DEPARTAMENTO DE CIÊNCIAS DA VIDA

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Bacterial Retropepsin-like Proteases: The Evidence from *Legionella pneumophila*

Paulo Alexandre Gonçalves Teixeira

2013

DEPARTAMENTO DE CIÊNCIAS DA VIDA

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Bacterial Retropepsin-like Proteases:
The Evidence from *Legionella pneumophila*

Dissertação apresentada à Universidade de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Bioquímica, realizada sob a orientação científica do Professor Doutor Euclides Pires (Universidade de Coimbra) e da Doutora Isaura Simões (CNC/Biocant)

Paulo Alexandre Gonçalves Teixeira

2013

“Eles não sabem, nem sonham,

que o sonho comanda a vida,

que sempre que um homem sonha

o mundo pula e avança

como bola colorida

entre as mãos de uma criança.”

Pedra Filosofal

António Gedeão

Agradecimentos

Como os sonhos não os realizamos sozinhos, quero deixar alguns agradecimentos ás pessoas que tornaram possível este caminho.

Ao Professor Doutor Carlos Faro e ao Professor Doutor Euclides por tornarem possível fazer-se ciência no Biocant tão bem como podemos ter o privilégio de fazer.

Á Doutora Isaura Simões, pela dedicação que mostra a cada passo dos dias que passam. Por contribuir de forma tão correcta e rigorosa para a minha formação e por me ensinar a procurar o importante em cada detalhe. Pela disponibilidade em receber cada dificuldade e trasforma-la numa solução. Pela liderança entusiasta que a todos motiva a ir mais além e a não desistir face a dificuldades. Agradeço então pela sabedoria e atitude que levo comigo destes últimos dois anos.

A todos na Unidade de Biotecnologia Molecular, ao Doutor Pedro Castanheira, á Carla, á Marisa, ao Rui que tudo resolve, ao André, á Joana, ao Pedro, á Rita e á Ana.

Á minha mãe, que sempre me ajudou em tudo o que preciso e que sempre me motivou e inspirou para ser tudo o que sou. Que me ensinou a correr por aquilo que queremos e acreditamos.

Á minha irmãzinha pela felicidade que mostra, pelo amor que tem por mim, e por me lembrar o quão importante é sermos sempre crianças: curiosos, sinceros e sonhadores.

Aos meus tios, que sempre me apoiaram nas mais diversas situações. Com os quais sempre partilhei não só preocupações mas também ambiente de festa.

Á minha prima que nunca me deixou faltar nada, que me transmitiu todo o espírito de Coimbra e nunca me deixou faltar á Queima!

Á minha avó que sempre se preocupou, que me viu crescer e a quem posso agradecer pelo mundo com quem descobri em criança.

Á minha querida Paula agradeço de forma muito especial, por me aquecer o coração nos piores e melhores momentos. Por me mostrar o quão longe a vida nos pode levar e como podemos encontrar a felicidade nas pequenas coisas que nos identificam, porque “nós” é muito mais que “eu e tu”.

Agradeço também aos pais da Paula: Carlos e Antónia, que calorosamente me acolheram como família e sempre se disponibilizaram para ajudar.

Deixo um grande abraço aos meus grandes amigos bioquímicos, em que tudo começou nos *after-lunch* no tropical! Á Daniela que muito me aturou neste último ano e com quem partilhei inúmeros cafés e conversas aleatórias, ao Ricardo Oliveira por manter a festa sempre no ar,á Susana, á Mikki e ao Beta pelos fantásticos amigos que são. Ao indispensável e inseparável trio *nerd* das noites de *League of Legends* e *Magic:* Caramelo, pela visão única do mundo que mais ninguém consegue ter; Tiago, pelas discussões intermináveis sobre qualquer tema que ocorresse; e Ricardo, que me apadrinhou como caloiro mal sabendo das demandas épicas que surgiriam mais tarde!

E como agradecimento final, deixo a todos os meu amigos que me tornaram quem sou e aos meus colegas de Bioquímica com quem partilhei estes anos de faculdade.

Table of Contents

Abbreviations

A280nm - Absorbance at 280nm

AIDS - Acquired Immune Deficiency Syndrome

AP - Aspartic Protease

CDR1 - Constitutive Disease Resistance 1

DABCYL - 4- (dimethylaminoazo)benzene-4-carboxylic acid

DMSO - Dimethyl sulfoxide

DNA - Deoxyribonucleic Acid

DNP - 2,4-Dinitrophenyl

DTT - Dithiothreitol

E-64 - Trans-Epoxysucciny-L-leucyl-amido(4-guanidino) butane

EDANS - 5-[(2 -aminoethyl)amino]naphthalene - 1 - sulfonic acid

EDTA - Ethylenediaminetetraacetic acid

EIAV - Equine Infectious Anemia Virus

ER – Endoplasmic Reticulum

FPLC - Fast Protein Liquid Chromatography

FIV – Feline Immunodeficiency Virus

HA – Hemagglutinin

HIV – Human Immunodeficiency Virus

IMAC - Immobilized metal ion affinity chromatography

Ins – insoluble fraction

IPTG - Isopropyl β-D-1-thiogalactopyranoside

kDa - Kilodalton

LB – Luria Broth

LBkan - Luria Broth with Kanamycin

LCV – *Legionella*-Containing Vacuole

LegRP – Legionella retroviral-like aspartic protease

LegRPsd – Soluble Domain of the Legionella retroviral-like aspartic protease

MCA - (7-methoxycoumarin-4-yl)acetyl

MCS – Multiple Cloning Site

MES - 2-(N-morpholino)ethanesulfonic acid

min – Minutes

MLV – Murine Leukemia Virus

mut - mutant

OD600nm - Optical density at 600nm

PDVF - Polyvinylidene Fluoride

PCR – Polymerase Chain Reaction

pI – Isoelectric Point

RNA - Ribonucleic Acid

RP – Retropepsin

RP-HPLC - Reversed phase high-performance liquid chromatography

RSV – Rous Sarcoma Virus

s – seconds

sol – soluble fraction

SDS - Sodium dodecyl sulfate

SDS-PAGE - Sodium dodecyl sulfate polyacrylamide gel electrophoresis

SIV – Simian Immunodeficiency Virus

TB - Terrific Broth

TBS - Tris Buffered Saline

TBST - Tris Buffered Saline with Tween 20

TFA - Trifluoroacetic Acid

TMH – Transmemebrane Helix

V -Volts

WT - Wild type

XMRV - Xenotropic murine leukemia virus-related virus

Abstract

A2 family of aspartic proteases harbors mostly proteases found in retroviruses – the retropepsins. The evolution theories regarding these proteases usually state that these proteins are related to pepsin-like proteases from family A1 by two different hypotheses. By the first (and usually most accepted) theory, upon infection of a eukaryote cell by a retrovirus, the retropepsin gene would have integrated into the host's genome and undergone a gene duplication and fusion event, giving rise to the first pepsin-like protease. The second theory would be that upon a virus infection of a eukaryote cell, half of a pepsin gene would be accidentally captured by the virus, evolving and creating the first retropepsin-like protease. Both theories neglected the possibility that these enzymes would exist in prokaryotes. Recently, through bioinformatics, several putative retropepsin-like and pepsin-like protease sequences have been found in prokaryotes, which challenge the evolutionary theories presented so far. In fact, it has already been experimentally proved the existence of the first pepsin-like protease in a prokaryote (shewasin A) and successfully characterized as an active aspartic protease.

For this matter, the experimental validation regarding the existence of retropepsin-like aspartic proteases in prokaryotes needs to be addressed. Studies in our lab have already experimentally demonstrated the presence in *Rickettsia* of a highly conserved active aspartic protease with retropepsin-like signatures and features. However, this protease may not be the only one: bioinformatics analysis showed the existence of a large group of genes related to the *Rickettsia* retropepsin-like protease in several distinct prokaryote families, all revealing conserved retropepsin-like sequence features but with some interesting differences between them.

The peculiar intracellular pathogen *Legionella pneumophila* is one of the several bacteria species in which a retropepsin-like sequence can be found. In this study, the putative soluble domain from the retropepsin-like protease sequence from the *L. pneumophila* genome has been cloned and expressed in *E. coli*. The protein's soluble domain, termed LegRPsd, was characterized in terms of its enzymatic activity and oligomerization states.

The LegRPsd protein has been proved an active enzyme, with our results pointing towards its inclusion as new member of the aspartic protease family. This protein showed proteolytic activity towards a peptide usually used as a typical substrate for aspartic proteases at an optimum pH of 4.0 and the activity is inhibited by about 50% by the aspartic protease inhibitor pepstatin A. LegRPsd was also proved to undergo auto-processing in a similar way as other retropepsin-like proteases, maturating into a form that is no longer inhibited by pepstatin A but strongly inhibited by specific inhibitors of HIV-1 retropepsin. The protein also showed other characteristics close to retropepsins, like the formation of a weak homodimer structure, thought to be needed for the formation of the active center and therefore, fundamental for activity. Moreover, protease activity was shown to be impaired by mutation of the putative catalytic aspartate.

Altogether, these results provide strong experimental evidences regarding the existence of an active protease in *L. pneumophila* close to retropepsins not only in terms of sequence-structure homology, but also by its biochemical properties. This information becomes most valuable because LegRP, together with the *Rickettsia* aspartic protease, stands as a strong evidence of a novel family of aspartic proteases conserved in prokaryotes and related to retropepsins; this brings a redefinition of evolutionary theories regarding pepsins and retropepsins, suggesting a new point of view where these prokaryotic sequences represent the most ancestral state of retropepsins. Not least important is that the discovery of a conserved aspartic protease in *L. pneumophila* and other pathogenic bacteria raises the possibility of finding new pathogenic pathways and consequently new therapeutic targets for these difficult-to-treat infections as an alternative to the conventional antibiotics.

Keywords

Aspartic Protease; Bacteria; *Legionella pneumophila*; Retropepsin

Resumo

A familia A2 de proteases aspárticas é constituida maioritariamente por proteases encontradas em retrovírus – as retropepsinas. As teorias evolutivas inerentes a estas proteases normalmente referem que estarão relacionadas com proteases do tipo pepsina pertencentes à familia A1 de proteases aspárticas. Pela primeira teoria (geralmente a mais aceite), durante a infeção de uma célula eucariota por um retrovírus, o gene da retropepsina terá sofrido duplicação e fusão dando origem à primeira protease do tipo pepsina. A segunda teoria é defende que após a infeção de uma célula eucariota por um vírus, metade do gene de uma protease do tipo pepsina terá sido acidentalmente capturada pelo vírus, evoluindo e criando a primeira enzima do tipo retropepsina. No entanto, ambas as teorias negligenciam a possibilidade de que estas enzimas pudessem existir previamente em procariotas. Recentemente, com a descoberta de sequências putativas de proteases do tipo retropepsina e do tipo pepsina em procariotas, estas teorias têm sido abaladas. Efetivamente, a existência de uma protease ativa do tipo pepsina em procariotas foi recentemente documentada e provada experimentalmente.

Por isto, é necessária a validação experimental no que toca à existencia destas proteases aspárticas do tipo retropepsina em procariotas. Estudos no nosso laboratório identificaram experimentalmente em Rickettsia uma protease aspártica altamente conservada provada como sendo ativa. Esta protease partilha assinaturas e motivos com as proteases do tipo retropepsina. No entanto, esta protease não será provavelmente a única: estudos bioinformáticos mostram a existência de um grande grupo de genes relacionados evolutivamente com a protease do tipo retropepsina encontrada em Rickettsia. Estas sequências foram encontradas em várias espécies distintas de procariotas, mas todas elas mostram características em comum com as proteases do tipo retropepsina, mantendo, no entanto diferenças significativas entre elas.

O patogénio intracelular *Legionella pneumophila* é uma das espécies bacterianas que foram encontradas contendo estas sequências do tipo retropepsina no seu genoma. Neste estudo, o domínio solúvel putativo de uma sequência de uma protease do tipo retroviral do genoma de Legionella pneumophila foi clonada e expressa em *E. coli*. O domínio solúvel desta proteína, chamado LegRPsd, foi caracterizado em termos de atividade enzimática a estados de oligomerização.

Foi demonstrado que a proteína apresenta atividade proteolítica, e os resultados apontam para que estanova enzima pertença efectivamente à família das proteases aspárticas. A LegRPsd tem atividade proteolítica contra um substrato típico de proteases aspárticas a um pH ótimo de 4.0, atividade esta que é inibida em cerca de 50% na presença de pepstatina A. Foi também provado que a proteína sofre auto-processamento, mudando o seu perfil de inibição no sentido de perda de inibição por pepstatina, acompanhada de um efeito inibitório por parte de inibidores da protease retroviral do HIV-1. A proteína também revelou outras similaridades com proteases retrovirais, como por exemplo a formação de um homodímero lábil, que se pensa ser necessário para a formação do centro ativo, e como tal, necessário para a atividade catalítica.

Os resultados obtidos suportam a existência de uma protease de *L. pneumophila* evolutivamente relacionada com as retropepsinas não apenas em termos de homologia sequência-estrutura, mas também pelas suas características bioquímicas. Esta informação é extremamente importante por duas razões: primeiro, em conjunto com a protease aspártica de *Rickettsia*, constituem uma forte evidência de uma nova família de proteases aspárticas conservadas em procariotas relacionadas com as retropepsinas. Com isto é possível uma re-definição das teorias evolutivas no que toca a pepsinas e retropepsinas, com a apresentação de um novo ponto de vista onde estas sequências de procariotas representam uma forma mais ancestral das retropepsinas; em segundo lugar, a descoberta de uma protease aspártica em *L. pneumophila* tal como em outros organismos patogénicos levanta a possibilidade de encontrar novas vias de patogenecidade e consequentemente a descoberta de novos alvos terapêuticos para estas infeções resistentes, como alternativa ao tratamento por antibióticos convencionais.

1. Introduction

1.1. Overview on Proteolytic Enzymes

Proteases, also termed peptidases or proteinases, are enzymes that catalyse the hydrolysis of peptide bonds. One important fact is that these enzymes are responsible for proteolysis of large proteins to amino acids and small peptides, which is fundamental for nutrition and protein recycling, but they are also extremely important in protein post-translational processing. Post-translational processing by proteases consists in cleavage of specific peptide bonds of a translated protein in order to activate or inactivate that protein, usually an enzyme or a peptide hormone. Proteases are described as existing in, at least, seven catalytic types, depending on the nature of the nucleophile responsible for the catalytic reaction. Therefore we have aspartic, cysteine, glutamic, asparagine, serine and threonine and metallo proteases. In metallo proteases the nucleophile is not an amino acid but a metal ion coordinated in the active site. Proteases can be grouped into families if they can be shown to be related by sequence comparison. These families can be grouped into clans if similarity is found by comparing structures (between some proteins of the same clan the sequences are so distantly related that no relations can be verified by sequence comparison). There are over 40 protease Clans and over 250 protease Families (Rawlings & Barrett 1993; Rawlings et al. 2010).

1.2. Aspartic Proteases

Aspartic proteases are proteolytic enzymes that use two aspartic acid residues responsible for the hydrolysis of the peptide chain in the active centre. In general, aspartic proteases have a water molecule supported by the Asp residues that is used as the nucleophile in the catalysis. According to the MEROPS database (Rawlings et al. 2010), aspartic proteases are distributed among at least five clans and 16 families. The biggest clan among aspartic proteases is the AA clan, consisting of endopeptidases with close related structural features (which will be exposed later in the text). It is possible to recognize a conserved sequence motif containing the active site Asp residues in all the families of the clan – hydrophobic – hydrophobic – **Asp** - Ser/Thr-Gly - Ala/Ser/Thr. This clan compromise the two biggest known families of aspartic proteases, family A1 (pepsin-like) and family A2 (retropepsin-like) with 181 and 52 identified proteins respectively (Rawlings et al. 2010).

1.3. Retropepsin-like Proteases

The main focus of this report will be on a specific group of aspartic proteases, the retropepsin-like proteases (RPs). These are part of family A2 (belonging to clan AA), specifically sub-family A2A, considering that other sub-families consist in transposon and retrotransposon proteases. Proteins of this family are also termed retroviral-like proteases. Retropepsins (RPs) were discovered in the late 1980's due to their essential function in processing proteins of the human immunodeficiency virus (HIV), responsible for AIDS progression after human infection. Indeed, during the 1990's, retroviral proteases were the most widely studied proteins by crystallographic methods and several complex structures of retroviral proteins from HIV-1, HIV-2 and SIV were studied and published (Wlodawer & Gustchina 2000). The ultimate goal was to find a specific inhibitor for these proteases which could be applied clinically in AIDS treatment, which would later become a reality, as it would be explained further.

1.3.1. The Gag-Pol Gene of a Retrovirus

Retroviral proteases are encoded as a part of the *pol* gene of retrovirus. The sequence encoding the protease is located between the *gag* gene, which encodes structural proteases for the virus, and other enzymes in the *pol* gene, such as reverse transcriptase and integrase. The RNA of these viruses is replicated through a DNA intermediary which is synthesised by the virus-encoded reverse transcriptase. This enzyme has no proof-reading and so errors in the replication are frequent, for example, in HIV-1 at least one nucleotide substitution occurs on average on each round of replication. This is why there is no "wild-type" HIV-1 protease, but a large number of sequences with a wide variation of mutations (B. Dunn et al. 2002).

Translation of the *gag-pol* mRNA produces, in most cases, a 55 kDa Gag protein. But when a translational frameshift occurs upstream of the protease gene, the stop codon after the gag sequence is no longer in frame, so a Gag-Pol fusion polyprotein is translated. Then, the protease (PR) cleaves itself by cutting peptide bonds at either ends of its sequence. The protease also cleaves additional bonds in the remaining Gag-Pol polyprotein to produce the reverse transcriptase and integrase enzymes, both essential for the virus replication. Cleavage of the fusion protein occurs at nine different sites with different affinity, causing the cleavage to occur in an ordered way, leading to virion maturation in an organized way (Freed 2001; M. M. Goodenow et al. 2002).

1.3.2. Structural Features of a Retropepsin-like Protease

These enzymes share many features with family A1 aspartic proteases, such as sensitivity to Pepstatin (at least partially for some enzymes) and inactivation by mutation of the catalytic Asp residues. But, unlike pepsin homologues, which are composed of more than 300 amino acids in a single monomer with two topologically similar domains, retropepsins are much smaller. These are active as a symmetric homodimer with a single active site originated by similar residues from both identical monomers (Wlodawer & Gustchina 2000)

Several structures from crystallography and NMR are currently available for retroviral-like proteases. Structures of more than 500 complexes of RPs from HIV-1, HIV-2 and simian immunodefiency virus (SIV) are available online, along with structures of retroviral proteases from equine infectious anaemia virus (EIAV), feline immunodeficiency virus (FIV), human T-cell leukemia virus and rous sarcoma virus (RSV) (Wlodawer & Gustchina 2000).

The secondary structure of RPs follows the same structural template as many other non-viral aspartic proteases. According to this template, each monomer is formed by a duplication of four structural elements: a hairpin (A_1), a wide loop (B_1, containing the catalytic Asp residue), an α-helix (C_1) and a second hairpin (D_1), repeating in $A_{2,}$ $B_{2,}$ C_2, and D_2 (Fig. 1). However, in all RPs for which the structure is known, the D_2 hairpin is replaced by a β-strand. The α-helix C_1 only exists in EIAV RP and appears as a single helical turn in RSV and FIV RPs. In HIV-1, HIV-2 and SIV, this α-helix C_1 is replaced by a loop. The length of the loop, as well as the small sequences connecting these structural elements varies according to the different RPs. The duplication of these elements (termed pseudodyad) may not arise from an evolutionary significant event, as it could be seen only as a need to support the folding of the protein. There are some important structural features conserved in all RPs for which structure is known. The flexible β loop D_1 (known as "flap" in some non-viral aspartic proteases) is functionally very important because it forms numerous interactions with ligands, caused by a change of orientation upon binding to them (these "flaps" may be considered to be in an "open" conformation when no ligand is present, moving downwards when the ligand is bound). As the protein is formed by symmetric dimers, two of these "flaps" exist and form interactions with the substrate. Also, the N- and C- termini of each polypeptide chain form together a four-stranded β-sheet interface (Wlodawer & Gustchina 2000; B. Dunn et al. 2002)

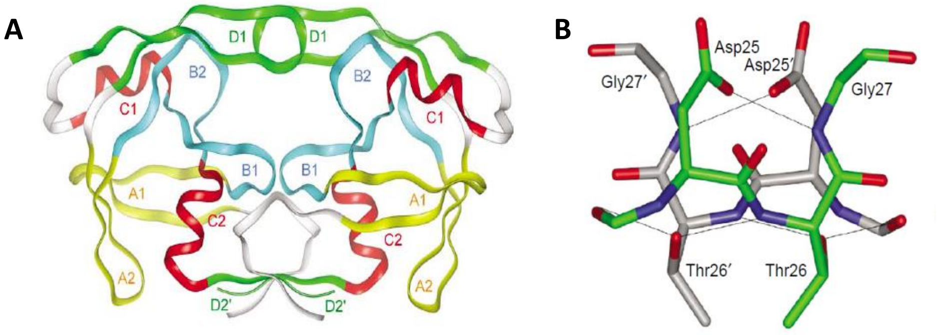


Figure 1. A) Structural template for retropepsins (RPs). In the symmetrical RP dimer, loops A_1 and A_2 are shown in yellow in each monomer, loops B1 and B2 containing the catalytic Asp are shown in blue. In red there are represented the helical segments C1 and C2. Finally, loop D1 in the retroviral monomers provides a double flap structure, whereas the 'half loops' D2' provide the four strands that form a β sheet at the bottom of the dimer. B) Representation of the 'fireman's grip', a stereotypical rigid network structure involving the Asp-Thr-Gly signature sequence in the RPs. The catalytic aspartic acid residue (Asp) is hydrogen-bonded to the backbone NH group of the glycine (Gly) of the same chain. In addition, the OH groups of threonine (Thr) are hydrogen-bonded to the backbone NH group of the threonine and to the carbonyl oxygen of the residue before the catalytic aspartic acid (adapted from Dunn, et al., 2002).

The active site of RPs contains a pair of aspartate residues (Asp25 and Asp25') which are fundamental for catalytic activity. These Asp are bridged by a water molecule, located within hydrogen-bond distance of the Oxygen atoms of their carboxyl groups. This water is probably the water molecule involved in the catalytic mechanism of peptide chains hydrolysis. The active site is therefore composed of three highly conserved amino acids (repeated in both the symmetric chains), these being Asp25, Thr26 (replaced by Ser38 in RSV) and Gly27. These are located in loop B_1 and the structure is stabilized by a network of hydrogen bonds. This network is quite rigid as a result of a series of interactions called the "fireman's grip" (Fig. 1) in which the main-chain NH of Gly27 of each chain accepts an hydrogen bond from one side-chain carboxyl oxygen of the catalytic Asp25 of the same chain. Also the Oγ of each Thr26 accepts an hydrogen bond from the main-chain NH group of the Thr26 of the opposite loop and this same Thr26 also donates an hydrogen bond to the carbonyl group of residue 24 of the opposite chain (Wlodawer & Gustchina 2000; B. Dunn et al. 2002)

1.3.3. Activity, Substrates and Inhibitors of Retropepsin-like Proteases

Substrate specificity studies of RPs have been carried out mostly by analysis of the cleavage sites in the Gag-Pol polyprotein, as this is the "natural" substrate for retroviral proteases. Although, based on the sites of processing it is not possible to reach a clear consensus sequence for the substrate (Tözsér, 2010). Even so, studies were made based on similarities between the amino acid sequences in cleavage sites. Analysis of a broad range of retroviral protease cleavage site sequences using synthetic peptides suggested two types of cleavage sites: the first type indicates that the cleavage of the polypeptide chain occurs between two hydrophobic residues (excluding Pro); the second type indicates the presence of an aromatic residue and a Pro defined as P1 and P1' positions respectively. A preference on P2 and P2' position has also been documented, for example: in HIV-1 RP, Val and Glu were found to be the optimal residues for P2 and P2' positions respectively in the hydrophobic*hydrophobic junction type (Pettit et al. 1991; Griffiths et al. 1992). However, natural occurrence of cleavage sites that are not compatible with any of the two types described is a fact. Later studies indicated that the simple definition of the two junction types is an oversimplification and that presence of the P2 and P2’ preference seems to be a function of the residues outside the P2-P2’ region (Tözsér et al. 1997). Furthermore, the optimum pH values for proteolytic activity of retropepsins usually ranges between pH 4.0 and pH 6.0 for most of the identified retropepsins, as this is the case of proteases from HIV (Ido et al. 1991), MLV (Fehér et al. 2006), SIV (Grant et al. 1991) and EIAV (Rawlings and Salvesen 2013).

Pepstatin inhibits proteases in family A1 and A2, as this is considered a characteristic of these families. Also, due to the existence of a large number of research groups involved in AIDS therapy research, some potent small molecule inhibitors of HIV RP have been developed in the past decades. Used together it is possible to create inhibitor cocktails generally effective in treating HIV infections in AIDS patients (Brower, E. et al., 2008). Some of these inhibitors include Saquinavir, Ritonavir, Indinavir, Nelfinavir, Amprenavir, Lopinavir, Fosamprenavir, Atazanavir, Tipranavir and Darunavir (Flexner, C. 2007).

1.3.4. Orthologues, Paralogues and Evolutionary Theories for Retropepsins

Comparison of structures between family A1 pepsins (a structure with two structurally similar lobes, with each lobe containing one of the active site aspartates) and family A2 retropepsins led to the hypothesis that an ancestral gene duplication of a retropepsin had taken place, followed by gene fusion and originating this new family known as the pepsin-like (X. L. Lin et al. 1992; Tang et al. 1978).

Many retroviral-like proteases found in living organisms (non-virus) are usually from endogenous retrovirus. However, some predicted retroviral-like aspartic proteases that are not embedded within endogenous retroviral elements have been described for both eukaryotes and prokaryotes. The *Saccharomyces cerevisiae* protein Ddi1 (DNA-damage inducible protein 1), which is not a peptidase, has been shown to contain a domain with a retropepsin-like structure, and the protein also has to dimerize to be active. This protein has been found by bioinformatics methods (Krylov & Koonin 2001) to be a eukaryote paralogue of retroviral proteases, and the analysis of this gene identified a group of related mammalian aspartic proteases typified by the mouse neuron specific nuclear receptor interacting protein NIX1. Also, the analysis identified homologues of retroviral-like proteases in all three sequenced genomes of alpha-Proteobacteria — *Rickettsia prowazekii, Mesorhizobium loti* and *Caulobacter crescentus* and in two species of gamma- Proteobacteria. These homologues in alpha-Proteobacteria may suggest the possibility that this protease has originally evolved in the alpha-Proteobacterial lineage, and since this bacterial lineage was the one that gave rise to mitochondria, the gene could have been transferred from the protomitochondrial genome to the ancestral eukaryotic nuclear genome, being lately acquired by the retrovirus from the host (Andersson & Zomorodipour 1998; Krylov & Koonin 2001).

Later, Bernard and colleagues in 2005 characterized for the first time a retroviral-like aspartic protease specifically expressed in human epidermis, which they called SASPase. This protease revealed to be active as a homodimer and contained a putative transmembrane sequence. The soluble domain was expressed in *E. coli* and an activation process was observed at pH 5, at which the enzyme (28kDa) cleaved itself into a 14kDa activated form. This protease was insensitive to Pepstatin but inhibition of auto-activation was verified in the presence of Indinavir, a specific HIV protease inhibitor. This provides new arguments in favour of an ancestral eukaryotic protease being at the origin of retroviral proteases (Bernard et al. 2005; Matsui et al. 2006).

In 2011, the crystal structure of the protease encoded by xenotropic murine leukemia virus–related virus (XMRV) was determined. The dimer interface of this retroviral protease, despite its overall similarity to other retropepsins, more closely resembles the structure of monomeric pepsin-like proteases. Certain structural features, like the existence of only one “flap” favours this resemblance with pepsin-like proteases. Also, analysis of certain parameters of this protease, such as its dimerization mode, makes it closer to the RP domain of Ddi1 rather than other retroviral proteases (Li et al. 2011). This is important by the fact that it can provide a link between the retropepsin-like proteases homologues found in eukaryotes and proteases found in retrovirus, or it can indicate the presence of a new evolutionary branch of retropepsins, also contributing to this evolution enigma.

Later, in 2012, a protein from *Leishmania major,* homologue to the *Saccharomyces cerevisiae* protein Ddi1, was shown to be an active aspartic protease (containing the typical DSG active site sequence) with affinity towards substrates used by retroviral proteases from family A2 at low pH values. Also, the protein was confirmed to be active in the form of an homodimer and suffered inhibition of the proteolytic activity by pepstatin A by about 70% and Nelfinavir by 60% (Perteguer et al. 2012).

Through several of these studies, it has been generally assumed that retroviral proteases represent the ancestral state of pepsins and that the ancestral half-pepsin had to dimerize to be active. However, it is also possible that retroviral proteases were derived from a normal bilobed pepsin when a virus captured half of its host's gene. Solving this question can be possible by searching more clear evidences for the presence of retropepsin-like enzymes in prokaryotes.

1.4. Aspartic Proteases in Prokaryotes

Relatively few APs have been described in bacteria and archaea. According to the MEROPS database (Rawlings et al. 2010), characterized prokaryote APs are distributed between distant families, belonging to different clans due to great differences in sequence motifs. Until recently, characterized APs in prokaryotes were distributed through families A8, A24, A25, A31, A26, A5 and A36. None of these families belongs to clan AA, belonging to different clans (A8 – clan AC; A24 – clan AD; A25 and A31 – clan AE; A26 – clan AF; A5 and A36 – unassigned clan). With these known APs in prokaryotes, no valuable information could be provided to support either of the evolutionary theories involving pepsins and retropepsins. That was until 2009, when Rawlings and Bateman found homologues of pepsin in the completed genomic sequences from seven species of bacteria (Rawlings & Bateman 2009). Later, in 2011, Simões and colleagues expressed in *E. coli* the aspartic protease gene from one of those seven identified bacterial species, *Shewanella amazonensis* (Simões et al. 2011).

Family type proteases and number of prokaryote identified proteins of these families are represented in Table 1.

Table 1. Prokaryote aspartic protease distribution among AP clans and families according to the MEROPS database.

Clan	Family	Sub-Family	No. of Identified Proteases	Type Protease
AA	A1	-	1	pepsin (*Homo sapiens*)
AC	A8	-	1	signal peptidase II (*Escherichia coli)*
AD	A24	A24A	9	type 4 prepilin peptidase 1 (*P. aeruginosa*)
		A24B	2	preflagellin peptidase (*Methanococcus maripaludis*)
AE	A25	-	1	gpr peptidase (*Bacillus megaterium*)
	A31	-	6	HybD peptidase (*Escherichia coli*)
AF	A26	-	6	omptin (*Escherichia coli*)
Unassigned clan	A5	-	1	thermopsin (*Sulfolobus acidocaldarius*)
	A36	-	1	sporulation factor SpoIIGA (*Bacillus subtilis*)

1.4.1. Bacterial Pepsin-Like Proteases

As been told, putative homologues of pepsin have been found by Rawlings and Bateman in the genome of seven species of bacteria: *Colwellia psychrerythraea, Marinomonas sp. MWYL1, Shewanella amazonensis, Shewanella denitrificans, Shewanella loihica, Shewanella sediminis* and *Sinorhizobium medicae.* These bacteria are all members of the class Gammaproteobacteria, and all except *Marinomonas* are members of the order Alteromonadales (*Marinomonas* is a member of the order Oceanospirillales) (Rawlings & Bateman 2009). These sequences showed the requisite hallmark motifs for pepsin-like aspartic proteases, providing the first strong evidence of pepsin-like aspartic proteases in prokaryotes. This led Simões et al. in 2011 to prove this experimentally. The pepsin-like protease gene from *Shewanella amazonensis* was expressed in *E. coli* and the recombinant protein (termed shewasin A) was purified and characterized. Shewasin A exists as a monomer, exhibits activity at acidic pH against a well-documented AP substrate (cleaves its substrates preferentially between hydrophobic amino acids), it is inhibited by pepstatin and does have the sequence motifs characteristic of family A1 APs. This provides a strong experimental evidence that family A1 aspartic proteases are not only confined to eukaryotes and that the bacterial protein shewasin A much probably belongs to the A1 family, and it is probably expressed as an active protein. The absence in shewasin A of a propeptide or signalling sequences and comparison with its closest eukaryotic homologues (most of these eukaryotic homologues require propeptides for correct folding) can lead to the conclusion that these bacterial pepsin-like APs may represent ancestral versions of eukaryotic A1 proteases, and hardly can be originated from horizontal gene transfers from eukaryotes to bacteria (Simões et al. 2011).

1.4.2. Prokaryote Retropepsin-Like Proteases

The finding of true retroviral-like protease homologues in prokaryotes could provide some useful information to support either of the evolutionary theories involving the A1 and A2 family of aspartic proteases. The knowledge of non-viral retroviral-like protease homologues may provide the necessary clues for evolutionary studies of these proteins. So far only SASPase and the *Leishmania major* aspartic protease are considered as true non-viral retroviral protease homologues. No other aspartic protease known so far has been proved as being a retroviral-like protease, furthermore, no retropepsin homologue was ever been undoubtfuly characterized in prokaryotes. However, some strong evidences make us believe in their existence as active proteases.

SpoIIGA

In 2008, Imamura and colleagues characterized a protease necessary for processing of pro-σ^E to σ^E factor in endospore formation of *Bacillus subtilis*, which was previously termed SpoIIGA, as a novel type of aspartic protease. The study insisted that SpoIIGA shared structural characteristics with the HIV protease in the C-terminus region and that this domain was probably active as a dimer (Imamura et al. 2008). However, previous studies classified this enzyme as a possible serine protease due to the presence of certain sequence motifs, which are hard to refute (Peters & Haldenwang 1991). The enzyme contains five putative transmembrane segments, which are stated by Imamura et al. as probable binding sites for SpoIIR signal sequence, proposing SpoIIGA as a novel signal transducing peptidase. Thus far, SpoIIGA has been accepted in the MEROPS database as a new unassigned family of aspartic proteases (A36). Due to the possibility of this enzyme being a serine protease, it is difficult to consider it as a strong evidence of retroviral proteases in prokaryotes.

PerP peptidase

This protease was identified by Chen and colleagues in 2006 as the responsible for proteolysis of PodJ (a polar factor that recruits proteins required for polar organelle biogenesis to the correct cell pole) to a form with altered activity in *Caulobacter crescentus'* cell cycle. It was termed after being recognized as a periplasmic protease, PerP meaning periplasmic protease of PodJ. The enzyme was identified as containing a putative signal sequence or membrane anchor at its N-terminus and a conserved aspartic protease motif in its periplasmic domain (Chen et al. 2006). No more information about the protease characterization was published, and these aspects show no relevance in Chen's article. However, on a simple overview of the sequence we can notice the existence of a single catalytic aspartate residue in a conserved Leu-Val-Asp-Thr-Gly-Ala sequence, proving evidence of this protease belonging to the AA clan of aspartic proteases. This fact has been documented on the MEROPS database. The protease is classified as a novel family of aspartic protease – family A32 of the AA clan; however, its sequence and the fact that only one catalytic Asp is present in the sequence, probably requiring the formation of a homodimer for catalytic activity, resembles characteristics of retroviral-like proteases. Indeed, a detailed characterization of this enzyme or structure determination could provide the evidences needed for its classification as a retroviral-like protease homologue.

Rickettsia's Aspartic Protease

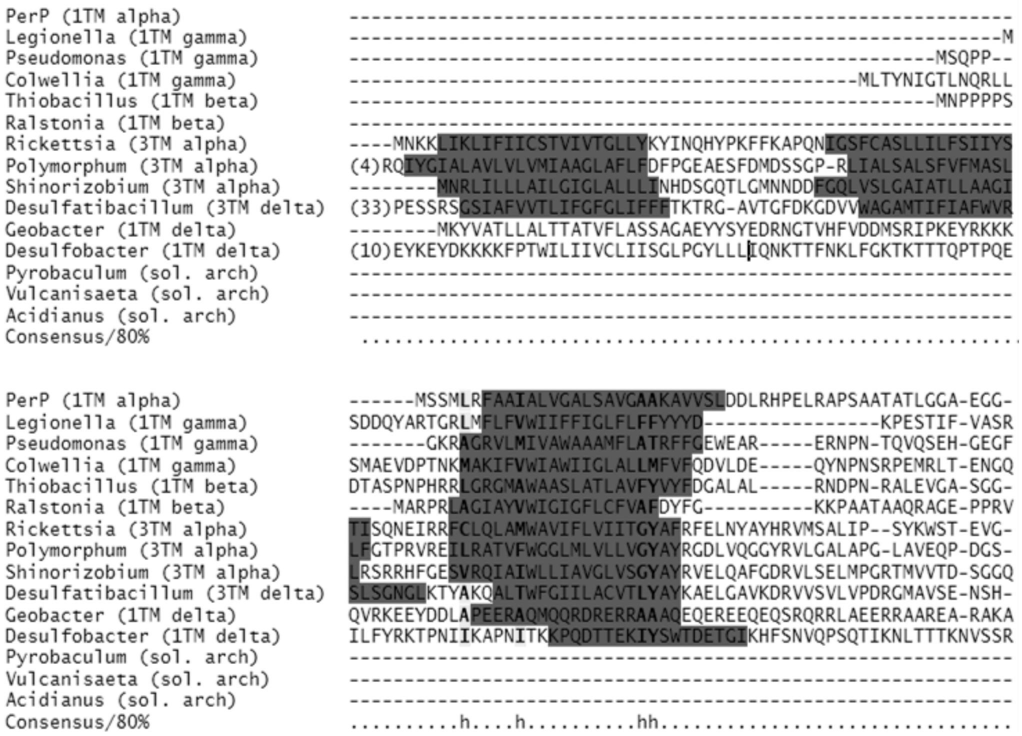
In our lab, a novel aspartic protease highly conserved in Rickettsia was found to share many characteristics with retroviral-like proteases, such as the existence of a single catalytic Asp residue in the sequence, present in a DTG motif, requiring the formation of a homodimer for catalytic activity. The sequence, composed of 231 aa, contains three putative transmembrane helixes. This aspartic protease (both the full-length sequence and the putative soluble domain) was already expressed in *E.coli*, purified and characterized in our laboratory. The soluble domain was verified to encode an active enzyme which undergoes multi-step processing, this processing being affected by pepstatin and EDTA. These characteristics provide strong evidence that this new AP from *Rickettsia* is indeed an active enzyme. Furthermore, the proteolytic activity of the enzyme has shown to be inhibited by retroviral protease inhibitors like Indinavir. This represents the first report on a novel retroviral-like AP from a gram-negative obligate intracellular species, and probably the first confirmation of active retroviral-like proteases present in prokaryotes (Cruz R., Simões I. Unpublished data).

1.4.3. Bioinformatics Search of Retropepsin-Like Proteases in Prokaryotes

With the finding of *Rickettsia's* aspartic protease, it was very tempting to search for other homologues of this protease in prokaryotes. Using the protein BLAST search (Altschul, et al., 1990) in non-redundant protein sequences (nr) database with the Blastp (protein-protein BLAST) algorithm, we obtained many interesting results about retroviral-like proteins distribution in prokaryotes. A BLAST search of *Ricketsia*'s AP sequence in bacteria provided many close-related putative protein sequences in various *Rickettsia* species, as it was confirmed by previous studies in our lab (unpublished data). We also found sequences of putative retroviral-like APs in other alpha-proteobacteria such as *Polymorphum gilvum, Sinorhizobium fredii* and *Mesorhizobium amorphae* with considerable values of similarity. All these sequences showed a motif with three putative transmembrane helixes (3TMH) on the N-terminus. To conclude about transmembrane helixes, TMpred (Hofmann and Stoffel, 1993) and TMHMM v2.0 (Krogh, et al., 2001) were used. Close to this sequence was also a sequence from a putative retroviral-like AP from *Desulfatibacillum alkenivorans*, this being noted for not being an alpha-proteobacteria, but a delta-proteobacteria. *Desulfatibacillum alkenivorans* sequence also shared the three putative transmembrane helixes motif on the N-terminus. A BLAST search of this last sequence in deltaproteobacteria revealed only another sequence close related, this time from *Plesiocystis pacifica,* also sharing the 3TMH motif and identified as a putative retroviral-like AP. However, this last one was composed of 397 aa, opposed to previous found sequences with a usual length of 200-250 aa.

When we performed a BLAST search of *Rickettsia's* AP in gamma-proteobacteria, we couldn't find any related sequence sharing the 3TMH motif. Instead, it retrieved many related putative retroviral-like AP sequences sharing a single transmembrane helix (1TMH) in the N-terminus. These sequences were found in many important gamma-proteobacteria such as *Pseudomonas aeruginosa, Pseudomonas putida* and *Legionella pneumophilla*, which are pathogenic bacteria. *Legionella pneumophilla* is an interesting bacteria to study as it is an intracellular parasite of eukaryote unicellular organisms and macrophages. Besides the gamma-proteobacteria matches, a BLAST search of these 1TMH sequences provided sequences of similar proteins with this same single transmembranar helix domain in beta-proteobacteria (such as *Thiobacillus denitrificans* and *Ralstonia sp. 5_7_47FAA*) and delta-proteobacteria (such as *Syntrophus aciditrophicus* and *Desulfobacter postgatei*). It should be noted that all the found sequences for both the 3TMH and the 1TMH belong to gram-negative Proteobacteria.

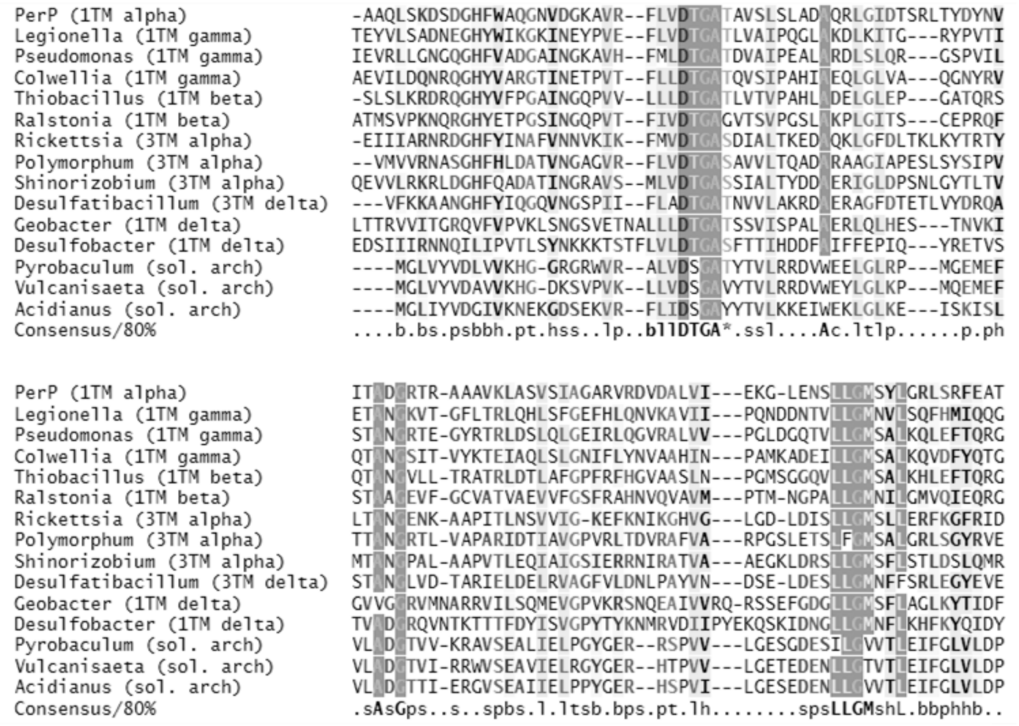
Neither the 3TMH nor the 1TMH sequence motif was found when a BLAST search was performed in archaea. The close-related sequences found share the retroviral-like motifs but do not contain putative transmembrane helixes. Instead, these archaea sequences were much shorter than the previous (about 100-120 aa) and are probably soluble enzymes. These results are very interesting as these sequences were mostly found in thermophilic organisms such as *Vulcanisaeta distributa, Acidianus hospitalis* and *Sulfolobus islandicus*. BLAST searches of these soluble retroviral-like APs in bacteria found some related sequences in a variety of different bacteria, such as *Desulfotomaculum kuznetsovii* (a gram-positive sulfate-reducing bacteria) and *Chlorobium chlorochromatii* (a green-sulfur bacteria). These sequences were aligned using the ClustalW algorithm included in the MEGA5 software package (Figure 2 and 3).



```
PerP (1TM alpha)              ------------------------------------------------------------
Legionella (1TM gamma)        -----------------------------------------------------------M
Pseudomonas (1TM gamma)       -----------------------------------------------------MSQPP--
Colwellia (1TM gamma)         ----------------------------------------------MLTYNIGTLNQRLL
Thiobacillus (1TM beta)       -----------------------------------------------------MNPPPPS
Ralstonia (1TM beta)          ------------------------------------------------------------
Rickettsia (3TM alpha)        ----MNKKLIKLIFIICSTVIVTGLLYKYINQHYPKFFKAPQNIGSFCASLLILFSIIYS
Polymorphum (3TM alpha)       (4)RQIYGIALAVLVLVMIAAGLAFLFDFPGEAESFDMDSSGP-RLIALSALSFVFMASL
Shinorizobium (3TM alpha)     --------MNRLILLLAILGIGLALLLINHDSGQTLGMNNDDFGQLVSLGAIATLLAAGI
Desulfatibacillum (3TM delta) (33)PESSRSGSIAFVVTLIFGFGLIFFFTKTRG-AVTGFDKGDVVWAGAMTIFIAFWVR
Geobacter (1TM delta)         --------MKYVATLLALTTATVFLASSAGAEYYSYEDRNGTVHFVDDMSRIPKEYRKKK
Desulfobacter (1TM delta)     (10)EYKEYDKKKKFPTWILIIVCLIISGLPGYLLLIQNKTTFNKLFGKTKTTTQPTPQE
Pyrobaculum (sol. arch)       ------------------------------------------------------------
Vulcanisaeta (sol. arch)      ------------------------------------------------------------
Acidianus (sol. arch)         ------------------------------------------------------------
Consensus/80%                  ...........................................................

PerP (1TM alpha)              ------MSSMLRFAAIALVGALSAVGAAKAVVSLDDLRHPELRAPSAATATLGGA-EGG-
Legionella (1TM gamma)        SDDQYARTGRLMFLFVWIIFFIGLFLFFYYYD----------------KPESTIF-VASR
Pseudomonas (1TM gamma)       -------GKRAGRVLMIVAWAAAMFLATRFFGEWEAR-----ERNPN-TQVQSEH-GEGF
Colwellia (1TM gamma)         SMAEVDPTNKMAKIFVWIAWIIGLALLMFVFQDVLDE-----QYNPNSRPEMRLT-ENGQ
Thiobacillus (1TM beta)       DTASPNPHRRLGRGMAWAASLATLAVFYVYFDGALAL-----RNDPN-RALEVGA-SGG-
Ralstonia (1TM beta)          ----MARPRLAGIAYVWIGIGFLCFVAFDYFG----------KKPAATAAQRAGE-PPRV
Rickettsia (3TM alpha)        TISQNEIRRFCLQLAMWAVIFLVIITGYAFRFELNYAYHRVMSALIP--SYKWST-EVG-
Polymorphum (3TM alpha)       LFGTPRVREILRATVFWGGLMLVLLVGYAYRGDLVQGGYRVLGALAPG-LAVEQP-DGS-
Shinorizobium (3TM alpha)     LRSRRHFGESVRQIAIWLLIAVGLVSGYAYRVELQAFGDRVLSELMPGRTMVVTD-SGGQ
Desulfatibacillum (3TM delta) SLSGNGLKTYAKQALTWFGIILACVTLYAYKAELGAVKDRVVSVLVPDRGMAVSE-NSH-
Geobacter (1TM delta)         QVRKEEYDDLAPEERAQMQQRDRERRAAQEQEREEQEQSRQRRLAEERRAAREA-RAKA
Desulfobacter (1TM delta)     ILFYRKTPNIIKAPNITKKPQDTTEKIYSWTDETGIKHFSNVQPSQTIKNLTTTKNVSSR
Pyrobaculum (sol. arch)       ------------------------------------------------------------
Vulcanisaeta (sol. arch)      ------------------------------------------------------------
Acidianus (sol. arch)         ------------------------------------------------------------
Consensus/80%                 ..........h....h..........hh................................
```

Figure 2. Alignments of putative transmembrane sequences in putative retroviral-like aspartic proteases present in bacteria and archaea. The sequences were aligned using the ClustalW algorithm included in the MEGA5 software package. Putative transmembranar sequences are highlighted. (Proteases are named after the bacteria's genus. "TM" = transmembranar helix; "sol." = soluble; "alpha" is an abbreviation for alphaproteobacteria, as much for delta, beta and gamma; "arch" = archaea).



```
PerP (1TM alpha)              -AAQLSKDSDGHFWAQGNVDGKAVR--FLVDTGATAVSLSLADAQRLGIDTSRLTYDYNV
Legionella (1TM gamma)        TEYVLSADNEGHYWIKGKINEYPVE--FLVDTGATLVAIPQGLAKDLKITG---RYPVTI
Pseudomonas (1TM gamma)       IEVRLLGNGQGHFVADGAINGKAVH--FMLDTGATDVAIPEALARDLSLQR---GSPVIL
Colwellia (1TM gamma)         AEVILDQNRQGHYVARGTINETPVT--FLLDTGATQVSIPAHIAEQLGLVA---QGNYRV
Thiobacillus (1TM beta)       -SLSLKRDRQGHYVFPGAINGQPVV--LLLDTGATLVTVPAHLADELGLEP---GATQRS
Ralstonia (1TM beta)          ATMSVPKNQRGHYETPGSINGQPVT--FIVDTGAGVTSVPGSLAKPLGITS---CEPRQF
Rickettsia (3TM alpha)        -EIIIARNRDGHFYINAFVNNVKIK--FMVDTGASDIALTKEDAQKLGFDLTKLKYTRTY
Polymorphum (3TM alpha)       --VMVVRNASGHFHLDATVNGAGVR--FLVDTGASAVVLTQADARAAGIAPESLSYSIPV
Shinorizobium (3TM alpha)     QEVVLRKRLDGHFQADATINGRAVS--MLVDTGASSIALTYDDAERIGLDPSNLGYTLTV
Desulfatibacillum (3TM delta) ---VFKKAANGHFYIQGQVNGSPII--FLADTGATNVVLAKRDAERAGFDTETLVYDRQA
Geobacter (1TM delta)         LTTRVVITGRQVFVPVKLSNGSVETNALLLDTGATSSVISPALAERLQLHES---TNVKI
Desulfobacter (1TM delta)     EDSIIIRNNQILIPVTLSYNKKKTSTFLVLDTGASFTTIHDDFAIFFEPIQ---YRETVS
Pyrobaculum (sol. arch)       ----MGLVYVDLVVKHG-GRGRWVR--ALVDSGATYTVLRRDVWEELGLRP---MGEMEF
Vulcanisaeta (sol. arch)      ----MGLVYVDAVVKHG-DKSVPVK--LLVDSGAVYTVLRRDVWEYLGLKP---MQEMEF
Acidianus (sol. arch)         ----MGLIYVDGIVKNEKGDSEKVR--FLIDSGAYYTVLKKEIWEKLGLKE---ISKISL
Consensus/80%                 ....b.bs.psbbh.pt.hss..lp..bllDTGA*.ssl....Ac.ltlp......p.ph

PerP (1TM alpha)              ITADGRTR-AAAVKLASVSIAGARVRDVDALVI---EKG-LENSLLGMSYLGRLSRFEAT
Legionella (1TM gamma)        ETANGKVT-GFLTRLQHLSFGEFHLQNVKAVII---PQNDDNTVLLGMNVLSQFHMIQQG
Pseudomonas (1TM gamma)       STANGRTE-GYRTRLDSLQLGEIRLQGVRALVV---PGLDGQTVLLGMSALKQLEFTQRG
Colwellia (1TM gamma)         QTANGSIT-VYKTEIAQLSLGNIFLYNVAAHIN---PAMKADEILLGMSALKQVDFYQTG
Thiobacillus (1TM beta)       QTANGVLL-TRATRLDTLAFGPFRFHGVAASLN---PGMSGGQVLLGMSALKHLEFTQRG
Ralstonia (1TM beta)          STAAGEVF-GCVATVAEVVFGSFRAHNVQVAVM---PTM-NGPALLGMNILGMVQIEQRG
Rickettsia (3TM alpha)        LTANGENK-AAPITLNSVVIG-KEFKNIKGHVG---LGD-LDISLLGMSLLERFKGFRID
Polymorphum (3TM alpha)       TTANGRTL-VAPARIDTIAVGPVRLTDVRAFVA---RPGSLETSLFGMSALGRLSGYRVE
Shinorizobium (3TM alpha)     MTANGPAL-AAPVTLEQIAIGSIERRNIRATVA---AEGKLDRSLLGMSFLSTLDSLQMR
Desulfatibacillum (3TM delta) STANGLVD-TARIELDELRVAGFVLDNLPAYVN---DSE-LDESLLGMNFFSRLEGYEVE
Geobacter (1TM delta)         GVVGGRVMNARRVILSQMEVGPVKRSNQEAIVVRQ-RSSEFGDGLLGMSFLAGLKYTIDF
Desulfobacter (1TM delta)     TVADGRQVNTKTTTFDYISVGPYTYKNMRVDIIPYEKQSKIDNGLLGMNFLKHFKYQIDY
Pyrobaculum (sol. arch)       VLADGTVV-KRAVSEALIELPGYGER--RSPVV---LGESGDESILGVVTLEIFGLVLDP
Vulcanisaeta (sol. arch)      VLADGTVI-RRWVSEAVIELRGYGER--HTPVV---LGETEDENLLGTVTLEIFGLVLDP
Acidianus (sol. arch)         VLADGTTI-ERGVSEAIIELPPYGER--HSPVI---LGESEDENLLGVVTLEIFGLVLDP
Consensus/80%                 .sAsGps..s..spbs.l.ltsb.bps.pt.lh........spsLLGMshL.bbphhb..
```

Figure 3. Alignments of predicted soluble domain sequences of putative retroviral-like aspartic proteases present in bacteria and archaea. The sequences were aligned using the ClustalW algorithm included in the MEGA5 software package. Conserved motifs and aminoacids are highlighted. (Proteases are named after the bacteria's genus. "TM" = transmembranar helix; "sol." = soluble; "alpha" is an abbreviation for alphaproteobacteria, as much for delta, beta and gamma; "arch" = archaea).

The sequences show a clear difference between the 1TMH and the 3TMH sequences (highlighted in Figure 2), being the last transmembranar helix conserved and present in all retroviral-like proteases in bacteria (not in archaea). It is also very clear the presence of the conserved hydrophobic-hydrophobic-Asp-Thr/Ser-Gly-Ala motif as well as the Ala-small-Gly and the Leu-Leu-Gly-Met motifs highly conserved downstream of the active site.

Next, these aligned sequences were analysed through a neighbour-joining phylogenetic tree also included in the MEGA5 package (Figure 4). The distance between the archaeal soluble form of the protein and the bacterial sequences is evident. Also, it is possible to distinguish four main clusters in the tree, possibly showing four main categories of these bacterial retroviral-like APs. The first, cluster A, compromise the sequences without transmembranar sequences from archaeal genomes. Next, the bacterial sequences divide into 3 clusters. The first, cluster B, is composed only of sequences from deltaproteobacteria with 1TMH. The next branch divides into 2 other clusters, C and D. Cluster C are mainly alphaproteobacteria with 3TMH, although the PerP sequence is included, which is a 1TMH sequence from an alphaproteobacteria. Also included is the sequence from *Desulfatibacillum alkenivorans*, which is a 3TMH sequence from a deltaproteobacteria. The last cluster, cluster D, are only gammaproteobacteria and betaproteobacteria with 1TMH.

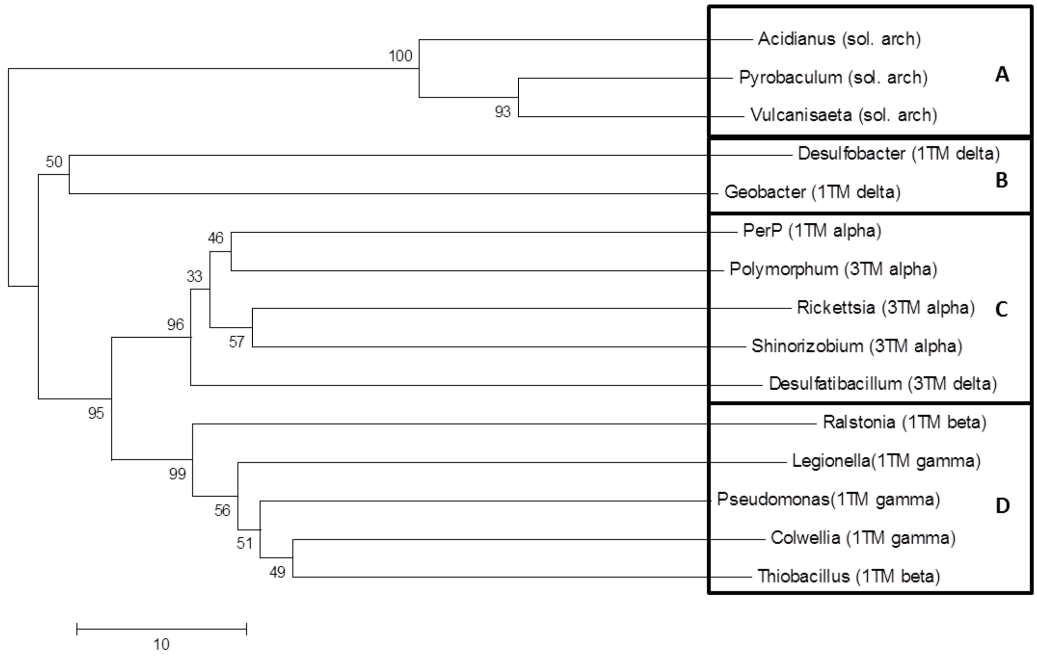


Figure 4. Neighbour-Joining phylogenetic tree of putative retroviral-like aspartic proteases present in bacteria and archaea. Bootstrap values are shown. It is possible no notice the presence of four main clusters: A) Only the soluble form present in archaea; B) Sequences from Deltaproteobacteria composed with 1TMH; C) Mainly sequences from alphaproteobacteria with 1TMH, an alphaproteobacteria with 1TM and a deltaproteobacteria with 3TMH are also present. D) Sequences from gammaproteobacteria and betaproteobacteria with 1TMH present.

This tree is congruent with the phylogenetic tree of Proteobacteria classes (Figure 5), showing that probably the retroviral-like AP genes have evolved mainly through clonal evolution and that has been conserved through the evolution of the bacterial species. This is a very important aspect, because it shows us that these proteases are probably very important for the bacterial cells, as they are conserved even in genomes with high selection.

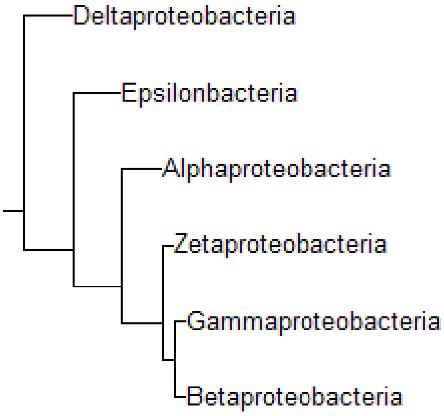


Figure 5. The phylogeny of Proteobacteria. The different classes of Proteobacteria are shown in a phylogenetic tree built through analyses of the 16S rRNA sequence. (Krieg, et al., 2005)

1.5. The *Legionella pneumophila's* Putative Retropepsin-Like Protease

Although it was seen that putative retroviral-like protease sequences in prokaryotes are a strong argument over a new consideration of the before mentioned evolutionary theories, additional experimental evidences are needed to prove that these sequences are coding for active enzymes and to elucidate what are the similarities and differences with other known homologues. This way, we have decided to study a particular retroviral-like AP from bacterial origin.

The chosen sequence was the homologue from *Legionella pneumophila* due to several interesting features of these bacteria.

1.5.1. *Legionella pneumophila* and the Legionnaire's Disease

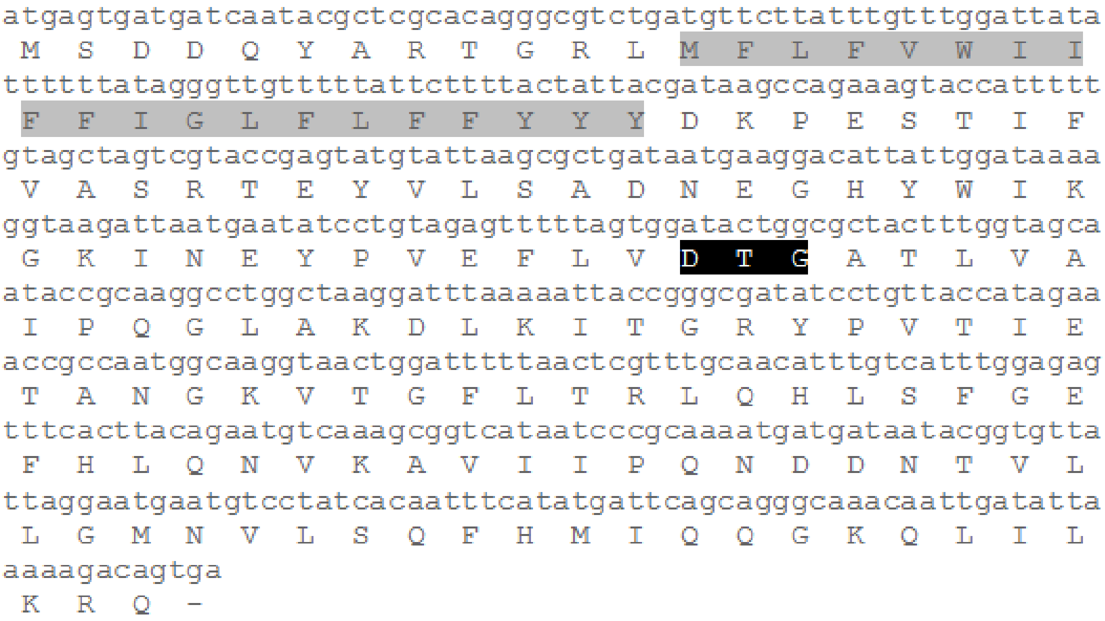
Legionella pneumophila is an intracellular gamaproteobacteria that naturally inhabits aqueous environments like ponds and lakes, living inside protozoa. This bacterium has the ability to survive phagocytosis by organisms like amoebas and other phagocytic protozoa. It happens that when *L. pneumophila* is phagocytized by protozoa, it activates its unique Dot/Icm type IV protein secretion system, which is responsible for the secretion of specific effector proteins through the phagosome into the host cell's cytosol (J. Costa et al. 2010)

These effector proteins quickly interfere with endosome fusion and vesicle trafficking, therefore avoiding the digestion of *L. pneumophila* cells. Furthermore, these effector proteins transiently fuse the phagosome with mitochondria and redirect vesicular traffic in order to intercept ER vesicles and send them to the phagosome containing the *L. pneumophila* cells to form the *Legionella*-containing-Vacuole (LCV). In about 4 or 5 hours the LCV completely changes its membrane into a rough-ER-like membrane, recognized by the host cell as a part of itself and supplying nutrients to the *L. pneumophila* cells found inside the LCV. The *L. pneumophila* cells then, protected by the LCV, radically change their phenotype into a replicative state, focused on replication inside the host cell. In about 10 hours, when cell density becomes high and nutrient supply is no longer enough, the *L. pneumophila* changes again its phenotype to the virulent state, which causes the lysis of the host cell and promotes the formation of flagella, turning the *L. pneumophila* cells into highly motile cells, free to be engulfed by another host and repeating the cycle (Newton et al. 2010; Rolando M, Buchrieser C. 2012).

Although this bacterium inhabits aqueous environments and generally only lives inside amoebas, in the recent decades legionella have become in increased contact with humans due to the creation of several artificial aquatic environments suited for Legionella's living. Legionella are capable of living in water circuits if these are not properly treated (either inside protozoa, or in the form of biofilms). These environments include large hotel boilers, showers, cooling towers, humidifiers and air conditioners among others. When water particles containing *Legionella pneumophila* cells are inhaled, the bacterium can infect alveolar macrophages the same way they do with protozoa, evading phagocytosis and initializing the infectious process. This leads to a clinical condition called the Legionnaire's Disease, which is a severe pneumonia with mortality rates reaching 30% of infected patients (Nguyen, T. M. et al. 2006).

Legionella pneumophila is a bacteria with a complex cell cycle and infectious process and few molecular pathways are well understood in these bacteria. Several effectors have been identified as needed for infection, however, much information regarding their function, regulation, synthesis and processing is not yet understood (Newton et al. 2010). Furthermore, *Legionella* are highly adaptive bacteria as they inhabit several different environmental hosts, in fact, as *L. pneumophilla* is not transmissive from human to human, the theory regarding its ability to infect humans is that alveolar macrophages are similar to evironmental protozoa hosts. It is hypothesized that host cycling in the environment maintains *L. pneumophila* as a generalist, probably due to purifying selection against mutations that diminish fitness in any of several naturally encountered protozoan hosts (Ensminger et al. 2012)

1.5.2. *Legionella pneumophila*'s Retropepsin-Like Protease Homologue



```
atgagtgatgatcaatacgctcgcacagggcgtctgatgttcttatttgtttggattata
 M  S  D  D  Q  Y  A  R  T  G  R  L  M  F  L  F  V  W  I  I
tttttatagggttgttttattcttttactattacgataagccagaaagtaccattttt
 F  F  I  G  L  F  L  F  F  Y  Y  Y  D  K  P  E  S  T  I  F
gtagctagtcgtaccgagtatgtattaagcgctgataatgaaggacattattggataaaa
 V  A  S  R  T  E  Y  V  L  S  A  D  N  E  G  H  Y  W  I  K
ggtaagattaatgaatatcctgtagagttttagtggatactggcgctactttggtagca
 G  K  I  N  E  Y  P  V  E  F  L  V  D  T  G  A  T  L  V  A
ataccgcaaggcctggctaaggatttaaaaattaccgggcgatatcctgttaccatagaa
 I  P  Q  G  L  A  K  D  L  K  I  T  G  R  Y  P  V  T  I  E
accgccaatggcaaggtaactggattttttaactcgtttgcaacatttgtcatttggagag
 T  A  N  G  K  V  T  G  F  L  T  R  L  Q  H  L  S  F  G  E
tttcacttacagaatgtcaaagcggtcataatcccgcaaaatgatgataatacggtgtta
 F  H  L  Q  N  V  K  A  V  I  I  P  Q  N  D  D  N  T  V  L
ttaggaatgaatgtcctatcacaatttcatatgattcagcagggcaaacaattgatatta
 L  G  M  N  V  L  S  Q  F  H  M  I  Q  Q  G  K  Q  L  I  L
aaaagacagtga
 K  R  Q  -
```

Figure 6. Retroviral-like aspartic protease homologue from *Legionella pneumophila* pneumophila (strain Philadelphia 1). Gene sequence (accession number lpg2007) embedded in the translated protein sequence. The putative transmembranar helix, highlighted in gray, spans about 20 aminoacids between Met13 and Tyr32 (TMpred and TMHMM 2.0). The conserved active site motif (Asp-Thr-Gly) is highlighted in black.

The putative retroviral-like aspartic protease gene from *Legionella pneumophila* genome translates into the protein displayed in Figure 6. The protein displays a putative transmembranar helix as the probability plot in figure 7 shows, with a predicted location in the inner membrane of the bacterium (using the PSORTb server algorithm), although the prediction of the direction of this transmembrane helix did not retrieve confident results when compared between different algorithms.

It was also possible to confirm the presence of the conserved motifs explored in the previous sections, with special attention to the putative active-site motif (hydrophobic-hydrophobic-D-T-G). The presence of a single DTG motif clearly suggests that the protein should need to dimerize to have proteolytic activity. To confirm this signature, the sequence was submitted to the PROSCAN algorithm server by NPS@ (Combet, C. et al. 2000) and a 100% Identity was obtained for the "Eukaryotic and viral aspartyl proteases active site" signature [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-{GQ}-[LIVM FSTNC]-{EGK}-[LIVMFGTA].

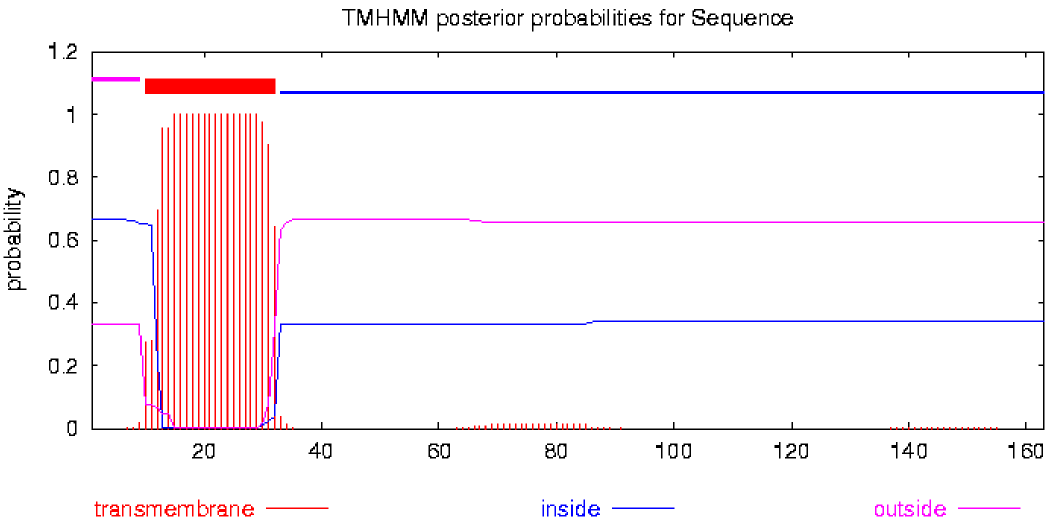


Figure 7. Plot showing the calculated probability of transmembrane helixes, as well as the probability of a sequence being in the outside or inside of the membrane. The plot was calculated by TMHMM 2.0 and was obtained by calculating the total probability that a residue sits in helix, inside, or outside summed over all possible paths through the model.

1.6. Objectives

Considering that a bioinformatics analysis reveals the presence of retropepsin-like sequences in several distinct species of prokaryotes, this study proposes to obtain new experimental evidences that support the existence of retropepsin-like proteins in bacteria. Together with the results from *Rickettsia* retropepsin-like protease, we aim to provide experimental evidences supporting that these sequences code for active enzymes, related with true retropepsins and other retropepsin-like proteases. These experimental evidences will allow us to clarify the position of these proteins in the evolution of pepsin and retropepsins and argument towards the postulated theory that these enzymes probably exist long before retroviruses, and represent the evolutive ancestral form that gave origin to the retropepsins.

For that purpose, from the group of prokaryotic gene sequences related to the *Rickettsia* retropepsin-like protease, we propose to study in detail the gene from *Legionella pneumophila* predicted to code for a retropepsin-like protein with different features than the *Rickettsia* retropepsin-like protease. For this, the recombinant retropepsin-like protease from *Legionella pneumophila* will be produced in *E. coli* in lab scale so it can be possible to perform a biochemical characterization of this enzyme.

A biochemical characterization represents a fundamental step in understanding these new enzymes as they probably represent a new family of proteases with possible unique and exclusive features. Biochemical characteristics such as features of the enzyme's proteolytic activity will probably allow the validation of the enzyme as an aspartic protease. Other characteristics, such as oligomerization states and inhibition by retroviral therapy drugs will allow us to compare it with other known retropepsins and retropepsin-like proteases.

Beyond the presented results, the optimization of the enzyme's production and characterization will allow the fast development of future studies such as structure determination by X-ray crystallography or even functional studies to determine *in vivo* relevance of the enzyme for the bacteria cells.

2. Materials & Methods

2.1. Materials

Most reagents in this study were obtained either from Sigma-Aldrich or Merck.

2.2. Methods

2.2.1. Cloning

The gene sequence of the putative aspartic protease *from L. pneumophila* (accession number: lp2007 from KEGG database) was optimized and chemically synthesized for recombinant expression in *E. coli* by GenScript according to the OptimumGene™ Codon Optimization Analysis algorithm. The considered soluble domain was then amplified by PCR from the delivered synthetic gene DNA. The primers used for the amplification were 5’ **CCATGG**GCGATAAACCGGAAAGCACCATT 3’ (Forward, includes a *Nco*I restriction sequence in the 5’ end, highlighted in bold) and 5’ **CTCGAG**TTGGCGTTTCAGAATCAGTTG 3’ (Reverse, includes a *Xho*I restriction sequence in the 5’ end, highlighted in bold). The amplified sequence (named LegRPsd) (containing the *Nco*I and *Xho*I sequences in the 5’ and 3’ end respectively) was confirmed by 1% agarose gel electrophoresis and cloned in a pGEM®-T Easy Vector (Promega) by the standard protocol. The ligation product was transformed in DH5α *E.* coli cells and positive colonies were selected by blue-white selection and confirmed by DNA sequencing made by Macrogen.

pGEM®-T Easy with correct LegRPsd sequence insert was digested with *Nco*I and *Xho*I restriction enzymes and the insert was purified by agarose gel electrophoresis followed by extraction of the DNA from the agarose gel using NZYGelpure kit (NZYTech). pET28a (Novagen) was also digested with *Nco*I and *Xho*I restriction enzymes. Both digestion products were incubated with DNA Ligase. The resulting vector was transformed in TOP10F’ *E. coli* cells and positive clones were selected by restriction analysis and confirmed by sequencing.

For generation of the active-site mutant, the Quickchange™ Site-Directed Mutagenesis kit was used using the resulting pET28a vector with the LegRPsd sequence as the template DNA and the primers 5’ CCGGTCGAATTTCTGGTGGCTACCGGTGCCACGCTGGTC 3’ (Forward) and 5’ GACCAGCGTGGCACCGGTAGCCACCAGAAATTCGACCGG 3’ (Reverse) (mutation underlined).

To clone the LegRPsd-HIS and LegRPsd-HA sequences in the pRSFDuet™-1 DNA, the soluble domain was amplified by PCR from the delivered optimized synthetic gene. To generate LegRPsd-HIS, the primers used for the PCR were 5’ **CCATGG**GAGATAAACCGGAAAGCACCATT 3’ (Forward, includes a *Nco*I restriction sequence in the 5’ end, highlighted in bold) and 5’ **GCGGCCGC**TTAGTGGTGATGATGGTGATGTTGGCGTTTCAGAATCAGTTGT 3’ (Reverse, includes a 6His coding sequence, underlined, a stop codon and a *Not*I restriction sequence in the 5’ end, highlighted in bold). To generate LegRPsd-HA, primers used for the PCR were 5’ **CATATG**GATAAACCGGAAAGCACCATT 3’ (Forward, includes a *Nde*I restriction sequence in the 5’ end, highlighted in bold) and **CTCGAG**TTATGCATAATCTGGAACATCGTATGGATATTGGCGTTTCAGAATCAGTTGT (Reverse, includes a HA tag coding sequence, underlined, a stop codon and a *Xho*I restriction sequence in the 5’ end, highlighted in bold). The amplified sequences were confirmed by 1% agarose gel electrophoresis and cloned in a pGEM®-T Easy Vector (Promega) by the standard protocol. The ligation product was transformed in DH5α *E. coli* cells and positive colonies were selected by blue-white selection and confirmed by sequencing.

pGEM®-T Easy with correct LegRPsd-HIS sequence insert was digested with *Nco*I and *Not*I restriction enzymes and the insert was purified by agarose gel electrophoresis followed by extraction of the DNA from the agarose gel using NZYGelpure kit (NZYTech). pRSF Duet was also digested with *Nco*I and *Not*I restriction enzymes. Digested insert and vector products were incubated with DNA Ligase. The resulting vector was transformed in TOP10F’ *E. coli* cells and positive clones were selected by restriction analysis and confirmed by sequencing. The process was repeated with LegRPsd-HA and *Nde*I and *Xho*I restriction enzymes to clone LegRPsd-HA in the MCS2 of the pRSF vector containing LegRPsd-HIS in MCS1. Positive clones were confirmed by DNA sequencing.

2.2.2. Expression of the recombinant LegRPsd, D41A mutant and LegRPsd-HIS/LegRPsd-HA

2.2.2.1. Expression Screening

For the expression screening of LegRPsd, the pET28a expression construct was transformed in BL21 star (Invitrogen) and C41 (Lucigen) competent *E. coli* cells. The transformed cells were inoculated in 5 mL Luria Broth with 50 µg/mL Kanamycin overnight at 37ºC. The next day the pre-inoculum was transferred to 10 mL LB or TB medium with 50µg/mL Kanamycin to an initial OD600nm of 0.02. The cells were incubated at 37ºC with constant rotation to an OD600nm of 0.7 and induced with 0.05mM or 0.1mM IPTG concentration. The culture was incubated at 37ºC with constant rotation for 3h for protein expression.

Expression of the D41A mutant construct at small scale for comparison with the expression of the wild-type LegRPsd was made by first inoculating transformed BL21 star cells in 5 mL Luria Broth with 50 µg/mL Kanamycin overnight at 37ºC. The next day the pre-inoculums were transferred to 10 mL LB medium with 50µg/mL Kanamycin to an initial OD600nm of 0.02. The cells were incubated at 37ºC with constant rotation to an OD600nm of 0.7 and induced with 0.05mM IPTG concentration. The cultures were incubated at 37ºC with constant rotation for 3h for protein expression.

For the co-expression screening of LegRPsd-HIS/LegRPsd-HA, the pRSF co-expression construct was transformed in BL21 star (Invitrogen) competent *E. coli* cells. The transformed cells were inoculated in 5 mL Luria Broth with 50 µg/mL Kanamycin overnight at 37ºC. The next day the pre-inoculum was transferred to 10 mL LB or TB medium with 50µg/mL Kanamycin to an initial OD600nm of 0.02. The cells were incubated at 37ºC with constant rotation to an OD600nm of 0.7 and induced with 0.05mM or 0.1mM IPTG concentration. The culture was incubated at 37ºC or 30ºC with constant rotation for 3h for protein expression.

Protein extraction at small scale was made using BugBuster Protein Extraction

Reagent (Novagen). 1 mL of cell culture was harvested for the use of the protein extraction kit. The insoluble fraction was also ressuspended in 200µL PBS for insoluble protein analysis by Western-blot.

2.2.2.2. Expression for Production of Recombinant Protein

The best conditions for scale-up expression of LegRPsd and D41A mutant and for the co-expression of LegRPsd-HIS/LegRPsd-HA were determined to be the same. The transformed BL21star cells were pre-inoculated in an Erlenmeyer flask overnight with LBkan medium (LB medium with 50µg/mL Kanamycin). LBkan medium was divided by Fernbach flasks in order that each flask contains 1 L LBkan each. Each flask was inoculated with 20 mL pre-inoculum. The cells were incubated at 37^{o}C with constant rotation to an OD600nm of 0.7 and then induced with 0.05mM IPTG. The cultures were then incubated at 37^{o}C with rotation for 3h.

For harvesting the cells from the 1 L cultures, cultures were centrifuged at 6000g for 20 min at 4^{o}C and the culture pellet was ressuspended in 20 mL per 1 L of culture in 20 mM sodium phosphate buffer pH 7.5, 10 mM Imidazole, 0.5 M NaCl. Lysozyme was added and the ressuspension was frozen at -20^{o}C.

2.2.3. Purification by IMAC-Ni^{2+} and Cation Exchange Chromatography

To obtain a soluble total protein extract, harvested cells expressing wt LegRPsd or LegRPsd D41A mutant were defrosted at room temperature to cause cell lysis. After that, cell lysates were incubated with DNAse and $MgCl_2$ at 4^{o}C for 2h to remove the large amount of DNA present. The resulting lysates were then centrifuged at 12.000g at 4^{o}C for 20min. The resulting soluble fraction was filtered through 0.2µm filters and applied to a Histrap HP 5mL column (Amersham Biosciences).

2.2.3.1. Optimization

The Histrap HP 5 mL column was equilibrated and washed with in 20 mM sodium phosphate buffer pH 7.5, 10 mM Imidazole, 0.5 M NaCl. For elution, 20 mM sodium phosphate buffer pH 7.5, 500 mM Imidazole, 0.5 M NaCl was prepared, and the adequate percentage of elution buffer was mixed with the equilibration buffer to create the desired Imidazole concentrations for the elution steps of 50 mM, 100 mM and 500 mM. Column was operated in an AKTA FPLC system (GE Healthcare Life Sciences) always at a flow of 5 mL/min. Fractions of 5 mL were collected through the elution steps.

Protein containing fractions were pooled and dialyzed using 3.5kDa cutoff dialysis membranes. The protein fractions were dialyzed against 5L of 20mM MES buffer pH 6.0 with 50mM NaCl overnight.

Next, the dialyzed pool was filtered through 0.2 µm filters and applied to a cation exchange Mono-S 5/50 GL column (GE Healthcare). The column was equilibrated and washed with 20 mM MES buffer pH 6.0 and eluted with a linear gradient of the elution buffer (20 mM MES buffer pH 6.0 1 M NaCl). Column was operated in an AKTA FPLC system (GE Healthcare Life Sciences) always at a flow of 0.75 mL/min. Fractions of 1 mL were collected through the elution steps.

2.2.3.2. Purification Process

The Histrap HP 5 mL column was equilibrated and washed with in 20 mM sodium phosphate buffer pH 7.5, 10 mM Imidazole, 0.5 M NaCl. For elution, 20 mM sodium phosphate buffer pH 7.5, 500 mM Imidazole, 0.5 M NaCl was prepared, and the adequate percentage of elution buffer was mixed with the equilibration buffer to create the desired Imidazole concentrations for the elution steps of 50 mM, 200 mM and 500 mM. Column was operated with an AKTA FPLC system (GE Healthcare Life Sciences) at constant a flow of 5 mL/min. Fractions of 5mL were collected through the elution steps.

Protein containing fractions were pooled and dialyzed using 3.5kDa cutoff dialysis membranes. The protein fractions were dialyzed against 5 L of 20 mM MES buffer pH 6.0 with 50mM NaCl overnight.

Next, the dialyzed pool was filtered through 0.2 µm filters and applied to a Cation Exchange Mono-S 5/50 GL column (GE Healthcare). The column was equilibrated and washed with 20 mM MES buffer pH 6.0 and eluted with an increasing gradient mixing the washing buffer with the elution buffer (20 mM MES buffer pH 6.0, 1 M NaCl). The Mono-S Column was operated in an AKTA FPLC system (GE Healthcare Life Sciences) always at a flow of 0.75 mL/min. Fractions of 1 mL were collected through the elution steps.

2.2.4. Auto-activation

To assess enzyme auto-activation properties, selected protein fractions from the cation exchange chromatography with larger amount of pure LegRPsd-15 (about 3 mg/ml of protein) were diluted in 50 mM sodium acetate buffer pH 3.0, 4.0, 5.0 or 6.0 with 100 mM NaCl in a 1:2 proportion (1 volume of sample for 2 volumes of sodium acetate buffer) in order to have the protein at pH 3.0, 4.0, 5.0 or 6.0. The pH values of resulting dilutions were confirmed using pH test strips. The resulting dilutions were then incubated at 37^{o}C in a water bath up to 36h and samples were collected at different time points of incubation (0, 6h, 12h, 24h and 36h).

2.2.5. Proteolytic Activity Assays

Proteolytic activity was tested towards several different fluorescent substrates: a Typical AP substrate [MCA-K]-K-P-A-E-F-F-A-L-[K-DNP] (Genscript), the CDR1 substrate [MCA-K]-L-H-P-E-V-L-F-V-L-E-[K-DNP] (Genscript), the HIV-1 protease substrate R-E(EDANS)-S-G-N-Y-P-I-V-Q-K(DABCYL)-R (Sigma), the Rickettsia retroviral-like protease substrate [MCA- K]- A-L-I-P-S-Y-K-W-S-[K-DNP] (Genscript) and the Plasmepsin V susbtrate DABCYL-L-N-K-R-L-L-H-E-T-Q-EDANS (Genscript). These substrates were tested in 96-well plates with a final concentration of 10 µM in 50 mM sodium acetate pH 4.0, 5.0 and 6.0, 100 mM NaCl. 50 µg of purified LegRPsd was added to each assay. To read the plates a Spectramax Gemini EM (Molecular Devices) was used, configured to excite the molecules at 340 nm and read emission at 485 nm (For plasmepsin V and

HIV-1 protease susbtrates) or excite at 328 nm and read emission at 393 nm (for all other susbtrates) at 37^{o}C. To determine cleavage of oxidized insulin β-chain, one volume of enzyme at 0,5 mg/mL was incubated with 5 volumes of oxidized insulin β-chain (1 mg/mL) in 50 mM sodium acetate buffer pH 4.0 or 6.0 and incubated in a water bath for 24h at 37^{o}C. The resulting peptides from the cleavage of oxidized insulin β-chain were then evaluated by Reverse Phase HPLC.

2.2.5.1. Assays at Different pH values

To determine the effect of pH on the proteolytic activity of LegRPsd towards the typical AP substrate, the reaction mix consisted of 5µL of substrate at 0,3mg/mL (final concentration of 5 µM), 50 µL enzyme at 1 mg/mL (50 µg of enzyme) and 145 µL of 50 mM sodium acetate buffer with 100 mM NaCl at pH 3.0, 3.5, 4.0, 4.5, 5.0 or 6.0 or 50 mM sodium citrate buffer with 100 mM NaCl for pH 2.5. The reaction was carried on 96-well plates read by a Spectramax Gemini EM (Molecular Devices) fluorimeter set to 37^{o}C during 3h (ex: 328 nm, em: 393 nm)

2.2.5.2. Assays in the Presence of Inhibitors

To test the effect of inhibitors the reaction mix consisted of 5µL of substrate at 0,3 mg/mL (final concentration of 5 µM), 50µL enzyme at 1 mg/mL (50µg of enzyme) and the volume completed with of 50 mM sodium acetate buffer with 100 mM NaCl at pH 4.0 to a final volume of 200 µL. The reaction was carried on 96-well plates read by a Spectramax Gemini EM (Molecular Devices) fluorimeter set to 37^{o}C during 3h (ex: 328 nm, em: 393 nm). The inhibitors used, as well as the final concentration are displayed in Table 2.

For the inhibitors dissolved in DMSO, a final concentration of 5% DMSO in the assay was required to keep its solubility. Due to the loss of activity of LegRPsd to about 50% in the presence of 5% DMSO, the amount of enzyme was doubled, so that the reaction mixture would consist in 5 µL of substrate at 0,3 mg/mL (final concentration of 5 µM), 50µL enzyme at 2 mg/mL (100µg of enzyme), DMSO added to 5% and the volume completed with of 50 mM sodium acetate buffer with 100 mM NaCl at pH 4.0 to a final volume of 200 µL. The reaction was again carried on 96-well plates read by a Spectramax Gemini EM (Molecular Devices) fluorimeter set to 37^{o}C during 3h (ex: 328 nm, em: 393 nm).

Table 2. Inhibitors used in the enzyme assays, respective concentration in the assay and solvent used for storage.

Inhibitor	Solvent	Final Concentration
Pepstatin	MeOH	1 uM
Bestatin	EtOH	10 uM
EDTA	Water	5 mM
E64	Water	10 uM
PEFABLOC	Water	1 mM
Indinavir	Water	1 mM
Darunavir	Water	25 uM
Saquinovir	DMSO	0.1 mM
Nelfinavir	DMSO	0.1 mM
Atazanavir	DMSO	0.5 mM
Ritonavir	DMSO	0.5 mM
Lopinavir	DMSO	0.5 mM
Amprenavir	DMSO	0.5 mM

2.2.5.3. Assays at Different Temperature Values

To determine temperature dependence it was used a Spex Fluoromax 3 Spectrofluorimeter (HORIBA JOBIN YVON) connected to a hot bath for temperature control. The reaction mixture was composed of 12µL substrate at 0.3 mg/mL (final concentration of 5 µM), 80 µL enzyme at 3 mg/mL (240 µg enzyme) and 50 mM Sodium acetate pH 4.0, 100 mM NaCl to a final volume of 500 µL. Fluorescence was measure for each temperature value for 20 min (ex: 328 nm, em: 393 nm).

2.2.5.4. Assays with Varying Substrate Concentration

For the kinetic parameters determination, the reaction mixture contained 80 µL enzyme at 3 mg/mL (240 µg enzyme), 50mM sodium acetate pH 4.0, 100mM NaCl to a final volume of 500 µL, and different volumes of substrate for final concentrations ranging from 0.19µM to 30.75µM. Spex Fluoromax 3 Spectrofluorimeter (HORIBA JOBIN YVON) connected to a hot bath at 37^{o}C was used and fluorescence was measure for each substrate concentration for 20 min (ex: 328 nm, em: 393 nm).

2.2.6. Reverse-Phase HPLC

Proteolytic activity of LegRPsd towards oxidized insulin β-chain was analyzed by Reverse-Phase HPLC. Upon incubation, TFA was added to the samples to a final concentration of 0.6% and these were centrifuged for 5 min at 12.000g (room temperature). The resultant soluble fraction was injected in a KROMASIL C18 column (Teknokroma) using a Prominence Shimadzu HPLC system and the resulting peptides were separated by RP-HPLC. The column was equilibrated and washed with 0.1% TFA and elution was carried out by a gradient of 0 to 80% acetonitrile in 0.1% TFA at a flow rate of 1 mL/min. Elution of the peptides was monitored at 220nm.

2.2.6. Analytical-Size Exclusion Chromatography

Native molecular weight of protein samples were analyzed by analytical-size exclusion chromatography. The samples were injected in a Superdex 200 HR 10/30 (GE Healthcare) column using a Prominence Shimadzu HPLC system. The column was equilibrated with 20 mM phosphate buffer pH 7.5 or 50mM sodium acetate pH 4.0 with 100 mM NaCl. To estimate the molecular weight the column was calibrated with protein standards: aprotinin (6.5kDa), ribonuclease (13.7kDa), carbonic anhydrase (29kDa), ovalbumin (43kDa), conalbumin (75kDa), aldolase (158kDa), ferritin (440kDa).

2.2.7. Purification by HisMag Sepharose Ni Beads

LegRPsd-HIS and LegRPsd-HA co-expressed in pRSF Duet vector were co-purified using HisMag Sepharose Ni (GE Healthcare) magnetic beads. The harvested cells expressing LegRPsd-HIS and LegRPsd-HA were defrosted at room temperature to cause cell lysis. After that, cell lysates were incubated with DNAse and $MgCl_2$ at 4^oC for 2h to remove the large amount of DNA present. The resulting lysates were then centrifuged at 12.000g at 4^oC for 20 min. In order to compare the purification between a crosslinked and a non-crosslinked fraction, the resulting soluble fraction was divided in two equal 10mL fractions. To one of the fractions it was added 1 mL Glutaraldehyde at 2.3% as a crosslinking agent and it was incubated for 5 min at 37^oC. Reaction was stopped with 1 mL 1 M Tris pH 8.0.

200 µL of HisMag Sepharose Ni beads were added to each fraction (cross-linked and non-crosslinked cell extracts) and the resulting suspension was incubated at 4^oC for 2h with constant rotation. Then, 1mL of each incubated fractions was transferred to an eppendorf tube and the magnetic beads were collected using a MagRack (GE Healthcare). The process was repeated using the same tube until all the beads have been collected. After this, 500 µL of washing buffer (20mM phosphate buffer pH 7.5, 10 mM Imidazole, 0.5 M NaCl) was added, mixed and the beads were collected. The washing process was repeated for two more times. Then, 150 µL of elution buffer (20 mM phosphate buffer pH 7.5, 500 mM Imidazole, 0.5 M NaCl) was added, mixed and the beads were collected. The elution process was repeated for two more times.

2.2.8. SDS-PAGE and Western-Blot

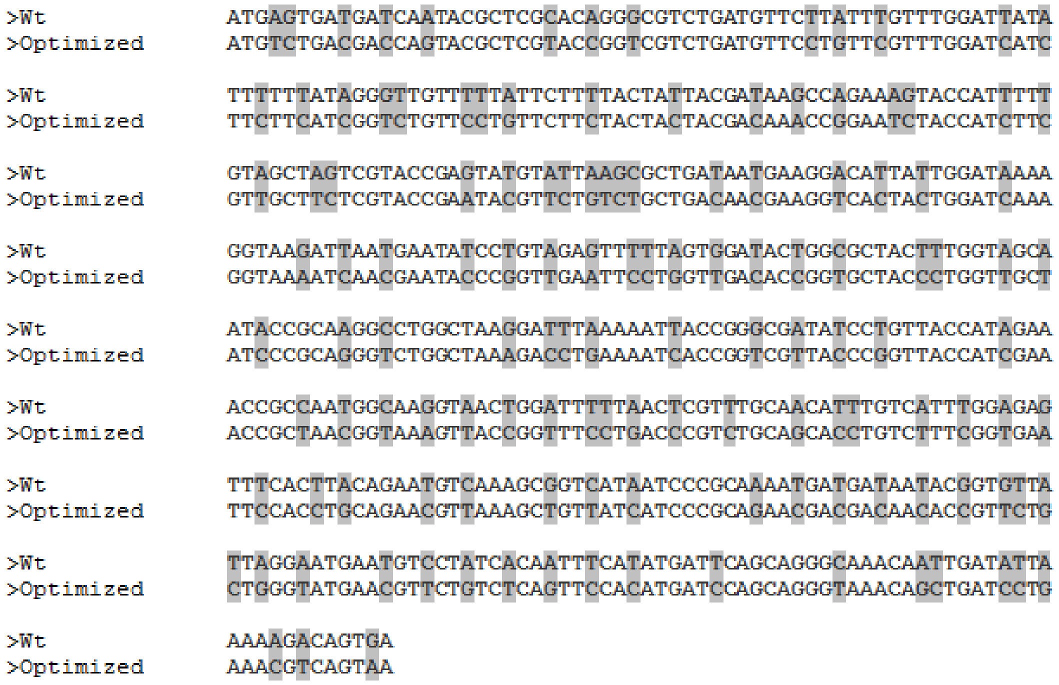
Protein samples were separated by SDS-PAGE in 12.5% polyacrylamide gels. Samples applied to gels were denatured by incubation for 10 min at 95^oC with 6x loading buffer (0.35 M Tris-HCl/0.28%SDS buffer pH 6.8, 30% Glicerol, 10% SDS, 0.6 M DTT and 0.012% Bromophenol Blue). The gels were run in a MiniProtean 3 system (Bio-Rad) at room temperature at 150 V (running buffer with 100 mM Tris, 100 mM Bicine, 0.1% SDS). The gels stained with Coomassie Brilliant Blue (50% methanol, 10% acetic acid and 0.2% Brilliant Blue R) were then incubated with 25% methanol and 5% acetic acid for removal of excess dye.

For Western blot analysis the proteins were separated by SDS-PAGE and transferred to a PDVF membrane (previously activated in methanol). The electrotransference were performed using a Trans-BlotR Electrophoretic Transfer cell (Bio-Rad) overnight at 40 V at 4 °C, using 25 mM Tris, 192 mM glycine and 20% methanol as transference buffer. After transfer, PDVF membranes were blocked for 1 hour with TBST buffer (150 mM NaCl, 10 mM Tris, pH 8.0, 0.1% Tween 20) containing 5% milk. Membranes were then incubated with primary antibody (anti-His (GenScript) or anti-HA with a dilution of 1:10000) in TBST buffer containing 0.5% milk. After incubation with the primary antibody, the membrane was washed at least 5 times for 5 min in TBST buffer containing 0.5% milk and incubated for 1h with the secondary antibody Anti-Mouse IgG + IgM alkaline phosphatase linked whole antibody (from goat; Amersham Biosciences), also with a dilution of 1:10000 in 0.5% milk. Membrane labeling was revealed with ECF™ substrate (GE Healthcare) in contact with the membrane for 5min and then scanned in a Molecular Imager FX (Bio-Rad).

3. Results and Discussion

3.1. The Recombinant Retropepsin-Like Protease from *Legionella pneumophila*

The gene for the retroviral-like protease from *Legionella pneumophila* (LegRP) was synthesized with the codon usage optimized for its heterologous expression in *E. coli* cells (Figure 8) delivered inserted in pUC57 vector.



```
>Wt            ATGAGTGATGATCAATACGCTCGCACAGGGCGTCTGATGTTCTTATTTGTTTGGATTATA
>Optimized     ATGTCTGACGACCAGTACGCTCGTACCGGTCGTCTGATGTTCCTGTTCGTTTGGATCATC

>Wt            TTTTTTATAGGGTTGTTTTTATTCTTTTACTATTACGATAAGCCAGAAAGTACCATTTTT
>Optimized     TTCTTCATCGGTCTGTTCCTGTTCTTCTACTACTACGACAAACCGGAATCTACCATCTTC

>Wt            GTAGCTAGTCGTACCGAGTATGTATTAAGCGCTGATAATGAAGGACATTATTGGATAAAA
>Optimized     GTTGCTTCTCGTACCGAATACGTTCTGTCTGCTGACAACGAAGGTCACTACTGGATCAAA

>Wt            GGTAAGATTAATGAATATCCTGTAGAGTTTTTAGTGGATACTGGCGCTACTTTGGTAGCA
>Optimized     GGTAAAATCAACGAATACCCGGTTGAATTCCTGGTTGACACCGGTGCTACCCTGGTTGCT

>Wt            ATACCGCAAGGCCTGGCTAAGGATTTAAAAATTACCGGGCGATATCCTGTTACCATAGAA
>Optimized     ATCCCGCAGGGTCTGGCTAAAGACCTGAAAATCACCGGTCGTTACCCGGTTACCATCGAA

>Wt            ACCGCCAATGGCAAGGTAACTGGATTTTTAACTCGTTTGCAACATTTGTCATTTGGAGAG
>Optimized     ACCGCTAACGGTAAAGTTACCGGTTTCCTGACCCGTCTGCAGCACCTGTCTTTCGGTGAA

>Wt            TTTCACTTACAGAATGTCAAAGCGGTCATAATCCCGCAAAATGATGATAATACGGTGTTA
>Optimized     TTCCACCTGCAGAACGTTAAAGCTGTTATCATCCCGCAGAACGACGACAACACCGTTCTG

>Wt            TTAGGAATGAATGTCCTATCACAATTTCATATGATTCAGCAGGGCAAACAATTGATATTA
>Optimized     CTGGGTATGAACGTTCTGTCTCAGTTCCACATGATCCAGCAGGGTAAACAGCTGATCCTG

>Wt            AAAAGACAGTGA
>Optimized     AAACGTCAGTAA
```

Figure 8. Alignment between the wild-type gene coding for retropepsin-like aspartic protease (accession number lpg2007) from *Legionella pneumophila* pneumophila (strain Philladelphia 1) and the optimized sequence for heterologous expression in *E. coli*. Nucleotide differences are highlighted.

Due to the presence of very hydrophobic sequences, transmembranar domains are usually difficult to express in *E. coli* cells and even more difficult to purify and study in terms of proteolytic activity. As transmembranar domains usually do not affect the proteolytic activity or stability of proteases, only the putative soluble domain (from Asp33 to Gln163, hereby designated LegRPsd) (see Figure 6) was selected for cloning and expression.

Using ProtParam (Gasteiger et al. 2005) it was possible to calculate the theoretical pI and molecular weight of the recombinant LegRPsd with the fused His-tag. The protein presents a theoretical pI of 6.79 and a putative molecular weight of 16 kDa. These parameters will be useful for the protein's first analysis of expression and purification.

3.1.1. Cloning the LegRP Soluble Domain

The gene was amplified through PCR using primers allowing the amplification of the optimized gene sequence coding for the soluble domain with the addition of restriction sequences *Nco*I and *Xho*I at the 5' and 3' ends respectively. These restriction sites were used to clone the soluble domain sequence in a pET-28a vector (Novagen) in-frame with the vector's His-tag at the C-terminus. The sequences used also allowed for the removal of the N-terminus his-tag sequence present in the vector.

The pET28a is a vector constructed for heterologous expression of recombinant proteins. The coding sequence is regulated by a strong T7 promoter, which is in turn regulated by a lacI gene. The lacI allows the inhibition of the protein expression in the absence of lactose or other molecular homologues. This turns pET28a into an inducible vector, meaning that the protein expression can be induced when the user decides the time is right by the addition of IPTG, a lactose homologue.

The vector was therefore constructed to allow expression of the soluble domain with a His-tag only on the C-terminus. The His-tag allows not only an easier purification through the use of Immobilized Metal Ion Chromatography (IMAC), but it allows also the detection of the protein through the use of an anti-His specific antibody. The choice to insert the His-tag on the C-terminus was due to previous results from our lab concerning the *Rickettsia* retroviral-like protease, in which was observed auto-processing at the N-terminus. This way, and anticipating a similar behavior, the His-tag was included in the C-terminus so it could be present in both possible precursor and mature forms.

3.2. Expression and Purification of the Recombinant LegRP soluble domain

3.2.1. Screening of Optimal Expression Conditions

To express the recombinant LegRP soluble domain (LegRPsd), it was first necessary to assess the optimal conditions for its heterologous expression in *E. coli* cells in a small-scale system. For the first test the vector containing the gene was transformed in two different *E. coli* strains (C41 and BL21star), which were tested for protein expression in two different culture media (Luria's Broth and Terrific Broth) and they were compared using two different concentrations of IPTG at the moment of induction (0.1 mM and 0.05 mM). In all cases, cells were grown to OD600nm = 0.7 before induction at 37ºC in a volume of 50 mL with 50 µg/mL of Kanamycin, and were kept at 37ºC for 3h after induction for protein expression).

Samples taken at the end of the 3h of expression were lysed and analyzed by SDS-PAGE followed by Western-Blot. Both soluble and insoluble fractions were analyzed through immunodetection using an anti-His antibody (Genscript) as the primary antibody. This allowed the detection of the expressed rLegRPsd due to its fusion with a C-terminus His-Tag (Figure 9).

At first sight we can affirm that overall expression in BL21 star offers better results. Furthermore the different conditions using this strain gave similar results as they all present good amount of protein expressed in soluble form. In this way, the selected expression condition was BL21 star strain in Luria's Broth media and inducing the expression with 0.05mM IPTG.

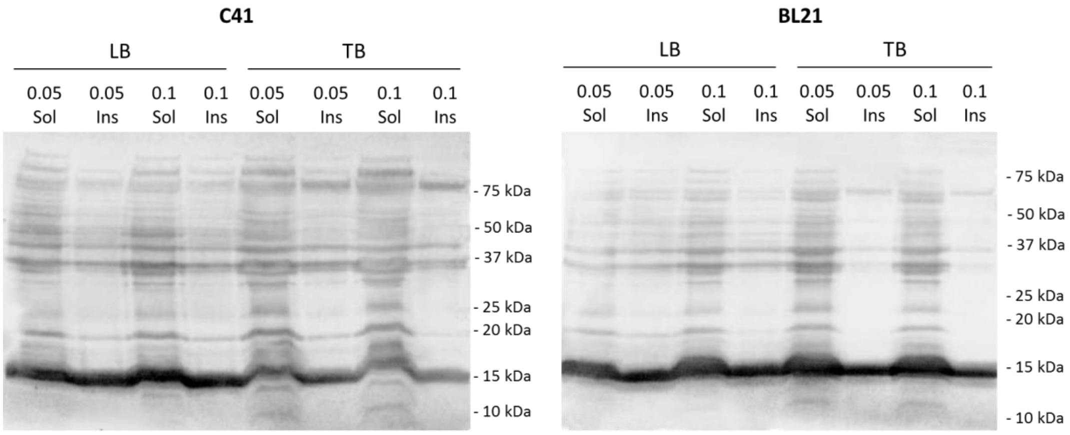


Figure 9. Western-Blot analysis of the recombinant LegRPsd expression screening. After protein expression in the tested conditions, cell extracts were analyzed by Western-Blot using an anti-His antibody. The expression of the recombinant LegRPsd was tested in two different *E. coli* strains: C41 (left) and BL21 star (right). For each strain it was tested growth and expression on Luria's Broth (LB) and Terrific Broth (TB) and induction made with two different concentration of IPTG: 0.05 (mM) and 0.1 (mM). For each harvested sample the soluble (Sol) and Insoluble (Ins) fractions were analyzed. Samples were loaded in 12.5% polyacrylamide gels and transferred to a PVDF membrane probed with anti-His antibody (dilution 1:10000)

3.2.2. Expression of the Recombinant LegRP Soluble Domain

After optimization of the expression conditions, the process was scaled-up with the objective of producing enough protein for purification and biochemical characterization. The BL21 star cells properly transformed with the previously expression construct of rLegRPsd were cultivated at 37ºC in separate flasks, each containing 1L of LB culture medium with 50 µg/mL Kanamycin. At an OD600nm of 0.7, protein expression was induced by addition of IPTG for 3 hours at 37ºC.

After expression, the cells were harvested and frozen at -20ºC overnight with the purpose of lysing *E. coli* cells by a freeze-thaw cycle.

3.2.2. Purification of the Recombinant LegRP Soluble Domain

In order to purify the recombinant LegRPsd, the C-terminal His-tag was used. As the histidine residues have the ability to coordinate with divalent metal ions, an Immobilized Metal-Ion Affinity Chromatography (IMAC) was planned as the first purification step. After this first step, other chromatographic techniques were considered for higher degree of purification, like ion-exchange chromatography.

3.2.2.1 Optimization of the Purification Process

After breaking the cells and extracting the soluble cytoplasmic fraction resulting from recombinant protein expression, the extract needed to be purified in order to obtain a final pool with only the recombinant LegRP present. For this, the soluble fraction of the cell extract was injected in a HisTrap™ HP 5mL column, an IMAC column containing bound Ni^{2+} which are capable of complexing with Histidine residues. After sample application, the column was washed with 20 mM sodium phosphate buffer pH 7.5 containing 10 mM Imidazole and 0.5 M NaCl. Imidazole competes for the ligation to the IMAC-Ni^{2+} column, therefore preventing protein non-specific binding. Elution of the bound protein was then performed by three steps of increasing Imidazole concentrations – 50mM, 100 mM and 500 mM. The principle of the elution using Imidazole is again based on a competition with the metal ion complexes. At high Imidazole concentrations the immobilized metal ion binds to imidazole molecules instead of binding to the His-tag. Protein presence was detected by measuring A280nm. Figure 10 shows the chromatogram obtained for the IMAC purification step and the respective concentration of Imidazole.

It is seen that a large amount of protein is eluted in two main steps. First, a lot of protein is eluted from the “shoulder” present in the 50 mM step. However, the protein is mostly eluted at 100 mM Imidazole, where a large peak is present.

The eluted fractions were analyzed by SDS-PAGE followed by Coomassie-blue staining. Coomassie-blue stains all the proteins present in a polyacrylamide gel, allowing the identification of the different proteins present in the samples by analyzing their molecular weights (Figure 11).

It was seen before that the recombinant LegRPsd with the fused His-tag would have a putative molecular weight of 16 kDa. In the SDS-PAGE analysis it is clear that a highly expressed protein is present at ~15 kDa in fractions 25 to 46 (late in the 50 mM Imidazole elution step) and two proteins at ~15 kDa and ~14 kDa proteins are present in fractions 47 to 60 (in the beginning of the 100 mM Imidazole elution step). Due to the high amount of protein it was thought that these could be two different forms of the recombinant LegRPsd, probably, the ~15kDa form (hereby designated LegRPsd-15) could correspond to the expressed LegRPsd (predicted to have a MW of 16kDa), and the ~14kDa form (hereby designated LegRPsd-14) would be a maturated form of LegRPsd-15, likely due to a proteolytic processing during the heterologous expression in *E. coli* cells.

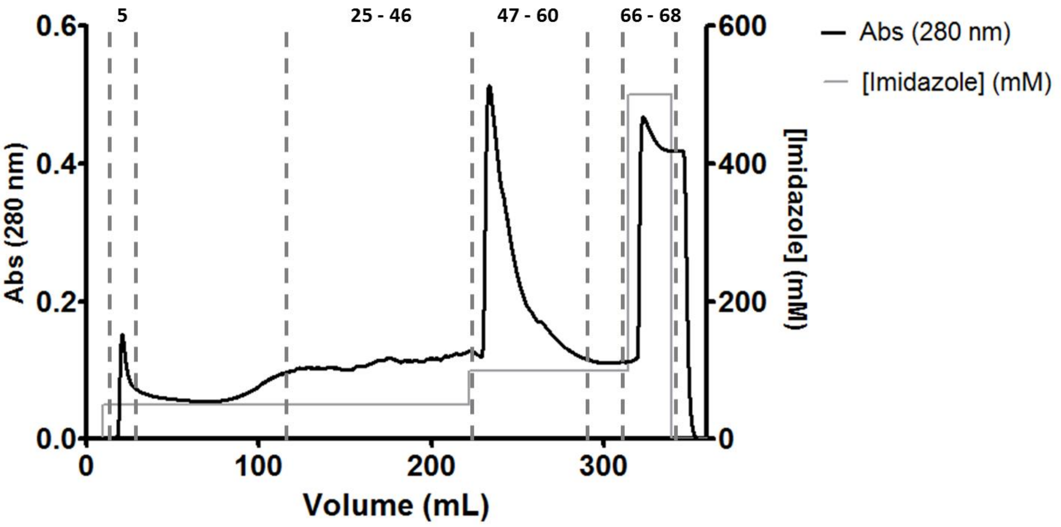


Figure 10. Purification of the recombinant LegRPsd by IMAC-Ni^{2+}. A280nm and Imidazole concentration are shown in function of volume. The column was equilibrated and washed with 10mM Phosphate Buffer pH 7.4, 0.5M NaCl, 10mM Imidazole. Protein was eluted using 10mM Phosphate Buffer pH 7.4, 0.5M NaCl with three different steps of Imidazole concentration (50 mM, 100 mM and 500 mM) at a flow rate of 5 mL/min and detected by its A280nm. Collected fractions numbers are shown at the top.

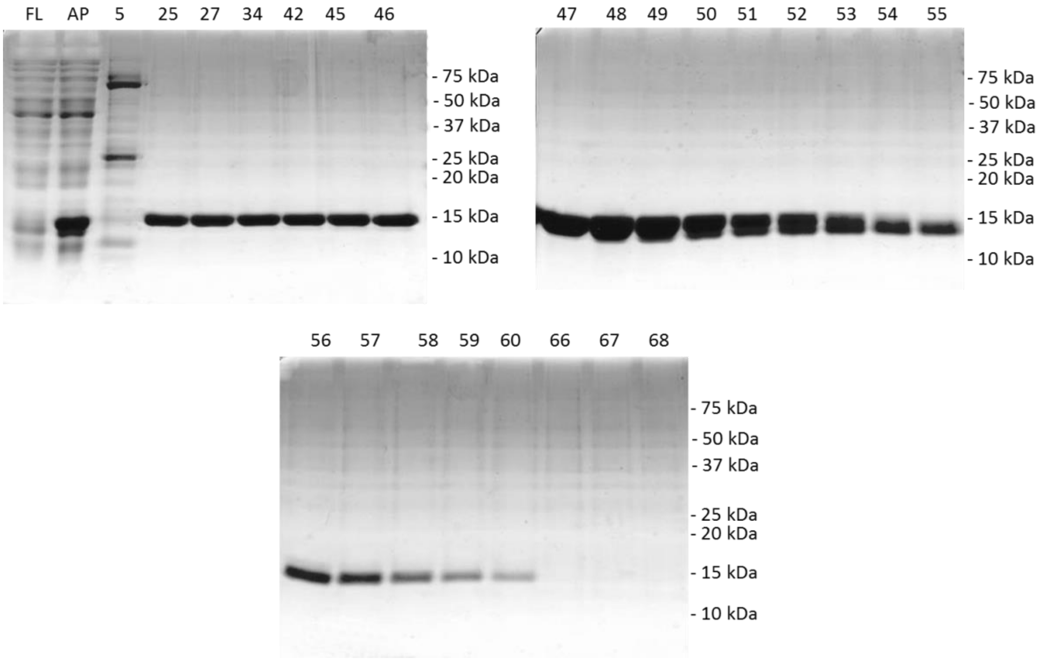


Figure 11. SDS-PAGE analysis of the fractions collected in purification of the recombinant LegRPsd by IMAC-Ni^{2+}. The fraction numbers correspond to the numbers displayed in Figure 8, AP refers to the total soluble cell extract applied to the column, while FL refers to the flowthrough sample that was collected after column injection (protein not-bound to the column).

To try to further separate these two different forms of the protein, and to remove any additional impurities, the eluted fractions were grouped in two different pools: fractions 25 to 46 were grouped in a first pool (from now on designated “50mM Pool”), and fractions 47 to 60 were grouped in a second pool (from now on designated “100mM Pool”). Then, these pools were dialyzed overnight against 20 mM MES buffer pH 6.0 with 50 mM NaCl and loaded in a Cation Exchange Mono-S 5/50 GL column (GE Healthcare). As the rLegRPsd presents a theoretical pI of 6.79, the column was equilibrated using 20mM MES Buffer at pH 6.0. At this pH, recombinant LegRPsd will present itself with a positive charge, therefore allowing its binding to the negatively charged column. After sample application, the column was washed with the same buffer and was eluted with a gradient of NaCl (20mM MES Buffer pH 6.0, 0 - 1M NaCl). The principle behind this elution is that the increase in the ionic force resulting from increasing NaCl concentration leads to the release of the protein bound to the column. This is due to the competition between charges that is increased by increasing the amount of salt ions present in the buffer. As with the HisTrap HP column, the presence of protein was detected by measuring the A280nm of the eluted volume. In Figure 12 A1 and B1 we can see the chromatograms obtained for the 50 mM imidazole and 100 mM imidazole pools, respectively. Conductivity increases as salt concentration increases so the conductivity values are a method to monitorize the concentration of NaCl present in the eluted volume.

The eluted fractions from both purifications were then analyzed by SDS-PAGE followed by Coomassie-blue staining. While Figure 12 A2 refers to the eluted fractions from the chromatogram shown in A1 (from the purification of the "50mM" Pool), Figure 12 B2 refers to the eluted fractions from the chromatogram shown in B1 (from the purification of the "100mM" Pool).

It is clear the presence of three major peaks in each chromatogram (in A1: corresponding to fractions 17 to 23, fraction 25 and fractions 27 and 28; in B1: corresponding to fractions 17 to 22, fraction 23 and fractions 24 to 29), being the first one clearly more intense than the other two. Although in the purification of the "50mM" Pool, the second and third peaks are much less intense than in the purification of the "100mM" Pool, when analyzed by SDS-PAGE followed by Coomassie Blue staining it is possible to see that both pools follow approximately the same profile of elution and that the first peak contains mostly LegRPsd-15, while the second and third peak contain both LegRPsd-15 and LegRPsd-14. No fraction displays only the LegRPsd-14 form.

The first trial for the purification resulted in purification of large amounts of protein and we could obtain fractions containing only the LegRPsd-15, which seems to be the predicted LegRPsd which MW was around 16kDa, so it is possible to say that the purification of this form was successful. However, the first purification step using IMAC-Ni^{2+} was clearly not optimized and it was necessary a reformulation of the elution steps. So a new and more efficient protocol needed to be elaborated before moving on with further protein characterization studies.

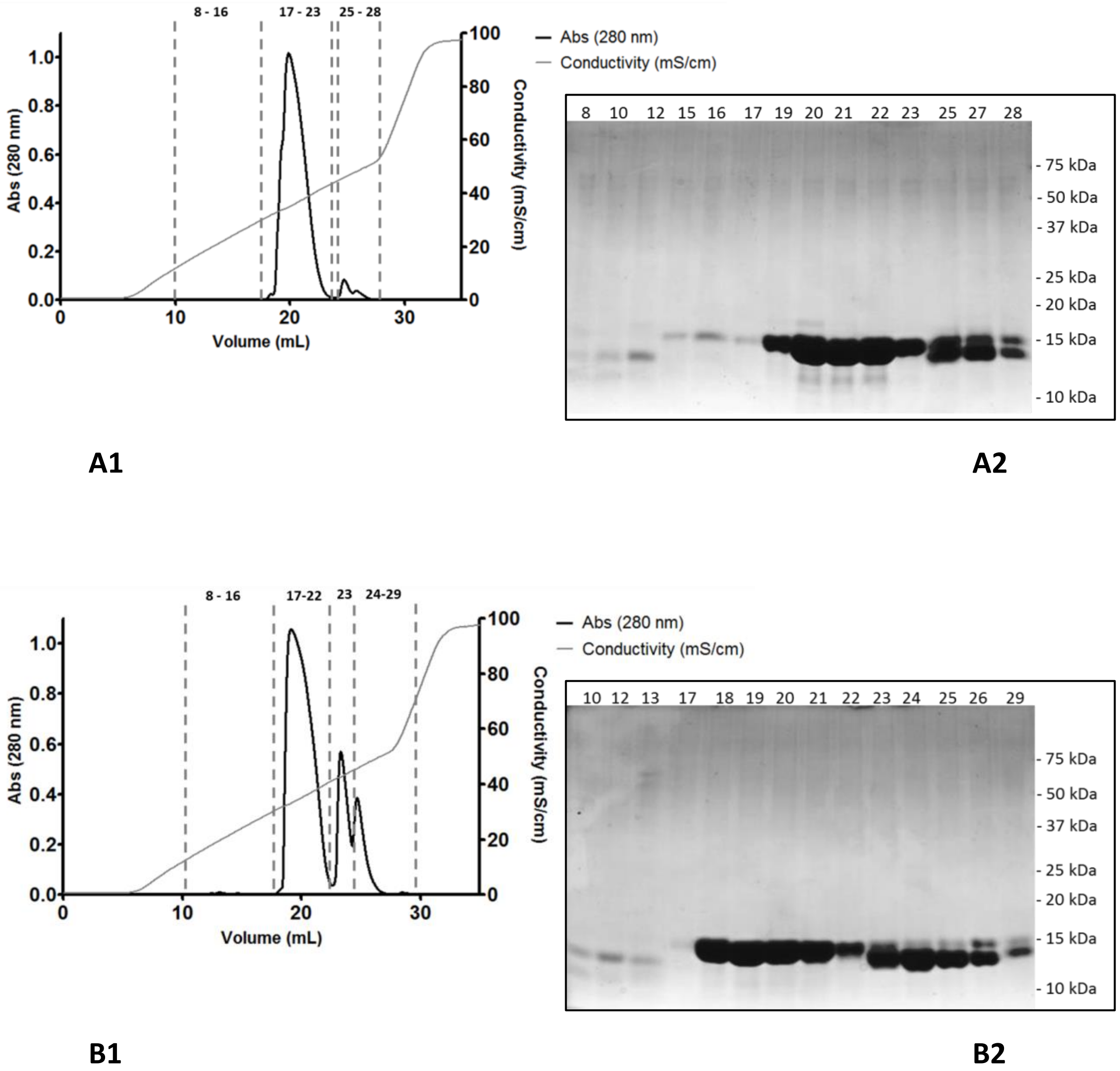


Figure 12. Purification of the recombinant LegRPsd by cation exchange chromatography in a Mono S 5/50 GL column . A1 shows the elution profile from the purification of the “50mM” Pool and A2 shows the SDS-PAGE analysis of the fractions eluted in A1. B1 shows the elution profile from the purification of the “100mM” Pool and B2 shows the SDS-PAGE analysis of the fractions eluted in B1. The MonoS column was equilibrated and washed using 20mM MES Buffer pH6.0 and protein was eluted from the column using 20mM MES Buffer pH 6.0 with a NaCl gradient from 0 to 1M at a flow rate of 0.75 mL/min and detected by its A280nm. The conductivity values are proportional to the concentration of NaCl. The 12.5% polyacrylamide gel was stained with Coomassie Blue. The fraction numbers in A2 and B2 correspond to the fraction numbers displayed at the top of A1 and B1 respectively.

3.2.2.2 Purification of the Recombinant LegRPsd

The first purification step of the recombinant LegRPsd (the IMAC-Ni^{2+}) was optimized and the 100 mM Imidazole elution step was substituted by a 200 mM imidazole step. Therefore, the optimized purification process used the same methodologies as in the previous section, with the elution steps of 50 mM, 200 mM and 500 mM Imidazole concentration. The resulting chromatogram is shown in Figure 13 (chromatogram representations use the same parameters described in the previous section).

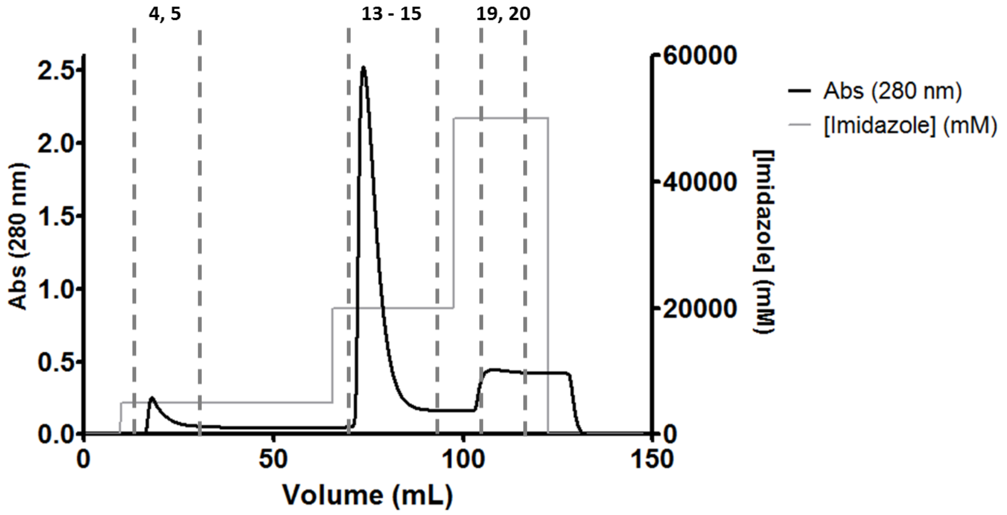


Figure 13. Purification of the recombinant LegRPsd by IMAC-Ni^{2+} upon optimization of elution steps. A280nm and Imidazole concentration are show in function of volume. The column was equilibrated and washed with 10mM Phosphate Buffer pH 7.4, 0,5M NaCl, 10mM Imidazole. Protein was eluted using 10mM Phosphate Buffer pH 7.4, 0,5M NaCl with three different steps of Imidazole concentration (50mM, 200mM and 500mM), at a flow rate of 0.75 mL/min and detected by its A280nm. Collected fractions numbers are shown at the top.

The fractions collected from the IMAC-Ni^{2+} were the analyzed by SDS-PAGE followed by Coomassie Blue staining to detect total proteins present and also analyzed by SDS-PAGE followed by Western Blot using an anti-His antibody to specifically detect the recombinant LegRPsd fused with the His-tag (Figure 14)

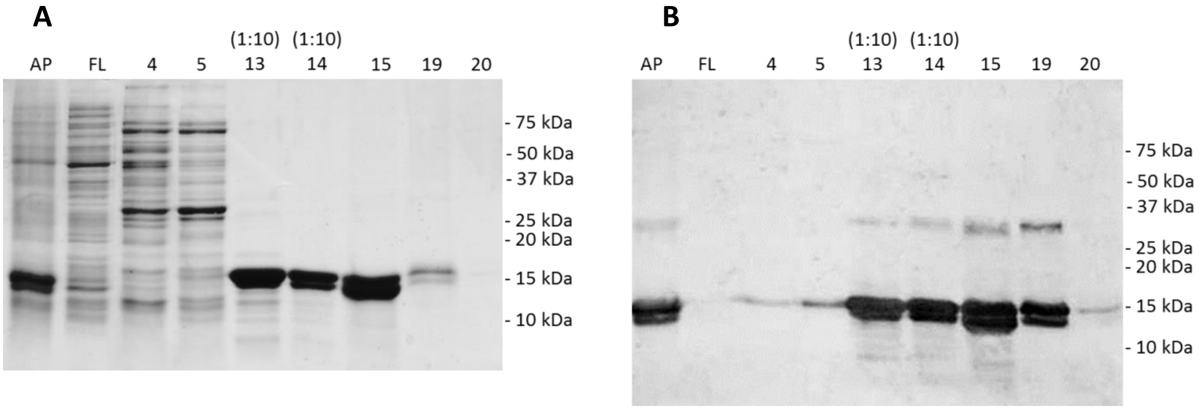


Figure 14. SDS-PAGE and Western-Blot analysis of fractions eluted from the purification of the recombinant LegRPsd by optimized IMAC-Ni^{2+}. The fraction numbers correspond to the numbers displayed in Figure 13, AP refers to the total soluble cell extract applied to the column, while FL refers to the flowthrough sample that was collected after column injection (protein not-bound to the column). Samples diluted are indicated with the respective dilution.

A) SDS-PAGE analysis (12.5% polyacrylamide) stained with Coomassie Blue. 5 µL of each sample denatured with 6x loading buffer was loaded in the gel

B) Western-Blot analysis using an anti-his-tag antibody (dilution 1:10000). 3 µL of each sample denatured with 6x loading buffer was loaded in the gel for the SDS-PAGE prior to electrotransference.

From the analysis of the chromatogram in Figure 13, it is clear that a large amount of protein is eluted at 200 mM Imidazole, and disperse through some few fractions, in contrast with the previous strategy where the protein was eluted in a large buffer volume and disperse through different elution steps. By Western Blot it is possible to confirm that it is mainly the recombinant LegRPsd due to the presence of a strong binding of the anti-His antibody at ~15kDa and ~14kDa. It is also possible to detect some labeling on the fractions eluted from the 50mM Imidazole step and the 500 mM Imidazole step, but the amount of protein present is much lower. This allowed the collection of the eluted fractions during the 200 mM Imidazole step to be joined in a single Pool containing large quantity of both LegRPsd-15 and LegRPsd-14.

It is important to note the presence of a band in the Western blot analysis close to 30kDa in the fractions containing large quantity of recombinant LegRPsd (fraction 13-15) and also in fraction 19. The band clearly is labeled by the anti-His antibody and may indicate the presence of dimers, constituting the first evidence that this retroviral-like AP is present in the form of homodimers, the structure needed to form the catalytic active center of the protease, as it is observed for retropepsins and other retropepsin-like proteases.

The pooled fractions (200 mM) were then dialyzed against 20 mM MES buffer pH 6.0, 50 mM NaCl in a 3.5kDa cutoff membrane to remove salt and imidazole, which would hinder the binding of the protein to the next column used for purification, the Cation Exchange Mono-S 5/50 GL column.

After dialysis the protein pool from the IMAC-Ni^{2+} purification was loaded in a cation exchange Mono-S 5/50 GL column with the same methodologies as before. The elution process was also kept the same since it presented good results. The column was equilibrated and washed with 20 mM MES Buffer pH 6.0 and eluted with a NaCl gradient from 0 to 1 M.

The chromatogram resulting from the MonoS column elution gradient is shown in Figure 15 (chromatogram representations use the same parameters described in the section before).

The fractions collected were analyzed by SDS-PAGE followed by Coomassie Blue staining to detect total proteins present and also analyzed by SDS-PAGE followed by Western Blot using an anti-His-tag antibody to specifically detect the recombinant LegRPsd fused with the His-tag (Figure 16).

It is possible to verify what was seen during the optimization in the previous section. The purification follows a profile showing three major peaks. Analysis of these peaks shows that the first one contains mainly LegRPsd-15 in high amounts while the two other peaks contain less total protein and both LegRPsd-15 and LegRPsd-14 are present.

Again it is possible to observe in the Western blot analysis, although much more faint, the presence of a band at around 30kDa in the eluted fractions (present in both fractions containing only LegRPsd-15 and a mix of LegRPsd-15 and LegRPsd-14, marked in the figure with a "*"), indicating probably the presence of homodimer formation.

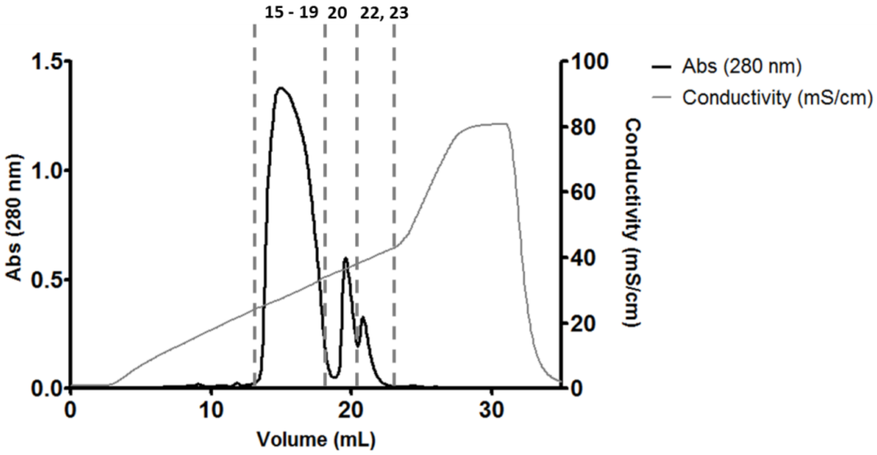


Figure 15. Purification of the recombinant LegRPsd by cation exchange Chromatography. The MonoS column was equilibrated and washed using 20mM MES Buffer pH6.0 and protein was eluted from the column using 20mM MES Buffer pH 6.0 with a NaCl gradient from 0 to 1M, at a flow rate of 0.75 mL/min and detected by its A280nm. The conductivity values are proportional to the concentration of NaCl. Collected fractions numbers are shown at the top.

With the modification introduced in the IMAC chromatography it was possible to accomplish a successful purification of the recombinant LegRPs-15. We can conclude that the purification of recombinant LegRPsd was a success, retrieving high yield of pure protein suitable for biochemical characterization. From this point forward we established the necessary conditions to study its biochemical properties. Moreover, fractions enriched in LegRPsd-14 (SDS-PAGE analysis in Figure 16), which is probably the result of a processing step, were also subjected to subsequent characterization.

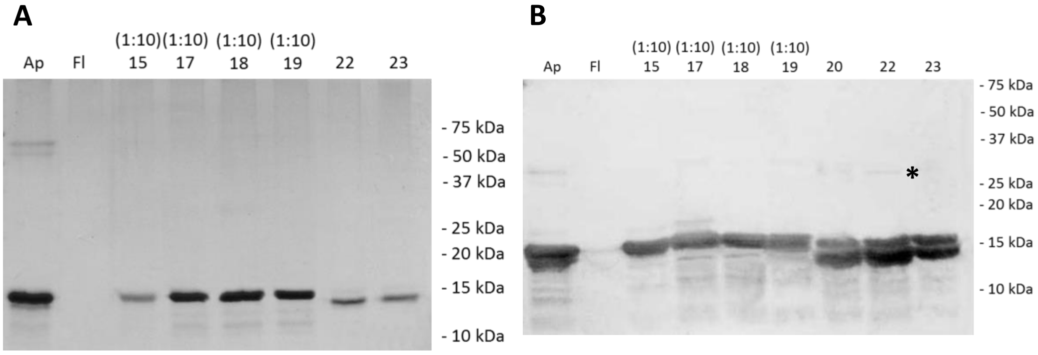


Figure 16. SDS-PAGE and Western-Blot analysis of fractions eluted from the purification of the recombinant LegRPsd by cation exchange chromatography. The fraction numbers correspond to the numbers displayed in Figure 15, AP refers to the total soluble cell extract applied to the column, while FL refers to the flowthrough sample that was collected after column injection (protein not-bound to the column). Samples diluted are indicated with the respective dilution.

A) SDS-PAGE analysis (12.5% polyacrylamide) stained with Coomassie Blue. 5 µL of each sample denatured with 6x loading buffer was loaded in the gel

B) Western-Blot analysis using an anti-his-tag antibody (dilution 1:10000). 3 µL of each sample denatured with 6x loading buffer was loaded in the gel for the SDS-PAGE prior to electrotransference. A dim band at ~30kDa is marked with a "*".

3.3. Biochemical Characterization

3.3.1. LegRPsd-15 and LegRPsd-14: Auto-Processing

During the purification it was possible to observe clearly that LegRPsd was present in two forms with different molecular weights. It was possible to detect by SDS-PAGE and Western-blotting that the recombinant LegRPsd was present in a ~14kDa form (LegRPsd-14) and in a ~15kDa form (LegRPsd-15). The first hint was that LegRPsd-14 would be a product of LegRPsd-15, likely formed due to auto-proteolytic processing of LegRPsd-15 after translation.

As will be focused in the next section, aspartic proteases tend to have their maximum proteolytic activity at lower pH values. Because some of these retropepsin-like are able to undergo autoproteolytic processing *in vitro* under acidic conditions, as it was previously explored when regarding proteases like SASPase (Bernard et al. 2005), the *Rickettsia* retropepsin-like protease (unpublished data) and PerP (Chen et al. 2006), we decided to evaluate the autoprocessing activity of LegRP15 at different pH values. In order to evaluate its autoprocessing ability, purified fractions of LegRPsd-15 were incubated at pH values ranging from pH 3.0 to 6.0 at 37°C up to 36h, and a timecourse analysis was performed at 6h, 12h, 24h and 36h. Conversion between LegRPsd-15 into the LegRPsd-14 form was observed at pH 6.0, but not at lower pH values. Figure 17 shows the SDS-PAGE followed by Coomassie Blue staining and Western-Blot analysis of samples collected over different incubation times at pH 6.0.

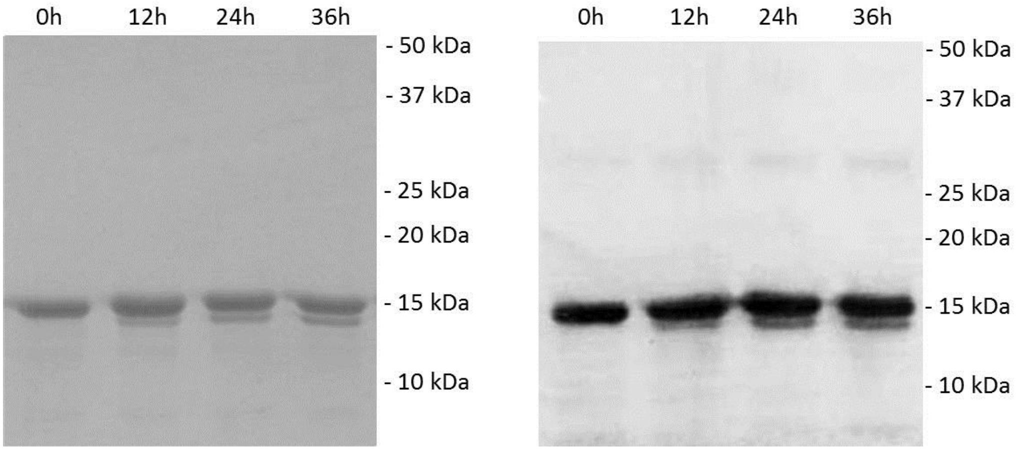


Figure 17. SDS-PAGE and Western-Blot analysis of the samples collected during incubation of LegRPsd-15 at pH 6.0. Samples were incubated in 50 mM sodium acetate buffer, 100 mM NaCl, pH 6.0 at 37°C and samples were collected at the beginning, at 12h, 24h and 36h.

These results suggest that this auto-processing is probably the phenomena responsible for the presence of both forms in the cytoplasm of *E. coli* expressing the recombinant protein. However, the *E. coli* cytoplasm probably offers more suitable conditions for the protein’s processing since the *in vitro* processing is apparently slower that what is seen when the protein is expressed for 3 hours.

Auto-proteolytic processing activity is a fundamental idea when we are talking about retropepsins, as retrovirus needs this specific function of its protease to mature. When other retropepsin-like enzymes are analyzed, this is a characteristic always present, even when only the soluble domain is analyzed. This *in vitro* autoprocessing was seen for SASPase (Bernard et al. 2005), PerP (Chen et al. 2006) and the *Rickettsia* retropepsin-like protease (Cruz R., Simões I., unpublished results).

3.3.2. Proteolytic Activity and Specificity of LegRPsd

The first step to prove that the expressed recombinant protein is an aspartic protease is obviously to verify its proteolytic activity towards several peptide substrates. As stated before, there isn't a specific substrate cleavage site sequence common to all retroviral-like aspartic proteases. The mainly recognized cleavage site sequences compromise in sites P1 and P1' large hydrophobic residues (much like Pepsin-like proteases) or an aromatic and proline residues

In this study we focused on checking the proteolytic activity of LegRPsd towards oxidized insulin β-chain and a diverse set of fluorogenic substrates available in the laboratory, shown to be cleaved by different aspartic proteases (displayed in table 4).

The fluorescent substrates tested were: a typical aspartic protease substrate due to general affinity of APs towards Phe-Phe cleavage sites (Dunn et al. 1986), a specific substrate to a plant AP named CDR1 (Simões et al. 2007), a substrate used to test the activity of HIV-1 protease (Sigma), a specific substrate designed for the retroviral-like aspartic protease from *Rickettsia conorii* and a substrate for plasmepsin V (an AP from *Plasmodium falciparum*) (Russo et al. 2010).

Table 4. List of substrates used to test the proteolytic activity of the purified recombinant LegRPsd (both the 14kDa and the 15kDa forms cleave the same substrates). In the Efficiency of Cleavage column, C refers to cleaved substrates while NC refers to non-cleaved substrates.

Substrate	Sequence	Efficiency of Cleavage
Typical AP	[MCA-K]-K-P-A-E-F-F-A-L-[K-DNP]	C
CDR1	[MCA-K]-L-H-P-E-V-L-F-V-L-E-[K-DNP]	NC
HIV-1 Protease	R-E(EDANS)-S-G-N-Y-P-I-V-Q-K(DABCYL)-R	NC
Rickettsia Retroviral-like Protease	[MCA- K]- A-L-I-P-S-Y-K-W-S-[K-DNP]	NC
Plasmepsin V	DABCYL-L-N-K-R-L-L-H-E-T-Q-EDANS	NC
Oxidized Insulin β-Chain	F-V-N-Q-H-L-C-G-S-H-L-V-E-A-L-Y-L-V-C-G-E-R-G-F-F-Y-T-P-K-A	C

Activity of LeRPsd towards these substrates was tested by measuring fluorescence at 37^{o}C for 3 hours and at different pH values. The fluorogenic substrates contain one fluorophore and a quencher group. When the substrate molecule is cleaved, the fluorophore does no longer suffer a quenching effect and emits fluorescence. This way the amount of fluorescence emitted by time unit is directly proportional to protease activity. The oxidized insulin β-chain was incubated with the recombinant protein samples at 37^{o}C for 24h at pH 4.0 and 6.0 and the incubations were then analyzed by reverse-phase HPLC. Figure 18 shows the resulting chromatograms of Insulin degradation, comparing the proteolytic activity towards this substrate at pH 4.0 and 6.0 and comparing the activity of LegRPsd-15 and LegRPsd-14.

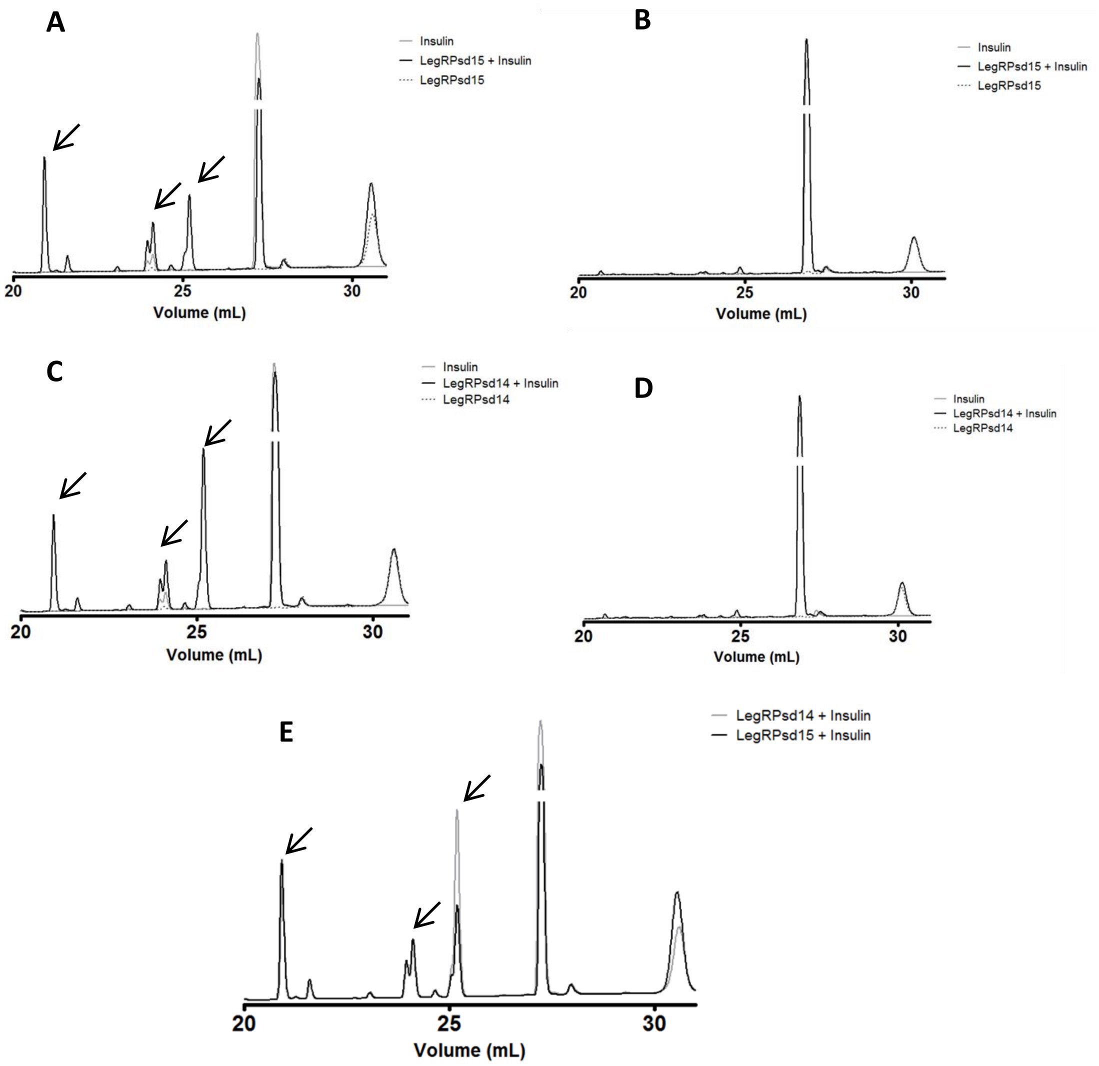


Figure 18. Reverse-phase HPLC elution profiles of hydrolyzed Oxidized Insulin Beta-Chain after incubation with LegRPsd. Oxidized beta insulin Chain was incubated with LegRPsd-14 or LegRPsd-15 for 24h at 37^{o} at pH 4.0 or pH 6.0. Upon precipitation with 0.6%TFA, the incubated samples were loaded on a RP-HPLC column equilibrated with 0.1% TFA. The peptides were then eluted with a gradient of 0.1% TFA, 0 - 80% CH_3CN, at a flow rate of 1 mL/min. Peptide presence was detected by its A220nm. Controls of oxidized insulin β-chain (incubated without LegRPsd) and LegRPsd were performed for each condition. Peaks resulting from Insulin cleavage are marked with arrows.

A) Proteolysis of Oxidized Insulin β-Chain by LegRPsd-15 at pH 4.0 and controls;

B) Proteolysis of Oxidized Insulin β-Chain by LegRPsd-15 at pH 6.0 and controls;

C) Proteolysis of Oxidized Insulin β-Chain by LegRPsd-14 at pH 4.0 and controls;

D) Proteolysis of Oxidized Insulin β-Chain by LegRPsd-14 at pH 6.0 and controls;

E) Comparison between the digestion profile of oxidized insulin β-chain by LegRPsd-14 and LegRPsd-15 at pH 4.0.

Both LegRPsd-15 and LegRPsd-14 showed the same preference for substrates, although with slightly different reaction rates. Of the tested fluorescent substrates, both forms only showed proteolytic activity towards the typical AP substrate, a substrate usually preferred by pepsin-like enzymes. LegRPsd-15 showed a specific activity of 1.11×10^{-5} $nmol.min^{-1}.\mu g^{-1}$ while LegRPsd-14 showed a specific activity of 1.81×10^{-5} $nmol.min^{-1}.\mu g^{-1}$ towards this substrate (substrate concentration 4.75 µM, pH 4.0).

Concerning oxidized insulin β-chain cleavage, it is obvious that the proteolytic activity at pH 4.0 is much higher in both forms than at pH 6.0. Comparing both forms at pH 4.0, we can observe that the specificity seems to be similar for both forms, however, LegRPsd-15 seems to be more efficient in generating one of the cleavage products and less with one of the others, suggesting slight differences in enzyme processivity. Identification of the cleavage sites will allow us to know more about its specificity, as this work is being processed using Mass Spectrometry by the Proteomics unit at CNC/Biocant.

3.3.3. Effects of the pH and Temperature on the Proteolytic Activity

Having a fluorogenic peptide substrate cleaved by the recombinant LegRPsd allows for a biochemical study of its proteolytic activity towards this substrate. These results are important to evaluate the protein's similarities and differences to other bacterial proteases as well as to evaluate similarities to retropepsin-like proteases.

In general, aspartic proteases have a tendency for having good proteolytic activity at acidic pH values, especially pepsin-like enzymes, which usually have very low optimum pH values. This mostly happens due to the features of the catalytic center that presents specific residues that need to be in specifically reduced forms to proceed to electron transfer in order to catalyze acid-base reactions needed for proteolysis (Andreeva 2003). Previous studies proved the optimum pH values for retroviral-like aspartic proteases are also usually around pH 4.0 or 5.0 but can go as high as pH 6.0

(Fehér A. et al. 2006; Ido E. et al. 1991; Rawlings & Salvesen 2013)

The next step in this study was to determine the optimum pH values for the hydrolysis of the typical AP substrate by LegRPsd-15 and LegRPsd-14. For that, both forms were tested in a fluorimeter with the typical AP substrate, at 37^{o}C for 3h, in 50 mM sodium acetate buffer, 100mM NaCl, with pH ranging from 3.0 to 6.0, or 50mM sodium citrate buffer, 100mM NaCl for pH 2.5. The resulting profile is shown in figure 19.

It is interesting to note that both forms display a similar pH dependence profile towards this substrate. The optimum pH value for both forms is pH 4.0, although LegRPsd-14 is more active than LegRPsd-15 at lower pH values, with LegRPsd-15 displaying around half of the activity of the LegRPsd-14 at pH 3.

At pH values above 5.0 none of the forms shows any proteolytic activity towards these substrates. This is an extremely curious result when compared to LegRPsd-15 auto-processing results. Auto processing does not occur at lower pH values, and it only occurs at pH 6.0. This could indicate that the mechanism of proteolysis towards the fluorescent substrate and the mechanism of auto-proteolysis probably are not the same. However, it could also happen due to conformational changes not only between the two forms LegRPsd-15 and LegRPsd-14, but also due to the differences in the protein's conformation between pH 4.0 and pH 6.0, that could influence, for example, the availability of the N-terminal cleavage site.

Analysis of these results in understanding possible *in vivo* functions can make us speculate that this protease probably won't be proteolytically active when it is located in the Legionella cytoplasm, due to the pH values near 6.0. However, the cytoplasm would probably be a favorable location for processing after translation. However, there is the probability that LegRPsd can be exported to the periplasm. PerP, a periplasmic protease from *Caulobacter* (described previously on the Introduction section) and an homologue of the LegRPsd was described as being a periplasmic protease (Chen et al. 2006). If LegRPsd is in fact only active as a proteolytic enzyme at pH lower than 5.0, then it will probably be located in the periplasm after being processed.

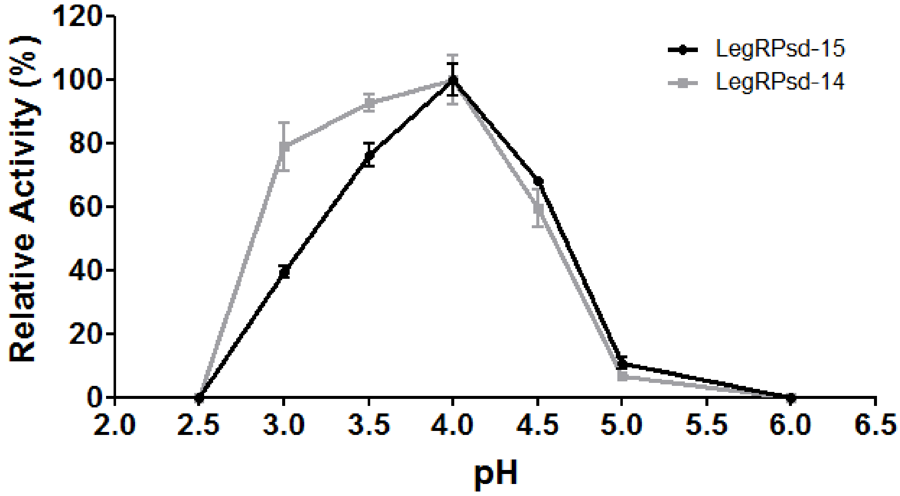


Figure 19. Effect of pH on the activity of recombinant LegRPsd-15 and LegRPsd-14. The activity was tested towards the fluorogenic typical substrate for APs [MCA-K]-K-P-A-E-F-F-A-L-[K-DNP]. The assays were performed by incubating purified recombinant LegRPsd at 37 ° C with 50 mM sodium acetate buffer,100 mM NaCl, between 3.0 and 6.0 or 50 mM sodium citrate buffer, 100 mM NaCl for pH 2.5 and measuring emitted fluorescence per

The periplasm has been shown in *E. coli* to be sensitive to extracellular pH variation, with very slow recovery rates (Wilks & Slonczewski 2007). Being also a gram-negative bacteria, we can extrapolate due to the similarity of the membrane, that the periplasm in *Legionella pneumophila* could also be sensitive to external pH. This means that the LegRP proteolytic activity could be triggered when the bacteria are phagocytized and the host starts to lower the phagosome's pH in attempt to digest the Legionella. Clearly, expression and localization studies in vivo are required to further elucidate this hypothesis.

Using the same substrate in 50 mM sodium acetate,100 mM NaCl, pH 4.0 it was possible to determine a temperature dependence curve for LegRPsd-15 (Figure 20). The temperature profile shows an increase in activity up to 55 Celsius degrees, and it starts to decline above that temperature. This profile has no significant value considering infection and phagocytic cycles. However, considering the growth conditions of *Legionella pneumophila* in water reservoirs, the temperature dependence acquires increased relevance. It is know that *Legionella pneumophila* are capable of living and proliferating in water reservoirs in the form of planktonic cells or biofilms up to 55 Celsius degrees, having its best growth conditions around 40 Celsius degrees (Dennis et al. 1984). The dependence on temperature by LegRPsd-15 could be related to the temperature tolerance of *Legionella*, therefore, it is possible that the protein can also be related to cell replication outside a phagocytic host. If the protein is related to fundamental cellular functions, it is expected that its temperature dependence profile is in agreement with the bacterium's preferred growth conditions or cell stress response.

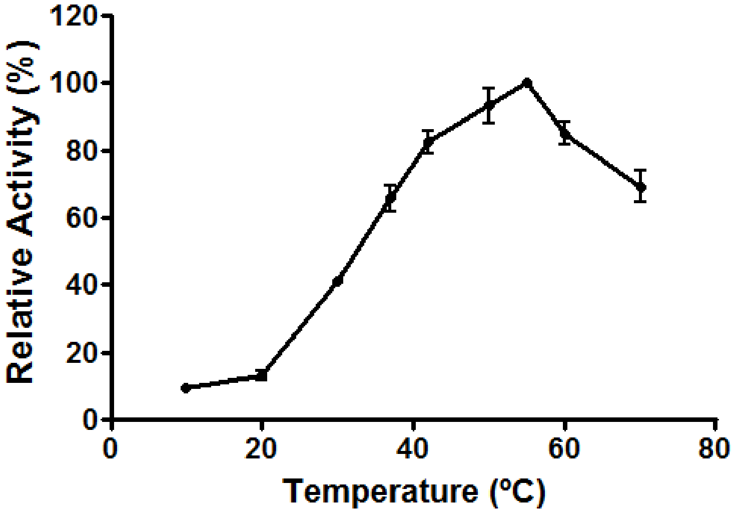


Figure 20. Effect of temperature on the activity of recombinant LegRPsd-15. The activity was tested towards the fluorogenic typical substrate for APs [MCA-K]-K-P-A-E-F-F-A-L-[K-DNP]. The assays were performed by incubating purified recombinant LegRPsd at different temperatures with 50 mM sodium acetate, 100 mM NaCl, pH 4.0 and measuring emitted fluorescence per second.

3.3.4. The Effect of Classical and Retropepsin Inhibitors

To characterize this enzyme as an aspartic protease, the enzyme should be tested with general inhibitors of the main protease classes. For this, the activity of LegRPsd towards the typical AP substrate was tested in the presence of several inhibitors. First, the activity was tested in the presence of: Pepstatin A, a typical

Inhibitor of aspartic proteases, usually its inhibition levels are very considerable in pepsin-like family members (A1 family) and in some members of the A2; bestatin: an inhibitor of metalloproteases, specifically aminopeptidases; pefabloc SC: an irreversible serine protease inhibitor; E-64: a specific cysteine protease inhibitor; EDTA: a metalloprotease inhibitor due to its chelating properties, generally proteins that need metal ions for its activity/folding are affected by this molecule.

Both forms were incubated with the inhibitors for 10 minutes at room temperature before proceeding with the enzyme activity assay as described before. Figure 21 shows the effect of the described inhibitors on each of the LegRPsd forms.

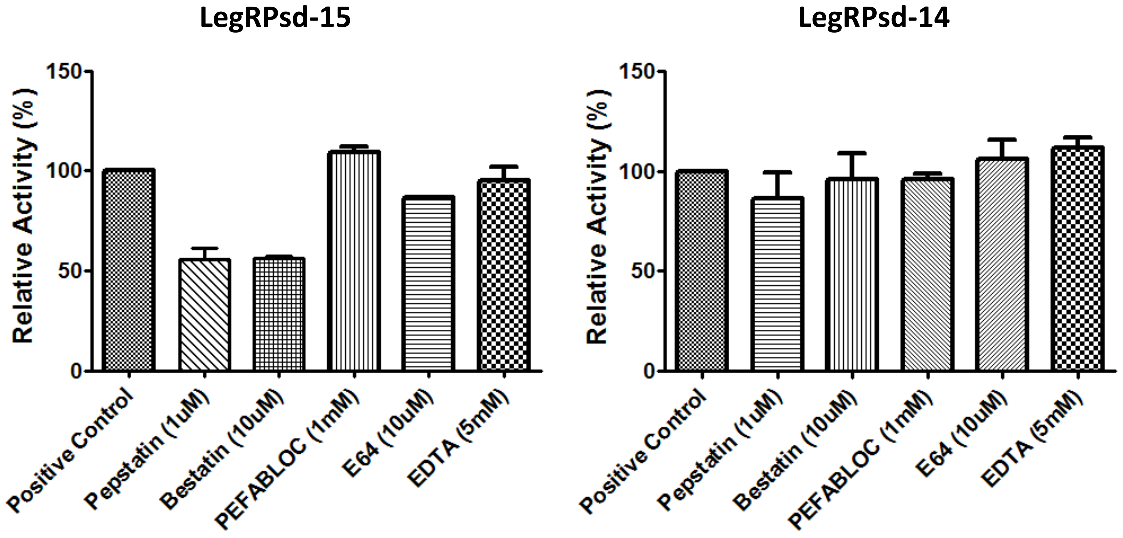


Figure 21. Inhibition profile of purified recombinant LegRPsd-15 and LegRPsd-14. The enzyme was incubated for 10 min at room temperature with each inhibitor and activity was tested towards the fluorogenic typical substrate for APs [MCA-K]-K-P-A-E-F-F-A-L-[K-DNP]. The assays were performed by incubating both forms of purified recombinant LegRPsd in 50 mM sodium acetate,100 mM NaCl, pH 4.0 and measuring emitted fluorescence per second.

LegRPsd-15 enzymatic activity is affected by the presence of pepstatin A. Even if this effect only reduces the activity to about 50%, it is a favorable argument in validating its characterization as an aspartic protease. Also, the lack of a full inhibition is not surprising as retroviral proteases are not consistent with the rate of Pepstatin inhibition (for example, HIV-1 Protease is strongly inhibited by pepstatin, while that is not true for the XMRV Protease) (Matúz, K. et al. 2012; Tyagi, S.C. 1992). Surprisingly, bestatin also affects LegRPsd-15 activity by about 50%. This is an inhibitor of aminopeptidases, therefore leaving an open chance for this protein to have some kind of aminopeptidase-like activity. On the other hand, recombinant LegRPsd-14 does not seem to be affected in large extent by any of the tested inhibitors. This can be probably justified by a change in conformation of the active site pocket during the autoprocessing, which necessarily affects activity and inhibitor binding.

One of the main objectives in this study is to assess possible similarities between this protease and retropepsins. In that matter, studying the effect of retroviral aspartic protease inhibitors becomes an essential step. The following inhibitors of retropepsins (all used in the treatment of HIV infections by inhibition of the HIV-1 and/or HIV-2 retropepsins) were used: Saquinovir, Nefinavir, Atazanavir, Ritonavir, Lopinavir, Amprenavir, Indinavir and Darunavir. As all these inhibitors, except for Indinavir and Darunavir, needed 5% DMSO in the assay buffer, which lead us to double the amount of enzyme in the assays in order to compensate for the loss of LegRPsd activity in the presence of DMSO (LegRPsd-14 and LegRPsd-15 both lose about 50% activity in the presence of 5% DMSO). All values are reported to the correspondent control assay performed in the absence of each inhibitor, in 5% DMSO. Figure 22 shows the inhibition profile of both forms in the presence of the anti-retroviral drugs.

Overall, LegRPsd-14 is more affected by retroviral aspartic protease inhibitors than LegRPsd-15. This is the contrary to what is verified with classical protease inhibitors (Figure 21). However, some retroviral inhibitors have the same effect on both forms, like Nelfinavir, Indinavir or Darunavir, not displaying a significant effect on the protease's activity. Amprenavir seems to have a considerable inhibitory effect on both forms, while Saquinovir has a significantly different effect on both forms (while it reduces the LegRPsd-15 activity only by about 25%, it promoted a reduction of about 70% on LegRPsd-14 activity). Strong inhibition by amprenavir is a characteristic shared with the HIV and the MLV retropepsins, suggesting close structural similarity between LegRPsd active site pocket and these retropepsins (Fehér et al. 2006).

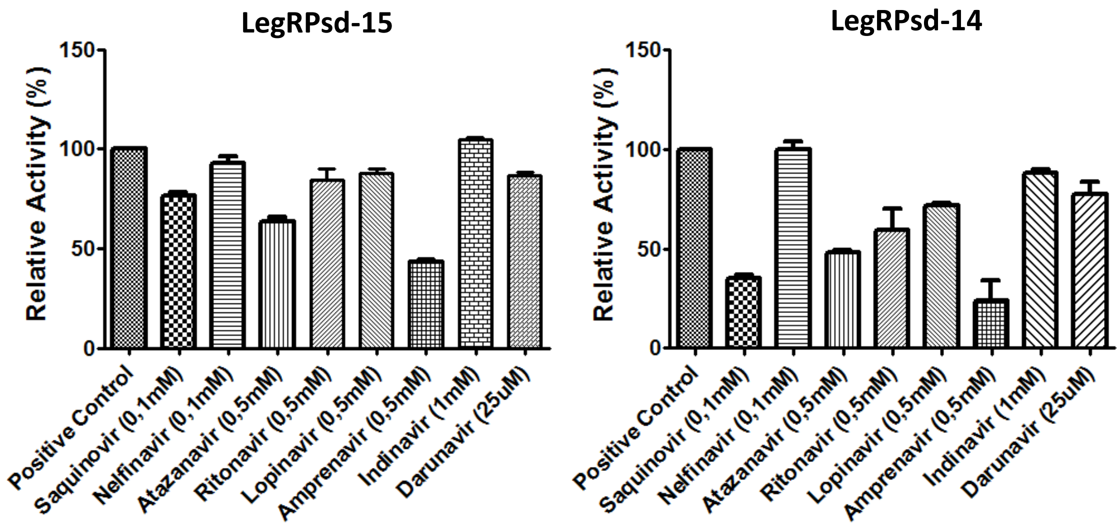


Figure 22. Inhibition profile of purified recombinant LegRPsd-15 and LegRPsd-14 by retropepsin inhibitors. Both forms were incubated for 10min at room temperature with each inhibitor and the activity was tested towards the fluorogenic typical substrate for APs [MCA-K]-K-P-A-E-F-F-A-L-[K-DNP]. The assays were performed by incubating purified recombinant LegRPsd in 50 mM sodium acetate buffer,100 mM NaCl, pH 4.0 and measuring emitted fluorescence per second. For activity assays of Saquinovir, Nelfinavir, Atazanavir, Ritonavir, Lopinavir and Amprenavir, 5% DMSO was kept in the assay buffer and enzyme quantity was doubled to compensate activity loss.

From these results, it appears that LegRPsd acquires a more “Retroviral-Like” profile when maturing from 15kDa to the 14kDa form. It loses its sensitivity to general and typical protease inhibitors while it is much more affected by specific retroviral protease inhibitors. The strong inhibition of LegRPsd by some retroviral protease inhibitors clearly suggests that the binding of these drugs to the LegRPsd active site may be similar to their binding to the HIV protease active center, which leads us to speculate that the catalytic pocket of LegRPsd must somehow resemble the one of retropepsins, known to be highly specific and unique. This is strong evidence supporting the similarity of these proteins. Altogether, the partial inhibition by pepstatin and the strong effect of specific retropepsin inhibitors like Saquinovir and Amprenavir provides strong evidences supporting the nature of LegRP as a novel member of the retropepsin familiy.

3.3.4. Kinetic Parameters of LegRPsd-15

Enzyme kinetics is a fundamental study in the characterization of any enzyme. With the information provided by the study of an enzyme's kinetic parameters it is possible to conclude numerous features of the protein. In a simple approach it is possible to compare reaction rates and enzyme affinity towards a substrate in absolute values comparable to other known enzymes, under similar conditions. Later, if the reaction has a favorable rate and speed, it is even possible to conclude about the nature of a specific inhibition and calculate the number of active sites (Nelson, D., Cox, M. & Lehninger, A. 2008)

LegRPsd-15 could be purified in enough quantity to perform some kinetic studies. Assays were performed varying the substrate concentration while keeping the enzyme concentration stable. By measuring the initial velocity of the reaction with each concentration of substrate, it was possible to plot the values with a regression model of a Michaelis-Menten kinetics (figure 23).

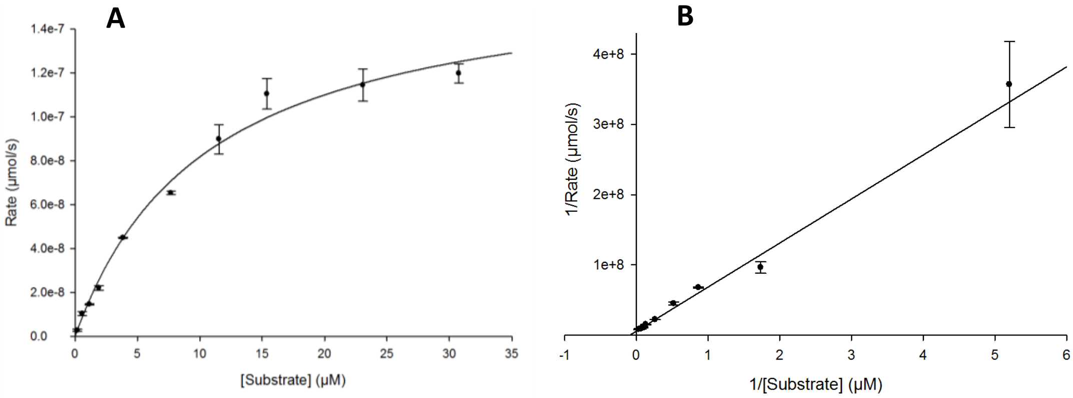


Figure 23. Effect of substrate concentration (typical AP substrate) on the initial velocity of recombinant LegRPsd. LegRPsd-15 activity was measured in the presence of different substrate concentrations. In A the initial velocity of the reaction is plotted in function of the substrate concentration. The points are fitted to a Michaelis-Menten regression model with a R^2 of 0,993. The plot shown is B is the Lineweaver-Burk plot of the same points, which is a transformation of the Michaelis-Menten plot into a linear regression. The calculated Km value is 10.53 ± 1.45 µM

It is possible to confirm that LegRPsd-15 closely follows a Michaelis-Menten

Kinetic model with high degree of confidence (R^2=0,993) with a Km of 10,53 µM and a V_{max} of 1.68e-7 µmol/s.

3.4. The D41A Active-Site Mutant of LegRPsd

As seen before, the proteolytic activity of an aspartic protease is only possible due to the presence of aspartate residues in the active site. These aspartate residues are fundamental to the electron transfer needed for the proteolysis. For this matter, an essential procedure in evaluating the nature of a protease activity (if it is an aspartic protease or if it belongs to another class) is to generate an active-site mutant, in which the putative catalytic aspartate is replaced by another amino-acid. In this case, the putative catalytic Asp41 was replaced by an alanine residue through site-directed mutagenesis of the pET28a vector containing the wt LegRPsd sequence. For the procedure it was used the Quickchange™ Site-directed Mutagenesis kit from Stratagene, with the mutation included in specific primers.

After generation of the mutated vector and confirmed by DNA sequencing, this was transformed in BL21 star *E. coli* cells. A small scale expression was performed in order to compare the expression of the mutated protein with the recombinant wild-type. Samples collected at the end of 3h expression at 37ºC (the optimal conditions for the recombinant LegRPsd expression were used in the expression of both mutated and wild-type) were analyzed by SDS-PAGE followed by Western blot (Figure 24).

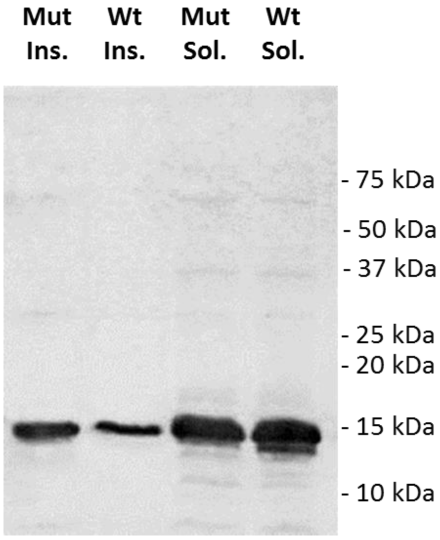


Figure 24. Western-blot analysis comparing recombinant LegRPsd (Wt) and D41A active-site mutant (Mut) expression in *E. coli*. Cells were grown to OD600nm = 0.7 and induced with 0.05mM IPTG, then expression was carried out at 37ºC for 3h. Cell extracts were separated in cytoplasmic soluble (Sol.) and Insoluble (Ins.) fractions and analyzed by Western-Blot using an anti-His-Tag antibody.

By this simple analysis it was possible to see that the mutation does not change considerably the levels of expressed protein in the soluble cytoplasmic fraction. Also, it was clear that the presence of the lower molecular weight at 14kDa corresponding to LegRPsd-14 is much more faint in the mutant expression sample.

In order to conclude about the loss of activity by mutating the catalytic aspartate, it was necessary to purify the protein. The procedures used to purify the mutated protein were the same as with the optimized purification of the wild type recombinant LegRPsd.

The soluble fraction of the cell extract was injected in an HisTrap™ HP 5mL column (IMAC-Ni^{2+}), and the column was washed with 20 mM sodium phosphate buffer pH 7.5, 10 mM Imidazole, 0.5 M NaCl. Protein was eluted by three steps of increasing Imidazole concentrations – 50mM, 200 mM and 500 mM. Figure 25 shows the chromatogram obtained for the His Trap column.

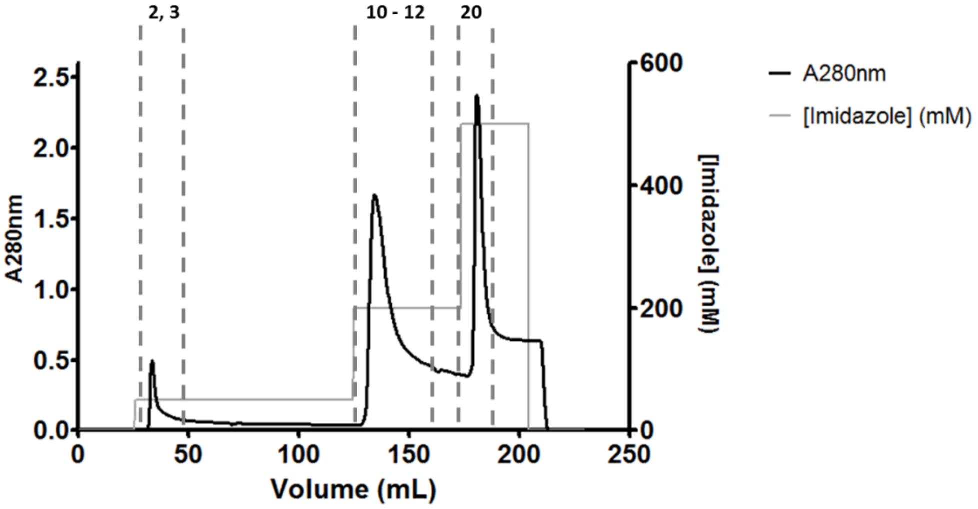


Figure 25. Purification of recombinant LegRPsd D41A mutant by IMAC-Ni$^{2+.}$. A280nm and Imidazole concentration are shown in function of volume. The column was equilibrated and washed with 10 mM phosphate buffer pH 7.4, 0,5 M NaCl, 10 mM Imidazole. Protein was eluted using 10 mM phosphate buffer pH 7.4, 0,5 M NaCl with three different steps of Imidazole concentration (50mM, 200mM and 500mM) and detected by its A280nm. Collected fractions numbers are shown at the top.

The fractions collected from the IMAC-Ni^{2+} chromatography were then analyzed by SDS-PAGE followed by Coomassie Blue staining to detect total proteins present and also analyzed by SDS-PAGE followed by Western Blot using an anti-His antibody to specifically detect the LegRPsd D41A mutant fused with the His-tag (Figure 26).

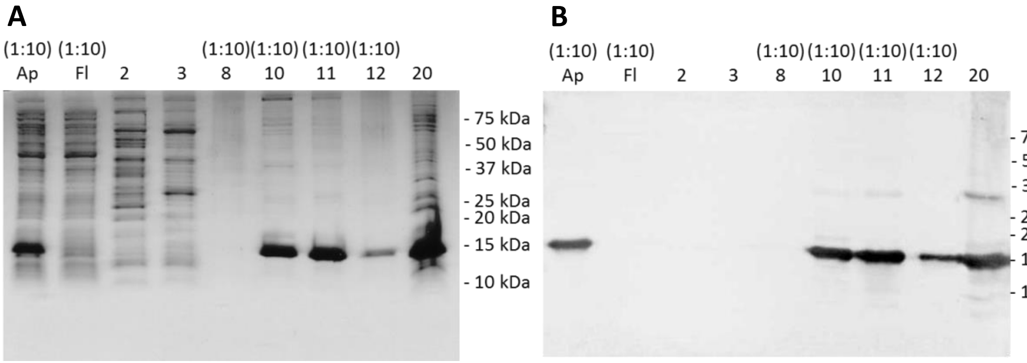


Figure 26. SDS-PAGE and Western-Blot analysis of the fractions elutedfrom the purification of LegRPsd D41A mutant by optimized IMAC-Ni^{2+}. The fraction numbers correspond to the numbers displayed in Figure 25, AP refers to the total soluble cell extract applied to the column, while FL refers to the flowthrough sample that was collected after sample injection (protein not-bound to the column). Samples diluted are indicated with the respective dilution.

A) SDS-PAGE analysis (12.5% polyacrylamide) stained with Coomassie Blue. 5 µL of each sample denatured with 6x loading buffer was loaded in the gel

B) Western-Blot analysis using an anti-his-tag antibody (1:10000). 3 µL of each sample denatured with 6x loading buffer was loaded in the gel for the SDS-PAGE prior to electrotransference.

It is possible to see that this chromatography profile is slightly different from the purification of the wild-type. On the mutant purification, a large pool of protein is eluted in the 500 mM Imidazole step. However, this pool is highly contaminated with other proteins in comparison with the pool eluted at 200 mM Imidazole. So, fractions eluted in the 200 mM Imidazole step were grouped and dialyzed against 20 mM MES pH 6.0, 50 mM NaCl: the same procedure carried in the purification of the wild-type.

The dialyzed pool was then applied to a MonoS column as in the wild-type purification and the same buffers were used for washing (20mM MES Buffer pH 6.0) and eluting (20mM MES buffer pH 6.0, with a gradient from 0 to 1M NaCl). The resulting chromatogram is shown in figure 27.

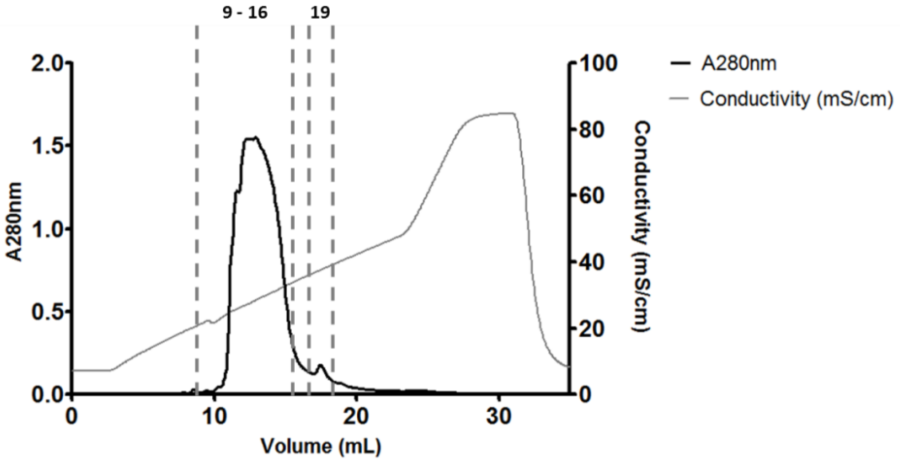


Figure 27. Chromatogram showing the elution phase of the LegRPsd D41A mutant purification by cation exchange chromatography. The column was equilibrated and washed using 20 mM MES buffer pH6.0 and protein was eluted from the column using 20 mM MES buffer pH 6.0 with a NaCl gradient from 0 to 1 M and detected by its A280nm. The conductivity values are proportional to the concentration of NaCl. Collected fractions numbers are shown at the top.

The fractions collected from the cation exchange chromatography were then analyzed by SDS-PAGE followed by Coomassie Blue staining to detect total proteins present and also analyzed by SDS-PAGE followed by Western Blot using an anti-His antibody to specifically detect the LegRPsd D41A mutant fused with the His-tag (Figure 28).

From the purified fractions one of the immediate observations is the absence of the 14kDa form. The only fraction containing a small amount of this lower molecular weight form is fraction 19, which compared to the wild type, is much less intense. Surprisingly, in fractions 13 and 14 we can see the clear presence of a band at 16kDa, clearly detected by the anti-His antibody as it is possible to see in the Western blot. When these fractions are analyzed in an SDS-PAGE stained with Coomassie-Blue in parallel with wild-type fractions, it was clear that this “form” of ~16kDa has higher molecular weight than LegRPsd-15 (Figure 29).

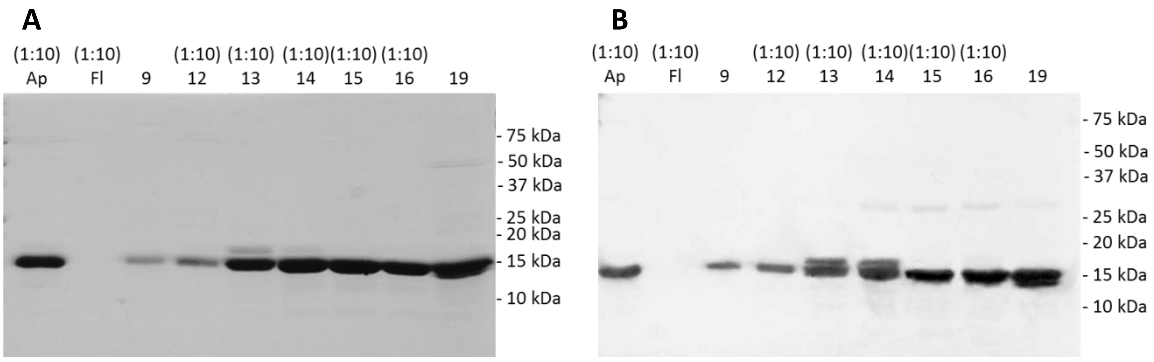


Figure 28. SDS-PAGE and Western-Blot analysis of the fractions from purification of the LegRPsd D41A mutant by cation exchange chromatography. The fraction numbers correspond to the numbers displayed in Figure 27, AP refers to the total soluble cell extract applied to the column, while FL refers to the flowthrough sample that was collected after column injection (protein not-bound to the column). Samples diluted are indicated with the respective dilution.

A) SDS-PAGE analysis (12.5% polyacrylamide) stained with Coomassie Blue. 5 µL of each sample denatured with 6x loading buffer was loaded in the gel

B) Western-Blot analysis using an anti-his antibody. 3 µL of each sample denatured with 6x loading buffer was loaded in the gel for the SDS-PAGE prior to electrotransference.

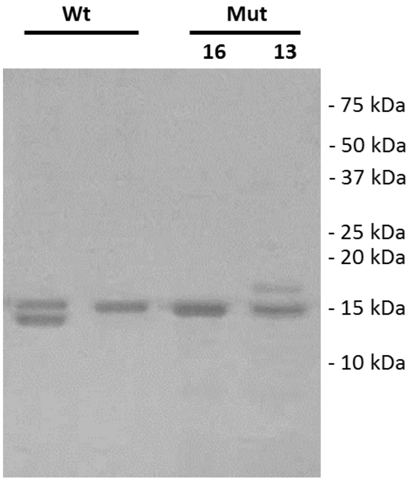


Figure 29. SDS-PAGE analysis comparing fractions eluted from the cationic exchange during the purification of the wild-type LegRPsd (Wt) and during the purification of the D41A mutant (Mut). Samples loaded from left to right are: sample of LegRPsd-14 + LegRPsd-15 (Wt); sample of LegRPsd-15 (Wt); sample from fraction 16 of mutant purification; sample from fraction 13 of mutant purification. Mutant samples numbered according to Figure 28. 5 µL of each sample diluted 1:20 denatured with 6x loading buffer was loaded in the gel.

The activity of the fractions 13 and 16 collected during the purification of the mutant was analyzed. The same amount of protein was added in every test and the activities of the fractions containing the mutated protein were compared to the activity of LegRPsd-15 (Figure 30).

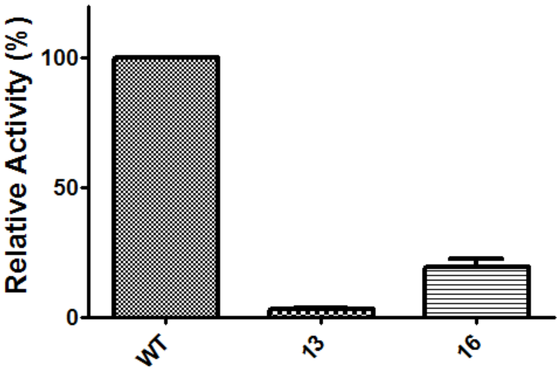


Figure 30. Activity comparison between LegRPsd-15 (WT) and fractions 13 and 16 collected from the cation exchange chromatography of the LegRPsd D41A mutant purification. The activity was tested towards the fluorogenic typical substrate for APs [MCA-K]-K-P-A-E-F-F-A-L-[K-DNP]. The assays were performed by incubating purified samples at 37^{o}C with 50 mM sodium acetate buffer,100 mM NaCl, pH 4.0 and measuring emitted fluorescence per second.

It is curious to observe that fraction 16, containing only the band correspondent to the 15kDa form, still retains about 20% of proteolytic activity towards the fluorescent substrate. However, fraction 13, containing both the 15kDa form and the newly detected 16kDa form, has no activity towards the substrate. Either way, it is clear that LegRPsd relies heavily on Asp41 for its proteolytic activity providing further strong evidence that this enzyme is an active retropepsin-like protease. Nevertheless, one cannot exclude that the activity present in fraction 16 may result from different catalytic phenomena not related (at least directly) to the active site catalytic triad.

In the D41A mutant purification, the amount of LegRPsd-14 D41A is only residual when compared to the wild-type. This means that the auto-processing capacity of the protein was indeed affected by the mutation of the active site, and this can be an argument against the existence of other kind of active center responsible for the autoprocessing, leading towards the possibility of conformational changes leading to different properties. Moreover, there is the presence of a new form at 16kDa that was not detectable in the purification of the wild-type LegRPsd. This 16kDa form (from now on referred as LegRPsd-16) is likely the precursor of LegRPsd-15. However, the processing from LegRPsd-16 to LegRPsd-15 may occur immediately after translation, or may have a very fast processing rate, therefore not being present when wt protein is extracted. With the mutation of the active-site, this processing rate was probably affected and this precursor form was still detected. Nevertheless, LegRPsd-15 is the most prevalent when the mutant is purified, clearly suggesting that the processing rate is still very high, and that this apparently first processing step appears to be highly independent of the active-site. Therefore, determining the N-terminal sequence of each of these protein products - LegRPsd-16, LegRPsd-15, LegRPsd-14, will be critical to further understand LegRP processing.

3.5. Native Molecular Weight and Oligomerization States

Retroviral-like aspartic proteases are known for the need to form homodimer structures in order to form its catalytic center. The bacterial retroviral-like aspartic protease homologues contain only one DTG triad and aspartic proteases catalytic center needs two aspartate residues in the catalytic center in order to catalyze peptide proteolysis, therefore, these bacterial retropepsin-like proteases would also need to dimerize in order to have proteolytic activity by a similar mechanism. For this matter, an essential part of LegRPsd characterization would be the analysis of native molecular weight and formation of homodimer structures. After recombinant LegRPsd purification by IMAC-Ni^{2+}, in fractions containing large amounts of LegRPsd-15 and LegRPsd-14, it was possible to detect a faint band labeled by the anti-His antibody around 30 kDa. This lead to the first observation that homodimer structures were present and for some reason (maybe the large amount of protein present), the structures were resistant to denaturing agents.

3.5.1. Molecular Weight Determination by Analytical Size-Exclusion Chromatography

With the previous data, samples of purified LegRPsd-15 and LegRPsd-14 were analyzed by analytical size-exclusion chromatography (Figure 31). According to the calibration performed to the column with standard molecular weight proteins, the estimated molecular weight of the proteins present in the samples were respectively 15 kDa and 14 kDa, showing no presence of protein at 30 kDa. This would mean that LegRPsd was mainly found in the monomeric form, regardless of the processing products.

It isn't surprising that the homodimer structure is a labile structure, probably only forming itself to promote proteolytic activity, much like it has been described for retropepsins (Tang et al. 1978). For this, pepstatin A was added to the sample and the elution buffer and again homodimer presence was checked through the same method. The logic behind this was that pepstatin A inhibits aspartic proteases by competing with the substrate by binding itself to the active site, and as an inhibitor of LegRPsd-15, it would bind to the active site and favor the formation of the homodimer structure. However, no protein at 30kDa was detected either way (Figure 31).

Facing this challenge, a more robust approach was needed to verify the presence of the homodimer structures. For that matter, the devised strategy was to express the LegRPsd sequence fused with an His-tag (LegRPsd-HIS) as it was being done, but at the same time co-express the same LegRPsd sequence but fused with a different tag, an HA-tag (LegRPsd-HA). By co-expressing LegRPsd-HIS and LegRPsd-HA and co-purifying LegRPsd-HIS using methods for specific purification of His-tag fused proteins, it should be possible to detect LegRPsd-HA molecules during purification if a dimeric structure would be present.

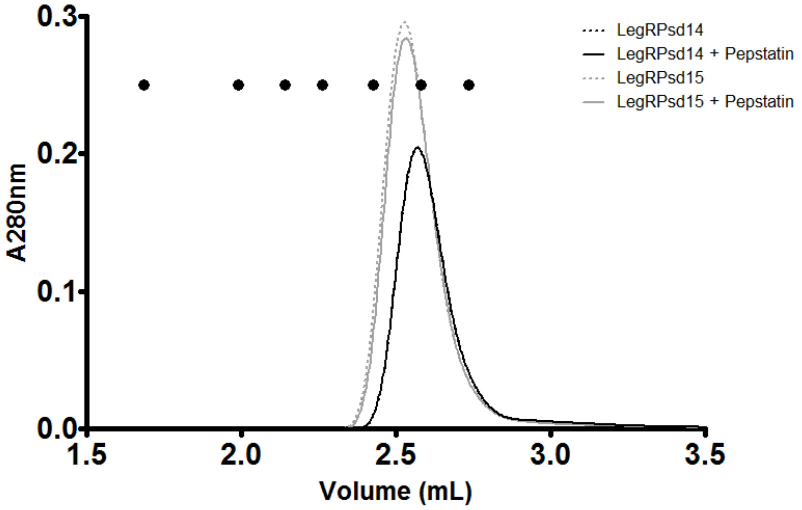


Figure 31. Analytical size-exclusion chromatography of LegRPsd-15 and LegRPsd-14 purified samples. The samples were run in a Superdex 200 10/30 in 50 mM sodium acetate buffer pH4.0, 100 mM NaCl. The fractions including Pepstatin were incubated with 1 µM Pepstatin A and 1 µM Pepstatin A was added to the running buffer. The black dots refer to the standard molecules for molecular weight determination, from right to left: Aprotinin 6.5kDa, Ribonuclease 13.7kDa, Carbonic Anhydrase 29kDa, Ovalbumin 43kDa, Conalbumin 75kDa, Aldolase 158kDa, Ferritin 440kDa.

3.5.2. Cloning LegRPsd in a Co-Expression vector with Different Fusion Tags

In order to co-express LegRPsd with the two different fusion tags it is useful to use a vector that allows the co-expression of two different sequences. For this, the pRSF Duet™-1 DNA (EMD Millipore) was used. This vector encodes two multiple cloning sites (MCS) each of which is preceded by a T7 promoter, lac operator, and ribosome binding site (rbs). To include the HIS-tag or HA-tag followed by a stop codon in the C-terminus of the sequence, the LegRPsd optimized sequence was amplified by PCR with the necessary tag sequence in the Reverse primer. With this, LegRPsd-HIS amplified sequence was cloned in the first MCS using *Nco*I and *Not*I restriction sites and LegRPsd-HA amplified sequence was cloned in the first MCS using *Nde*I and *Xho*I restriction sites.

3.5.3. Screening the Co-Expression of LegRPsd-HIS and LegRPsd-HA

The pRSF-Duet vector containing the two sequences was transformed in BL21star cells, and as this was a new expression vector it would be necessary to test which would be the best conditions for expressing both forms of the protein. For that, cells were grown to OD600nm = 0.7 in LB medium with 50 µg/mL Kanamycin and then expression for 3h was compared using two different temperatures (30^{o}C and 37^{o}C) and two different IPTG concentrations for induction (0.1 mM and 0.05 mM). Samples extracted from the cytoplasm of the cells expressing the recombinant proteins were analyzed in a SDS-PAGE followed by Western-Blot using two different primary antibodies: anti-His and anti-HA (Figure 32).

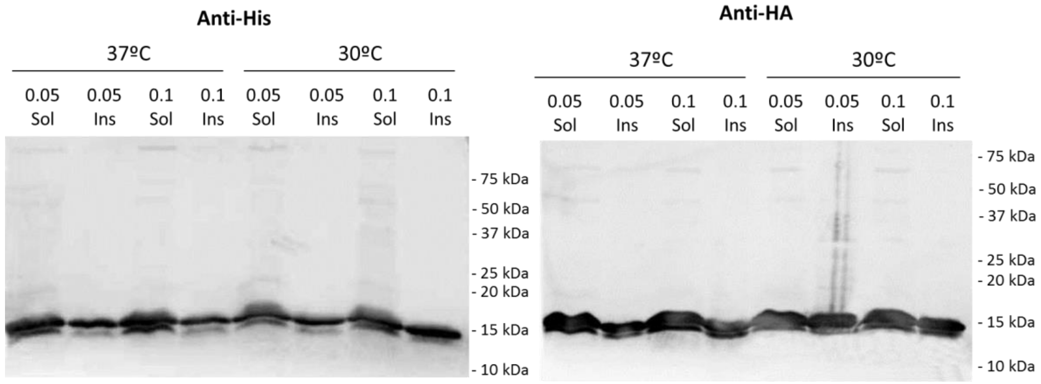


Figure 32. Western-Blot analysis of recombinant LegRPsd-HIS and LegRPsd-HA co-expression screening. After protein expression during 3h in LB medium in the tested conditions, cell extracts were analyzed by Western-Blot using an anti-His (left) and anti HA (right) antibodies. The co-expression was tested at two different temperatures (30^{o}C and 37^{o}C) and the induction was made with two different concentration of IPTG: 0,05 (mM) and 0,1 (mM). For each harvested sample the soluble (Sol) and Insoluble (Ins) fractions were analyzed. (Ant-His and Anti-HA diluted 1:10000)

The chosen conditions for the scale-up expression were the use of 0.05mM IPTG at 37^{o}C, the same conditions used for expressing the LegRPsd sequence in pET28a.

3.5.4. Co-Expression and Co-Purification of LegRPsd-HIS and LegRPsd-HA by Standard Chromatographic Methods

The BL21 star cells properly transformed with the pRSF vector containing the two sequences for expression of LegRPsd-HIS and LegRPsd-HA were cultivated at 37ºC in separate flasks, each containing 1 L of LB culture medium with 50 µg/mL Kanamycin. At an OD600nm of 0.7 the protein expression was induced by addition of IPTG and protein was expressed for 3 hours at 37ºC.

After expression, the cells were harvested and frozen at -20ºC overnight with the purpose of lysing the *E. coli* cells by a freeze-thaw cycle.

The first attempt for co-purifying LegRPsd-HIS and LegRPsd-HA was done by the same chromatographic techniques used in the standard purification of LegRPsd. The first purification step was made using an IMAC-Ni^{2+} using the 50mM, 200mM and 500mM Imidazole elution steps. The resulting chromatogram is shown in Figure 33. Collected fractions were analyzed by SDS-PAGE followed by Western-Blot (Figure 34). For the Western-blot two different antibodies were used for labeling: anti-His and anti-HA. This way it would be possible to monitorize whether the LegRPsd-HA molecules were co-purified with the LegRPsd-HIS molecules or not.

It is possible to see by the chromatogram that all the protein was eluted either at 50mM Imidazole or at 200 mM Imidazole. However, by the Western-blot it is clear that most of the LegRPsd-HA was lost in the flowtrough, not bound to the LegRPsd-HIS molecules that were bound to the IMAC-Ni^{2+} column. Even considering this great loss of LegRPsd-HA, it is possible to distinguish some faint bands labeled with anti-HA antibody in the fractions where LegRPsd-HIS was eluted, indicating that some LegRPsd-HA molecules were eluted with the LegRPsd-HIS. It is also surprising to see a large amount of LegRPsd-HIS in the flowthrough fraction. As the A280nm values of the peak in which the protein was eluted are not as high as in the standard purification procedure, the column does not seem to be saturated. Maybe LegRPsd-HIS was bound to LegRPsd-HA and therefore the His-tag would be less exposed to bind to the column.

This can be considered an evidence of LegRPsd-HIS/LegRPsd-HA dimer formation.

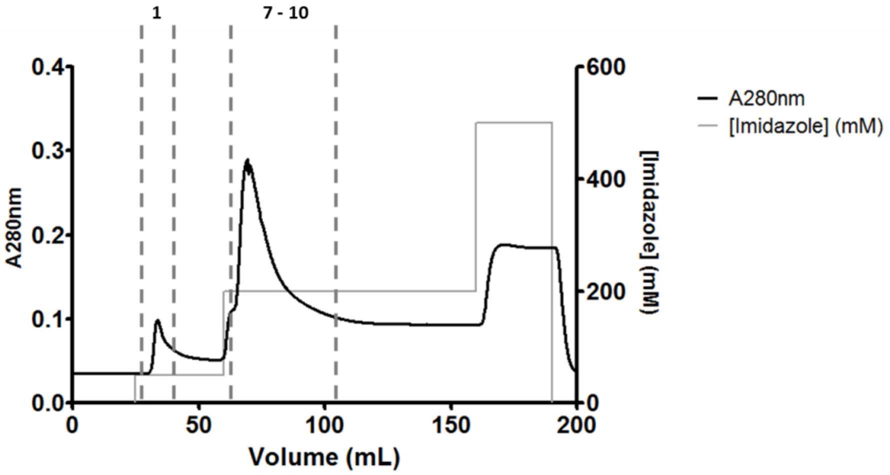


Figure 33. Purification of LegRPsd-HIS and LegRPsd-HA co-expression by IMAC-Ni^{2+}. A280nm and Imidazole concentration are shown in function of volume. The column was equilibrated and washed with 10mM phosphate buffer pH 7.4, 0.5 M NaCl, 10 mM Imidazole. Protein was eluted using 10mM phosphate buffer pH 7.4, 0.5 M NaCl with three different steps of Imidazole concentration (50mM, 200mM and 500mM) at a flow rate of 5 mL/min and detected by its A280nm. Collected fractions numbers are shown at the top.

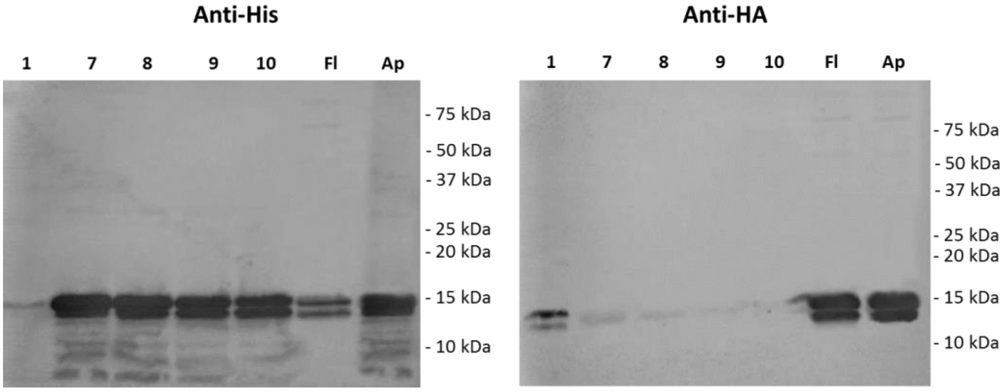


Figure 33. Western-Blot analysis of the fractions eluted in the purification of LegRPsd-HIS and LegRPsd-HA by IMAC-Ni^{2+}. The fraction numbers correspond to the numbers displayed in figure 32, AP refers to the total soluble cell extract applied to the column, while FL refers to the flowthrough sample that was collected after sample injection (protein not-bound to the column). Samples diluted are indicated with the respective dilution. Two different primary antibodies were used for labeling: anti-His-Tag (left) and anti HA-tag (right) antibodies. 3 µL of each sample diluted 1:10 denatured with 6x loading buffer was loaded in the gel for the SDS-PAGE prior to electrotransference. (Anti-His and Anti HA diluted 1:10000)

The LegRPsd-HA presence in eluted LegRPsd-HIS fractions and the presence of LegRPsd-HIS in the flowthrough clearly suggests the formation of dimeric structures between the two molecules. However, it would be needed a stronger evidence for this interaction. In an attempt to concentrate these fractions and verify the presence of the homodimers, the collected pools from the 200mM Imidazole eluted fractions were dialyzed and injected in a cation exchange column just like what was done in the purification of the recombinant LegRPsd for biochemical studies. The chromatography was done by the same procedure as in the purification of the recombinant LegRPsd and the resulting chromatogram in shown in Figure 35.

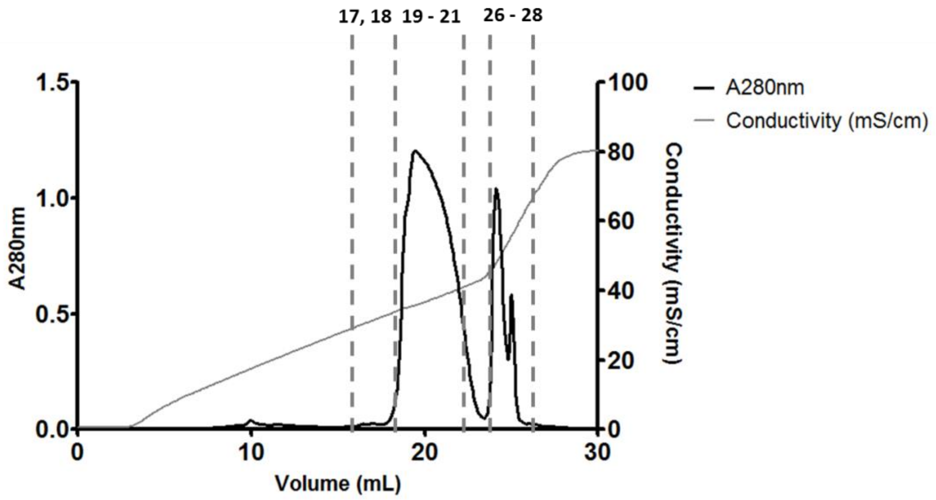


Figure 35. Chromatogram of LegRPsd-HIS and LegRPsd-HA co-purification by cation exchange chromatography. The column was equilibrated and washed using 20mM MES buffer pH6.0 and protein was eluted from the column using 20mM MES buffer pH 6.0 with a NaCl gradient from 0 to 1M at a flow rate of 0.75 mL/min and detected by its A280nm. The conductivity values are proportional to the concentration of NaCl. Collected fractions numbers are shown at the top.

Collected fractions were analyzed by SDS-PAGE followed by Western-Blot (Figure 36). For the Western-blot again two different antibodies were used for labeling: anti-His and anti-HA.

From the analysis of the Western-blot and the chromatogram it was possible to conclude that although there are fractions where both LegRPsd-HIS and LegRPsd-HA are co-eluted (fractions 17, 18 and a faint signal in fraction 19), most fractions corresponding to LegRPsd-HIS display no signal for LegRPsd-HA.

As this would not constitute a strong evidence to prove the existence of homodimer structures of LegRPsd-HIS and LegRPsd-HA, other methods were used.

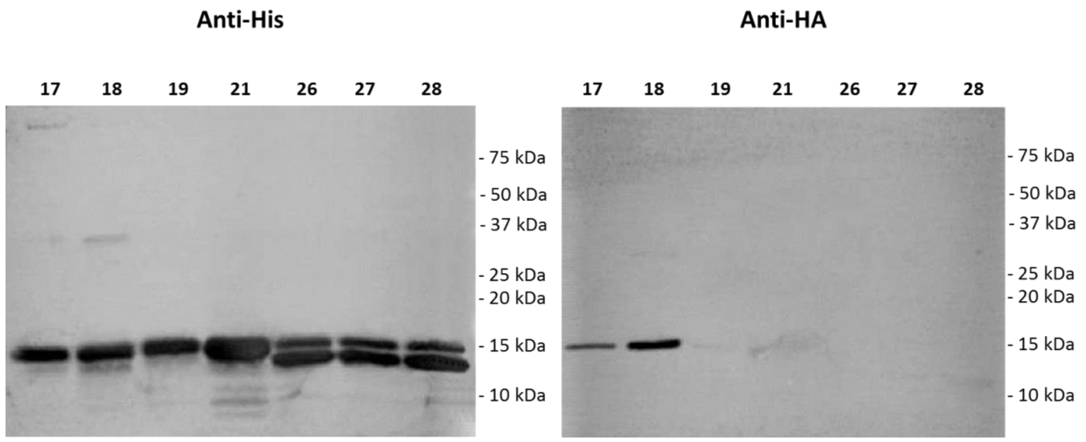


Figure 36. Western-blot analysis of fractions eluted from the LegRPsd-HIS and LegRPsd-HA co-purification by cation exchange chromatography. The fraction numbers correspond to the numbers displayed in Figure 35, AP refers to the total soluble cell extract applied to the column, while FL refers to the flowthrough sample that was collected after sample injection (protein not-bound to the column). Two different primary antibodies were used for labeling: anti-His (left) and anti HA (right) antibodies. 3 µL of each sample diluted 1:10 denatured with 6x loading buffer was loaded in the gel for the SDS-PAGE prior to electrotransference. (Anti-His and anti-HA were diluted to 1:10000)

3.5.5. Co-Purification of LegRPsd-HIS and LegRPsd-HA by HisMag Sepharose Ni Magnetic Beads

As IMAC-Ni^{2+} was not fully informative regarding co-purification of LegRPsd-HIS/LegRPsd-HA dimers, another method was tested. For this experiment, HisMag Sepharose Ni^{2+} Magnetic Beads (GE Healthcare) were used. These beads contain immobilized Ni^{2+} Ions and rely on the same principle as the IMAC-Ni^{2+}. However, this is a small scale high sensitivity method, therefore more appropriate for co-purification of LegRPsd-HIS/LegRPsd-HA dimers, if present. For this method, the co-expression was made as previously described.

As the homodimer structures formed by LegRPsd seem to be very labile, a control reaction was performed in which half the volume of the cytoplasmic fraction was incubated at 37ºC for 5 min with glutaraldehyde. Glutaraldehyde is a protein crosslinking agent, which usually proves itself very useful in detecting weak protein interactions. If the proteins are at a distance lower than ~7Å, glutaraldehyde is capable of binding the proteins by covalent interactions (Salem et al. 2010; Crisona & Cozzarelli 2006).

Both the cross-linked and non-cross-linked fractions were then incubated with the HisMag Sepharose Ni^{2+} Magnetic Beads for 2h at 4ºC. The beads were then washed with 20mM sodium phosphate buffer pH 7.5 , 10mM Imidazole, 0.5M NaCl for three times and then eluted with 20mM sodium phosphate buffer pH 7.5 , 500mM Imidazole, 0.5M NaCl also for three times. All washing and eluted fractions were analyzed by SDS-PAGE followed by Western-Blot using anti-His and anti-HA primary antibodies (Figure 37).

It is clear the presence of a band at 30kDa in the eluted fractions of the crosslinking experiment, immunodetected with both antibodies. These results confirm the presence of LegRPsd-HIS/LegRPsd-HA dimeric structures. Even in the non-crosslinked eluted fractions it is possible to find the homodimer structures, although the labeling with anti-HA on the non-crosslinked fractions is of very low intensity. It was also very interesting to note a band in the crosslinked eluted fractions at 60kDa labeled with both anti-His and anti-HA antibodies. This could suggest the crosslinking of tetramer structures, which was already reported for the HIV-1 protease, although thought to be a nonspecific association due to the presence of high concentration of protein and with no relevance for *in vivo* function (Holzman et al. 1991).

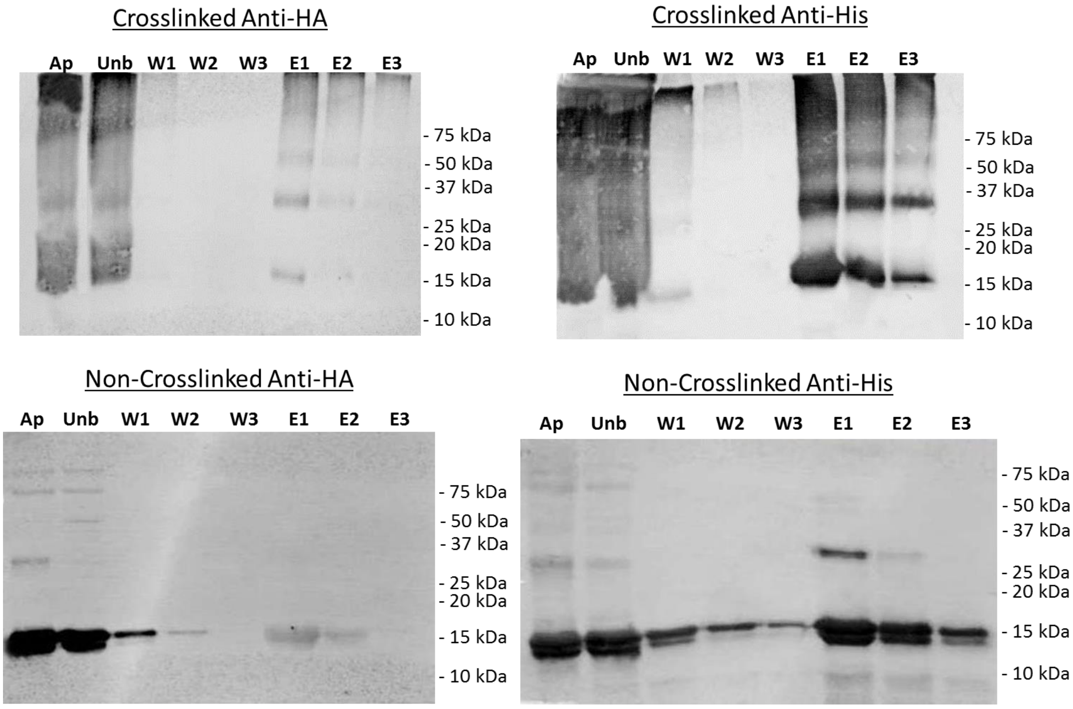


Figure 37. Western-Blot analysis of the washing and elution fractions collected during the co-purification of LegRPsd-HA and LegRPsd-HIS by HisMag Sepharose Ni^{2+} Magnetic Beads. Ap refers to the total soluble cell extract applied to the magnetic beads, while Unb refers to the protein not-bound to the beads. Cell extracts where the crosslinking reagent glutaraldehyde was used are labeled as crosslinked . W1, W2 and W3 refer to the first, second and third washing steps respectively while E1, E2 and E3 refers to the first, second and third elution steps respectively. Two different primary antibodies were used for labeling: anti-His (left) and anti-HA (right) antibodies diluted 1:10000. 3 µL of each sample diluted 1:10 denatured with 6x loading buffer was loaded in the gel for the SDS-PAGE prior to electrotransference.

To confirm the molecular weight of the structures detected by Western-Blot, elution fractions of crosslinked and non-crosslinked (E1) assayswere analyzed by analytical size-exclusion chromatography in a Superdex 200 10/30 (resulting chromatogram shown in Figure 37). In the non-crosslinked sample only one peak corresponding to the elution of a protein at 15kDa was observed, consistent with the presence of LegRPsd monomers. This may be due to the weak interaction between the LegRPsd molecules to form homodimeric structures. However, the crosslinked sample showed a different elution profile with the presence of a peak corresponding to the monomeric 15kDa form as well as a broad shoulder corresponding at the elution of a ~30kDa structure. This shoulder is of difficult perception likely due to the presence of several protein complexes with higher molecular weights, due to non-specific crosslinking (as visible in the Western-blot analysis).

With this complementary data, there is strong evidence that LegRPsd exists in the form of a homodimer structure like it was predicted, with similarity to retropepsins, which suggest that this structure is necessary for the protease activity.

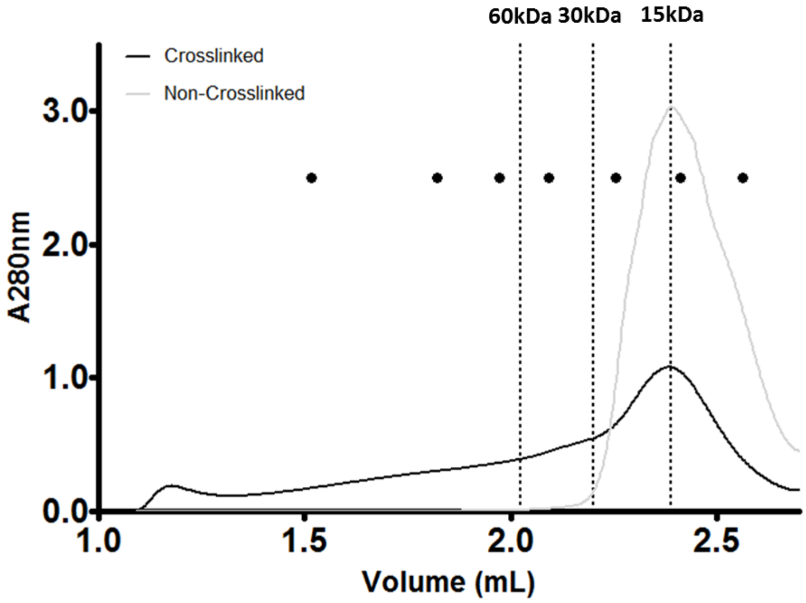


Figure 37. Analytical size-exclusion chromatography of elution samples (E1) from HisMag Sepharose Ni Beads purification. The samples were applied in a Superdex 200 10/30 equilibrated in 20mM phosphate buffer pH 7.5. The black dots refer to molecular weight of protein standards used for column, from right to left: Aprotinin 6.5kDa, Ribonuclease 13.7kDa, Carbonic Anhydrase 29kDa, Ovalbumin 43kDa, Conalbumin 75kDa, Aldolase 158kDa, Ferritin 440kDa. The dotted lines correspond to the extrapolated molecular weight values of 60kDa, 30kDa and 15kDa corresponding to the tetramer, dimer and monomer structures of LegRPsd.

4. Conclusions

The soluble domain of the retropepsin-like aspartic protease homologue from *L. pneumophila* proved to be a protein with high yields of heterologous expression in *E. coli*, allowing the purification of large amounts of pure protein in useful time for protein-quantity demanding experiments, like structure determination by X-ray diffraction. Furthermore, the expressed protein was shown to be an active enzyme allowing its biochemical characterization.

During purification LegRPsd presented itself in at least two distinct forms with different molecular weights. Concerning this fact, it was verified that the enzyme is capable of *in vitro* auto-processing at pH6.0. In fact, the results from the D41A mutant suggest the possibility of a multi-step auto-processing, which would start with a ~16 kDa form (LegRPsd-16) which would be processed to a ~15 kDa form (LegRPsd-15) inside the *E. coli* cell quickly after translation, and then maturating to a ~14 kDa form (LegRPsd-14) through a slower reaction. Although we cannot exclude the participation of an endogenous *E. coli* protease in the processing of LegRPsd, the existence of a transmembranar retropepsin homologue with multi-step auto-processing would not be an exclusive feature of LegRP, as this phenomenon is also described for SASPase (Bernard et al. 2005) and the *Rickettsia* retropepsin-like protease (unpublished data), even when only the soluble domain of these proteins is expressed. From this we could start anticipating LegRPsd similarity with retropepsins, as the autoprocessing ability is one of the most important characteristics to consider about retropepsins.

When tested towards fluorescent substrates, both LegRPsd-15 and LegRPsd-14 were able to cleave a typical AP substrate, but not any other tested fluorescent substrate, and although its auto-activation occurs at pH 6.0, proteolytic activity towards the typical AP substrate was only relevant at pH values between 3.0 and 4.5, with an optimum pH value of 4.0, a value among the usual optimum pH values for retropepsins and retropepsin-like proteases, which usually range between pH 4.0 and 6.0. However, both forms showed different inhibition profiles, and while LegRPsd-15 was inhibited by about 50% in the presence of pepstatin (by which LegRPsd-14 activity was not affected), LegRPsd-14 was much more sensitive to retropepsin-specific inhibitors, especially saquinovir and amprenavir. This inhibition for HIV protease specific inhibitors suggests a structural proximity between the active site of retropepsins like HIV and MLV and LegRPsd. The change in the inhibition profile seen by the maturation of LegRPsd-15 to LegRPsd-14 indicates that the cleavage of the LegRPsd-15 N-terminal probably leads to a structural change, which somehow affects the activity of the protein. However, it does not seem to represent a significant change in cleavage site preferences, as LegRPsd-14 retains the same substrate preference, even when the peptides resulting from cleavage of oxidized insulin β-chain are analyzed. It should be noted that concerning the evolutive relations between these new retropepsin-like enzymes and retropepsins, these changes in activity through the multi-step auto-processing has significant evolutionary implication. In this case, auto-processing seems to be necessary for more than just cleavage of the transmembrane sequence at the N-terminus, which should be made in the first processing step (from LegRPsd-16 to LegRPsd-15). The processing of LegRPsd-15 to LegRPsd-14 seems to bring a more functional change, much like the changes observed through the maturation of HIV retropepsin.

Partial pepstatin inhibition of LegRPsd-15 and lost of proteolytic activity by the D41A mutant seem to indicate that LegRPsd is an aspartic protease, and biochemical properties similar to retropepsins and other studied retropepsin-like enzymes bring LegRPsd close to the A2 family. Furthermore, LegRPsd was shown by different methods to be able to form homodimeric structures. The formation of homodimers is a fundamental characteristic in any retropepsin or retropepsin-like protease, as this structure is needed for the formation of an active center.

Together with experimental evidences of other retropepsin-like enzymes, LegRP brings a new light to the emerging theory of pepsin and retropepsin evolution. These new evidences help building-up the theory that retropepsin-like proteins exist long before the retropepsins found in retrovirus, and they evolved with bacteria into eukaryotes, only to be captured by virus later on.

However, validation of a protease is not done with some few results, and some issues concerning biochemical characterization still need to be addressed. It is fundamental do undergo studies in order to determine LegRPsd cleavage specificity, in order to determine the preference of cleavage sites, its changes along maturation, and compare it to other retropepsin and retropepsin-like proteases. This will provide important information as substrate specificity is closely related to catalytical pocket structure, which is another important question to be addressed. The possibility to produce pure LegRPsd in high yields makes it a great advantage for crystallography and structure determination by X-ray diffraction. Determination of these prokaryote retropepsin-like enzymes structure will allow us to understand their structural proximity with eukaryotic retropepsin-like enzymes and viral retropepsins, providing most necessary information to address questions regarding retropepsin evolution.

Other analyzes of most importance regarding these evolutionary theories would consist in studying the molecular evolution of the legRP gene by analyzing the gene sequence through several *L. pneumophila* strains, both environmental and pathogenic. The purpose of this would be to address whether this protein was being conserved by the *Legionella* or if it is undergoing selective pressure. A coherent molecular evolution study could exclude the possibility of a “recent” horizontal gene transfer from eukaryotes or retrovirus.

Another important issue to be addressed in the terms of retropepsin-like proteases would be to determine the function of these bacterial retropepsin-like proteases in the bacteria cells. We have seen that these sequences seem to be conserved, even in species with high genome selectivity like *Rickettsia*, however, finding the function of these proteins would provide invaluable information towards determining the bacteria’s need for them. Also, these proteins could be involved in new bacterial pathways and may be related to the bacteria’s life cycle, infection, or any other cellular function. Understanding the function of these retropepsin-related proteins in bacteria can be of most importance for the fields of bacteriology, enzymology and even to develop new specific therapies for dangerous infections like Legionelosis.

To dissect the *in vivo* function of LegRP, knockout studies will be fundamental in order to assess the importance of the enzyme for the *Legionella* viability and ability to infect hosts. Whether the protein is related to infection or not, understanding its expression and mRNA presence through different *Legionella pneumophila* strains, both pathogenic and non-pathogenic, and through the different phases of its life cycle could also be very interesting, as it would allow us to understand its role in cell regulation and its regulation by other bacterial cell pathways.

Bibliography

Altschul, SF et al., 1990. Basic local alignment research tool. *The Journal of molecular biology*, 215(3), pp.403-10.

Andersson, S. & Zomorodipour, A., 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, 396, pp.133–143.

Andreeva, N.S. 2003. Special Features of the Three-Dimensional Structure Defining Properties of Aspartic Proteases. *Russian Journal of Bioorganic Chemistry* 29, 453-456.

Bernard, D. et al., 2005. Identification and characterization of a novel retroviral-like aspartic protease specifically expressed in human epidermis. *J Invest Dermatol.*125(2):278-87

Brower, E.T., et al. Inhibition of HIV-2 protease by HIV-1 protease inhibitors in clinical use. *Chemical Biology & Drug Design.* 71;298-305

Chen, J.C. et al., 2006. Cytokinesis signals truncation of the PodJ polarity factor by a cell cycle-regulated protease. *The EMBO journal*, 25(2), pp.377–86.

Combet, C. et al., 2000. NPS@: Network Protein Sequence Analysis. *TIBS,* 25(291) 147-150

Costa, J. et al., 2010. Molecular evolution of Legionella pneumophila dotA gene , the contribution of natural environmental strains. *Environmental Microbiology*. 12(10):2711-29

Crisona, N.J. & Cozzarelli, N.R., 2006. Alteration of Escherichia coli topoisomerase IV conformation upon enzyme binding to positively supercoiled DNA. *The Journal of biological chemistry*, 281(28), pp.18927–32.

Dennis, P.J. et al. 1984. A note on the temperature tolerance of Legionella *Journal of Applied Microbiology,* 56(2) 349-350

Dunn, B. et al., 1986. A systematic series of synthetic chromophoric substrates for aspartic proteinases *Biochem J.* 237(3): 899–906

Dunn, B. et al., 2002. Retroviral proteases. *Genome Biology*, 3(4), pp.reviews3006.1–reviews3006.7.

Ensminger, A.W. et al., 2012 Experimental Evolution of Legionella pneumophila in Mouse Macrophages Leads to Strains with Altered Determinants of Environmental Survival. *PLoS Pathog* 8(5): e1002731

Fehér A et al. 2006 Characterization of the murine leukemia virus protease and its comparison with the human immunodeficiency virus type 1 protease. *J Gen Virol.* 87(Pt 5):1321-30.

Flexner, C., 2007. HIV drug development: the next 25 years. *Nature Reviews* 6(12):959-66

Freed, E.O., 2001. HIV-1 replication. *Somatic cell and molecular genetics*, 26(1-6), pp.13–33.

Gasteiger E. et al., 2005 Protein Identification and Analysis Tools on the ExPASy Server; (In) John M. Walker (ed): *The Proteomics Protocols Handbook*, Humana Press. pp. 571-607

Goodenow, M.M. et al., 2002. Naturally occurring amino acid polymorphisms in human immunodeficiency virus type 1 (HIV-1) Gag p7(NC) and the C-cleavage site impact Gag-Pol processing by HIV-1 protease. *Virology*, 292(1), pp.137–49.

Grant, S.K. et al., 1991. Purification and biochemical characterization of recombinant simian immunodeficiency virus protease and comparison to human immunodeficiency virus type 1 protease. *Biochemistry*, 30 (34), pp 8424–8434

Griffiths, J.T. et al., 1992. Different requirements for productive interaction between the active site of HIV-1 proteínase and substrates containing -hydrophobic*hydrophobic- or -aromatic*pro- cleavage sites. *Biochemistry*, 31(22), pp.5193–200.

Hofmann, A & Stoffel, W, 1993. TMbase - A database of membrane spanning proteins segments. *Biological Chemistry Hoppe-Seyler*, 374, p. 166.

Holzman, T.F. et al., 1991. Inhibitor stabilization of human immunodeficiency virus type-2 proteínase dimer formation. *The Journal of biological chemistry*, 266(29), pp.19217–20.

Ido E, et al. 1991 Kinetic studies of human immunodeficiency virus type 1 protease and its active-site hydrogen bond mutant A28S. *J Biol Chem*. 266(36):24359-66

Imamura, D. et al., 2008. Evidence that the Bacillus subtilis SpoIIGA protein is a novel type of signal-transducing aspartic protease. *The Journal of biological chemistry*, 283(22), pp.15287–99.

Krieg, N., et al., 2005. *Bergey's Manual of Systematic Bacteriology: The Proteobacteria*, Springer.

Krogh, A et al., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *The Journal of molecular biology*,305, pp. 567-80.

Krylov, D.M. & Koonin, E. V, 2001. A novel family of predicted retroviral-like aspartyl proteases with a possible key role in eukaryotic cell cycle control. *Current biology : CB*, 11(15), pp.R584–7.

Li, M. et al., 2011. Crystal structure of XMRV protease differs from the structures of other retropepsins. *Nat Struct Mol Biol.*, 18(2), pp.227–229.

Lin, X.L. et al., 1992. Enzymic activities of two-chain pepsinogen, two-chain pepsin, and the amino-terminal lobe of pepsinogen. *The Journal of biological chemistry*, 267(24), pp.17257–63.

Matúz, K. et al. 2012 Inhibition of XMRV and HIV-1 proteases by pepstatin A and acetyl-pepstatin. *FEBS J.* 279(17):3276-86

Matsui, T. et al., 2006. Mouse homologue of skin-specific retroviral-like aspartic protease involved in wrinkle formation. *The Journal of biological chemistry*, 281(37), pp.27512–25.

Nelson, D., Cox, M. & Lehninger, A. 2008 *Principles of Biochemistry*, 5th ed, New York: W. H. Freeman

Newton, H.J. et al., 2010. Molecular Pathogenesis of Infections Caused by Legionella pneumophila. *Clinical Microbiology Reviews*, 23(2), pp.274–298.

Nguyen, T. M. et al. 2006. A community-wide outbreak of Legionnaires disease linked to industrial cooling towers—how far can contaminated aerosols spread? *J. Infect. Dis.* 193:102–111.

Perteguer, M.J. et al., 2012. Ddi1-like protein from Leishmania major is an active aspartyl proteínase. *Cell stress & chaperones*. 18(2):171-81

Peters, H.K. & Haldenwang, W.G., 1991. Synthesis and fractionation properties of SpoIIGA, a protein essential for pro-sigma E processing in Bacillus subtilis. *Journal of bacteriology*, 173(24), pp.7821–7.

Pettit, S.C. et al., 1991. Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid. *The Journal of biological chemistry*, 266(22), pp.14539–47.

Rawlings, N.D. & Barrett, A.J., 1993. Evolutionary families of peptidases. *The Biochemical journal*, 290 (Pt 1, pp.205–18.

Rawlings, N.D., et al., 2010. MEROPS: the peptidase database. *Nucleic acids research*, 38(Database issue), pp.D227–33.

Rawlings, N.D. & Bateman, A., 2009. Pepsin homologues in bacteria. *BMC genomics*, 10, p.437.

Rawlings, N.D. & Salvesen, G. S., 2013.*Handbook of Proteolytic Enzymes*, 3rd ed., Academic Press

Ribera, E. et al., 2011. [Characteristics of antiretroviral drugs]. *Enfermedades infecciosas y microbiología clínica*, 29(5), pp.362–91.

Rolando M & Buchrieser C. 2012. Post-translational modifications of host proteins by Legionella pneumophila: a sophisticated survival strategy. 7(3):369-81

Russo, I. et al. 2010 Plasmepsin V licenses Plasmodium proteins for export into the host erythrocyte. *Nature* 463(7281):632-6

Salem, M. et al. 2010. Revisiting glutaraldehyde cross-linking: the case of the Arg-Lys intermolecular doublet. *Acta crystallographica. Section F, Structural biology and crystallization communications*, 66(Pt 3), pp.225–8.

Simões, I. et al. 2007. Characterization of Recombinant CDR1, an Arabidopsis Aspartic Proteinase Involved in Disease Resistance *The Journal of Biological Chemistry*,282(43), pp. 31358 –31365

Simões, I. et al., 2011. Shewasin A, an active pepsin homolog from the bacterium Shewanella amazonensis. *The FEBS journal*, 278(17), pp.3177–86.

Tang, J. et al. 1978. Structural evidence for gene duplication in the evolution of the acid proteases. *Nature* 271, 618 - 621

Tözsér, J. et al., 1997. Studies on the symmetry and sequence context dependence of the HIV-1 proteínase specificity. *The Journal of Biological Chemistry*, 272(27), pp.16807–14.

Tyagi, SC. et al. 1992 Inhibitors of human immunodeficiency virus-1 protease. *Biochem Cell Biol.,* 70(5):309-15.

Wilks, J.C. & Slonczewski, J.L., 2007. pH of the cytoplasm and periplasm of Escherichia coli: rapid measurement by green fluorescent protein fluorimetry. *Journal of bacteriology*, 189(15), pp.5601–7.

Wlodawer & Gustchina, 2000. Structural and biochemical studies of retroviral proteases. *Biochimica et biophysica acta*, 1477(1-2), pp.16–34.