving the changes in the product concentration. All the experiments were repeated at least twice, each one in triplicate.

Pepstatin IC_{50} value was calculated using GraphPad Prim version 5.00 software, USA. Michaelis-Menten constant (K_m) was calculated using Sigmaplot 12.0 software from Systat software Inc, USA. The constant was determined with help of EK module by fitting and additionally by analysis of Michaelis-Menten and Lineweaver-Burk plots.

2.5 Size-exclusion chromatography and multi-angle laser light scattering (SEC-MALS)

The molar mass (M_w) and M_w distributions of monomeric and dimeric forms of rProSmCD1 Δ CT were determined on an ÄKTA Pure chromatography system equipped with a Superdex 200 10/300 GL column (catalogue No. 28-9909-44). The sample (0.1 mL, 1 mg/mL) was applied onto the column at a flow rate of 0.7 mL \cdot min $^{-1}$ in a buffer consisting of 20 mM Tris-HCl pH 7.5, 200 mM NaCl, or 100 mM sodium acetate, 150 mM NaCl, pH 3.8. The MALS system was a Wyatt DAWN HELEOS II with an added Wyatt QELS dynamic light-scattering unit connected to a Wyatt Optilab T-rEX refractive-index detector. The data were analysed using the Wyatt *ASTRA* 6 software (Wyatt Technology).

2.6 Analytical Ultracentrifugation

For characterisation of the protein samples, sedimentation velocity (SV) scans were recorded for a series of dilutions, starting from either 1.0 or 0.7 mg/mL, for the neutral (20 mM Tris, 200 mM NaCl, pH 7.5) or acidic (100 mM sodium acetate, 150 mM NaCl, pH 3.8) condition, respectively. All experiments were performed at 45000 rpm, using a Beckman XL-I analytical ultracentrifuge with an An-50Ti rotor. Data were recorded using the absorbance (at 280 nm) and interference optical detection systems. The density and viscosity of the buffers were measured experimentally using a DMA 5000M densitometer equipped with a Lovis 200ME viscometer module. The partial specific volume for the protein was calculated from the amino acid sequence using the public domain software program SEDNTERP, developed by Hayes, Laue and Philo (<http://www.jphilo.mailway.com/download.htm#SEDNTERP>). Data were processed using SEDFIT [16], fitting to the sedimentation coefficients - c(s) model.

3. Results and Discussion

A truncated form of the proenzyme of SmCD1 (372 aa, 40.8 kDa) lacking the 43-residue C-terminal sequence that is not conserved among vertebrate and most invertebrate cathepsin Ds

was cloned into three vectors (pOPING, pOPINE and pOPINH) for expression in mammalian cells. The expression vectors pOPING and pOPINE add a C-terminal His-tag to the inserted sequence and either replace the native signal sequence with the μ phosphatase signal sequence which is resident in the vector (pOPING) or enable the native signal sequence to be retained (pOPINE). pOPINH introduces a N-terminal His-tag downstream of the μ phosphatase signal sequence (Supplementary figure 1). Of the three vectors tested, pOPING gave the highest level of expression of truncated rProSmCD1 Δ CT as assessed by western blotting of cell supernatants following transient expression in HEK 293T cells (Figure 1a). The doublet that was observed may be due to partial *N*-glycosylation at one or both predicted *N*-linked glycosylation sites in the enzyme at N109 and N200. Substituting the endogenous leader sequence for the native one was the key to obtaining successful secretion of the enzyme. The construct with C-terminal His-tag (pOPING) expressed significantly higher rProSmCD1 Δ CT amount than the N-terminal His-tag version (pOPINH).

rProSmCD1 Δ CT was purified from 1 L media of transiently transfected HEK 293T cells culture with a yield of approximately ~16 mg/L and purity of $\geq 95\%$, as assessed by SDS-PAGE densitometry (Figure 1b). It is notable that the yield of rProSmCD1 Δ CT obtained by transient expression in HEK cells was sixteen times higher than previously reported for production of the full-length form using the baculovirus/insect cell expression system of approximately 1 mg/L of cell culture [11]. As expected, incubation of the purified rProSmCD1 Δ CT at pH 3.8 was associated with a molecular size shift and release of an approximately 4 kD species as assessed by SDS-PAGE (Figure 1b). Incubation of haemoglobin with rProSmCD1 Δ CT at pH 3.8 resulted in digestion of the protein into low molecular weight fragments (Figure 1c).

The activity of rSmCD1 produced in HEK cells was further analysed using the fluorogenic substrate 7-methoxycoumarin-4-acetyl-GKPILFFRLK(DNP)-D-R-amide. Activity was only

1 observed at acidic pHs with optimum activity between pH 3-4 (Figure 2a). The percentage of
2 active enzyme within the preparation was determined to be 32.6%, by titration with pepstatin
3 (Supplementary Figure 2). The enzyme showed simple Michaelis-Menten kinetics with a K_m
4 and V_{max} for the substrate determined: $0.93 \pm 0.65 \mu M$ and of 520 ± 88 AU/min, respectively
5 (Figure 2b, c). The IC_{50} for inhibition by pepstatin was measured as 7.0 nM (Figure 2d).

6 The oligomeric state of rProSmCD1 Δ CT was investigated by SEC-MALS. At neutral pH (7.5),
7 the enzyme eluted with a molecular weight of 85.2 kDa, corresponding to the dimer. Whereas
8 incubation at acidic pH (3.8) for 1 hour resulted in the protein eluting with a molecular weight
9 of 42.1 kDa which would correspond to a monomer (Figure 3). Interestingly, a 5 min incubation
10 at acid pH resulted in a broad protein peak centred on 60 kDa indicative of a mixture of
11 monomers and dimers.

12 Sedimentation velocity experiments by AUC confirmed the dimerization of rProSmCD1 Δ CT
13 and effect of lowering pH. At three enzyme concentrations (2.39, 11.9 and 23.9 μM) and
14 neutral pH rProSmCD1 Δ CT ran as a species with a sedimentation coefficient of 5S and a
15 molecular weight consistent with that of a dimer (Figure 4a and Table 1). Under the acidic
16 conditions, the protein, appeared predominantly as a species at around 3-4 S with a minor
17 species at 7S. A concentration dependant shift in the sedimentation coefficient of the major
18 species, from a lower S-value towards a higher one was observed indicative of an equilibrium
19 between the monomer and a higher order oligomer (Figure 4b).

20 To date there have been no reports of that cathepsin D forms dimers, although they are a feature
21 of cathepsin E [17, 18] and retroviral aspartyl proteases [19, 20]. In the case of cathepsin E, a
22 disulphide bond links the two monomers to form the functionally active dimer [18]. There are
23 no equivalent cysteines in SmCD1 (Supplementary figure 3) that could form a disulphide
24 bridge between the monomers. It may be speculated that the observed pH dependent dimer-

monomer transition of SmCD1 is indicative of conformational change that occurs on activation of the enzyme through pro-peptide cleavage.

4. Summary

In this work we have reported the expression of recombinant C-terminal truncated cathepsin D from *S. mansoni* in mammalian HEK293T cells and recovery of enzyme from cell media in high yield (~16 mg/ L cell culture) by a combination of IMAC and SEC. The purified enzyme showed activity in degrading haemoglobin and cleaving a commercial aspartyl protease peptide substrate, as well being inhibited by pepstatin, a classic inhibitor of aspartyl proteases. Unexpectedly, rProSmCD1 Δ CT behaved as a non-covalent dimer in solution and resolved into a monomer on exposure to acidic conditions, presumably reflecting the conformational changes the protein undergoes on activation. This new expression system can contribute to further advancing the knowledge about this valuable antischistosomal target.

Acknowledgments

The authors would like to thank Brazilian funding agencies CNPq, CAPES, FAPERJ and FIOCRUZ for financial support and fellowships. This study received grants from FAPERJ, the Newton Fund/Research Councils UK (E-26/170.009/2015) and Newton Fund/Academy of Medical Sciences UK (ITPMZO55).

References

- [1] Brinkworth RI, Prociv P, Loukas A, Brindley PJ. 2001. Hemoglobin-degrading, aspartic proteases of blood-feeding parasites: substrate specificity revealed by homology models. *J Biol Chem* 276(42):38844-51.
- [2] Delcroix M, Sajid M, Caffrey CR, Lim KC, Dvorak J, Hsieh I, Bahgat M, Dissous C, McKerrow JH. 2006. A multienzyme network functions in intestinal protein digestion by a platyhelminth parasite. *J Biol Chem* 281(51):39316-29.
- [3] Szecsi PB. The aspartic proteases. *Scand J Clin Lab Invest Suppl.* 1992;210:5-22.
- [4] Dunn BM. Structure and mechanism of the pepsin-like family of aspartic peptidases. *Chem Rev.* 2002 Dec;102(12):4431-58.
- [5] Sauer MC, Senft AW. Properties of a proteolytic enzyme from *Schistosoma mansoni*. *Comp Biochem Physiol B.* 1972 Jun 15;42(2):205-20.
- [6] Becker MM, Harrop SA, Dalton JP, Kalinna BH, McManus DP, Brindley PJ. Cloning and characterization of the *Schistosoma japonicum* aspartic proteinase involved in hemoglobin degradation. *J Biol Chem.* 1995 Oct 13;270(41):24496-501. Erratum in: *J Biol Chem* 1997 Jul 4;272(27):17246
- [7] Morales ME, Rinaldi G, Gobert GN, Kines KJ, Tort JF, Brindley PJ. RNA interference of *Schistosoma mansoni* cathepsin D, the apical enzyme of the hemoglobin proteolysis cascade. *Mol Biochem Parasitol.* 2008 Feb;157(2):160-8. Epub 2007 Nov 1. PMID: 18067980
- [8] Morales ME, Kalinna BH, Heyers O, Mann VH, Schulmeister A, Copeland CS, Loukas A, Brindley PJ. Genomic organization of the *Schistosoma mansoni* aspartic protease gene, a platyhelminth orthologue of mammalian lysosomal cathepsin D. *Gene.* 2004 Aug 18;338(1):99-109.

- [9] Nawaratna, S. S., McManus, D. P., Moertel, L., Gobert, G. N., and Jones, M. K. Gene Atlasing of digestive and reproductive tissues in *Schistosoma mansoni*. PLoS Negl Trop Dis. 2011 Apr 26;5(4):e1043. doi: 10.1371/journal.pntd.0001043.
- [10] Cesari IM, Valdivieso E, Schrevel J. 1998. Biochemical characterization of cathepsin D from adult *Schistosoma mansoni* worms. Mem Inst Oswaldo Cruz 93 Suppl 1:165-8.
- [11] Brindley PJ, Kalinna BH, Wong JY, Bogitsh BJ, King LT, Smyth DJ, Verity CK, Abbenante G, Brinkworth RI, Fairlie DP *et al.*. 2001. Proteolysis of human hemoglobin by schistosome cathepsin D. Mol Biochem Parasitol 112(1):103-12.
- [12] Silva FP, Jr., Ribeiro F, Katz N, Giovanni-De-Simone S. 2002. Exploring the subsite specificity of *Schistosoma mansoni* aspartyl hemoglobinase through comparative molecular modelling. FEBS Lett 514(2-3):141-8.
- [13] Morales ME, Rinaldi G, Gobert GN, Kines KJ, Tort JF, Brindley PJ. 2008. RNA interference of *Schistosoma mansoni* cathepsin D, the apical enzyme of the hemoglobin proteolysis cascade. Mol Biochem Parasitol 157(2):160-8.
- [14] Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, Rahman N, Stuart DI, Owens RJ. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res. 2007;35(6):e45. Epub 2007 Feb 22.
- [15] Nettleship JE, Rahman-Huq N, Owens RJ. The production of glycoproteins by transient expression in Mammalian cells. Methods Mol Biol. 2009;498:245-63.
- [16] Schuck P. (2000). Size distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. Biophysical Journal 78:1606-1619.
- [17] Ostermann N, Gerhartz B, Worpenberg S, Trappe J, Eder J. Crystal structure of an activation intermediate of cathepsin E. J Mol Biol. 2004 Sep 17;342(3):889-99.

- 1 [18] Zaidi N, Kalbacher H. Cathepsin E: a mini review. Biochemical and Biophysical Research
2 Communication 367 (2008) 517-522.
- 3 [19] Sadiq SK, Noé F, De Fabritiis G. Kinetic characterization of the critical step in HIV 1
4 protease maturation. PNAS 2012, vol 19 No 50, 20449-20454.
- 5 [20] Darke PL, Jordan SP, Hall DL, Zugay JA, Shafer JA, Kuo LC. Dissociation and
6 association of the HIV 1 protease dimer subunits: equilibria and rates. Biochemistry
7 1994, 33, 98-105.
- 8

Table legends

Table 1. Hydrodynamic parameters

Species molecular weights from the c(s) analysis for rProSmCD1 Δ CT in both neutral (pH 7.5) and acidic conditions (pH 3.8). For each sample concentration the molecular weight of each species is shown, together with the best-fit frictional ratio for the distribution.

Figure legends

Figure 1. (a) Western blot of culture media analysed on SDS-PAGE 4-12% following expression of rProSmCD1 Δ CT in HEK 293T cells in three different vectors (lanes G, E and H). (b) SDS-PAGE 4-12% analysis of purified rProSmCD1 Δ CT (lane 1) and activated (lane 2) forms following 1 h incubation at pH 3.8 and 37 °C. The position of the pro-peptide is indicated by an arrow and the numbers refer to molecular weight standards. (c) SDS-PAGE 4-12% analysis of the haemoglobinolytic activity of rProSmCD1 Δ CT. Lane 1, activated enzyme. Lane 2, no enzyme control consisting of haemoglobin in activity buffer without recombinant protein.

Figure 2. Enzyme activity and inhibition of rProSmCD1 Δ CT. (a) Effect of pH on activity. (b) Reaction rate as a function of substrate concentration. (c) Lineweaver-Burke plot. (d) Inhibition by pepstatin.

Figure 3. SEC-MALS analysis of the rProSmCD1 Δ CT at acidic (3.8) and neutral (7.5) pH. Neutral buffer was composed of 20 mM Tris pH 7.5, 200 mM NaCl; acidic buffer was 100 mM sodium acetate pH 3.8, 150 mM NaCl. Samples were eluted at 60 min (or 5min where stated) during the run revealing the presence of a dimer at neutral pH that dissociated to a monomer at acidic pH.

Figure 4. Analytical ultracentrifugation of rProSmCD1ΔCT. Sedimentation velocity distributions of rProSmCD1ΔCT at (a) pH 7.5 and (b) pH 3.8. Initial sedimentation distributions were analysed in SEDFIT and subsequently fitted in SEDANAL.

Supplementary figure 1. Scheme depicting the constructs screened for expression of rProSmCD1ΔCT. While pOPING and H code for a heterologous leader sequence, pOPINE retains native leader sequence. RPTPmu: μ phosphatase signal sequence, K6His: KHHHHHH tag. 3C: cleavage site for 3C HRV viral protease.

Supplementary figure 2. Determination of active enzyme concentration ([Et]) by titration with tight-binding inhibitor pepstatin. Assays were performed at 37°C with 8 μ M of FRET substrate in 100 mM sodium acetate + 100 mM NaCl pH 3.5 buffer. Total protein used in each assay point was 21.5 nM and [Et] was determined to be 7.0 nM (32.6%).

Supplementary figure 3. Sequence alignment of cathepsins D and E from various organisms. Catalytic triad amino acid residues (DTG characters in red), and disulfide bridges (identical shades for pair of cysteines: green, purple and blue) marked. Cysteines for dimer formation in cathepsin E are shaded yellow. Conserved amino acid residues are shaded black. It is also possible to observe the C-terminal extension that is common to liver flukes: *Schistosoma mansoni* (P91802), *Schistosoma japonicum* (Q26515), *Opisthorchis viverrini* (Q45HJ6), *Clonorchis sinensis* (Q95VA2), *Fasciola gigantica* (AEE69372) and *Fasciola hepatica* (ACI04164). Remaining sequences aligned: *Homo sapiens* cathepsin D (P077339), *Mus musculus* cathepsin D (P18242), *Gallus gallus* cathepsin D (Q05744), *Clupea harengus* cathepsin D (Q9DEX3), *Xenopus laevis* cathepsin E (Q805F3), *Rana catesbeiana* cathepsin E

1 (Q800A0), *Oryctolagus cuniculus* cathepsin E (P43159), *Homo sapiens* cathepsin E (P14091),
2 *Mus musculus* cathepsin E (P70269), *Rattus rattus* cathepsin E (P16228).

3

4 **Supplementary figure 4.** Complete curve with all SEC-MALS data.

5

6

7

1 **Table 1**

Monomer MW (kDa)	Detection method	Concentration (mg/mL)	Major Species [#]		f/f ₀
			Peak 1 MW (kDa)	Peak 2 MW (kDa)	
Neutral					
41.7	Absorbance	1.0	80.5	-	1.20
		0.5	82.3	-	1.23
		0.1	86.3	-	1.27
	Interference	1.0	91.1	-	1.34
		0.5	90.2	-	1.35
		0.1	79.2	-	1.25
Acidic					
41.7	Absorbance	0.7	51.8	153	1.23
		0.35	48.3	133	1.22
		0.07	49.1	143	1.31
	Interference	0.7	53.2	130	1.30
		0.35	44.1	123	1.19
		0.07	28.9	90.7	0.95

2 [#]Both samples also contain other trace contaminants visible only in the interference data.

3

4

5

6

7

8

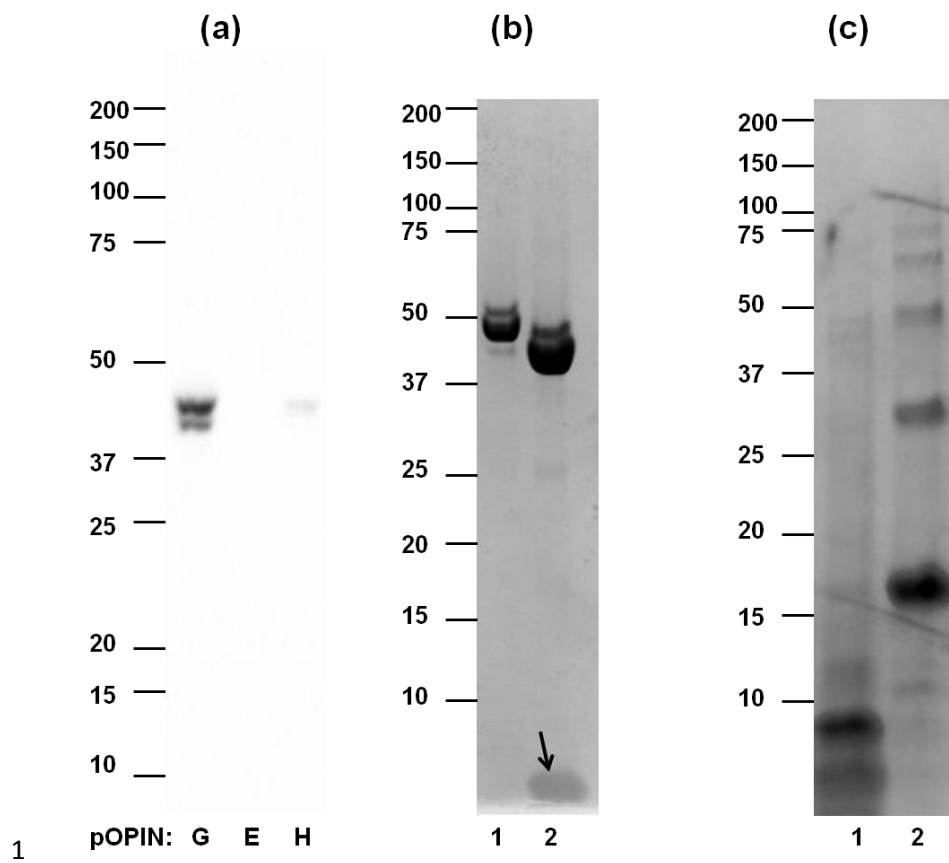


Figure 1.

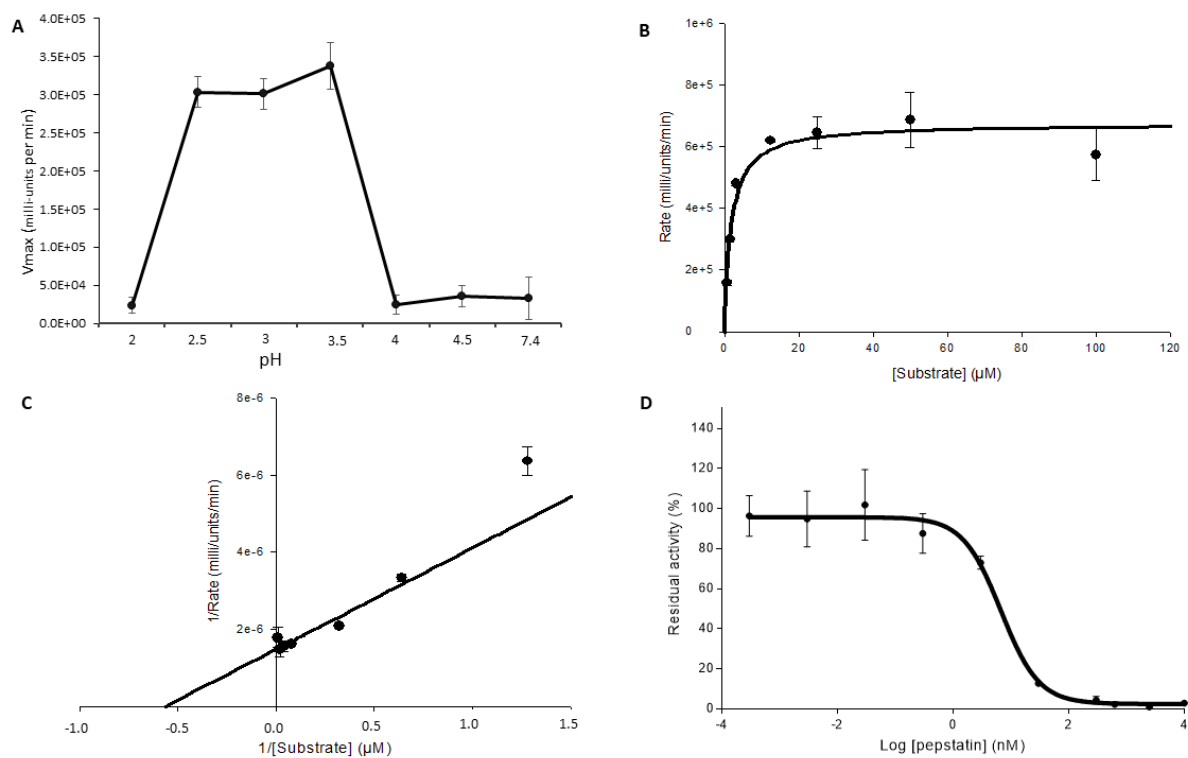


Figure 2.

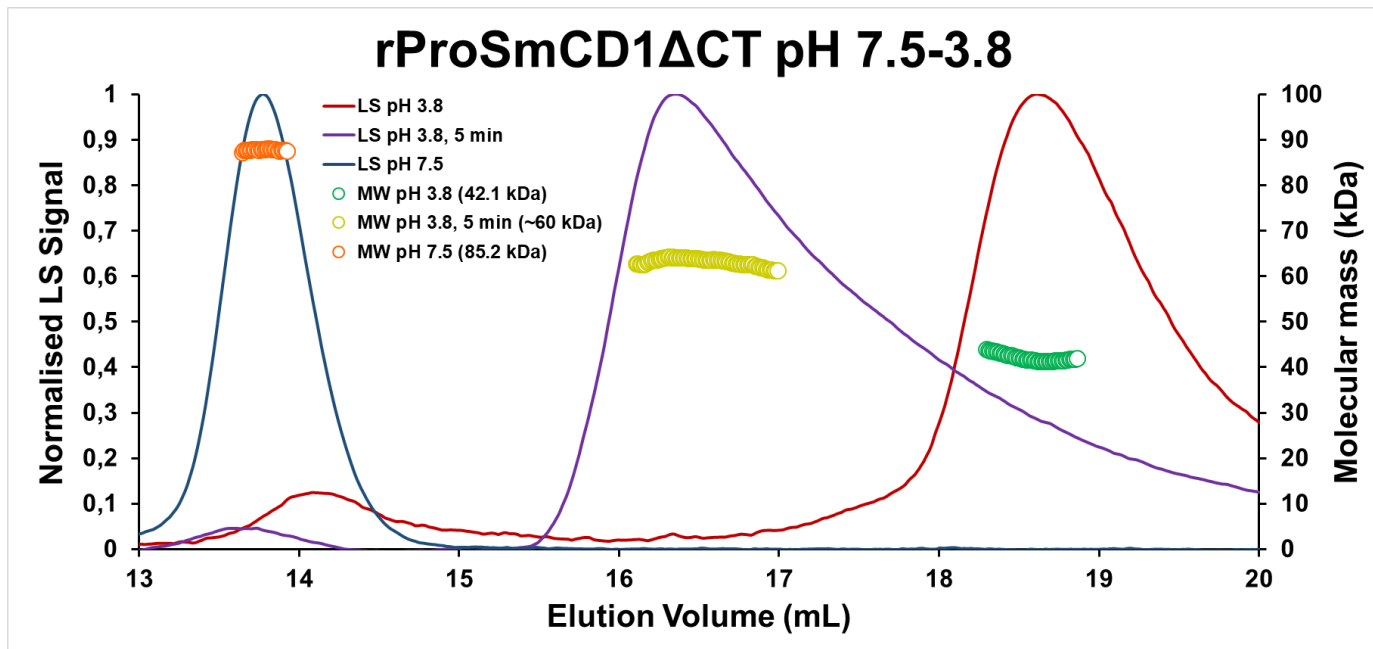


Figure 3.

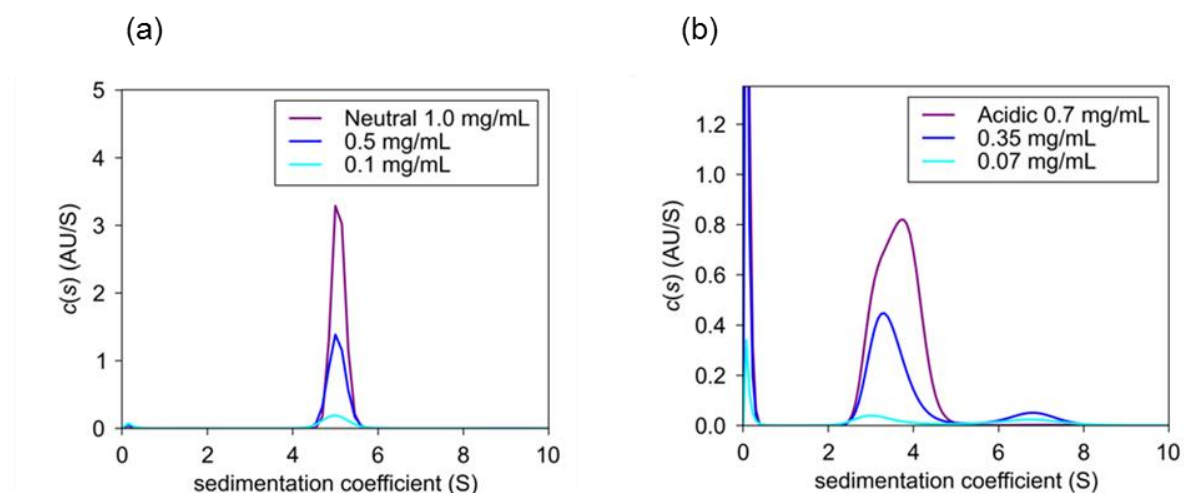
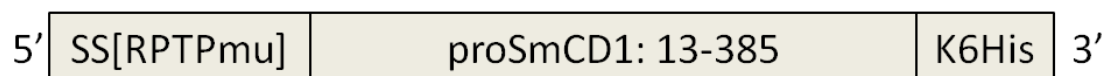
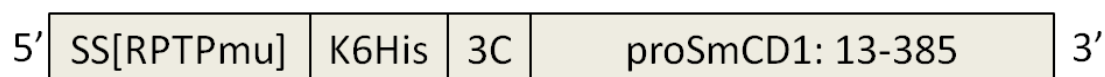


Figure 4.

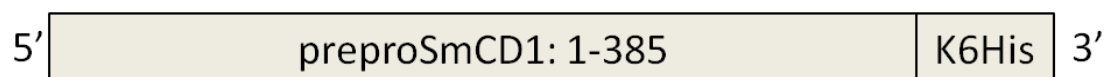
pOPING-proSmCD1ΔCT



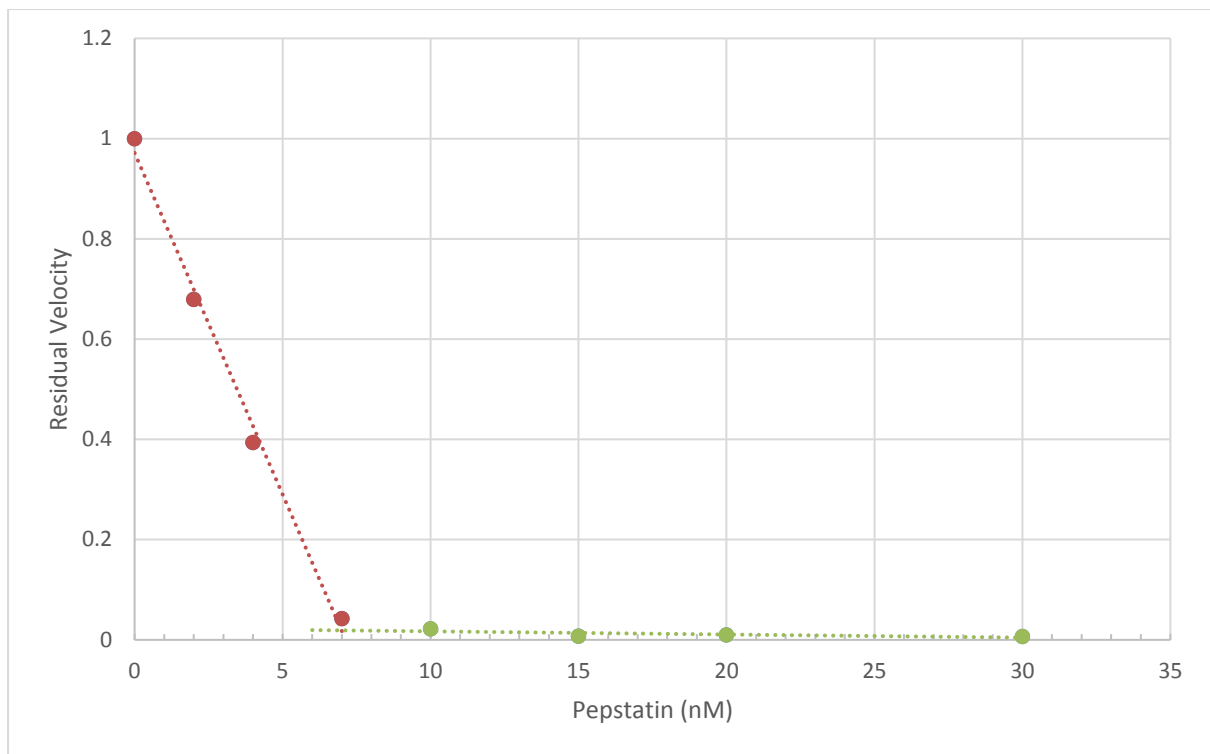
pOPINH-proSmCD1ΔCT



pOPINE-preproSmCD1ΔCT



Supplementary figure 1.



Supplementary figure 2.

SmCD1 : --EVVRIDLHPLKSAQRTIEFFETSLIV-----KKVWLSRVSG-VDPPHPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKYSYFDIAQLERKYSRSTSTVI : 106
 SjCD1 : --EVVRVLPVPLKSAQRTIEFFETSLVNV-----QKWFSSRFNS-VEPRFPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 106
 CATD_OPIVI : --SVIRIDLGGFKNVRRIMEVTPVQQL-----NFTSI-SFVG-NGSIFPRNNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 105
 ASPP_CLOSI : --SVIRIDLGGFKNVRRIMEVTPVQQL-----NFTSI-SFVG-NGSIFPRNNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 105
 CATD_FAGIG : --DVIRIRRPFKTTSCQISEYSLDWES-----SQRLFGKYAGRNGSIFPRNNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 107
 CATD_FAHEP : --DVIRIRRPFKTTSCQISEYSLDWES-----SQRLFGKYAGRNGSIFPRNNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 107
 CATD_HUMAN : --LVRIRLHKFTSIRMTSEVGSVEDLILKGPITKYSMQSSPKTTEPVSEPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 112
 CATD_MOUSE : --LVRIRLHKFTSIRMTSEVGSVEDLILKGPITKYSMQSSPKTTEPVSEPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 112
 CATD_CHICK : --LVRIRLHKFTSIRMTSEVSEIPDMNAITQFLKFKLG-FADLAETPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 111
 CATD_CLUHA : --LVRIRLHKFTSIRMTSEVSEIPDMNAITQFLKFKLG-FADLAETPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 111
 CATE_XENLA : --LVRIRLKRKQSIKTKEK-KLSHLWTQGGIDMVQVTDSSNDQAPSEPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 109
 CATE_RANCA : --LVRIRLKRKQSIKTKEK-KLSHLWTQGGIDMVQVTDSSNDQAPSEPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 109
 CATE_RABIT : --TLDRVLRQPSIKKRAQCQLSEFWKAKHVDVMQVTECTMEQSANPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 110
 CATE_HUMAN : --SLDRVLRQPSIKKRAQCQLSEFWKAKHVDVMQVTECTMEQSANPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 110
 CATE_MOUSE : QGALHVRIRRHQSIIKKRAQCQLSEFWKAKHVDVMQVTECTMEQSANPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 112
 CATE_RAT : QGVLRHVRIRRHQSIIKKRAQCQLSEFWKAKHVDVMQVTECTMEQSANPRNYDAVGGDTTGCPCHSIVFSSNLNWPSPKHSYFDIAQLERKYSRSTSTTV : 112

R pL r l g E L NY D Y G i iGtPpQ F V FDTgSSNLNWPSP C AC H S TY

SmCD1 : PNCETPSVHVGSGLSFSLSLQ-----LGSILGKCGTPEPATQCGVLNVMHFDGILGMAYPSISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 210
 SjCD1 : PNCETPSVHVGSGLSFSLSLQ-----LGSILGKCGTPEPATQCGVLNVMHFDGILGMAYPSISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 210
 CATD_OPIVI : PNCETPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 209
 ASPP_CLOSI : ANCTPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 203
 CATD_FAGIG : ANCTPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 211
 CATD_FAHEP : ANCTPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 211
 CATD_HUMAN : KNCETPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 227
 CATD_MOUSE : KNCETPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 225
 CATD_CHICK : KNCETPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 215
 CATD_CLUHA : KNCETPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 215
 CATE_XENLA : SNEPSPVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 213
 CATE_RANCA : SNEPSPVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 213
 CATE_RABIT : EVNTPSPVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 214
 CATE_HUMAN : QPCTPSVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 214
 CATE_MOUSE : EVNTPSPVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 216
 CATE_RAT : EVNTPSPVHVGSGVSGILSLIYVS-----VGTVTIKNCTPEAMKPEGIAVAFHFDGILGMGFRTISVGVTPFVNINICGLIESVVFVLSRNIASVIGEPHMI : 216

G F i YG Gs G D V Q FgE PG F a FDGILG v vF nm Q v Fg y G EL

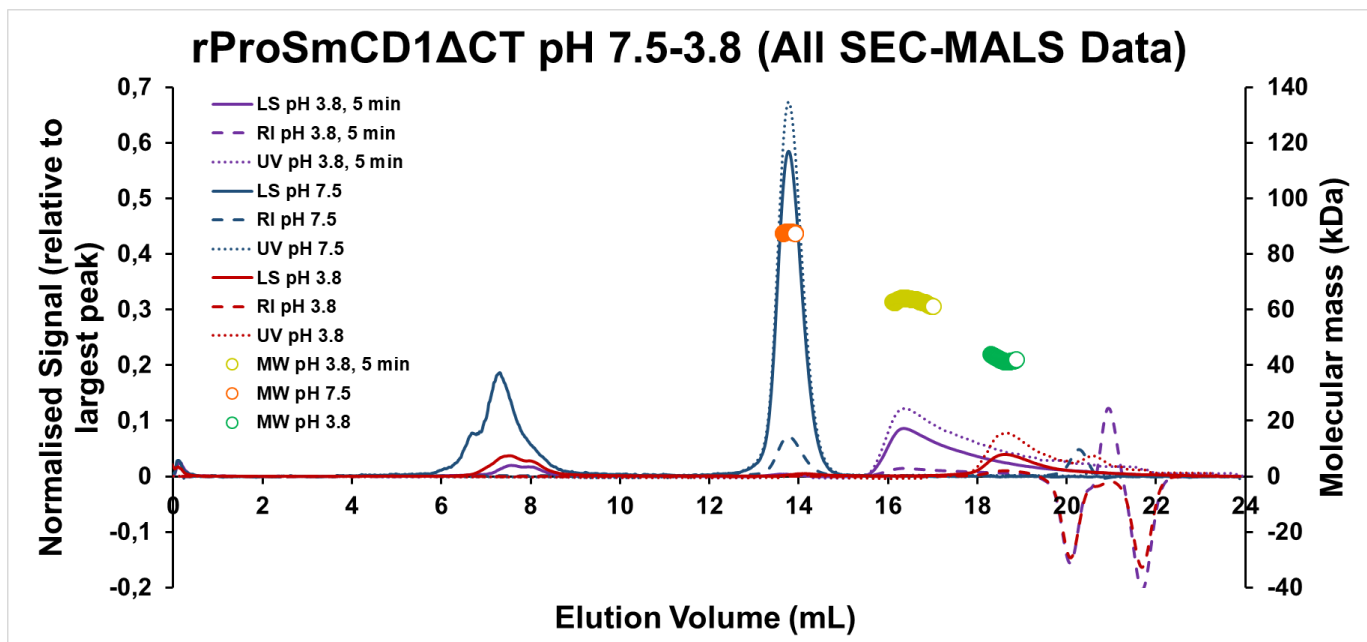
SmCD1 : GGLPKKYSFEINYNVDLQSSVLFKMLKLTIS-DMTAIPGCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 324
 SjCD1 : GGLPKKYSFEINYNVDLQSSVLFKMLKLTIS-DLSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 324
 CATD_OPIVI : GGLPKKYSFEILLWAPLTHEYVVKVSMNVG-GMKICENCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 323
 ASPP_CLOSI : GGLPKKYSFEILLWAPLTHEYVVKVSMNVG-GMKICENCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 317
 CATD_FAGIG : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 325
 CATD_FAHEP : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 325
 CATD_HUMAN : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 342
 CATD_MOUSE : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 340
 CATD_CHICK : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 330
 CATD_CLUHA : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 330
 CATE_XENLA : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 324
 CATE_RANCA : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 324
 CATE_RABIT : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 328
 CATE_HUMAN : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 328
 CATE_MOUSE : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 330
 CATE_RAT : GGLPKKYSFEILLWAPLTHEYVVKVRIEFP-GVSIQCCGIAHFTSMAGHDEIQINAKIKGTRLPGGIVTVSGNINNIITIDVINGKAMTTEPTILKYSKMG : 330

GG D G T YWq D C GC AI DTGTSII GP I Ga Y C p f g l Y G

SmCD1 : SEIGLTGEMLELLE-RKKLWILGDFICKKPIIFDMGKRAVGEAKVDPSSYHHTKVYSPMLRLFFAQSP-CAASETPNGVFAFSKLLSDVE : 415
 SjCD1 : SEIGLTGEMLELLE-RKKLWILGDFICKKPIIFDMGKRAVGEAKVDPSSYHHTKVYSPMLRLFFAQSP-CAASETPNGVFAFSKLLSDVE : 416
 CATD_OPIVI : KTLGLSGMGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 408
 ASPP_CLOSI : KTLGLSGMGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 402
 CATD_FAGIG : RTVGVTSIGLIVP-VGLLWILGDFVFGSNYVFDRLDNNVGGAEARL : 413
 CATD_FAHEP : RTVGVTSIGLIVP-VGLLWILGDFVFGSNYVFDRLDNNVGGAEARL : 413
 CATD_HUMAN : KTLGLSGMGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 392
 CATD_MOUSE : KTLGLSGMGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 390
 CATD_CHICK : KTLGLSGMGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 378
 CATD_CLUHA : KTLGLSGMGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 378
 CATE_XENLA : GGVCGSGGGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 381
 CATE_RANCA : GGVCGSGGGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 381
 CATE_RABIT : MQRGSGGGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 377
 CATE_HUMAN : MQRGSGGGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 377
 CATE_MOUSE : MQRGSGGGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 379
 CATE_RAT : MQRGSGGGLDIPFGAGLWLGDFVFGSNYVFDRLDNNVGGAEARL : 379

C gf G d gpLMIIGdvFI Y vFD rVG A

Supplementary figure 3.



1

2 **Supplementary figure 4.**