UPPSALA
UNIVERSITET

UPTEC X 19028

Examensarbete 30 hp
Juni 2019

Recombinant production of the *Giardia intestinalis* cysteine protease CP10217 in *Pichia pastoris*

Sofia Gräsberg

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Recombinant production of the *Giardia intestinalis* cysteine protease CP10217 in *Pichia pastoris*

Sofia Gräsberg

Giardia intestinalis is one of the leading causes of diarrheal diseases, affecting about 280 million people every year. By characterizing the virulence factors of *G. intestinalis*, new drug targets can be found to treat giardiasis. In addition to the adhesive disc and the variant surface proteins, cysteine proteases are some of the most interesting virulence factors in *Giardia*. In this project, one of the major secreted cysteine proteases, CP10217, was studied. The intent was to study the structure by modelling and to characterize CP10217 by expressing it in yeast cells and purifying the supernatant by using Immobilized Metal Affinity Chromatography (IMAC). The molecular modelling showed that the *Giardia* CP can be modelled on the structures of human and *Trypanosoma brucei* CPs. The process of expressing and purifying CP10217 in this manner proved difficult. The protease seemed to be very active when expressed, probably resulting in self-cleavage into its active form and later digestion of the whole protein, leading to a low protein yield from the purification. Two approaches were tested in order to increase the protein yield. First, expression at different pH ranges, and secondly, by re-cloning CP10217 with an extra 108 bp sequence at the 5' end. While changes in pH did not seem to affect the yield, sequencing results of the new vector showed that the cloning worked. More work on this new vector is needed to further analyse, and possibly characterize CP10217.

Handledare: Staffan Svärd
Ämnesgranskare: Erik Holmqvist
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC X 19028

Populärvetenskaplig sammanfattning

Rent vatten är en självklarhet för oss som bor i Sverige, men vad händer när våra vattenreningsverk slutar fungera? Sådant som vi inte tror finns, som vi är skyddade från med vårt rena vatten, kommer att infektera oss. Bakterier, virus, svampar och parasiter är allt sådant som renas bort och som kan orsaka stora epidemier när vi inte längre har rent vatten. Det är så det ser ut i många utvecklingsländer och en av de ledande orsakerna till att man får diarré av smutsigt vatten är den lilla, encelliga parasiten *Giardia intestinalis*. I sin cystform finns den i smutsigt vatten och tar sig ner till mag-och tarmkanalen. I tunntarmen spricker cystan upp och fyra s.k. trofozoiter simmar ut. Dessa fäster sig på tunntarmen och börjar replikera sig, ju fler de blir desto mer förstörs cellerna i tunntarmen. Detta gör att vatten börjar flöda mellan cellerna och leder till diarré. När trofozoiterna vandrar längre ner i tarmen börjar de anta sin cystform igen och med avföringen kommer de ut i naturen. Detta är livscykeln för *G. intestinalis*. Cystorna kan överleva upp till 4 veckor i naturen och närmare 280 miljoner människor blir årligen infekterade världen över.

Det finns läkemedel för att behandla en *Giardia*-infektion, men precis som med antibiotika för bakterieinfektioner, börjar även *Giardia* bli resistent. Just därför är det viktigt att upptäcka nya läkemedel. En av de viktiga beståndsdelarna när parasiten ska fästa sig till tunntarmen är de proteaser som den utsöndrar. Proteaserna klyver de proteiner som skyddar utsidan av tunntarmen, som består av mikrobiotan och ett slemlager, vilket skapar en passage för parasiten att simma ner igenom. Genom att studera vad det är som proteaserna klyver mer exakt kan man också få reda på hur man kan förhindra klyvningen. En typ av de proteaser som utsöndras är cysteinproteaser (CP). CP10217, som studerades i detta projekt, är en av de CP som utsöndras mest av parasiten och är därför ett intressant protein att undersöka vidare.

För att kunna studera CP10217 användes jästceller (*Pichia pastoris*) som uttrycker och kan utsöndra proteaset i större mängder. Eftersom CP10217 utsöndras i mediet behövde man endast mediet för att kunna rena fram CP10217. För att rena användes IMAC (immobilized metal affinity chromatography) som är en metod där proteaset fäster sig på nickelladdade gelkulor och sedan tvättas ut. Detta visade sig vara svårare än planerat då små mängder av CP10217 kunde fås fram, men inte nog för att vidare analysera det. Därför undersöktes det om olika pH-värden på mediet kunde påverka mängden slutprodukt. Resultatet visade ingen markant skillnad och därför testades det att klona om den vektor som hade använts för att uttrycka CP10217 i *P. pastoris*.

Detta genomfördes genom att lägga till en kort DNA-sekvens innan den kodande sekvensen för CP10217 i *P. pastoris* vektorn. Tanken med detta var att proteinet kommer att bli större och på så sätt mindre aktivt. Man kommer därför kunna få ut mer protein efter uttryck och rening. Det testades även att klona in CP10217 i en annan vektor för att kunna studera och lokalisera det i *G. intestinalis*. Sekvenseringsresultatet visade att kloningen lyckades för de båda vektorerna och kan därför fortsätta arbetas med. Då tiden för projektet tog slut kunde varken den extra DNA-sekvensens påverkan eller lokaliseringen av CP10217 i *G. intestinalis* undersökas vidare, vilket bör göras i framtida studier. Om den extra sekvensen fungerar, och mängden CP10217 blir markant högre, kan man gå vidare till karaktäriseringen av proteaset och på så sätt komma närmare nya läkemedel för att bekämpa *Giardia*-infektioner.

Table of Contents

Abbreviations

BSA	Bovine serum albumin
BMGY	Buffered glycerol-complex medium
BMMY	Buffered methanol-complex medium
CFS	Chronic fatigue syndrome
CP	Cysteine protease
DNA	Deoxyribonucleic acid
HCMP	High-cysteine membrane protein
IBS	Irritable bowel syndrome
IEC	Intestinal epithelial cell
IMAC	Immobilized metal affinity chromatography
MSA	Multiple sequence alignment
OE-PCR	Overlap-extension PCR
PCR	Polymerase chain reaction
SDS-PAGE	Sodium dodecyl sulphate–polyacrylamide gel electrophoresis
T_m	Melting temperature
UTR	Untranslated region
VSG	Variant surface glycoprotein
VSP	Variant surface protein

1 Introduction

Giardia intestinalis (syn. *G. lamblia* and *G. duodenalis*) is a non-invasive protozoan parasite infecting the human upper small intestine, causing diarrheal disease (giardiasis) in approximately 280 million people per year worldwide(Ankarklev *et al.* 2010). Although giardiasis is spread across the globe, many developing countries are considered endemic regions, where medication is limited, and the risk of contaminating water and food are higher (Einarsson *et al.* 2016a). It is not only humans that are infected with *G. intestinalis*, animals can also get giardiasis. To this date, there are eight groups, or assemblages, of *G. intestinalis* (assemblages A to H). Assemblages A and B infect both humans and animals, and the rest of the assemblages (C-H) are host-specific. Because of this, it is believed that some genotypes of assemblages A and B may have zoonotic potential. However, it has been shown that animals do not share identical multi-locus types with humans (Ryan & Cacciò 2013).

G. intestinalis has two major life cycle stages; the infectious cyst and the disease-causing, proliferating trophozoite. The route of infection is the oral-faecal route. Cysts are usually ingested via contaminated water or food, and the low dose of only 10 cysts are enough to cause infection (Ankarklev *et al.* 2010). Because of the low infectious dose, *Giardia* can cause large water-borne outbreaks (Efstratiou *et al.* 2017). When the cyst has been ingested, it excysts due to different stimuli, such as the acidic environment in the stomach, and bile and trypsin in the duodenum, and releases four trophozoites (Einarsson *et al.* 2016b). The trophozoites adhere to the epithelial cells in the upper small intestine, which can result in abdominal pain, nausea, vomiting, diarrhoea, malabsorption and weight loss. As the trophozoites migrate further down the small intestine, lipid starvation and the elevated pH in the environment signals the trophozoite to begin encystation (Einarsson & Svärd 2015). This will make the parasite ready for transmission into a new host as a cyst. The cyst will be able to survive out in the environment for several weeks until it is ingested by a new host (Einarsson *et al.* 2016a). The incubation time is usually 1-2 weeks and although the acute phase usually lasts 1-3 weeks, asymptomatic cases occur, showing no sign of giardiasis (Certad *et al.* 2017). It has been reported that post-infectious syndromes such as irritable bowel syndrome (IBS) and chronic fatigue syndrome (CFS) can occur after a *Giardia* infection. This was studied due to the large water outbreak in Bergen in 2004 where roughly 1200 people got infected by *Giardia*, and showed that 3 years later 46% had developed IBS (Litleskare *et al.* 2015). To treat giardiasis, 5-nitroimidazoles are used, such as metronidazole. Metronidazole causes DNA and protein damage to the parasite and can also target the redox enzyme, thioredoxin reductase, which causes oxidative stress when inactivated. Although this drug is effective to treat giardiasis, there are known cases of resistance to the drug and therefore new targets for drugs are needed (Einarsson *et al.* 2016a).

During host-cell interactions, the trophozoites upregulate several cell-associated mechanisms, including high-cysteine membrane proteins (HCMPs) and variant-specific surface proteins (VSPs) (Ringqvist *et al.* 2011). *G. intestinalis* also secretes a lot of proteins, including many proteases, upon host-cell contact (Ma'ayeh *et al.* 2017) and since *Giardia* is a non-invasive parasite, these proteins have been suggested to be involved in the pathogenesis of the disease (Einarsson *et al.* 2016a). Among the secreted proteins, cysteine proteases (CPs) seem to be important virulence factors due to their association with attenuated immune cell chemotaxis by cleaving chemokines (Cotton *et al.* 2014, Liu *et al.* 2018), intestinal epithelial cell (IEC) villin breakdown (Bhargava *et al.* 2015), disruption of the cellular junctions and increase of the intestinal permeability (Maia-Brigagão *et al.* 2012, Koh *et al.* 2013, Fisher *et al.* 2013, Halliez *et al.* 2016), effects on the intestinal normal flora and biofilm formation (Beatty *et al.* 2017), and growth inhibition of intestinal bacterial pathogens (Gerbaba *et al.* 2017). The current assembly of *G. intestinalis* WB (assemblage A) genome (www.giardiadb.org) contains 26 CP genes (Liu *et al.* 2019). Nine of these belong to the cathepsin B-like subfamily, including CP10217, CP14019, CP16160, CP16468, CP16779 and CP17516 which are the major secreted CPs from *G. intestinalis* during the interaction with the IECs (DuBois *et al.* 2006, Ma'ayeh *et al.* 2017, Dubourg *et al.* 2018). However, their roles in the molecular pathogenesis of *G. intestinalis* remains unclear for most of the CPs. Thus, by determining the substrate specificity of the major secreted CPs and searching for their targets accordingly, their roles could be determined.

CP10217 is a 912 bp (303 amino acid) long protein with a mass of about 33.5 kDa. CP10217 is synthesized as a proenzyme (syn. zymogen), *i.e.* an inactive enzyme, with a 19 amino acid long signal peptide followed by a pro-domain and the main domain for the mature, active enzyme (Figure 1). Proenzymes need to be cleaved in order to form the active enzyme. This happens when the pro-domain gets recognized by another active protease and cleaves the peptide bonds. After cleave, only the mature, active form of the enzyme remains (Nelson & Cox 2017, UniProt Consortium 2018). No current crystal structure of the protein exists, therefore a good structure prediction using the protein sequence is a good way to see how the protein could look like and how it could be affected when *e.g.* adding a tag to the C-terminal. As mentioned above, CP10217 (syn. CP1 and GL50803_10217) is one of the major secreted CPs and was one of the first to be detected (Ward *et al.* 1997, DuBois *et al.* 2006). This suggests that it is secreted in high amounts as well as being important for the first encounter with the IECs. The intent of this project was to characterize the cleavage specificity of CP10217. If proteins that are present in the human small intestine can be cleaved by this CP, a potential drug target to treat giardiasis could be characterized.

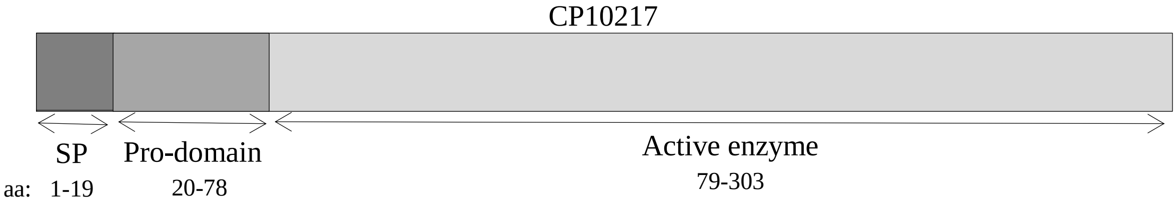


Figure 1. This figure summarizes the different domains of CP10217, starting with a 19 amino acid long signal peptide (SP), followed by the pro-domain, which will be cleaved away for CP10217 to be mature and active.

2 Materials and methods

2.1 Materials

A number of solutions and buffers were used to synthesise CP10217. Below they are listed with their names and contents.

2.1.1 Media

- *BMGY (Buffered Glycerol-complex Medium)*
  1% yeast extract, 2% peptone, 100 mM K_2HPO_4/KH_2PO_4 pH=6.0, 1.34% YNB, $4x10^{-5}$% biotin, 1% glycerol
- *BMMY (Buffered Methanol-complex Medium) with K_2HPO_4/KH_2PO_4 pH=6.0*
  1% yeast extract, 2% peptone, 100 mM K_2HPO_4/KH_2PO_4 pH=6.0, 1.34% YNB, $4x10^{-5}$% biotin, 0.5% methanol
- *BMMY with K_2HPO_4/KH_2PO_4 pH=7.0*
  1% yeast extract, 2% peptone, 100 mM K_2HPO_4/KH_2PO_4 pH=7.0, 1.34% YNB, $4x10^{-5}$% biotin, 0.5% methanol
- *BMMY with $C_2H_3NaO_2/CH_3COOH$ pH=4.0*
  1% yeast extract, 2% peptone, 100 mM $C_2H_3NaO_2/CH_3COOH$ pH=4.0, 1.34% YNB, $4x10^{-5}$% biotin, 0.5% methanol
- *BMMY with K_2HPO_4/KH_2PO_4 pH=8.0*
  1% yeast extract, 2% peptone, 100 mM K_2HPO_4/KH_2PO_4 pH=8.0, 1.34% YNB, $4x10^{-5}$% biotin, 0.5% methanol
- *Low salt LB medium pH=7.5 with Zeocin™*
  1% tryptone, 0.5% yeast extract, 0.5% NaCl, 25 µg/ml Zeocin™
- *Low salt LB agar, pH=7.5 with Zeocin™*
  1% tryptone, 0.5% yeast extract, 0.5% NaCl, 15 g/l agar, 25 µg/ml Zeocin™
- *LB medium with Ampicillin*
  LB medium, 50 µg/ml Ampicillin

2.1.2 Buffers

- *Washing buffer*
  20 mM imidazole, 500 mM NaCl, 20 mM NaH_2PO_4/Na_2HPO_4 pH=7.4
- *Elution buffer*
  500 mM imidazole, 500 mM NaCl, 20 mM NaH_2PO_4/Na_2HPO_4 pH=7.4
- *Exchange buffer*
  500 mM NaCl, 20 mM NaH_2PO_4/Na_2HPO_4 pH=7.4
- *Fixing solution*
  40% ethanol, 10% acetic acid
- *10x SDS-PAGE running buffer (TGS)*
  250 mM Tris, 1.92 M glycine, 1% SDS

- *TBST (Tris-buffered saline with tween)*
  TBS with 0.1% Tween 20
- *5% Milk blocking buffer*
  TBST with 5% non-fat (<0.1%) dry milk powder
- *Antibody dilution buffer with BSA*
  TBST with 5% Bovine Serum Albumin
- *1x Zymogram renaturing buffer*
  2.5% Triton X-100

2.2 Bioinformatic structure analysis of CP10217

Two different tools were used to perform multiple sequence alignment (MSA) on 6 orthologs to GL50803_10217. The orthologs are DHA2_150845, GLP15_4313, GL50581_3619, GSB_155190, GSB_153364 and GL50581_159. To perform the MSA, the online tools MUSCLE and CLUSTALΩ were used (Madeira *et al.* 2019). From the MSA, a phylogenetic tree could be built to study their relations. The online tool "One-click" phylogeny analysis was used to build the tree (Dereeper *et al.* 2010). To predict the secondary structure of CP10217, SPIDER2 (Yang *et al.* 2017) and JPred (Drozdetskiy *et al.* 2015) were used. And to predict the tertiary structure, SWISS-MODEL (Waterhouse *et al.* 2018) and Predictein (Ansell *et al.* 2019) were used. By using I-TASSER (Iterative Threading ASSEmbly Refinement), an overall analysis of CP10217 was done, including MSA, secondary structure, 3D-models, structurally close proteins and function predictions (Zhang 2008).

2.3 Expression of CP10217

Pichia pastoris was used for the recombinant production of CP10217. By having a yeast expression system, it is more likely that CP10217 will be folded correctly when synthesised in *P. pastoris*. Below it is described how this was done and what was tested to optimize the expression of CP10217.

2.3.1 Strain and expression vector

P. pastoris X-33 cells with the pPICZαA plasmid (Figure 2) was used to express CP10217. The yeast cells had been prepared by a PhD student before the project started according to the pPICZα A, B and C user manual (Invitrogen^{TM} 2010), with the exception that the sequence being inserted into the cloning site contained a polyhistidine (His$_6$) tag and a stop codon, and therefore the vector's own *c-myc* epitope, 6xHis tag and stop codon was not expressed.

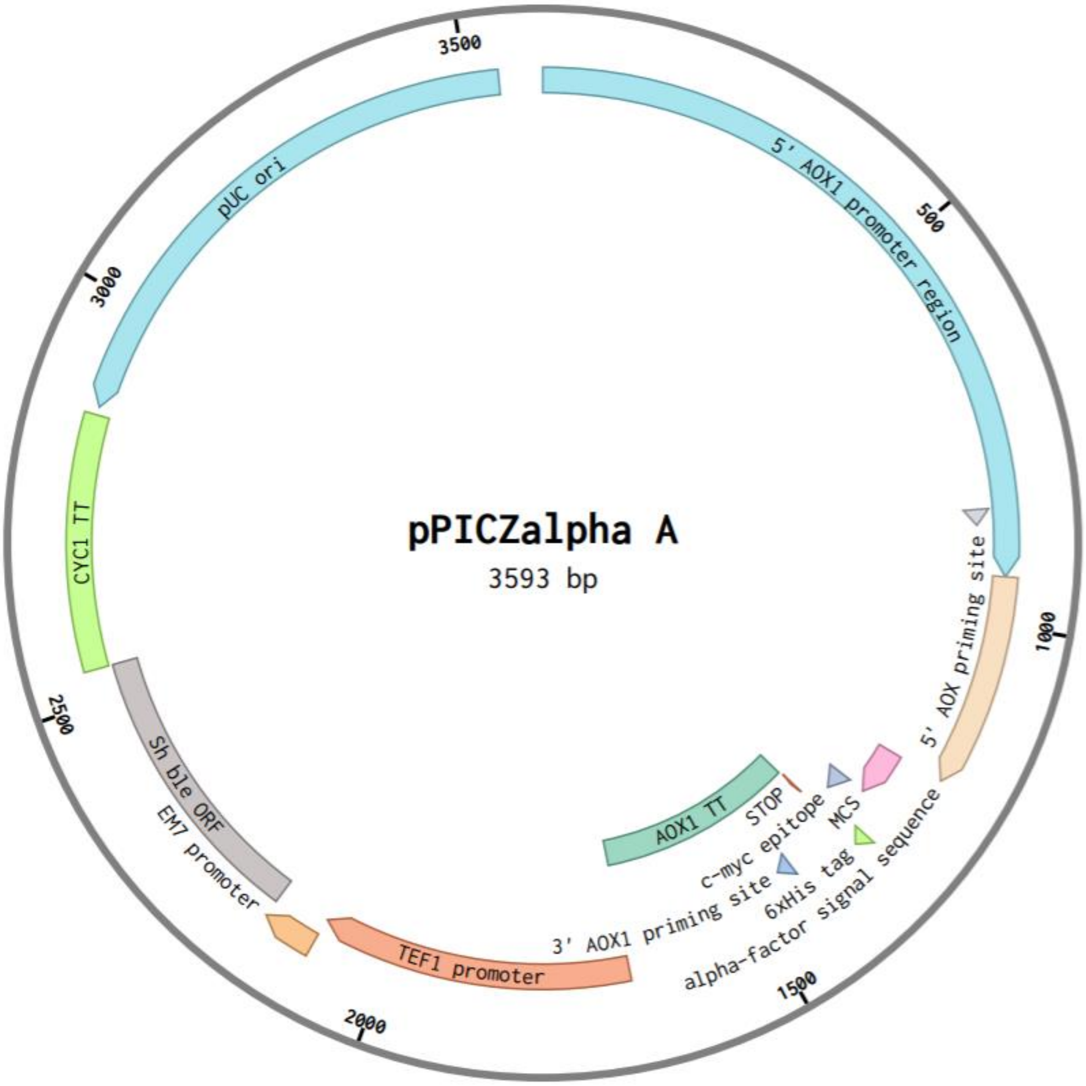


Figure 2. This figure summarizes the layout of the pPICZα A plasmid. The α-factor, which make it possible for *P. pastoris* to efficiently secrete the protein, is followed by a multiple cloning site to insert the gene of interest, a c-myc epitope, a polyhistidine (His$_6$) tag and a stop codon.

The CP10217 gene was first amplified by PCR from genomic DNA extracted from *G. intestinalis* WB isolate using the forward primer 10217-5': 5'-GGGAATTCGAGCTAAGGC GGATTCAGGC -3' and the reverse primer 10217-3': 5'-CCTCTAGATCAATGATGATGA TGATGATGGTCAAGATATACAGCATAGATTTCATCC-3'. The amplicons were cloned into the *EcoRI/XbaI* sites of the pPICZα A vector. It was then transformed into competent *Escherichia coli* TOP10 cells and grown on plates containing Zeocin^{TM}. Resistant strains were picked and sent for sequencing. The vector was linearized, and then transformed into competent *P. pastoris* X-33 cells by electroporation. Test-expression and test-purification was done to determine the optimal expression conditions.

2.3.2 Scale-up expression

Frozen *P. pastoris* X-33 cells expressing CP10217 was inoculated in 200 ml BMGY. The cells were incubated at 30°C with shaking at 150 rpm for 72 h. The cells were harvested and centrifuged at 3000*xg* for 5 min at room temperature. The pellet was resuspended in 1 litre BMMY to a final OD_{600}=1-1.5 to start induction. Cells were incubated at 30°C with shaking at 150 rpm for 72 hours. 100% methanol was added every 24 h to a final concentration of 0.5% to maintain secretion of protein. Cells were harvested by centrifuging at 3000*xg* for 15 min. Since the protein is secreted, the supernatant was saved for purification.

A second try to get more protein was done after the first purification. This time, pH 6 and pH 7 conditions were used. After 72 h inoculation in 200 ml BMGY and the harvesting of cells, one pellet was resuspended in 500 ml BMMY with K_2HPO_4/KH_2PO_4 pH=6.0 and the other pellet was resuspended in 500 ml BMMY with K_2HPO_4/KH_2PO_4 pH=7.0 to a final OD_{600}=1-1.5 and then incubated and harvested as before.

A third try to get more protein was done at the original condition of pH 6, but this time with an increased culture volume of 2 litre. After 72 h inoculation in 200 ml BMGY and harvesting of cells, the pellet was resuspended in 2 litre BMMY (pH=6.0) to a final OD_{600}=1-1.5 and incubated at 30°C with shaking at 125 rpm for 72 hours. 100% methanol was added every 24 h to a final concentration of 0.5% to maintain secretion of protein. Cells were harvested as before.

One last try with this construct was made with four different pH conditions (pH 4, 6, 7 and 8) to try to get more protein. After 72 h inoculation in BMGY, the pellets were resuspended in 4x500 ml BMMY, each containing $C_2H_3NaO_2/CH_3COOH$ pH=4.0 and K_2HPO_4/KH_2PO_4 pH=6, 7 and 8 buffers, respectively. The cells were incubated and harvested as before.

2.4 Purification of CP10217

After each expression of CP10217 in *P. pastoris*, the supernatant was collected and filtered, and protein purification using immobilized metal affinity chromatography (IMAC) was performed. Since the secreted CP10217 has a His_6-tag, this will bind to the Ni-NTA resin used for the IMAC (Block *et al.* 2009).

First, the supernatant was filtered with a Filtropur BT50, 500 ml, 0.2 µm bottle filter (Sarstedt). The HisPur™ Ni-NTA resin (Thermo Scientific) was centrifuged at 700*xg* for 2 min and the supernatant was removed carefully and discarded. The beads were washed 3 times in Washing buffer at 700*xg* for 2 min. After the washing, the beads were added to the filtered supernatant and the mix was incubated for 2 h at room temperature while stirring with a magnetic stirrer. The bottle was left overnight to allow the beads to settle.

For 1 litre supernatant-bead mix, 900 ml supernatant was taken out, leaving the beads in 100 ml. The 100 ml supernatant-bead mix was put on a column, containing filter paper to prevent bead migration. 40 ml Washing buffer was added to the column to wash the beads. Thereafter, 4 ml Elution buffer was added to the column to elute the protein. All flow through and wash was saved and stored at 4°C. The eluate was put on a Pierce protein concentrator PES, MWCO, 5-20 ml (Thermo Scientific) together with 20 ml Exchange buffer to concentrate the protein. The concentrator was centrifuged at 3273*xg* for 30 min at 4°C. The flow through was poured out and an additional 20 ml Exchange buffer was added to the concentrator and centrifuged again under the same conditions. The concentrated protein was aliquoted and stored at -80°C until use.

To verify if any protein had been acquired from the expression and purification, concentration measurements and SDS-PAGEs were performed for each of the four tries, as described below (Section 2.4.1 and 2.4.2). A western blot was performed when there had been no visible bands on the SDS-PAGE, since this method is more sensitive and can detect lower protein concentrations. Western blot was done for the first, second and fourth try as described below (Section 2.4.3). A zymogram was performed to check if the protein was active and could cleave gelatine. This was only done for the first try, as described below (Section 2.4.4).

2.4.1 Concentration measurement

To measure the protein concentration, Qubit™ 2.0 Fluorometer (Invitrogen™) was used. Three standards and two sample assay tubes were prepared according to Qubit™ Assays Quick reference card. The sample assay tubes contained 1 µl sample and 5 µl sample, mixed with 199 µl and 195 µl Qubit™ Working solution, respectively. The tubes were incubated for 15 min at room temperature before measuring the concentration.

2.4.2 SDS-PAGE

SDS-PAGE was performed to detect CP10217. 20 µl of the protein was mixed with 6.67 µl 4x Laemmli sample buffer (Bio-Rad). The sample was heated at 90-100°C for 5 min to denature the protein. 1x TGS running buffer was added to the inner and outer chamber of the Mini-PROTEAN tetra cell (Bio-Rad). The sample and 5 µl of PageRuler Prestained Protein ladder and 5 µl PageRuler Unstained Protein ladder (Thermo Scientific) was loaded onto an Any kD Mini-PROTEAN TGX stain-free precast protein gel, 10 well, 50 µl (Bio-Rad). The gel ran at 100V for 80-90 min. Since it is a stain-free gel, it is possible to take pictures of the gel directly after the run. A ChemiDoc XRS+ (Bio-Rad) machine was used to take pictures of the gel.

If no bands were visible on the stain-free gel picture, the gel was stained with QC Colloidal Coomassie staining (Bio-Rad). The gel was rinsed in deionized water and then it was fixed in 50 ml fixing solution for 15 min at room temperature with gentle agitation. After rinsing the gel in deionized water, the staining solution was added to cover the gel and it was incubated for 20 h at room temperature with gentle agitation. Thereafter, the gel was destained in deionized water for 3 h before taking pictures of it with the ChemiDoc.

Western blot

When there were no visible bands on the SDS-PAGE, even after staining, a western blot was performed. Same conditions for the SDS-PAGE as above was used. To transfer the gel onto a membrane, Tans-blot turbo Mini PVDF transfer pack and the Trans-blot turbo system (Bio-

Rad) was used. The pre-programmed protocol for MIXED MW (5-150 kD, 7 min, 2.5 A, up to 25 V) for 1 mini gel was run. After completed transfer, membrane was rinsed in deionized water followed by washing in TBST for 5 min at room temperature with gentle agitation. It was then incubated in 5% Milk blocking buffer for 1-2 hours at room temperature with gentle agitation. After rinsing the membrane in deionized water, it was washed in TBST for 3x5 min, before incubating it in 10 ml diluted primary antibody solution (1:5000 dilution of Anti-6x His-tag rabbit antibody (ab137839) from Abcom in TBST with 5% BSA) overnight at 4°C with gentle shaking. The membrane was washed in TBST for 3x5 min at room temperature with gentle agitation before incubating it in 10 ml diluted secondary antibody solution (1:5000 dilution of Anti-rabbit IgG, HRP-linked antibody (7074S) from Cell Signalling Technology in TBST with 5% BSA) for 2 hours. Thereafter, the membrane was washed in TBST for 3x5 min. Clarity Western ECL substrates (Bio-Rad) was mixed and added to the membrane. After 5 min incubation under aluminium foil, it was taken to the ChemiDoc for pictures.

2.4.3 Zymogram

A zymogram gel was used to check if the protein was active and could cleave gelatine. Sample was prepared by mixing 4 µl 5X Non-reducing lane marker sample buffer (Thermo Scientific) with 16 µl sample. After assembling the XCell SureLock Mini-Cell (Invitrogen™) the sample and 5 µl PageRuler Unstained Protein Ladder (Invitrogen™) was loaded onto the gel (Novex 10% Zymogram plus (gelatine) protein gel, 1.0 mm, 10 well, Invitrogen™). The gel ran at 125 V for 100 min. The gel was incubated in 1x Zymogram renaturing buffer for 30 min at room temperature with gentle agitation. Thereafter, 1x Zymogram developing buffer was added to the gel for 30 min. After decanting, fresh 1x Zymogram developing buffer was added and the gel was incubated at 37°C overnight without agitation. After discarding the developing buffer, the gel was carefully rinsed in deionized water and thereafter it was fixed in 50 ml fixing solution for 15 min at room temperature with gentle agitation. The gel was rinsed in deionized water and then QC Colloidal Coomassie staining solution was added to incubate the gel for 20 hours. After discarding the staining solution, the gel was destained in deionized water for 3 hours and then brought to the ChemiDoc to take pictures.

2.5 Cloning

To be able to optimize the expression and purification process of CP10217 it was decided to clone in an additional 108 bp from the 5’ UTR of CP16160, in front of CP10217 in the pPICZα A vector. Why this sequence was chosen was because it had accidentally been included when expressing and purifying CP16160 and it had resulted in a high protein yield. Also, CP10217, including its 5’ and 3’ UTR sites, was cloned into the pPacV-integ-Hehl vector to be able to visualize it in *G. intestinalis*. At the end of the 10217 gene, a His$_6$-tag and a stop codon were added before the 3’ UTR, for the pPacV-integ-Hehl vector. The sequences for the inserts used were of *G. intestinalis* WB genome origin and gathered from GiardiaDB. The vectors and the inserts (Figure 3) were visualized using the online tool Benchling (https://benchling.com/). To design the primers, Thermo Fisher Scientific’s T_m calculator was used to decide how long the primers were supposed to be for similar T_m values. The designed primers can be seen in Table 1.

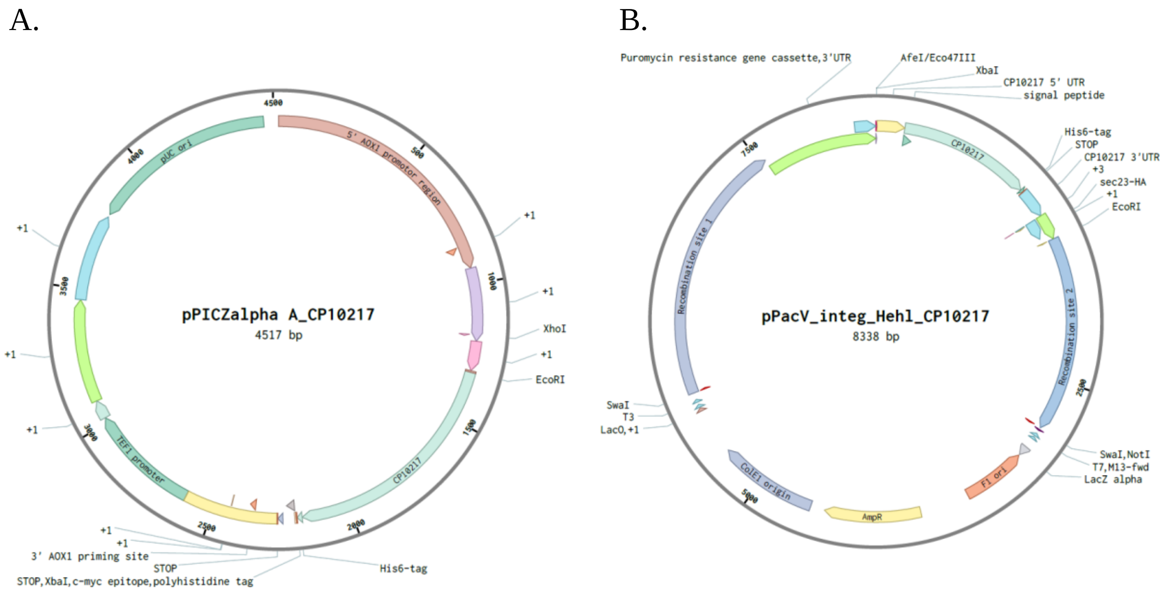


Figure 3. The pPICZα A vector containing the 16160 5’ UTR and the CP10217 gene with an His_6-tag (A). The pPacV-integ-Hehl vector containing the CP10217 gene with its 5’ and 3’ UTR and also a His_6-tag at before the 3’ UTR (B).

Table 1. The designed primers for cloning CP10217 into pPICZα A vector and pPacV-integ-Hehl vector. The restriction enzyme site in the sequence is showed in italics. The part of the sequence annealing to the template is showed in bold.

Name/vector	Sequence	Restriction enzyme
5’UTR 16160 fwd/pPICZαA	5’-TC*TCTCGAG*AAAAGAGAGGCTGAAGCT**TCTTTGTCCGCGCTACCTCTC** -3’	*XhoI*
5’ UTR 16160 rev./pPICZαA	5’- TC*GAATTC***TTTAATTTCGATTTTGCCCGTTTGT** -3’	*EcoRI*
5’UTR 10217 fwd/pPacV-integ-Hehl	5’- TC*TCTAGA***CGCAGCAGCATAGCTTTCTC** -3’	*XbaI*
10217 + his rev/pPacV-integ-Hehl	5’- GTGGTGATGGTGATGATG**GTCAAGATATACAGCATAGATTTCATCCT** -3’	-
His + 3’ UTR 10217 fwd/pPacV-integ-Hehl	5’- GACCATCATCACCATCACCACTAG**CGCCTGTCTTCATACATAGCTAC** -3’	-
3’UTR 10217 rev/pPacV-integ-Hehl	5’- GT*TTAATTAA***ACTGCTCTACATGTTCTGGGG** -3’	*PacI*

2.5.1 PCR amplification

To amplify the desired sequences, Thermo Scientific Phusion Hot Start II High-Fidelity DNA polymerase kit was used. The template DNA used was the *G. intestinalis* WB genome. The PCR reactions were prepared according to the manufacturer's instructions, using ~100 ng template DNA and 0.5 µM primer. Because of the product size difference, two PCR's ran in parallel and their specifications can be seen in Table 2.

Table 2. PCR cycling conditions.

PCR (16160 and 10217 3'UTR products)				**PCR (10217 5'UTR product)**			
Cycle step	***Temp. (°C)***	***Time***	***#Cycles***	***Cycle step***	***Temp. (°C)***	***Time***	***#Cycles***
Initial denaturation	98	30 s	1	Initial denaturation	98	30 s	1
Denaturation	98	10 s	35	Denaturation	98	10 s	35
Annealing	63.5	30 s		Annealing	63.9	30 s	
Extension	72	10 s		Extension	72	34 s	
Final extension	72 4	5 min ∞	1	Final extension	72 4	5 min ∞	1

2.5.2 Agarose gel electrophoresis

To verify that the PCR's had been successful, samples were run in a 1% agarose gel. Samples were mixed with 6X DNA loading dye (Thermo Scientific) before loaded onto the gel. GeneRuler 100 bp DNA ladder, ready-to-use and GeneRuler 1 kb DNA ladder, ready-to-use (Thermo Scientific) were also loaded to the gel. The gel ran at 100 V for 40 min. By using a UV-camera, the bands could be visualised. Thereafter, GeneJET Gel extraction kit (Thermo Scientific) was used to extract the PCR fragments from the agarose gel. The purified DNA was diluted in sterile water instead of elution buffer. Once the PCR fragments had been extracted and purified, the concentration of each sample was measured with a Nanodrop.

2.5.3 Overlap-extension PCR (OE-PCR)

Because there were two fragments for the CP10217 gene to be inserted into the pPacV-integ-Hehl vector, OE-PCR was used to fuse these fragments together. The different PCR conditions used can be seen in Table 3. PCR 1 was run to make the two fragments overlap and anneal to each other. The overlapping parts of the fragments were the His$_6$-tag. Here, everything except the primers were added to the PCR tubes and hence the tube volume was 45 µl. Then, for PCR 2, the primers were added to the tubes before running. A negative control was also added, having sterile water instead of the two fragments.

Table 3. OE-PCR cycling conditions.

PCR 1 (Overlap annealing)				PCR 2 (Primer annealing)			
Cycle step	***Temp. (°C)***	***Time***	***#Cycles***	***Cycle step***	***Temp. (°C)***	***Time***	***#Cycles***
Initial denaturation	98	30 s	1	Initial denaturation	98	30 s	1
Denaturation	98	10 s		Denaturation	98	10 s	
Annealing	63.8	30 s	10	Annealing	67	30 s	25
Extension	72	41 s		Extension	72	41 s	
Final extension	72 4	5 min ∞	1	Final extension	72 4	5 min ∞	1

As for the first PCR, OE-PCR products were run on a 1% agarose gel at 10 V for 40 min, and then put under UV-camera to visualise the product. Thereafter, the GeneJET Gel extraction kit was used to purify the DNA. Sterile water was used to dilute the DNA instead of elution buffer. Nanodrop was used to measure the concentration of the fragment.

2.5.4 pPICZαA_10217 plasmid purification

An overnight culture of *E. coli* TOP10 cells containing the pPICZαA_10217 vector was grown at 37°C in low salt LB medium with Zeocin™. The next day, the cells were harvested by centrifugation at 3270x*g* for 14 min. To purify the vector DNA, Thermo Fisher's GeneJET Plasmid miniprep kit was used. Sterile water was added instead of elution buffer, when eluting the vector DNA. Once purified, the concentration was measured using a Nanodrop.

2.5.5 Digestion

The 5' UTR 16160 PCR fragment, 10217 OE-PCR fragment and the pPICZαA_10217 vector were each mixed according to Table 4. Thereafter, the tubes were incubated at 37°C for 1h to digest the DNA. To verify that the digestion had worked, a gel electrophoresis was run at 80 V for 40 min. Using GeneJET Gel extraction kit, each digestion was purified out from the gel. As before, 50 µl sterile water was added instead of elution buffer to elute each digestion. The concentrations were measured using Nanodrop.

Table 4. The reaction mixture used for the digestion of 10217 from OE-PCR, 5'UTR 16160 fragment and pPICZαA_10217 vector.

Component	10217 from OE-PCR (µl)	5'UTR 16160 (µl)	pPICZαA_10217 vector (µl)
DNA	12.7 (1 µg)	30 (1 µg)	7.6 (1 µg)
Enzyme 1	1 (*XbaI*)	1 (*XhoI*)	1 (*XhoI*)
Enzyme 2	1 (*PacI*)	1 (*EcoRI*)	1 (*EcoRI*)
10X FastDigest® buffer	2	3	2
dH_2O	3.3	5	8.4
Total	20	40	20

2.5.6 Ligation

The digested DNA fragments were mixed according to Table 5, using a 3:1 ratio of insert to vector. The pPac-integ-Hehl vector had already been digested by a PhD student. Thermo Scientific T4 DNA Ligase and 10X T4 DNA ligase buffer was used. Thereafter, the mix was incubated at 22°C for 30 min.

Table 5. The reaction mixtures for the ligation of pPICZαA_10217 + 5'UTR 16160 and pPacV-integ-Hehl + 10217.

Component	pPICZαA_10217 + 5'UTR 16160 (µl)	pPacV-integ-Hehl + 10217 (µl)
Vector DNA	9.3 (90 ng)	2.6 (65 ng)
Insert DNA	0.8 (8.328 ng)	3 (37.01 ng)
10X T4 DNA ligase buffer	2	2
T4 DNA ligase	1	1
dH_2O	6.9	11.4
Total	20	20

2.5.7 Transformation into competent *E. coli* cells

For each ligated vector, 50 µl of *E. coli* DH5α cells were mixed with 5 µl ligated vector DNA. The bottom of the tube was gently flicked to mix the cells and the DNA. The mixture was incubated on ice for 30 min. Thereafter, the cells were heat-shocked for 30 sec at 42°C and then put on ice again. While the cells containing pPacV-integ-Hehl_10217 were on ice for 10 min, 250 µl LB medium was added to the cells with pPICZαA_10217+5'UTR 16160 and the mixture was incubated at 37°C with shaking at 200 rpm for 60 min. 50 µl of the pPacV-integ-Hehl_10217 mixture was spread on LA-plates containing 50 µg/ml ampicillin after 10 min incubation on ice. It was then incubated at 37°C overnight. 50 µl of the pPICZαA_10217+5'UTR 16160 mixture was spread on Low salt LB plates containing 25 µg/ml zeocin after 60 min incubation. The plate was then incubated at 37°C overnight. After the incubation, the colonies on the plates were counted. Six colonies from each of the two plates were each inoculated in 2 ml LB media. The *E. coli* colonies presumably containing the pPacV-integ-Hehl_10217 vector were inoculated in LB media with Ampicillin, and the *E. coli* colonies presumably containing the pPICZαA_10217+5'UTR 16160 vector were inoculated in low salt LB media with Zeocin. They were then incubated at 37°C, with shaking at 175 rpm, overnight. The next day, the cells were harvested by centrifugation at 3270x *g*, 22°C for 10 min. GeneJET Plasmid Miniprep kit was used to purify the vectors. Vector DNA was eluted with sterile water instead of elution buffer. The concentration of the 12 samples were measured with Nanodrop. Thereafter, a PCR was run to verify that the inserts had been ligated into the plasmid. For preparing the vector DNA for the PCR, the protocol for Thermo Scientific's PCR Master Mix (2X) was followed. Table 6 is showing the cycling conditions for the PCR runs. 5 µl of each of the PCR samples were diluted in 5 µl sterile water and 2 µl 6X loading dye. From this, 5 µl was loaded onto a 1% agarose gel, and the gel ran for 30 min at 100V. After the gel electrophoresis, the vector DNA was sent for Sanger sequencing, using Eurofins genomics Mix2Seq kit. 15 µl vector DNA was mixed with 2 µl primer. A total of four tubes were sent for sequencing; two tubes with the pPacV-int-Hehl_10217 vector DNA and two tubes with the pPICZαA_10217+ 5'UTR 16160 vector DNA, having forward and reverse primers in each separate tube. The results from the sequencing were analysed using the BioEdit tool (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and Benchling.

Table 6. The PCR cycling conditions for verifying that the inserts had been ligated into the vectors.

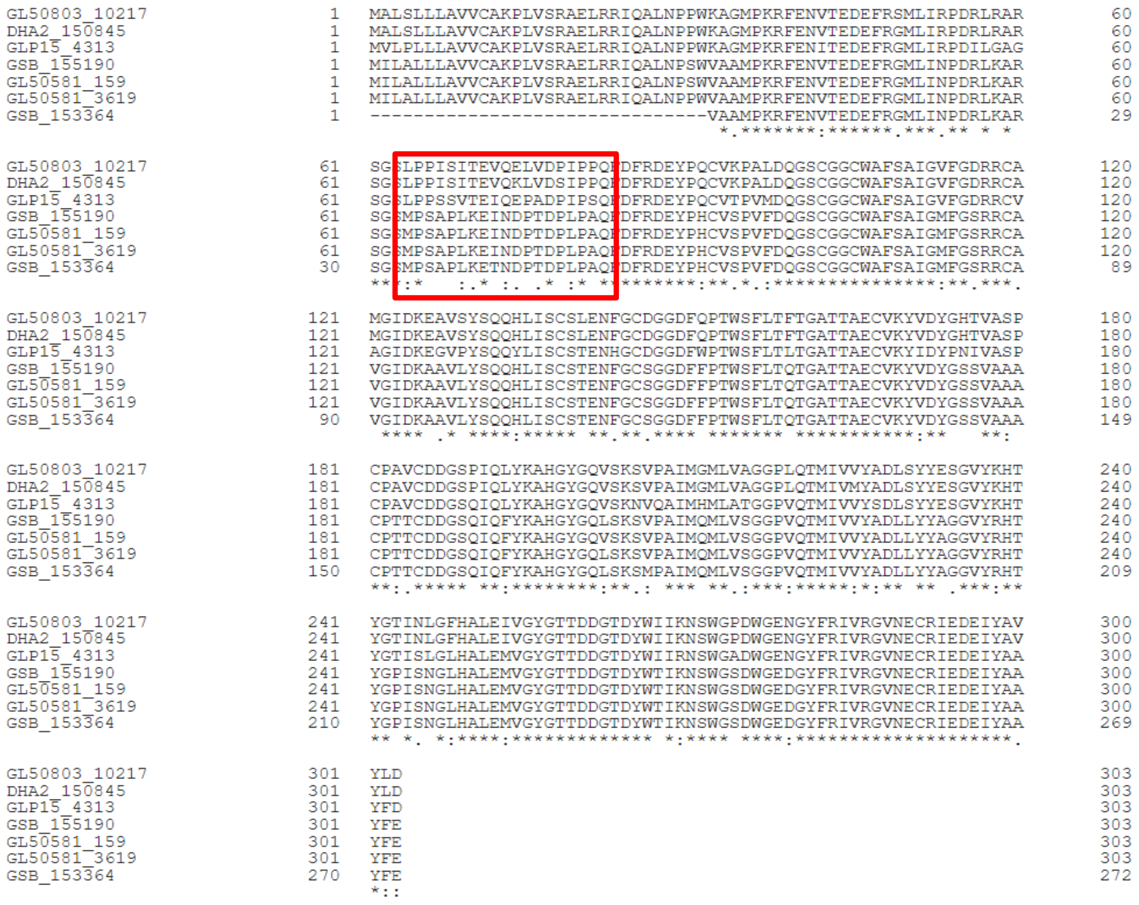
PCR (pPacV-integ-Hehl_10217)				**PCR (pPICZαA_10217+5'UTR 16160)**			
Cycle step	***Temp. (°C)***	***Time***	***#Cycles***	***Cycle step***	***Temp. (°C)***	***Time***	***#Cycles***
Initial denaturation	95	2 min	1	Initial denaturation	95	2 min	1
Denaturation	95	30 s	35	Denaturation	95	30 s	35
Annealing	53.9	30 s		Annealing	55.6	30 s	
Extension	72	1 min 20 s		Extension	72	10 s	
Final extension	72 4	15 min ∞	1	Final extension	72 4	15 min ∞	1

3 Results

3.1 Structure analysis

3.1.1 Multiple sequence alignment

Six orthologs to CP10217 were aligned using MUSCLE (Figure S1) and ClustalΩ (Figure 4). When studying the MSA, the results from the ClustalΩ alignment showed a sequence identity of 65%, with 197 identical positions and 45 similar positions out of 303 amino acids. It can also be seen that among the sequences, the assemblage E ortholog (GLP15_4313) is the one that varies the most from the others but is more like the assemblage A sequences than the assemblage B sequences. From the MSA, a phylogenetic tree was generated, which can be seen in Figure 5. The tree has significant values of 1, 0.99 and 0.89. When looking at the branches one can see that assemblage A sequences cluster together as do assemblage B sequences, whereas the assemblage E sequence is closer to assemblage A than B.



```
GL50803_10217        1    MALSLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENVTEDEFRSMLIRPDRLRAR      60
DHA2_150845          1    MALSLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENVTEDEFRGMLIRPDRLRAR      60
GLP15_4313           1    MVLPLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENITEDEFRGMLIRPDILGAG      60
GSB_155190           1    MILALLLAVVCAKPLVSRAELRRIQALNPSWVAAMPKRFENVTEDEFRGMLINPDRLKAR      60
GL50581_159          1    MILALLLAVVCAKPLVSRAELRRIQALNPSWVAAMPKRFENVTEDEFRGMLINPDRLKAR      60
GL50581_3619         1    MILALLLAVVCAKPLVSRAELRRIQALNPPWVAAMPKRFENVTEDEFRGMLINPDRLKAR      60
GSB_153364           1    -------------------------------VAAMPKRFENVTEDEFRGMLINPDRLKAR      29
                                                          *.*******:******.***.** * *

GL50803_10217       61    SGSLPPISITEVQELVDPIPPQFDFRDEYPQCVKPALDQGSCGGCWAFSAIGVFGDRRCA      120
DHA2_150845         61    SGSLPPISITEVQKLVDSIPPQFDFRDEYPQCVKPALDQGSCGGCWAFSAIGVFGDRRCA      120
GLP15_4313          61    SGSLPPSSVTEIQEPADPIPSQFDFRDEYPQCVTPVMDQGSCGGCWAFSAIGVFGDRRCV      120
GSB_155190          61    SGSMPSAPLKEINDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA      120
GL50581_159         61    SGSMPSAPLKEINDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA      120
GL50581_3619        61    SGSMPSAPLKEINDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA      120
GSB_153364          30    SGSMPSAPLKETNDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA      89
                          ***:*   :.* :. .* :* *********:**.*.:***************:**.***.

GL50803_10217      121    MGIDKEAVSYSQQHLISCSLENFGCDGGDFQPTWSFLTFTGATTAECVKYVDYGHTVASP      180
DHA2_150845        121    MGIDKEAVSYSQQHLISCSLENFGCDGGDFQPTWSFLTFTGATTAECVKYVDYGHTVASP      180
GLP15_4313         121    AGIDKEGVPYSQQYLISCSTENHGCDGGDFWPTWSFLTLTGATTAECVKYIDYPNIVASP      180
GSB_155190         121    VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA      180
GL50581_159        121    VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA      180
GL50581_3619       121    VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA      180
GSB_153364          90    VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA      149
                           **** .* ****:***** **.**.**** ******* ***********:**   **:

GL50803_10217      181    CPAVCDDGSPIQLYKAHGYGQVSKSVPAIMGMLVAGGPLQTMIVVYADLSYYESGVYKHT      240
DHA2_150845        181    CPAVCDDGSPIQLYKAHGYGQVSKSVPAIMGMLVAGGPLQTMIVMYADLSYYESGVYKHT      240
GLP15_4313         181    CPAVCDDGSQIQLYKAHGYGQVSKNVQAIMHMLATGGPVQTMIVVYSDLSYYESGVYKHT      240
GSB_155190         181    CPTTCDDGSQIQFYKAHGYGQLSKSVPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT      240
GL50581_159        181    CPTTCDDGSQIQFYKAHGYGQVSKSVPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT      240
GL50581_3619       181    CPTTCDDGSQIQFYKAHGYGQLSKSVPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT      240
GSB_153364         150    CPTTCDDGSQIQFYKAHGYGQLSKSMPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT      209
                          **:.***** **:********:**.: *** **.:***:*****:*:** ** .***:**

GL50803_10217      241    YGTINLGFHALEIVGYGTTDDGTDYWIIKNSWGPDWGENGYFRIVRGVNECRIEDEIYAV      300
DHA2_150845        241    YGTINLGFHALEIVGYGTTDDGTDYWIIKNSWGPDWGENGYFRIVRGVNECRIEDEIYAV      300
GLP15_4313         241    YGTISLGLHALEMVGYGTTDDGTDYWIIRNSWGADWGENGYFRIVRGVNECRIEDEIYAA      300
GSB_155190         241    YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA      300
GL50581_159        241    YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA      300
GL50581_3619       241    YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA      300
GSB_153364         210    YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA      269
                          ** *. *:****:************* *:**** ****:********************.

GL50803_10217      301    YLD                                                               303
DHA2_150845        301    YLD                                                               303
GLP15_4313         301    YFD                                                               303
GSB_155190         301    YFE                                                               303
GL50581_159        301    YFE                                                               303
GL50581_3619       301    YFE                                                               303
GSB_153364         270    YFE                                                               272
                          *::
```

Figure 4. ClustalΩ MSA of CP10217 and its orthologs. The red square indicates the part of the MSA where the sequences varied the most.

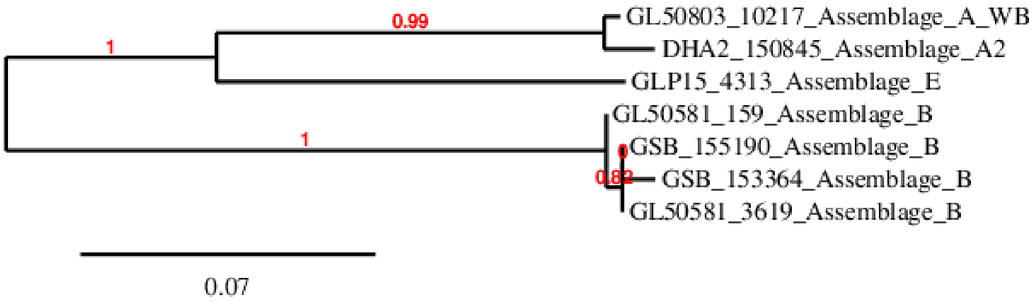


Figure 5. Phylogenetic tree of CP10217 and its orthologs. Assemblages A clustered together, and assemblages B clustered together. Assemblage E is closer to assemblage A than assemblage B.

3.1.2 Secondary and tertiary structure prediction

The prediction of the secondary structure of CP10217 using JPred showed that helices are the dominating structure for this protein, and that sheets are present but less frequently and at the end of the sequence (Figure 6). The confidence level for this structure is mostly high throughout the sequence. Most of the residues seems to be buried (B), and especially at the end of the sequence. Looking at the results from the SPIDER2 prediction (Figure 7), the secondary structure prediction seems to be similar to JPred. In this prediction, SPIDER2 has predicted how much of each residue’s surface area that is exposed (rASA) by the numbers 0-9. 0 meaning 10% surface area exposed and 9 meaning over 90% surface area exposed. Most of the residues seem to have low rASA numbers.

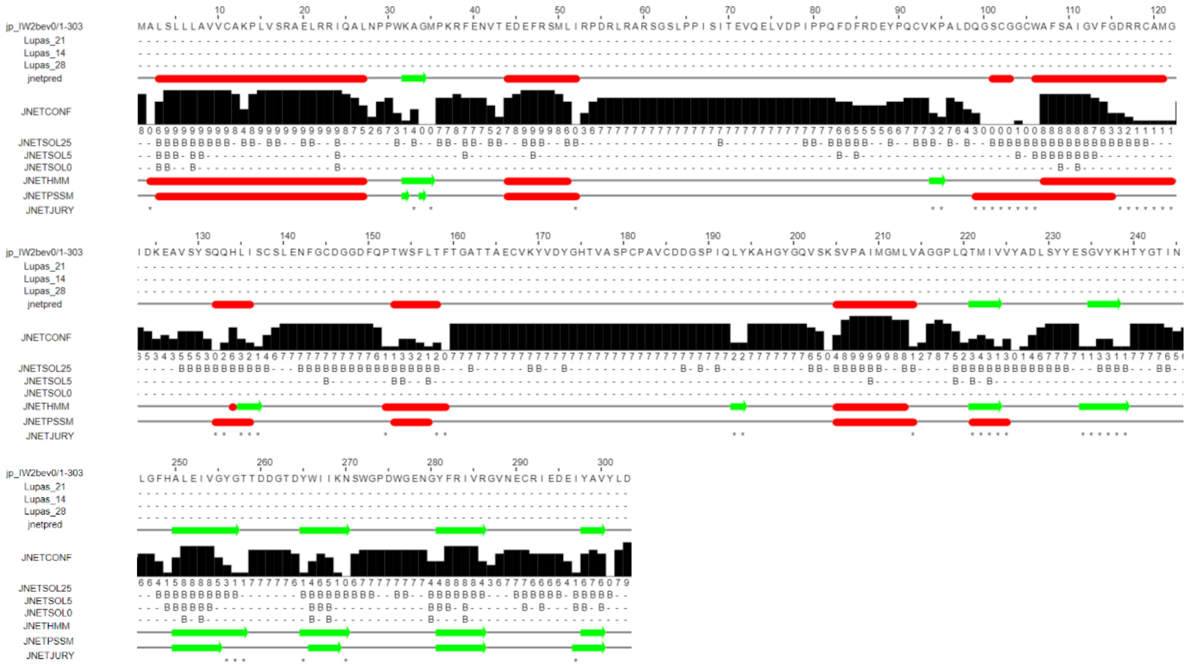
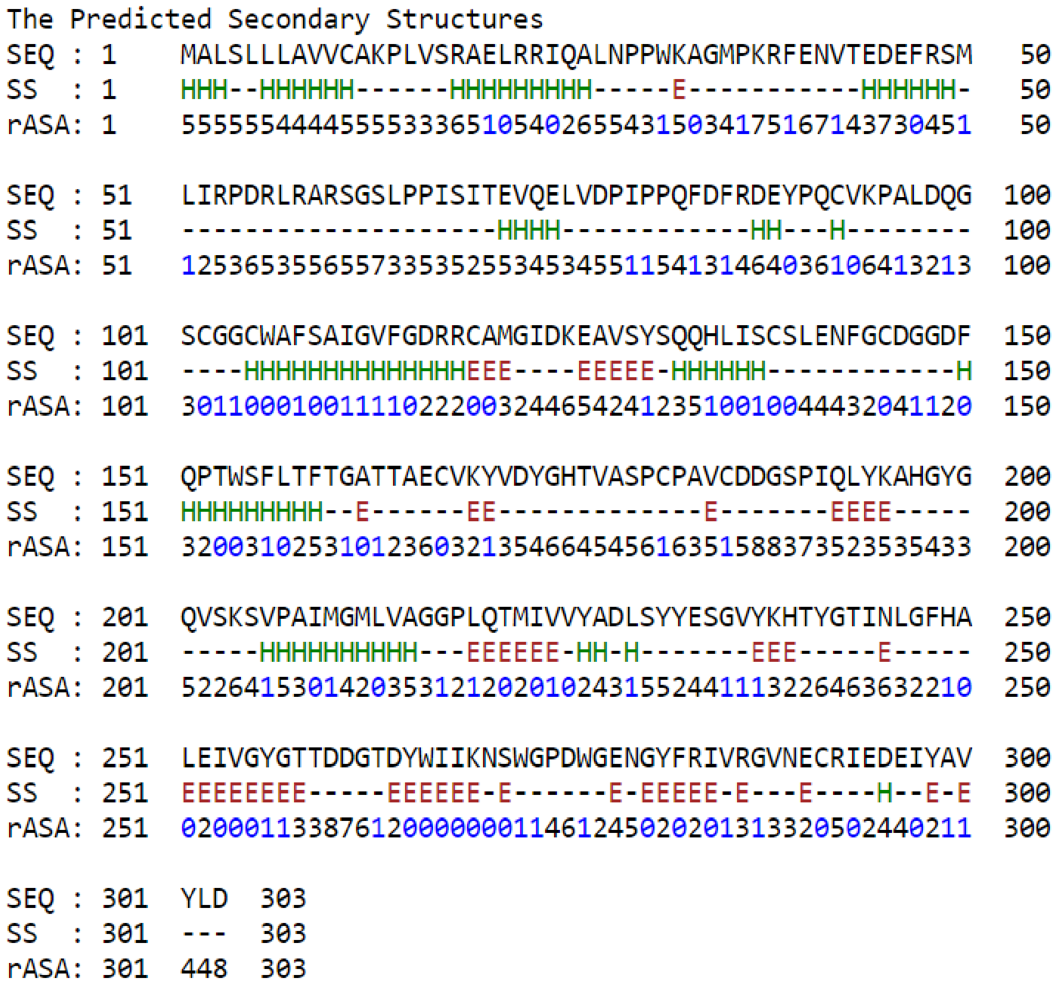


Figure 6. Secondary structure prediction using JPred. Red tubes indicate helices and green arrows indicate sheets. The JNETCONF row indicates the confidence level for the structure in that position. The JNETSOL rows indicate if the residue is buried (B) or not. The dominating structure is helices throughout most of the sequence, and sheets at the end of the sequence.



```
The Predicted Secondary Structures
SEQ : 1     MALSLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENVTEDEFRSM   50
SS  : 1     HHH--HHHHHH------HHHHHHHHH-----E-----------HHHHHH-   50
rASA: 1     55555544445555333651054026554315034175167143730451   50

SEQ : 51    LIRPDRLRARSGSLPPISITEVQELVDPIPPQFDFRDEYPQCVKPALDQG  100
SS  : 51    --------------------HHHH------------HH---H--------  100
rASA: 51    12536535565573353525534534551154131464036106413213  100

SEQ : 101   SCGGCWAFSAIGVFGDRRCAMGIDKEAVSYSQQHLISCSLENFGCDGGDF  150
SS  : 101   ----HHHHHHHHHHHHHHEEE----EEEEE-HHHHHH------------H  150
rASA: 101   30110001001111022200324465424123510010044432041120  150

SEQ : 151   QPTWSFLTFTGATTAECVKYVDYGHTVASPCPAVCDDGSPIQLYKAHGYG  200
SS  : 151   HHHHHHHHH--E------EE-------------E-------EEEE-----  200
rASA: 151   32003102531012360321354664545616351588373523535433  200

SEQ : 201   QVSKSVPAIMGMLVAGGPLQTMIVVYADLSYYESGVYKHTYGTINLGFHA  250
SS  : 201   -----HHHHHHHHHH---EEEEEE-HH-H-------EEE-----E-----  250
rASA: 201   52264153014203531212020102431552441113226463632210  250

SEQ : 251   LEIVGYGTTDDGTDYWIIKNSWGPDWGENGYFRIVRGVNECRIEDEIYAV  300
SS  : 251   EEEEEEEE-----EEEEEE-E------E-EEEEE-E---E----H--E-E  300
rASA: 251   02000113387612000000011461245020201313320502440211  300

SEQ : 301   YLD  303
SS  : 301   ---  303
rASA: 301   448  303
```

Figure 7. Secondary structure prediction of CP10217 using SPIDER2. The SS line represents the structure of each residue. 'H' represents helices, 'E' sheets and '–' coils. The rASA line tells the relative accessible surface area numbered 0-9, 0 meaning 10% surface area exposed and 9 meaning over 90% surface area exposed.

For the tertiary structure prediction, both tools generated a lot of models, where the best models with the highest coverage and sequence similarity were similar for both tools. Around 95% coverage, and around 30% sequence identity. The best SWISS-MODEL result can be seen in Figure 8. This model was based on the *Trypanosoma brucei* procathepsin B template (PDB: 4N4Z.1A), having a coverage of 93%, sequence similarity of 36% and a sequence identity of 33%. In Figure 8B one large sheet, one smaller sheet and 7 helices are visible in the cartoon model. The Predictein tertiary structure prediction is shown in Figure 9. This model was based on the human procathepsin B template (PDB: 2PBHA), having a coverage of 94% and a sequence identity of 32%. In Figure 9A the same observations as in Figure 8B can be seen and also, the N- (left) and C-terminal (right) is marked by red blobs.

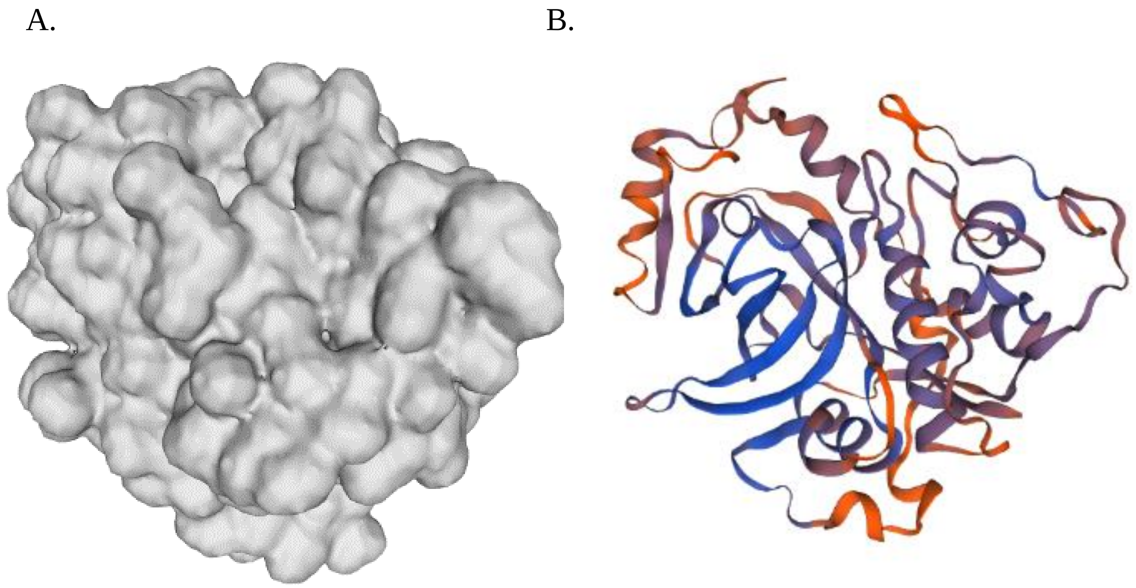


Figure 8. SWISS-MODEL tertiary structure prediction. Model template is the *Trypanosoma brucei* procathepsin B (PDB: 4N4Z.1A), having a coverage of 93%, sequence similarity of 36% and a sequence identity of 33%. Figure showing the surface (A) and the cartoon (B) structure of the protein. The blue colour indicates a correct structure and the more red, the further away from the correct structure it is.

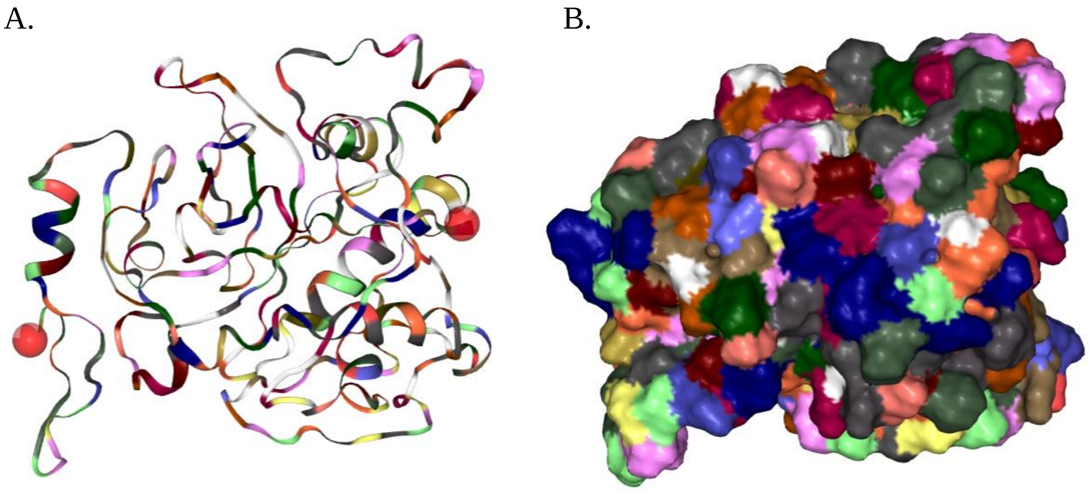
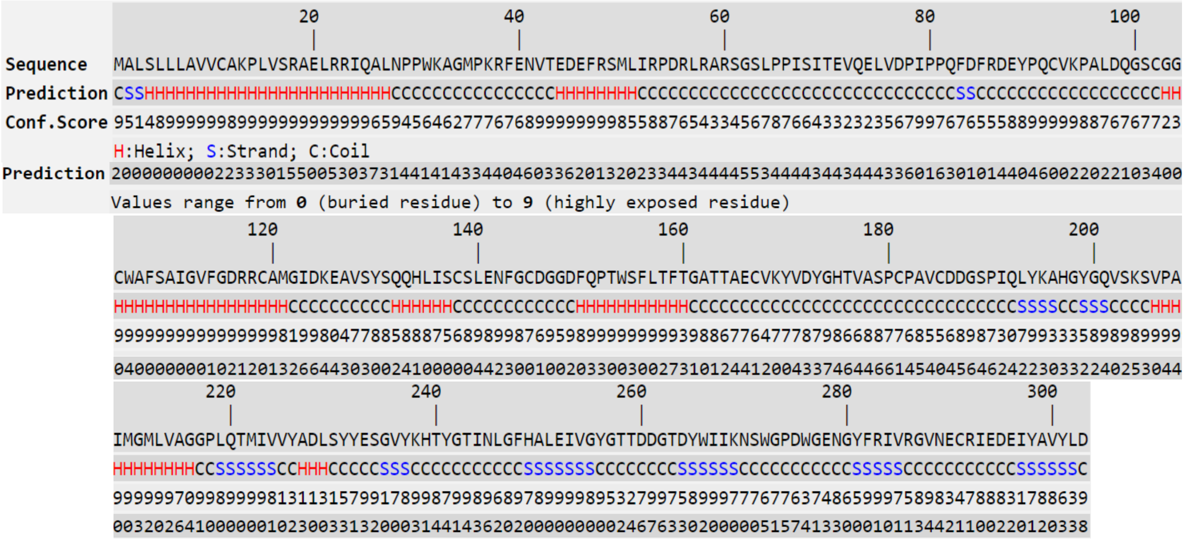


Figure 9. Predictein tertiary structure prediction. The colours represent the different residues. Model template is a human procathepsin B (PDB: 2PBHA), having a coverage of 94% and a sequence identity of 32%. Figure is showing (A) cartoon with N-terminal (left) and C-terminal (right) marked with the red blobs and (B) the surface structure of the protein.

3.1.3 I-TASSER analysis

The I-TASSER analysis predicted the secondary structure, tertiary models and protein functions, such as ligand binding sites Enzyme Commission (EC) number and active sites, and Gene Ontology (GO) terms based on the MSA to the PDB database.

The secondary structure prediction is very similar to the JPred and SPIDER2 prediction, showing a high confidence score (majority >5). The predicted solvent accessibility has an overall low score (<5), indicating that most of the residues are buried or at least not highly exposed to the surface (Figure 10). The tertiary structure prediction by I-TASSER resulted in 5 models, where model 1 had a confidence score of 0.95, corresponding to a model with correct global topology (Figure 11). The other models had confidence scores below -1.5, which does not correspond to a good prediction model.



```
                             20                  40                  60                  80                 100
                              |                   |                   |                   |                   |
Sequence   MALSLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENVTEDEFRSMLIRPDRLRARSGSLPPISITEVQELVDPIPPQFDFRDEYPQCVKPALDQGSCGG
Prediction CSSHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCSSCCCCCCCCCCCCCCCCCCHH
Conf.Score 95148999999899999999999999659456462777676899999999855887654334567876643323235679976765558899999887676772 3
           H:Helix; S:Strand; C:Coil
Prediction 20000000002233301550053037314414143344046033620132023344344445534444344344433601630101440460022022103400
           Values range from 0 (buried residue) to 9 (highly exposed residue)
                         120                 140                 160                 180                 200
                          |                   |                   |                   |                   |
           CWAFSAIGVFGDRRCAMGIDKEAVSYSQQHLISCSLENFGCDGGDFQPTWSFLTFTGATTAECVKYVDYGHTVASPCPAVCDDGSPIQLYKAHGYGQVSKSVPA
           HHHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCSSSSCCSSSCCCCHHH
           99999999999999998199804778858887568989987695989999999398867764777879866887768556898730799333589898999 9

           04000000010212013266443030024100000442300100203300300273101244120043374644661454045646242230332240253044
                     220                 240                 260                 280                 300
                      |                   |                   |                   |                   |
           IMGMLVAGGPLQTMIVVYADLSYYESGVYKHTYGTINLGFHALEIVGYGTTDDGTDYWIIKNSWGPDWGENGYFRIVRGVNECRIEDEIYAVYLD
           HHHHHHHHCCSSSSSSCCHHHCCCCCSSSCCCCCCCCCCCSSSSSSSCCCCCCCCSSSSSSCCCCCCCCCCCSSSSSCCCCCCCCCCCSSSSSSC
           99999970998999981311315799178998799896897899998953279975899977767763748659997589834788831788639
           00320264100000010230033132000314414362020000000000246763302000005157413300010113442110022012033 8
```

Figure 10. I-TASSER secondary structure prediction of CP10217. Confidence score (0-9) of the predicted structure and predicted solvent accessibility (0-9) for each residue.

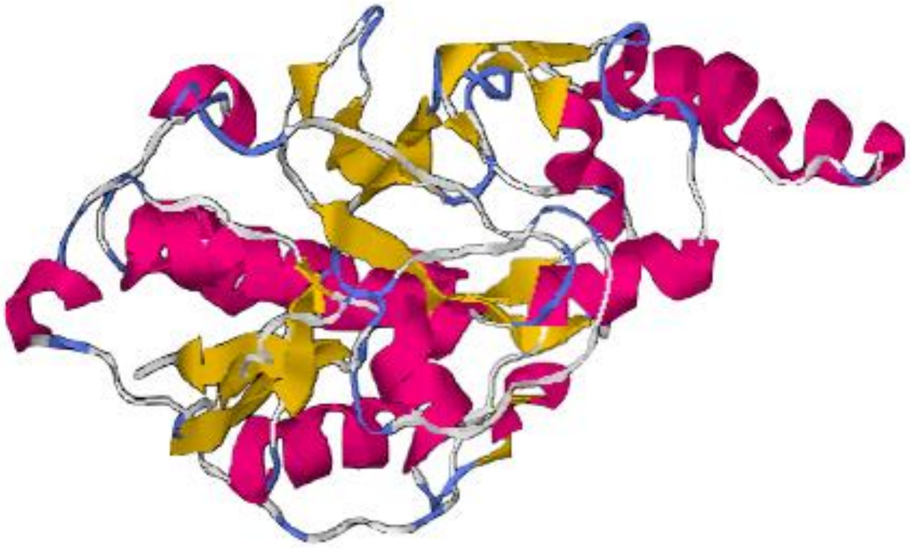


Figure 11. The best predicted tertiary structure by I-TASSER, with a confidence score of 0.95, indicating correct global topology for the model.

I-TASSER used the best predicted model (Figure 11) as a template when performing a structural alignment using the PDB library. This generated a table of the top 10 PDB hits (Table 7). The ranking is based on the TM-score, i.e. the structural alignment, and here we can see that all hits have a score above 0.8, indicating a high similarity between the query and template structure. In this table we can also see the identity and coverage of the alignment.

Table 7. The top 10 structurally close proteins in PDB to the first I-TASSER model. The TM-score indicate the structural similarity, higher TM-score means higher similarity. Identity is indicating the sequence identity in the aligned region. Coverage represent the coverage of the alignment.

PDB Hit	Rank	TM-score (%)	Identity (%)	Coverage (%)
5FPWA	1	91.6	31.3	93.7
2PBHA	2	89.6	32.2	94.4
3QJ3A	3	85.6	25.3	95.4
7PCKA	4	85.1	24.7	93.4
1MVVA	5	85.0	25.0	93.7
1CS8A	6	84.3	24.2	92.7
1M0HA	7	83.9	18.1	93.1
6CZKA	8	83.7	24.1	93.1
1PCJA	9	83.3	22.6	93.4
5EF4A	10	83.2	24.9	92.4

I-TASSER also predicted some functions of CP10217. Only one PDB hit had a good confidence score for the predicted ligand binding site, a human cathepsin K (1NLJA) with a confidence score of 0.85. The remaining four hits had scores below 0.10 but contained the same ligand binding site residues (Table 8). The EC (enzyme commission) number and active site prediction showed a confidence score around 0.6 for all the five PDB hits. The EC number specifies enzyme-catalysed reactions, so if two enzymes catalyses the same reaction, they will receive the same EC number. This result can be seen in Table 9. Here we can also see the predicted active site residues, showing residues 99, 105 and 270 for most hits. The consensus prediction of GO terms for the molecular function is a cysteine-type endopeptidase activity, for the biological process it is proteolysis, and for the cellular component it is an extracellular space or lysosome.

Table 8. Predicted ligand binding sites. The C-score represent the confidence score of the prediction, higher score indicates a more reliable prediction.

PDB Hit	Rank	C-score	Ligand binding site residues
1NLJA	1	0.85	99,103,104, 105, 106, 141, 143, 147, 148, 149, 150, 222, 225, 248, 249, 272, 296
4DMYB	2	0.10	99, 103, 105, 106, 141, 142, 143, 146, 147, 148, 149, 152, 222, 248, 250, 296
3LFYC	3	0.04	99, 105, 248, 249, 272
3N3GA	4	0.01	98, 99, 100, 230, 231, 233, 272, 275, 276
3C9EA	5	0.01	86, 259, 278, 279, 283

Table 9. EC number and active site residue prediction. C-score is the confidence score for the EC number prediction, where a higher score indicates more reliable prediction of the EC number. TM-score is the measure of globally structure similarity between query and template protein.

PDB Hit	Rank	C-score	TM-score (%)	Identity	Coverage	EC Number	Active Site Residues
1GECE	1	0.673	64.7	0.282	0.690	3.4.22.25	99, 105, 249, 270
2EFMA	2	0.661	65.5	0.292	0.700	3.4.22.51	99, 105, 249, 270
1BP4A	3	0.616	63.6	0.307	0.677	3.4.22.2	99, 105, 270
2OZ2C	4	0.605	65.5	0.292	0.700	3.4.22.51	99, 103, 105, 108, 144, 147, 249
1IWDA	5	0.590	64.6	0.279	0.686	3.4.22.-	99, 105, 270

3.2 CP10217 purification

Using the already complete pPICZαA vector, containing the CP10217 gene linked with a His$_6$-tag, expression in *P. pastoris*, inoculation in BMGY and then resuspension in BMMY was done to express and secrete the protein in the medium. After purifying with IMAC, the protein concentration was measured using Qubit Fluorometer 2.0 (Invitrogen™) to be 108 ng/µl. To verify that it was in fact CP10217, an SDS-PAGE was performed (Figure 12A). Here, only the ladder is visible and no band around 33 kDa, for the sample, can be seen. After staining the gel with Coomassie Blue (Figure 12B), still no band was visible.

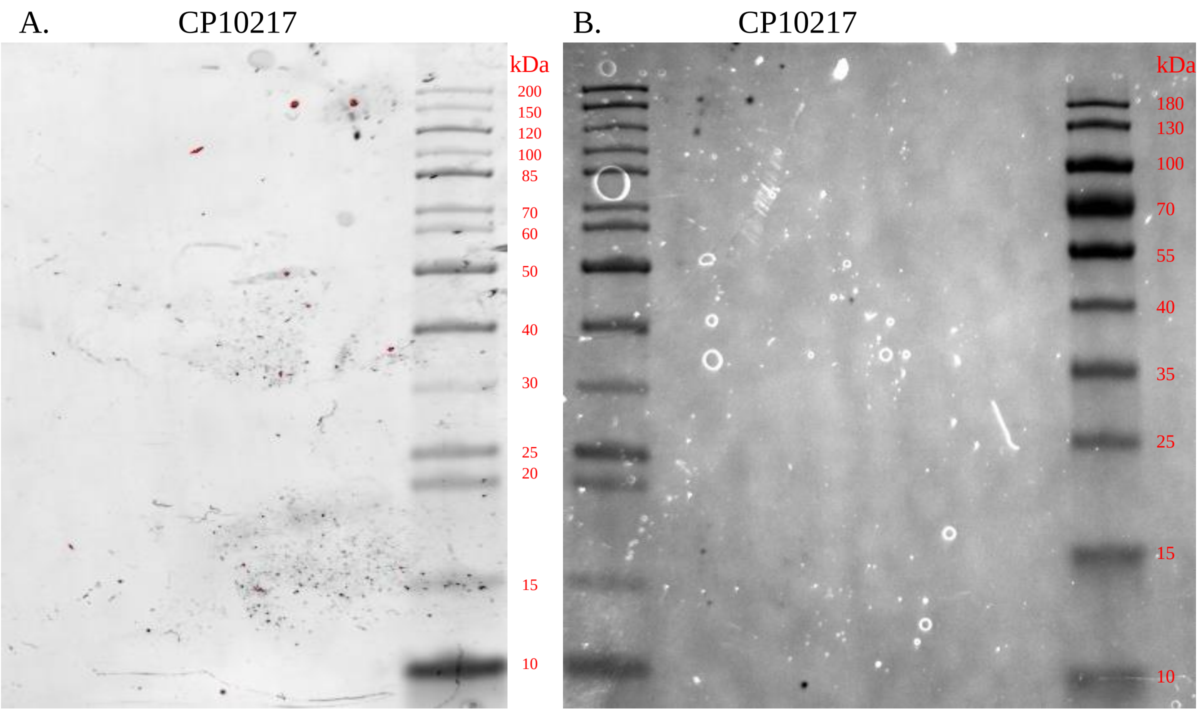


Figure 12. The results from the SDS-PAGE analysis. No band for CP10217 can be seen on the stain-free gel (A) nor on the Coomassie stained gel (B).

Because of these results, a Western blot was performed since it is much more sensitive than an SDS-PAGE. In Figure 13A, the results from the Western blot can be seen. Here a band is visible around 28.9 kDa, indicating that it is indeed CP10217, but in its active form. Since it now was positive that it indeed was CP10217, a zymogram gel was used to analyse the activity of the protein. As can be seen in Figure 13B, no gelatine cleavage can be seen, indicating no enzyme activity.

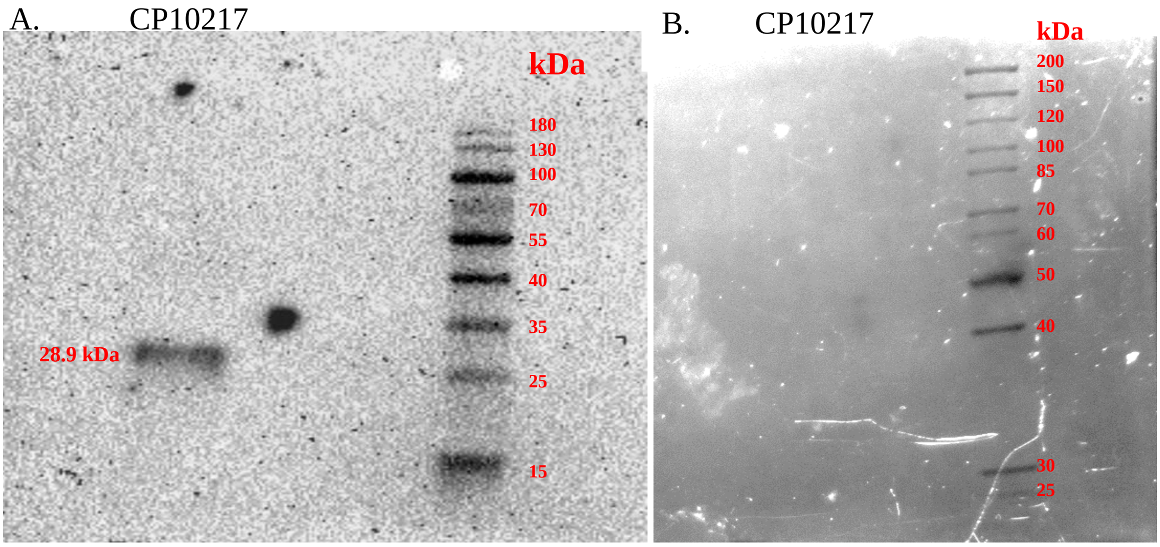


Figure 13. Western blot of CP10217 is showing a band around 28.9 kDa indicate that it is CP10217, but in its active form (A). Zymogram gel analysis of CP10217 showed no cleavage of gelatine (B).

Because of these results, it was decided to perform the expression and purification steps again, but with two different pH conditions; pH 6 and pH 7. By raising the pH to 7 in the medium, the pH would be further from the optimal activity condition for this CP. That would hopefully lead to less activity and a higher protein yield. The protein concentrations were measured to be 114 ng/µl for the pH 6 sample and 75.6 ng/µl for the pH 7 sample. Once again, no bands could be seen on the SDS-PAGE, even if stained with Coomassie blue (Figure S2). Therefore, a Western blot was done, and it showed only one band (29.6 kDa) for the sample having pH 6 and no band for the pH 7 sample (Figure 14A). Because of this, the Western blot was done once more and only a band (34.5 kDa) for the pH 7 sample and none for the pH 6 sample could be seen (Figure 14B). Another Western blot was done because of these results, and once again, only one (faint) band could be seen for the pH 6 sample at 29.8 kDa (Figure 14C).

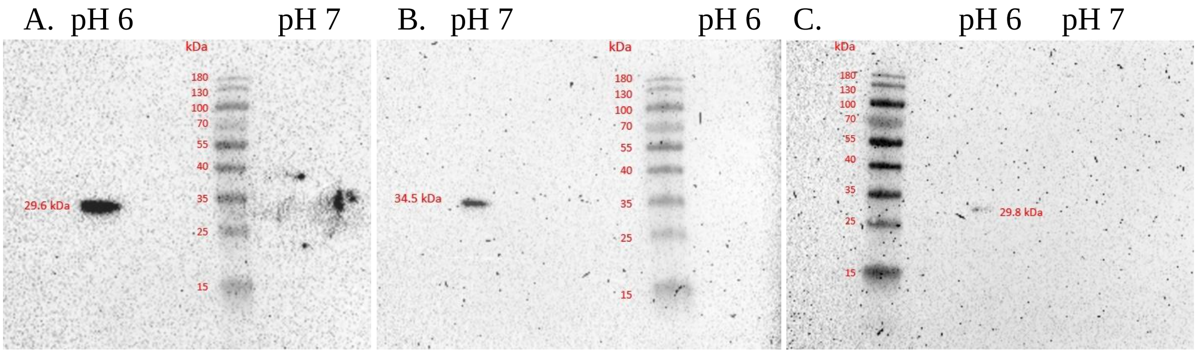


Figure 14. Western blot of CP10217 with pH 6 and pH 7. One band can be seen at 29.6 kDa for the pH 6 sample (A), one band at 34.5 kDa for the pH 7 sample (B) and one faint band at 29.8 kDa for the pH 6 sample (C).

Since it still seemed like the protein amount was low, a 2-litre culture with the pH 6 buffer was made. The Qubit measurement showed a protein concentration of 220 ng/µl this time, which was a bit higher than before. The SDS-PAGE did not show any band (Figure S3). The 2-litre culture was tried again. This time, all the previously purified protein samples were pooled together with the newly purified one and when measuring the protein concentration, it showed 270 ng/µl. The results of the SDS-PAGE still showed no band for CP10217 (Figure S4).

Because of all the previous results, optimization of the expression using four different pH conditions were attempted. Like before, pH 6 and pH 7 conditions were chosen, but also pH 4 and pH 8, to see whether CP10217 would be less active. The protein concentrations were 69.6 ng/µl (pH 4), 48.0 ng/µl (pH 6), 68.4 ng/µl (pH 7) and 93.0 ng/µl (pH 8). In Figure 15, the results from the SDS-PAGE and western blot can be seen. No bands are visible for any of the samples on the SDS-PAGE (Figure 15A) and only a band for the pH 8 sample is visible around 28 kDa (Figure 15B), but no bands for the other samples on the western blot. The one band that is visible is indicating that the CP10217 is in its active form.

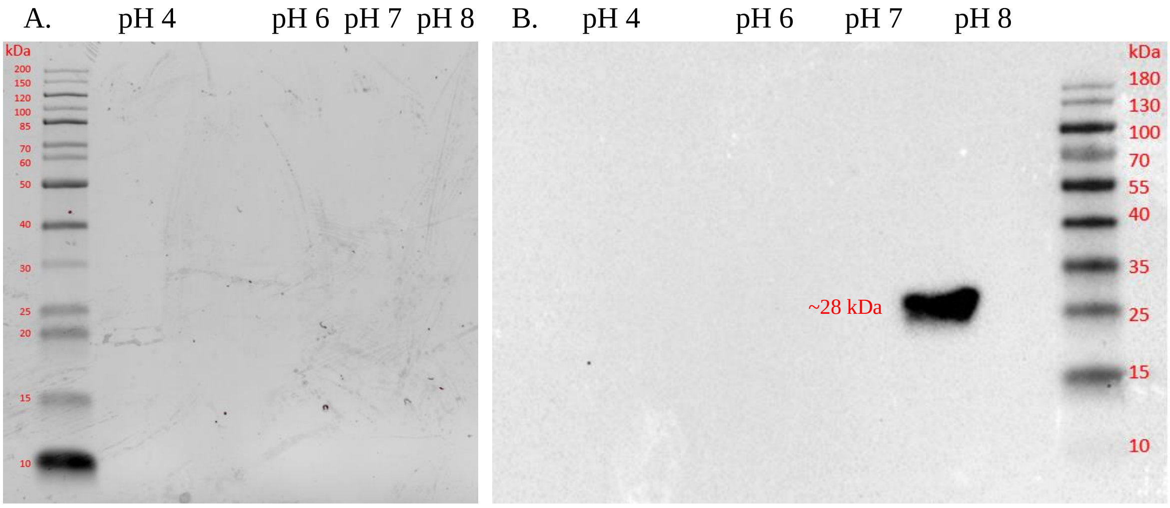


Figure 15. Results from the SDS-PAGE is showing no band for any of the different pH samples (A). Only one band, for the pH 8 sample, can be seen on the western blot around 28 kDa (B).

3.3 Cloning of 5’ UTR 16160 in front of CP10217 in pPICZαA vector and CP10217 in pPacV-integ-Hehl vector

Starting with PCR amplification, the designed primers could anneal to the DNA template, resulting in amplification of the wanted sequences (Figure 16A). Expected sizes were 143 bp, 234 bp and 1126 bp for the 5’ UTR 16160 fragment, His$_6$ + 3’ UTR 10217 fragment and 5’ UTR + 10217 + His$_6$ fragment, respectively. As can be seen in Figure 16A, bands around 150 bp, 250 bp and 1 kb is visible, indicating that the fragments are the correct ones. The measured concentrations for the fragments and the pPICZαA_10217 vector can be seen in Table 10. After the gel extraction, an OE-PCR was performed. The expected size of the product was 1339 bp for the CP10217 gene with its 5’ and 3’ UTR. As can be seen in Figure 16B, a band between 1 kb and 1.5 kb is visible, indicating that it is the CP10217 fragment. The measured concentration can be seen in Table 10. After the plasmid preparation of pPICZαA_10217, the measured concentration was 132 ng/ µl for the purified vector DNA.

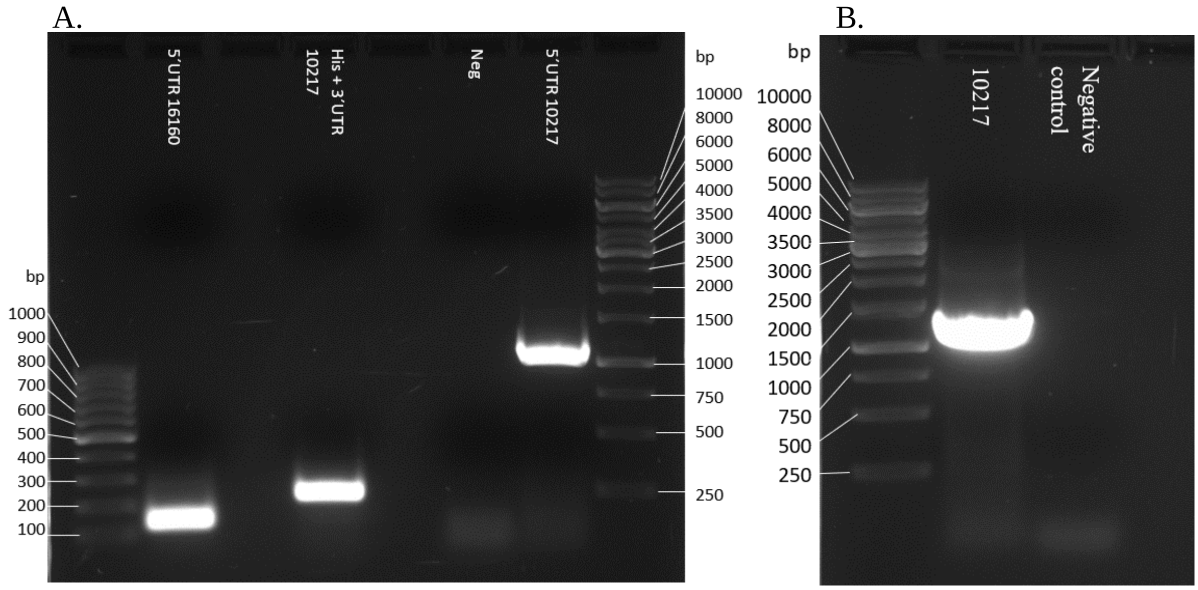


Figure 16. The results from the PCR amplification showing clear bands for each of the PCR fragments. For the 5' UTR fragment, a band around 150 bp can be seen. For the His$_6$ + 3' UTR 10217, a band around 250 bp can be seen. No band is visible for the negative control. A band around 1 kb can be seen for the 5' UTR + 10217 fragment (A). The results from the OE-PCR is showing a clear band between 1 kb and 1.5 kb, indicating that it is the CP10217 fragment. No band is visible for the negative control (B).

Table 10. Nanodrop measurements of the purified DNA fragments from the PCR, OE-PCR runs.

Sample	**Concentration (ng/µl)**
5' UTR 16160	33.0
His$_6$ + 3' UTR 10217	39.3
5' UTR + 10217 + His$_6$	28.8
10217	78.6

Now that all fragments had been amplified and purified, as well as having the pPICZαA_10217 vector purified, digestion was performed. The agarose gel UV image of the digested DNA (Figure 17) is showing bands around 4500 bp for pPICZαA_10217, 1400 bp for the 10217 fragment and below 250 bp for 5' UTR 16160, which is consistent to the expected sizes 4409 bp, 1329 bp and 136 bp. The concentration measured for each of the digestions can be seen in Table 11. The concentration for each of the digested samples were relatively low.

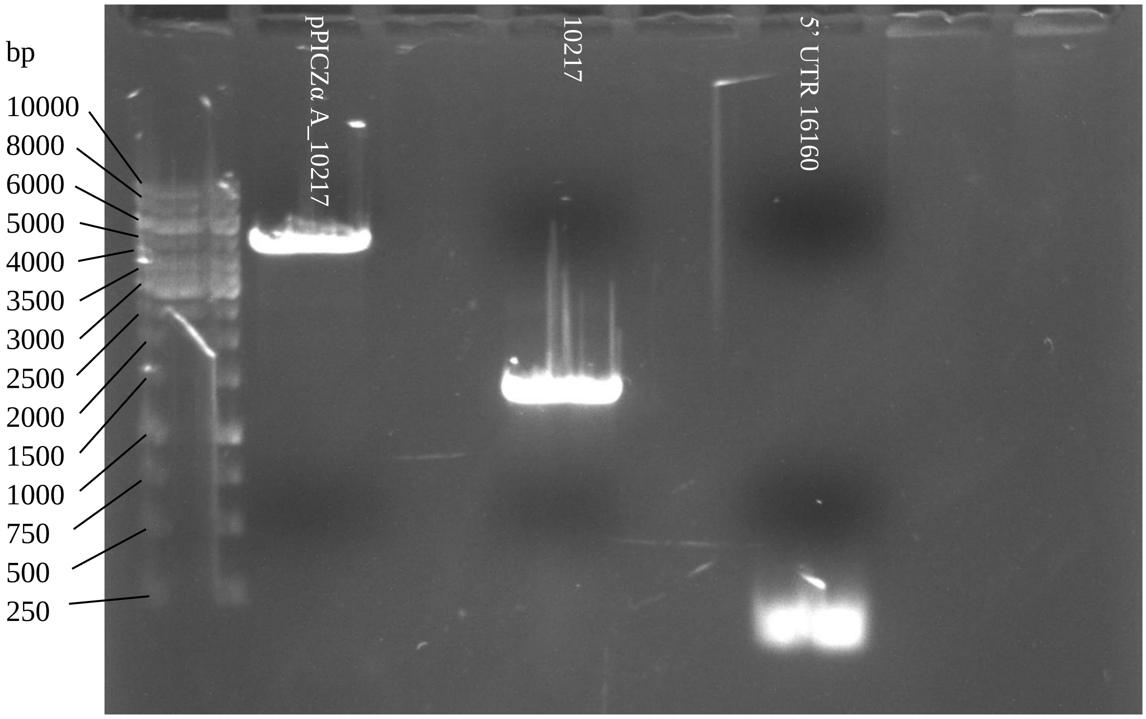


Figure 17. Agarose gel UV image of the digested pPICZαA_10217 vector, 10217 OE-PCR fragment and the 5’ UTR 16160 fragment. The result from the digestion is showing bands around 4500 bp, 1400 bp and below 250 bp, indicating the correct sizes for each fragment.

Table 11. Nanodrop measurements of the digested DNA fragments and vector.

Sample	Concentration (ng/µl)
Digested 5’ UTR 16160	10.5
Digested 10217	12.2
Digested pPICZαA_10217	9.7

After the digestion, the inserts and vectors were ligated and then transformed into *E. coli* DH5α cells. 29 colonies were counted for the pPacV-integ-Hehl_10217 containing cells and 691 colonies were counted for the pPICZαA_10217+5’UTR 16160 containing cells. After growing overnight cultures, the vectors were purified, and the concentrations of the samples were measured with Nanodrop (Table 12). Thereafter, a PCR was done to confirm that the ligation and transformation had worked. As can be seen in Figure 18, all colonies that were picked contained the inserts of interest. Bands are visible for all PCR samples around 1400 bp and 140 bp for each of the two inserts, which matches the expected sizes of 1339 bp and 143 bp. To verify that it was the correct inserts, vector DNA was sent for Sanger sequencing. The results showed that it indeed was the wanted sequences inserted into the vector. Although, when aligning the sequences, it was found that for CP10217 amino acid 294 (Glutamic acid - GAG), in the pPacV-int-Hehl vector, the first base had been changed to a T, resulting in a premature stop codon (TAG).

Table 12. Nanodrop measurements of the purified transformed vectors.

pPacV-integ-Hehl_10217 sample	Concentration (ng/µl)	pPICZαA_10217+ 5'UTR 16160 sample	Concentration (ng/µl)
1	361.9	1	108.7
2	359.8	2	61.9
3	199.5	3	55.9
4	342.5	4	71.2
5	278.5	5	89.5
6	316.6	6	86.5

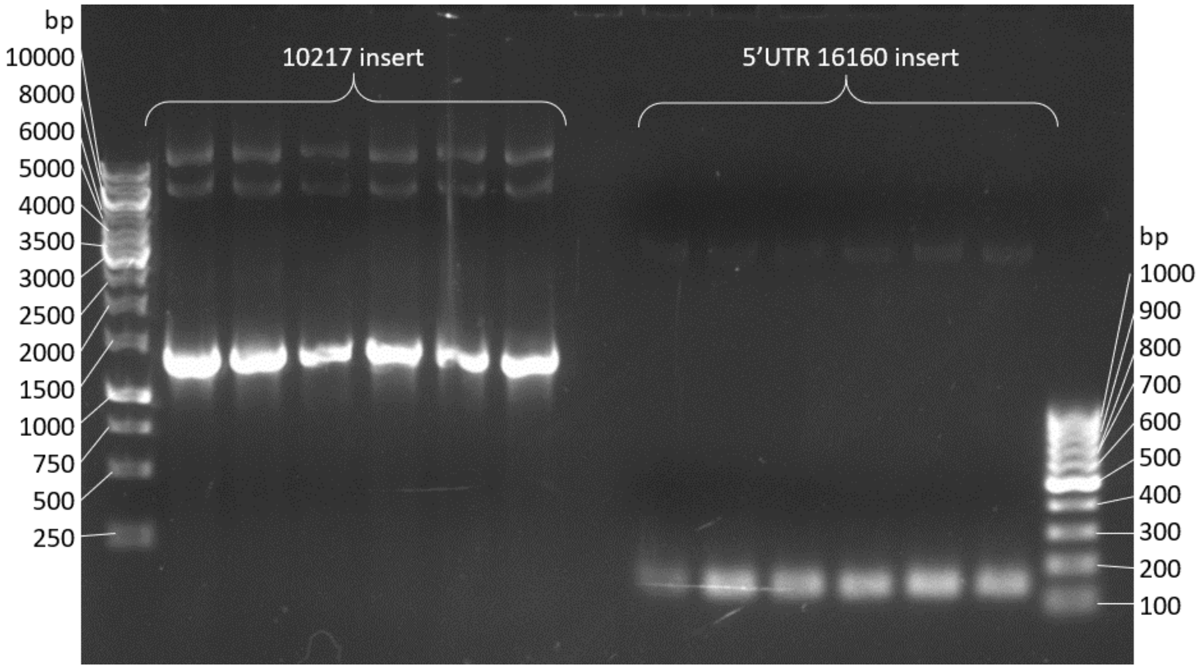


Figure 18. Agarose gel UV image verifying that the inserts had been ligated with the vectors. For the 10217 insert, bands between 1000 and 1500 bp is visible, indicating the correct insert size of 1339 bp. For the 5'UTR 16160 insert, bands between 100 and 200 bp are visible, indicating the correct insert size of 143 bp.

4 Discussion

As can be seen in Figure 4, the results from the MSA is showing that the ortholog sequences are very similar to each other. Only one small part of the MSA is varied. In this region, the sequences belonging to assemblage B (GSB_155190, GL50581_159, GL50581_3619 and GSB_153364) are the same but differ quite a lot from assemblage A (GL50803_10217 and DHA2_150845) and assemblage E (GLP15_4313) sequences. Assemblages A has only two amino acids differing between them, meaning that assemblage E is the odd one out, but still being more similar to assemblages A than B. This relation can be confirmed when looking at the phylogenetic tree (Figure 5). The generated tree has good significance on the branches, meaning that assemblage E is more closely related to assemblages A than B. This is interesting because assemblages A and B can both infect humans and animals, but assemblage E is only known to infect animals, like cattle, sheep, pigs and goats (Ryan & Cacciò 2013). Since the assemblage E ortholog seems to be more closely related to assemblage A, this might indicate that they can infect the same hosts and perhaps even assemblage E could infect humans zoonotically. Fantinatti *et al.* (2016) suggest that this might be true by showing that 15 out 44 nursery children in Rio de Janerio had the assemblage E *Giardia* genotype.

The secondary structure predictions from JPred (Figure 6), SPIDER2 (Figure 7) and I-TASSER (Figure 10) all showed similar results. Most of the residues are buried and not very exposed to the surface. They also show similar results for the number of helices and sheets, which in turn can be seen in the tertiary structure predictions (Figure 8, Figure 9, Figure 11). The SWISS-MODEL template (Figure 8) was procathepsin B from *T. brucei*, which also is an extracellular parasite using its variant surface glycoproteins (VSGs) to avoid and evade the immune system (Ponte-Sucre 2016). The Predictein template (Figure 9) was the human procathepsin B. Both *Trypanosoma* and human procathepsin B was recurring results when predicting the structure for CP10217, having high coverage for the global topology. Even though the sequence identity was low for all predictions, the coverage was usually high (>0.8), which would suggest good global topology (Table 7, Table 8, Table 9). When there is no crystal structure for a protein, it is good to have a reliable prediction when working with it experimentally in the laboratory. By knowing the predicted topology of the protein, it will be easier to *e.g.* design primers and tagging the protein, such that one can see if the structure of the protein would be heavily affected or not, and thus have a better chance for success.

The results from the purification clearly show that it is in fact CP10217 that is purified, but in very small amounts, only visible on a western blot (Figure 13) and not on a SDS-PAGE (Figure 12). The reason for this could be because the CP is too active when it is secreted, cleaving everything in its way, including other CP10217 molecules. Since we know that it is one of the major secreted CPs (Ward *et al.* 1997, DuBois *et al.* 2006) and perhaps one of the first to be secreted, this might not be surprising. If it is one of the first to be secreted, specific cleavage might not be wanted, but rather cleave as much as possible to ensure that the cell can evade the different obstacles, like the microbiota and mucus layer, to be able to attach to the epithelial cells in the small intestine. However, since the intent of this project was to characterize CP10217, a higher protein yield needed to be achieved from the expression and purification. Therefore, it was decided to try to optimize the expression by changing the pH conditions, hoping that the protein might be less active at another pH. The pH conditions that were chosen were based on the optimal pH conditions for activity of three other CPs (14019, 16779 and 16160) which are peaking at pH 5.5-6.0 (Liu *et al.* 2018). By going below or above these pH values, it could mean that the protein would be less active, and since *Pichia* can grow in pH ranges from 3 to 8 (Cregg *et al.* 1993, Jafari *et al.* 2011), pH 4, 6, 7 and 8 were chosen. The first try, by using pH 6 and 7, resulted in a band at 34.5 kDa (Figure 14B), which match the size of the full protein. But since the band for pH 7 is visible only when the band for pH 6 is not, and vice versa, it is hard to conclude anything from this analysis. It could easily have been a mix up when loading the samples. Therefore, it was further investigated by adding pH 4 and pH 8 expression conditions as well. The results from this try show low protein concentrations for all samples, as well as no visible bands on the SDS-PAGE (Figure 15A) and only showing a band for pH 8 on the western blot (Figure 15B). What can be concluded from this attempt is that (1) *P. pastoris* can indeed grow in medium with different pH ranges, and (2) CP10217 can be expressed in and purified from these different mediums, but with the same outcome as previous attempts.

Another approach to optimize the expression could have been by using an inhibitor in the media. There are commercially available CP inhibitors, both reversible and irreversible, that one can try to add to see whether it has a good effect on this protein. The reason for not trying this was (1) because it had already been tried by adding Pefabloc (Sigma-Aldrich) and no change in the activity had been noticed, and (2) it could be more efficient to try to re-clone the vector. It would still be interesting to try adding an inhibitor, but this time a CP inhibitor, rather than the serine inhibitor Pefabloc, like the irreversible inhibitor CA074, which is specific for cathepsin B proteins (Villalobo *et al.* 2003), or the reversible inhibitor leupeptin, which is also specific to cathepsin B (Knight 1980).

Since the expression and purification of CP10217 had resulted in too low protein yield, it was decided to modify the already existing pPICZαA_10217 vector by inserting 108 bp from the 5’ end of CP16160, which hopefully would affect the activity of CP10217. The reason for choosing the 108 bp from 16160 was because it had accidentally been cloned into the vector by a PhD student when expressing and purifying CP16160 and showed very good results. The sequence available on GiadiaDB for CP16160 has included a start codon 108 bp in front of the actual start codon, located in the 5’ UTR region of this CP. When examining this closer, it is believed that this 108 bp makes a significant difference for the activity of the CP when expressed in *Pichia*. The thought was that the effects would be same for CP10217, and if so, this might suggest a new method to clone, express and purify proteases. Also, by inserting CP10217 into the pPacV-int-Hehl vector, the protein could be localized in *G. intestinalis* cells. As Figure 18 shows, the amplified inserts show the correct sizes of 1339 bp for 10217, and 143 bp for 5’ UTR 16160, suggesting that it is the correct inserts. After sequencing the vector DNA, it was confirmed that it indeed was the wanted inserts, but it also showed a mutation in the CP10217 insert sequence in pPacV-int-Hehl. For CP10217 amino acid 294, the first base, G, had been changed to a T, resulting in a premature stop codon (TAG), i.e. there had occurred a point-nonsense mutation (Strachan *et al.* 2011). It is possible that this point-nonsense mutation came to be when performing one of the PCRs. Even though the Phusion polymerase has a very low error rate compared to the Taq polymerase (Li *et al.* 2006), the estimated percentage of the PCR products having an error is about 2 percent for this 1339 bp insert. If this error occurs in the first rounds of the amplification, all other PCR products will have this as well. Also, when looking at the chromatogram in the BioEdit tool, the signal is showing a peak for T, but also a low signalling peak for G, suggesting that it might be some vector DNA with the correct nucleotide, but that most of the vector DNA has the point-nonsense mutation. To further investigate this, more of the vector DNA could be sequenced to see if one of the six samples had more of the correct vector.

The next step to further work on CP10217 would be to first linearize the vector DNA, transform it into competent *P. pastoris* cells and express the protein. Hopefully, this would lead to a less active protein and a larger protein yield, than the previous attempts. Thereafter, characterization of the protein can start by performing *e.g.* phage display. This will hopefully result in finding the cleavage specificity of CP10217 and thus coming one step closer to fully characterizing it.

5 Acknowledgements

First of all, I would like to thank my supervisor Staffan Svärd for giving me this great opportunity and interesting project to work on. Thank you for letting me into your group and showing me the academic world of research. You are a great teacher and I have learned a lot from you. A special thank you to Erik Holmqvist for making time in your schedule, meeting with me to discuss the project and reviewing my work. A big thank you to Jana Grüttner for helping me in the lab and answering all my million questions that I had. Thank you for all the laughs and support you have given me. I would also like to thank all other members of the Giardia group, always being so nice to me and helping me when I needed it. Same goes to the other people in the Micro group in the corridor, thank you for everything.

Thank you Jan Andersson and Lena Henriksson for examining and evaluating my project. I would also like to thank you for all these years in the X-programme, making sure that we get a good education and are prepared for the future. The program would not be what it is without your hard work.

Lastly, I would like to thank my lovely family and friends for always supporting and encouraging me, I would not be where I am today if it was not for you. Thank you for always

asking questions about my project, even though it might be difficult to understand, and thank you for all the other times when I needed to take my mind off everything.

References

Ankarklev J, Jerlström-Hultqvist J, Ringqvist E, Troell K, Svärd SG. 2010. Behind the smile: cell biology and disease mechanisms of Giardia species. Nature Reviews Microbiology 8: 413–422.

Ansell BRE, Pope BJ, Georgeson P, Emery-Corbin SJ, Jex AR. 2019. Annotation of the Giardia proteome through structure-based homology and machine learning. GigaScience, doi 10.1093/gigascience/giy150.

Beatty JK, Akierman SV, Motta J-P, Muise S, Workentine ML, Harrison JJ, Bhargava A, Beck PL, Rioux KP, McKnight GW, Wallace JL, Buret AG. 2017. Giardia duodenalis induces pathogenic dysbiosis of human intestinal microbiota biofilms. International Journal for Parasitology 47: 311–326.

Bhargava A, Cotton JA, Dixon BR, Gedamu L, Yates RM, Buret AG. 2015. Giardia duodenalis Surface Cysteine Proteases Induce Cleavage of the Intestinal Epithelial Cytoskeletal Protein Villin via Myosin Light Chain Kinase. PLoS ONE, doi 10.1371/journal.pone.0136102.

Block H, Maertens B, Spriestersbach A, Brinker N, Kubicek J, Fabis R, Labahn J, Schäfer F. 2009. Chapter 27 Immobilized-Metal Affinity Chromatography (IMAC): A Review. I: Burgess RR, Deutscher MP (red.). Methods in Enzymology, s. 439–473. Academic Press,

Certad G, Viscogliosi E, Chabé M, Cacciò SM. 2017. Pathogenic Mechanisms of Cryptosporidium and Giardia. Trends in Parasitology 33: 561–576.

Cotton JA, Bhargava A, Ferraz JG, Yates RM, Beck PL, Buret AG. 2014. Giardia duodenalis Cathepsin B Proteases Degrade Intestinal Epithelial Interleukin-8 and Attenuate Interleukin-8-Induced Neutrophil Chemotaxis. Infection and Immunity 82: 2772–2787.

Cregg JM, Vedvick TS, Raschke WC. 1993. Recent Advances in the Expression of Foreign Genes in Pichia pastoris. Nature Biotechnology 11: 905–910.

Dereeper A, Audic S, Claverie J-M, Blanc G. 2010. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evolutionary Biology 10: 8.

Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Research 43: W389–W394.

DuBois KN, Abodeely M, Sajid M, Engel JC, McKerrow JH. 2006. Giardia lamblia cysteine proteases. Parasitology Research 99: 313–316.

Dubourg A, Xia D, Winpenny JP, Al Naimi S, Bouzid M, Sexton DW, Wastling JM, Hunter PR, Tyler KM. 2018. Giardia secretome highlights secreted tenascins as a key component of pathogenesis. GigaScience 7: 1–13.

Efstratiou A, Ongerth JE, Karanis P. 2017. Waterborne transmission of protozoan parasites: Review of worldwide outbreaks - An update 2011–2016. Water Research 114: 14–22.

Einarsson E, Ma'ayeh S, Svärd SG. 2016a. An up-date on Giardia and giardiasis. Current Opinion in Microbiology 34: 47–52.

Einarsson E, Svärd SG. 2015. Encystation of Giardia intestinalis—a Journey from the Duodenum to the Colon. Current Tropical Medicine Reports 2: 101–109.

Einarsson E, Troell K, Hoeppner MP, Grabherr M, Ribacke U, Svärd SG. 2016b. Coordinated Changes in Gene Expression Throughout Encystation of Giardia intestinalis. PLoS Neglected Tropical Diseases, doi 10.1371/journal.pntd.0004571.

Fantinatti M, Bello AR, Fernandes O, Da-Cruz AM. 2016. Identification of *Giardia lamblia* Assemblage E in Humans Points to a New Anthropozoonotic Cycle. Journal of Infectious Diseases 214: 1256–1259.

Fisher BS, Estraño CE, Cole JA. 2013. Modeling Long-Term Host Cell-Giardia lamblia Interactions in an In Vitro Co-Culture System. PLOS ONE 8: e81104.

Gerbaba TK, Green-Harrison L, Buret AG. 2017. Modeling Host-Microbiome Interactions in Caenorhabditis elegans. Journal of Nematology 49: 348–356.

Halliez MCM, Motta J-P, Feener TD, Guérin G, LeGoff L, François A, Colasse E, Favennec L, Gargala G, Lapointe TK, Altier C, Buret AG. 2016. Giardia duodenalis induces paracellular bacterial translocation and causes postinfectious visceral hypersensitivity. American Journal of Physiology-Gastrointestinal and Liver Physiology 310: G574–G585.

Invitrogen^{TM}. 2010. pPICZ A, B, C Pichia Vectors - Thermo Fisher Scientific. WWW-dokument 2010-07-07: https://www.thermofisher.com/order/catalog/product/V19520?SID=srch-srp-V19520. Hämtad 2019-01-28.

Jafari R, Sundström BE, Holm P. 2011. Optimization of production of the anti-keratin 8 single-chain Fv TS1-218 in Pichia pastoris using design of experiments. Microbial Cell Factories 10: 34.

Knight CG. 1980. Human cathepsin B. Application of the substrate *N* -benzyloxycarbonyl-l-arginyl-l-arginine 2-naphthylamide to a study of the inhibition by leupeptin. Biochemical Journal 189: 447–453.

Koh WH, Geurden T, Paget T, O'Handley R, Steuart RF, Thompson RCA, Buret AG. 2013. Giardia duodenalis Assemblage-Specific Induction of Apoptosis and Tight Junction Disruption in Human Intestinal Epithelial Cells: Effects of Mixed Infections. Journal of Parasitology 99: 353–359.

Li M, Diehl F, Dressman D, Vogelstein B, Kinzler KW. 2006. BEAMing up for detection and quantification of rare sequence variants. Nature Methods 3: 95.

Litleskare S, Wensaas K-A, Eide GE, Hanevik K, Kahrs GE, Langeland N, Rortveit G. 2015. Perceived food intolerance and irritable bowel syndrome in a population 3 years after a giardiasis-outbreak: a historical cohort study. BMC Gastroenterology, doi 10.1186/s12876-015-0393-0.

Liu J, Fu Z, Hellman L, Svärd SG. 2019. Cleavage specificity of recombinant Giardia intestinalis cysteine proteases: Degradation of immunoglobulins and defensins. Molecular and Biochemical Parasitology 227: 29–38.

Liu J, Ma'ayeh S, Peirasmaki D, Lundström-Stadelmann B, Hellman L, Svärd SG. 2018. Secreted Giardia intestinalis cysteine proteases disrupt intestinal epithelial cell junctional complexes and degrade chemokines. Virulence 9: 879–894.

Ma'ayeh SY, Liu J, Peirasmaki D, Hörnaeus K, Bergström Lind S, Grabherr M, Bergquist J, Svärd SG. 2017. Characterization of the Giardia intestinalis secretome during interaction with human intestinal epithelial cells: The impact on host cells. PLoS Neglected Tropical Diseases, doi 10.1371/journal.pntd.0006120.

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research, doi 10.1093/nar/gkz268.

Maia-Brigagão C, Morgado-Díaz JA, De Souza W. 2012. Giardia disrupts the arrangement of tight, adherens and desmosomal junction proteins of intestinal cells. Parasitology International 61: 280–287.

Nelson DL, Cox MM. 2017. Lehninger principles of biochemistry, Seventh edition. W.H. Freeman ; Macmillan Learning, New York : Houndmills, Basingstoke.

Ponte-Sucre A. 2016. An Overview of Trypanosoma brucei Infections: An Intense Host–Parasite Interaction. Frontiers in Microbiology, doi 10.3389/fmicb.2016.02126.

Ringqvist E, Avesson L, Söderbom F, Svärd SG. 2011. Transcriptional changes in Giardia during host–parasite interactions. International Journal for Parasitology 41: 277–285.

Ryan U, Cacciò SM. 2013. Zoonotic potential of Giardia. International Journal for Parasitology 43: 943–956.

Strachan T, Read AP, Strachan T. 2011. Human molecular genetics, 4th ed. Garland Science/Taylor & Francis Group, New York.

UniProt Consortium T. 2018. UniProt: the universal protein knowledgebase. Nucleic Acids Research 46: 2699–2699.

Villalobo E, Moch C, Fryd-Versavel G, Fleury-Aubusson A, Morin L. 2003. Cysteine Proteases and Cell Differentiation: Excystment of the Ciliated Protist Sterkiella histriomuscorum. Eukaryotic Cell 2: 1234–1245.

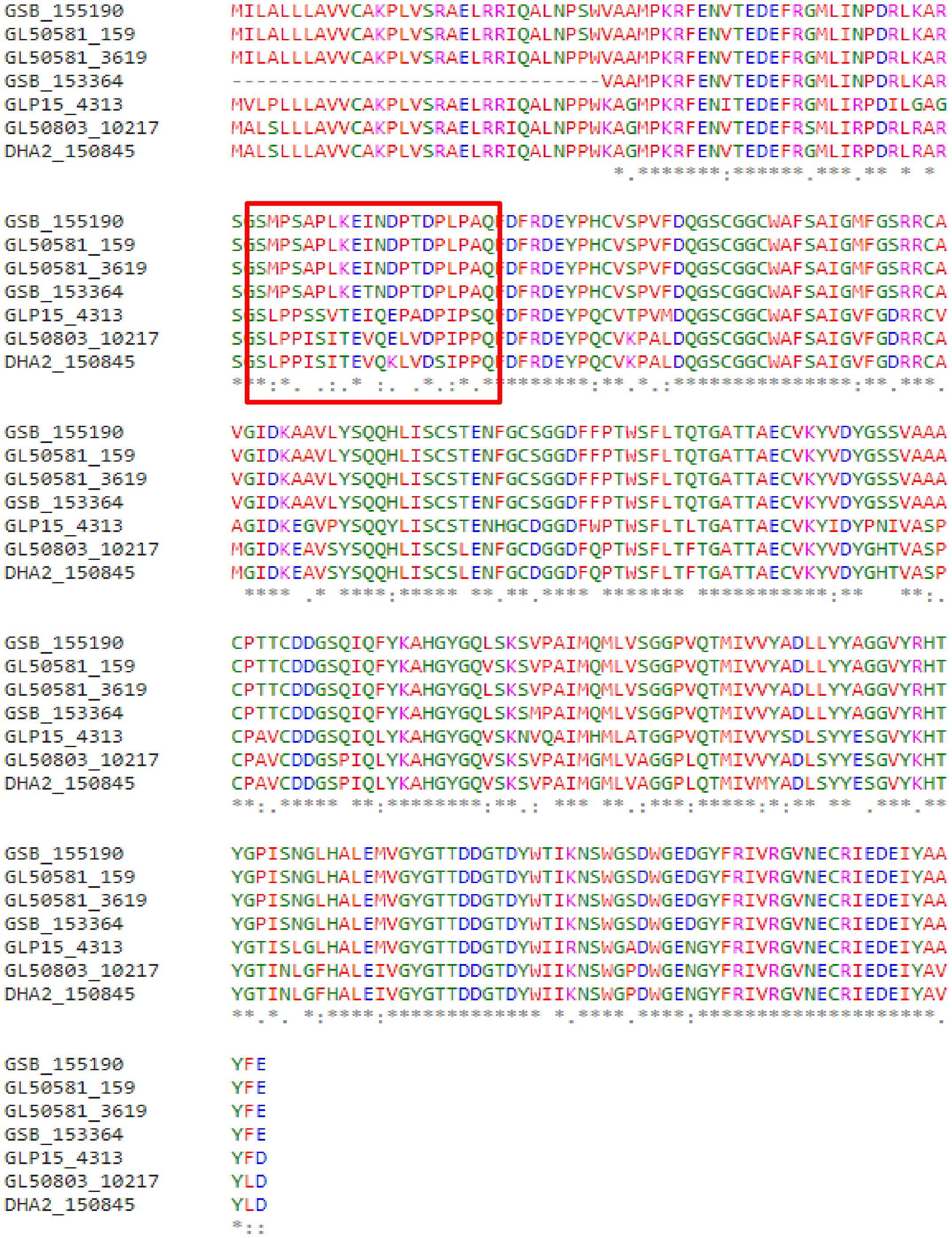
Ward W, Alvarado L, Rawlings ND, Engel JC, Franklin C, McKerrow JH. 1997. A Primitive Enzyme for a Primitive Cell: The Protease Required for Excystation of Giardia. Cell 89: 437–444.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research 46: W296–W303.

Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Zhou Y. 2017. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. Methods in Molecular Biology (Clifton, NJ) 1484: 55–63.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9: 40.

Supplements



```
GSB_155190       MILALLLAVVCAKPLVSRAELRRIQALNPSWVAAMPKRFENVTEDEFRGMLINPDRLKAR
GL50581_159      MILALLLAVVCAKPLVSRAELRRIQALNPSWVAAMPKRFENVTEDEFRGMLINPDRLKAR
GL50581_3619     MILALLLAVVCAKPLVSRAELRRIQALNPPWVAAMPKRFENVTEDEFRGMLINPDRLKAR
GSB_153364       -------------------------------VAAMPKRFENVTEDEFRGMLINPDRLKAR
GLP15_4313       MVLPLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENITEDEFRGMLIRPDILGAG
GL50803_10217    MALSLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENVTEDEFRSMLIRPDRLRAR
DHA2_150845      MALSLLLAVVCAKPLVSRAELRRIQALNPPWKAGMPKRFENVTEDEFRGMLIRPDRLRAR
                                                 *.*******:******.***.** * *

GSB_155190       SGSMPSAPLKEINDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA
GL50581_159      SGSMPSAPLKEINDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA
GL50581_3619     SGSMPSAPLKEINDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA
GSB_153364       SGSMPSAPLKETNDPTDPLPAQFDFRDEYPHCVSPVFDQGSCGGCWAFSAIGMFGSRRCA
GLP15_4313       SGSLPPSSVTEIQEPADPIPSQFDFRDEYPQCVTPVMDQGSCGGCWAFSAIGVFGDRRCV
GL50803_10217    SGSLPPISITEVQELVDPIPPQFDFRDEYPQCVKPALDQGSCGGCWAFSAIGVFGDRRCA
DHA2_150845      SGSLPPISITEVQKLVDSIPPQFDFRDEYPQCVKPALDQGSCGGCWAFSAIGVFGDRRCA
                 ***:*. .:.* :. .*.:*.*********:**.*.:***************:**.***.

GSB_155190       VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA
GL50581_159      VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA
GL50581_3619     VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA
GSB_153364       VGIDKAAVLYSQQHLISCSTENFGCSGGDFFPTWSFLTQTGATTAECVKYVDYGSSVAAA
GLP15_4313       AGIDKEGVPYSQQYLISCSTENHGCDGGDFWPTWSFLTLTGATTAECVKYIDYPNIVASP
GL50803_10217    MGIDKEAVSYSQQHLISCSLENFGCDGGDFQPTWSFLTFTGATTAECVKYVDYGHTVASP
DHA2_150845      MGIDKEAVSYSQQHLISCSLENFGCDGGDFQPTWSFLTFTGATTAECVKYVDYGHTVASP
                  **** .* ****:***** **.**.**** ******* ***********:**   **:.

GSB_155190       CPTTCDDGSQIQFYKAHGYGQLSKSVPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT
GL50581_159      CPTTCDDGSQIQFYKAHGYGQVSKSVPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT
GL50581_3619     CPTTCDDGSQIQFYKAHGYGQLSKSVPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT
GSB_153364       CPTTCDDGSQIQFYKAHGYGQLSKSMPAIMQMLVSGGPVQTMIVVYADLLYYAGGVYRHT
GLP15_4313       CPAVCDDGSQIQLYKAHGYGQVSKNVQAIMHMLATGGPVQTMIVVYSDLSYYESGVYKHT
GL50803_10217    CPAVCDDGSPIQLYKAHGYGQVSKSVPAIMGMLVAGGPLQTMIVVYADLSYYESGVYKHT
DHA2_150845      CPAVCDDGSPIQLYKAHGYGQVSKSVPAIMGMLVAGGPLQTMIVMYADLSYYESGVYKHT
                 **:.***** **:********:**.: *** **.:***:*****:*:** ** .***.**

GSB_155190       YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA
GL50581_159      YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA
GL50581_3619     YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA
GSB_153364       YGPISNGLHALEMVGYGTTDDGTDYWTIKNSWGSDWGEDGYFRIVRGVNECRIEDEIYAA
GLP15_4313       YGTISLGLHALEMVGYGTTDDGTDYWIIRNSWGADWGENGYFRIVRGVNECRIEDEIYAA
GL50803_10217    YGTINLGFHALEIVGYGTTDDGTDYWIIKNSWGPDWGENGYFRIVRGVNECRIEDEIYAV
DHA2_150845      YGTINLGFHALEIVGYGTTDDGTDYWIIKNSWGPDWGENGYFRIVRGVNECRIEDEIYAV
                 **.*. *:****:************* *.****.****:********************.

GSB_155190       YFE
GL50581_159      YFE
GL50581_3619     YFE
GSB_153364       YFE
GLP15_4313       YFD
GL50803_10217    YLD
DHA2_150845      YLD
                 *::
```

Figure S1. MUSCLE MSA of CP10217 and its orthologs. The red square indicates the part of the MSA where the sequences varied the most. The sequences seem to be very conserved between the orthologs. Assemblage E (GLP15_4313) is the one that varies the most from the others but are more like the assemblage A sequences (GL50803_10217 and DHA2_150845) than the assemblage B sequences.

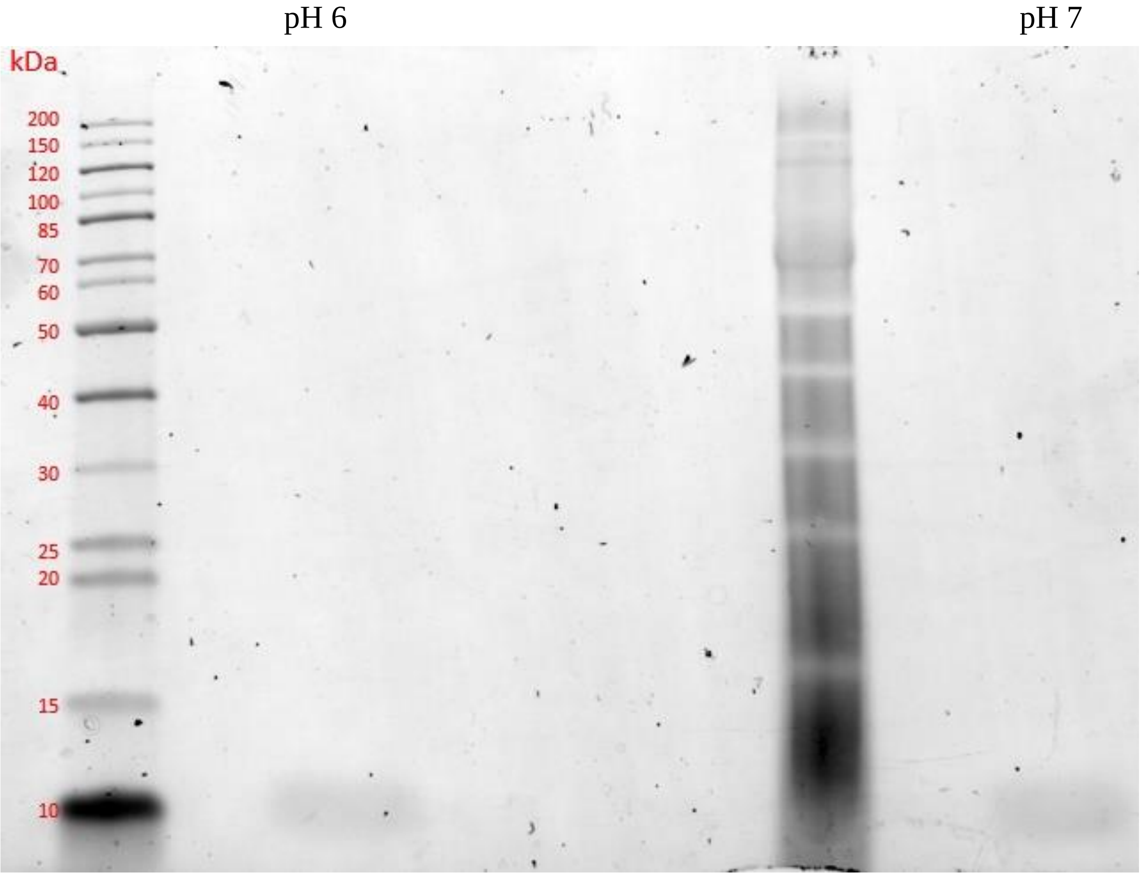


Figure S2. SDS-PAGE of the two different pH samples show no bands.

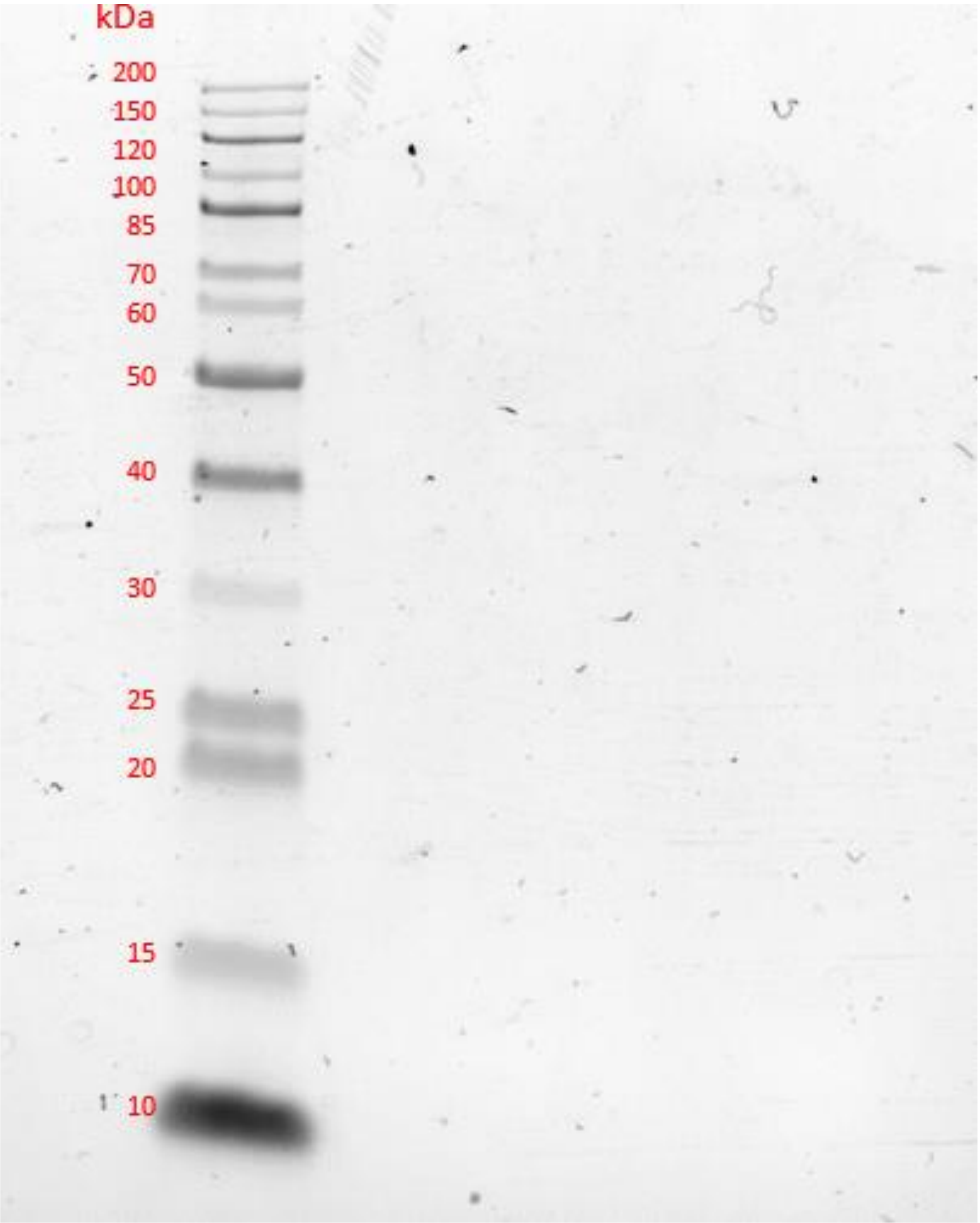


Figure S3. SDS-PAGE of the first try with the 2 litre culture, showing no band.

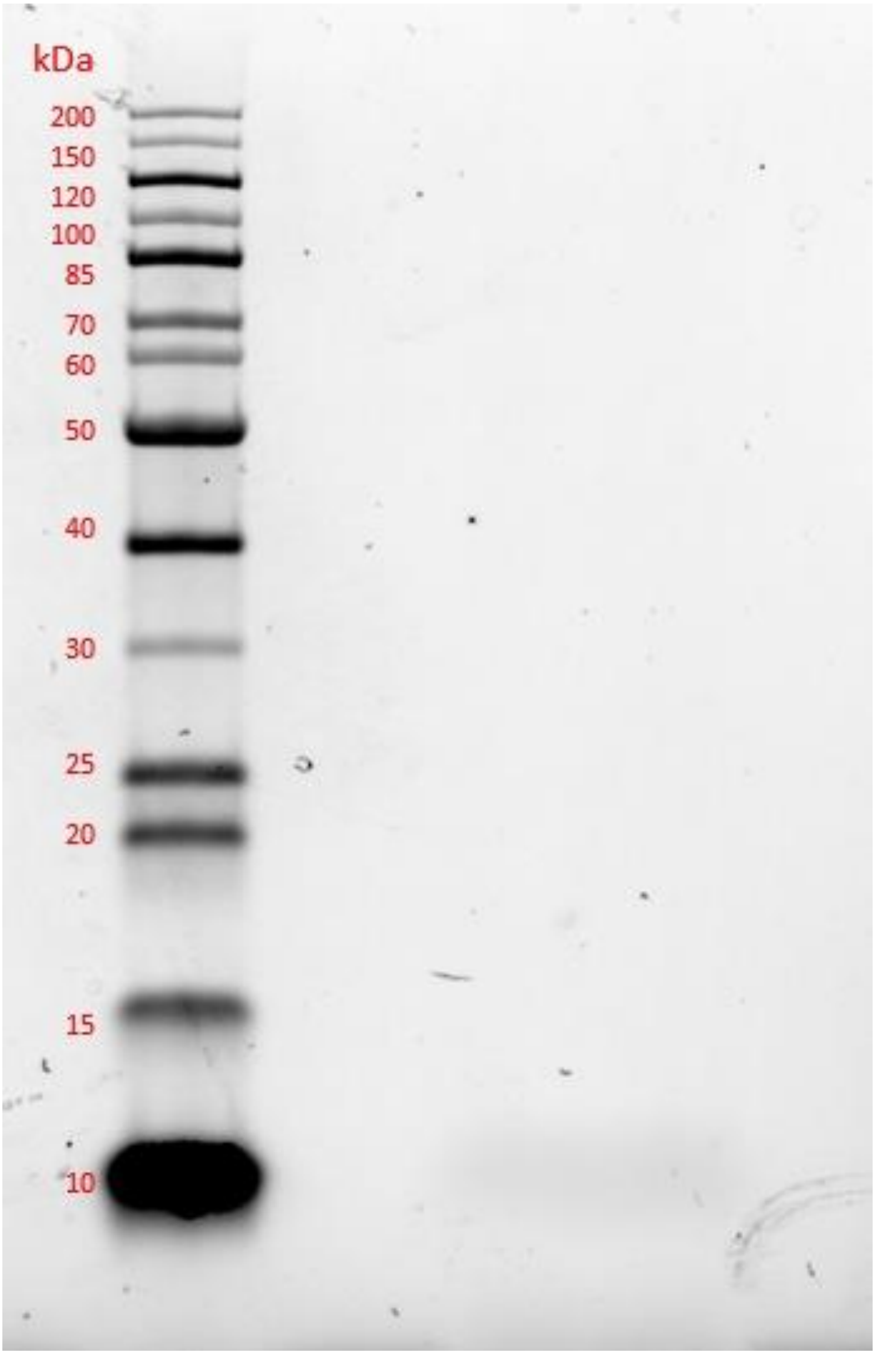


Figure S4. SDS-PAGE of the second try with the 2-litre culture, showing no bands.