Breakdown of keratin-laden biomass waste by the thermophilic bacterium *Fervidobacterium pennivorans* strain T

UNIVERSITY OF BERGEN
Faculty of Mathematics and Natural Sciences

Edoardo Mandolini

Supervisor: Prof. Nils-Kåre Birkeland

Co-supervisor: Rubén Javier Lòpez

Master's Thesis in Microbiology
Department of Biological Sciences
Faculty of Mathematics and Natural Sciences

University of Bergen
June 2020

Table of content

Abstract

Developing a more sustainable agro-industry has become a necessity in light of the current environmental crisis. Biocatalysts are already adopted in many industrial applications and have quickly optimized, and in some cases replaced, existing biochemical reactions within the modern agro-industry. Extremozymes, in particular, are valuable tools for processes requiring harsh industrial conditions where, for example, increased temperature may be beneficial for the bioavailability and solubility of organic compounds as well as for improvement in degradation of substrates. In this regard, alternatives to landfill disposal or incineration of keratinous materials such as feathers, wool, hides, hair etc. are emerging and efforts in exploiting thermo-stable keratinolytic biocatalysts have been attempted. Nonetheless, keratin degradation remains a complex process poorly understood and thus limiting the current toolbox of useful enzymes and organisms needed to meet all demands.

In this study, a newly isolated strain of an anaerobic, thermophilic microorganism belonging to the Thermotogae phylum, *Fervidobacterium pennivorans* strain T, was assessed for its capability of degrading native chicken feathers. By following a multi-omics approach, its proteolytic system was explored in the attempt to isolate new keratinase candidates. First, the physiology of *F. pennivorans* strain T was further investigated in batch cultures and the first growth curve of an organism of this species was described, showing a generation time of 150 minutes and a long stationary phase. Then, the complete genome of the organism was sequenced and analysed, revealing interesting molecular features, such as inverted genomic blocks,

when compared to its most closely related organisms: *F. pennivorans* DSM9078^T and *F. islandicum* AW-1. The strain T genome was slightly shorter (2002515 base pair) and had ANI values of 97.65 % and 80.90% to the compared organisms, respectively, but the same number of predicted protease-encoding genes (55) were found by gene mining analysis.

Next, feather degradation by the organism was up-scaled using a bioreactor to further evaluate its potential in industrial applications and cells were sampled for transcriptomics purposes. *F. pennivorans* strain T performed mediocrely in the fermenter, but RNA extraction was, however, not successful. From secretomics analysis of growing cultures, an extracellular serine protease named Peg_1025 was identified, showing high sequence conservation with the subtilisin type proteases, especially with subtilisin Ak1 from *Geobacillus stearothermophilus* strain AK1. By multiple sequencing alignment, the catalytic triad His, Asp, Ser, as well as a signal peptide and a propeptide domain were predicted. Three dimensional structural modelling using subtilisin Ak1 as template, showed Peg_1025 to possess several insertions of unknown functions compared to subtilisin Ak1, only one conserved Ca^{2+} binding site as well as lack of a disulphide bond in the active cleft. Nonetheless, important structural motifs remained conserved. The enzyme was successfully expressed in *E. coli* using N- and C-terminal His-tag and soluble proteins were active at 70°C in proteolytic activity assays that used casein as substrate. Phylogenetic analyses revealed that Peg_1025 belongs to a distinct clade of Thermotogae peptidases separated from fervidolysin and Ak1, and as such, it represents the first characterized member of this phylogenetic group. Although the specific role of the serine protease in feather degradation remains unclear, the general results from this study confirm that *F. pennivorans* strain T possesses a complex machinery with keratinolytic power. The biology of this extremophile remains an intriguing field of exploration, further encouraged by its biotechnological potential that is still left to unfold.

Introduction

Applications of biocatalysts to different industry sectors have been playing an increasingly role in the wealth and development of the society. In the face of a changing globalized economy, declining fossil energy resources, environmental pollution and climate crisis, the increase in development and large-scale use of microbial biotechnology is viewed as both an opportunity and a necessity as a strategy towards attaining a strengthened bio-based economy instead of, or to complement, traditional industrial and agricultural production processes.

It is surprising how the variety of fields in which microbial cells and their derived enzymes, altogether under the name of biocatalysts, have already quickly optimized, and in some cases replaced, existing biochemical reactions or entire processes within modern agro-industry. Their role is widespread from converting renewable resources, such as wastes and byproducts, into fine chemicals, biopolymers, biomaterial and biofuels (industrial biotechnology), to the production and processing of food and feed (food biotechnology), from the bioremediation of contaminated sites and water treatments (environmental biotechnology), to the exploitation of microorganisms to produce pharmaceuticals (medical biotechnology) (Kirk et al., 2002; Lorenz & Eck, 2005).

Whereas whole cells are often used for synthetic reactions that require cofactors which must be regenerated, their derived enzymes have raised great interest to biotechnology companies worldwide for being capable of accepting a wide range of complex molecules yet maintaining a remarkable substrate specificity (Schmid et al., 2001). The current industrial economy has already available thousands of biomolecules of estimated value of 5.9 billion USD in 2020 (Industrial Enzymes Market 2020), obtained from a diverse range of microorganisms.

Enzymes need to function sufficiently well according to several performance parameters (Table 1) (Lorenz & Eck, 2005). However, many of these enzymes do not withstand harsh industrial conditions, which can differ greatly from standard “physiological” conditions: moderate temperature (10-37°C), pH ~ 7, salinity ranging from 0.15 to 0.5 M NaCl, pressure 1 atm and sufficient water availability (Aguilar et al. 1998; Antranikian et al. 2005) making their activity and performance one of the major drawbacks.

Table 1 Biochemical properties to be taken into consideration when evaluating biocatalyst functionality. Kat, catalytic reaction rate; kcat, catalytic constant; Km, Michaelis constant; U, unit. Table adapted from Lorenz et al. (2005).

Activity	**Stability**
Turnover frequency (K_{cat})	Temperature stability
Specific activity (kat/kg, U/mg)	pH stability
Temperature profile	Ingredient/byproduct stability
pH profile	Solvent stability
Efficiency	**Specificity**
Space-time yield	Substrate range
Product inhibition	Substrate specificity (K_m,k_{cat}/K_m)
Byproduct/ingredient inhibition	Substrate regioselectivity and enantioselectivity
Producibility/expression yield	
	Substrate conversion (%), yield

Running certain industrial processes under unconventional conditions by using enzyme technology has already provided significant advantages to the industries and their increase can only be beneficial both to the society and to the environment. The increase of temperature has a significant influence on the bioavailability and solubility of organic compounds as well as in the decrease of viscosity and risk of contaminations allowing improvement in polymerization or degradation of substrates (Adams et al., 1995; Gerday & Glansdorff, 2007). The use of high pressure during processing and sterilization of food, for instance, can induce the formation of gels and also the denaturation or coagulation of proteins without affecting colour or flavour (Van den Burg, 2003). Acidic conditions are optimal for efficient extraction of metals such as copper and gold, whereas basic pH is favourable in the detergent industries. High concentration of salt may be preferred to avoid contaminations in reactions that require physiological temperature (Gerday & Glansdorff, 2007).

However, the present toolbox of biocatalysts is a limiting factor in industries and not diversified enough to sufficiently exploit the true potential for creating novel applications under these extreme conditions (Figure 1).

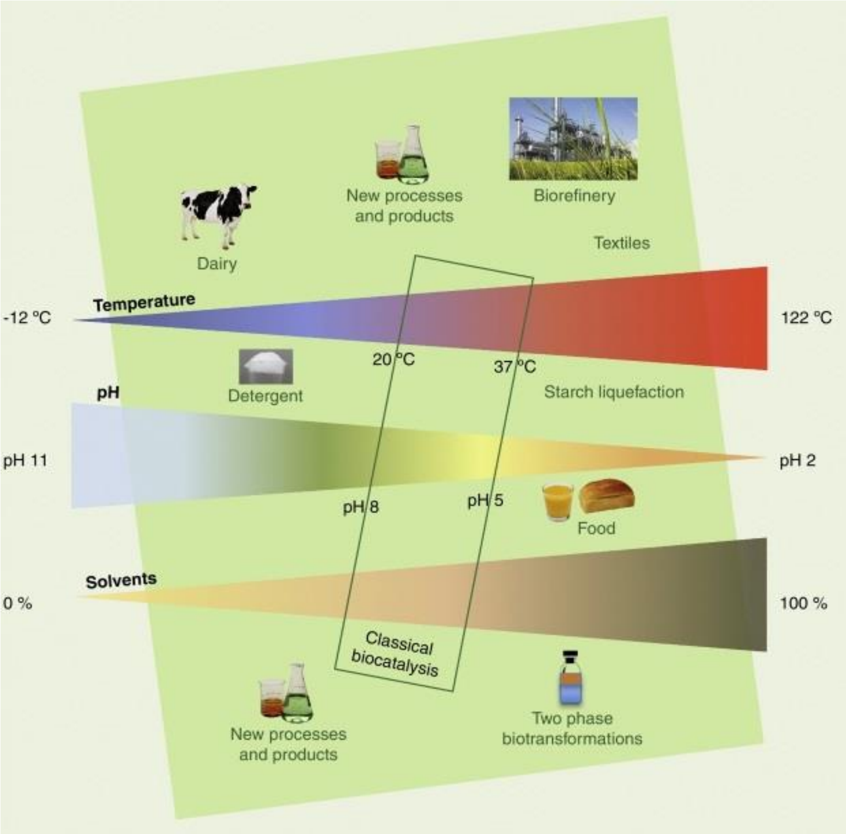


Figure 1 Classical biocatalysts can only be used in standard physiological conditions whose applications are limited if compared to the potential that extreme conditions can exploit in the industries (Elleuche et al., 2014).

As a result, industries all over the globe have been focusing on the discovery and characterization of new types of organisms, with their relevant enzymes, that may thrive in conditions that resemble best the ones found in extreme industrial processes (Herbert, 1992). Such conditions can be found in natural environments with elevated and low temperature (40-100 °C and < 15°C, respectively), high pressures (< 100 atm), high and low pH values (> 9 and < 5), high salinity (> 3.5 %), dry conditions and environments with high UV exposure environments as well as elevated concentrations of heavy metals and radioisotopes (Gerday & Glansdorff, 2007). Because of their generally inhospitable characteristics, these environments are said to be extreme and organisms adapted to these conditions are called extremophiles (Figure 2).

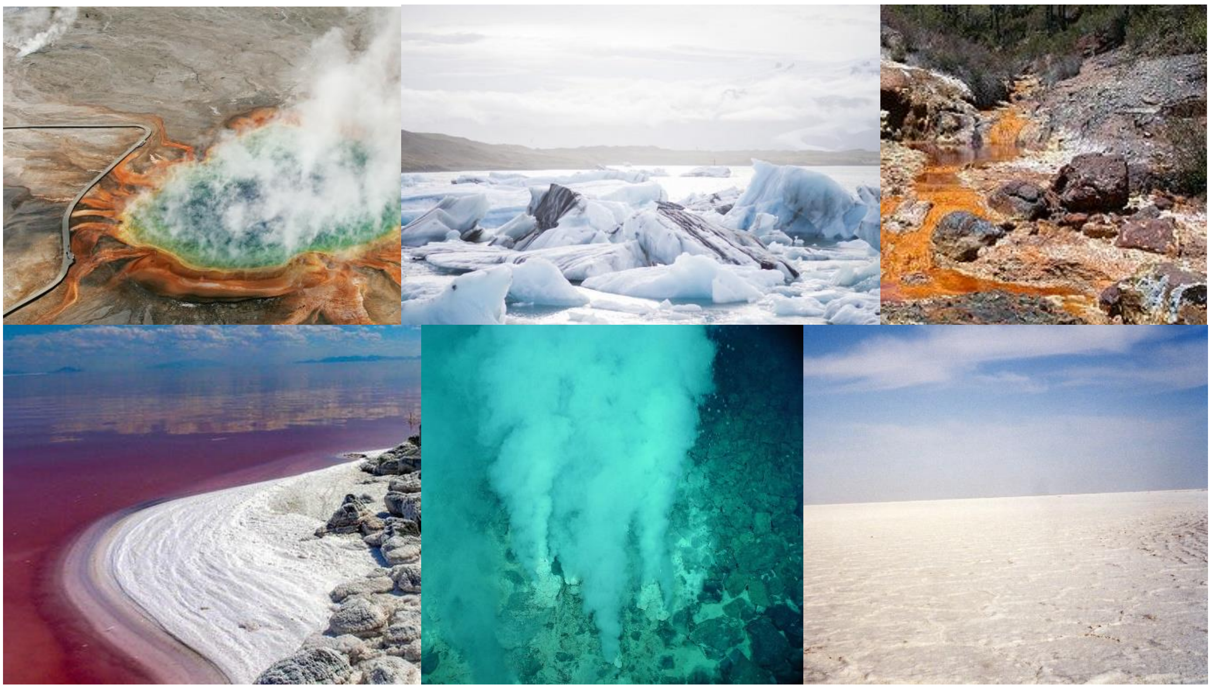
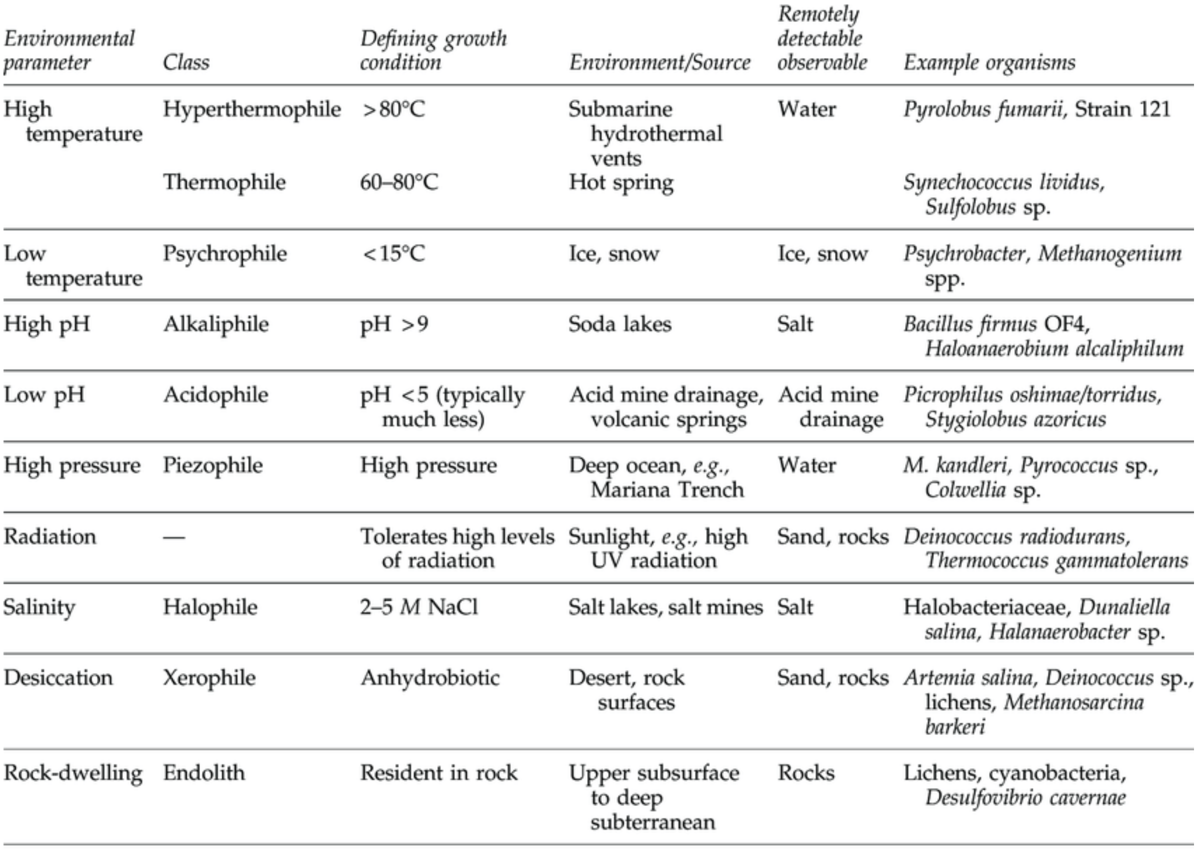


Figure 2 Examples of extreme environments harbouring microbial life. From left to right starting from the top: terrestrial hot pools, sea ice sheet, acid mine drainage site, soda lake, deep sea hydrothermal vents, high altitude saltern desert.

Extremophilic microorganisms are taxonomically widely distributed and are a functionally diverse group (Cowan et al., 2015). In general, they are divided into two categories: extremophiles, including those which require one or more extreme conditions to grow, and extremotolerant organisms, able to tolerate extreme conditions, although they grow optimally at mesophilic conditions (Canganella & Wiegel, 2011).

Extremophiles can be classified according to the conditions in which they grow (Table 2). Thermophiles and hyperthermophiles are organisms that grow optimally at temperatures of 45-80 °C and > 80 °C, respectively, typical of geothermal waters, hot springs, mud pots, fumaroles, geysers, deep-sea hydrothermal vents, volcanic environments, and also in engineered environments, such as compost facilities and anaerobic reactors (Gerday & Glansdorff, 2007; Orellana et al., 2018; Rampelotto, 2013). In contrast, it is easy to forget that ~70% of the Earth is mostly cold and that most ecosystems are exposed to temperatures that are permanently below 5°C. These comprise the oceans at depths < 1000 m and the Polar Regions, ice and snows covers as well as alpine zones and underground environments. Microorganisms living in such conditions are called psychrophiles and have been found physiologically active at temperatures as low as −20 °C, but in general having an optimal growth temperature of <15 °C (Feller & Gerday, 2003). Linked to the deep seas, there are the piezophiles, that is, organisms that require pressures as high as 130 MPa (Yayanos et al., 1982) for growth. Organisms optimally adapted to pH < 5 and pH > 9 are named acidophiles and alkaliphiles, respectively. The former can be found in environments where chemical oxidation of mineral species such as sulphur and sulphide minerals exists, e.g., in volcanic areas or hydrothermal vent systems, but also in mine drainage sites with high concentrations of pyrite (Fe_2S) or where biological processes that generate acidity occur, such as in stomachs, fermentation and nitrification (Canganella & Wiegel, 2011; Gerday & Glansdorff, 2007). On the other hand, high pH values are found in naturally occurring environments such as soda lakes and underground alkaline water but also in relatively small alkaline niches such as intestines of insects. They can also be found in artificial alkaline environments such as liquid of indigo fermentation and in bio-wastes of food-processing industries (Canganella & Wiegel, 2011; Gerday & Glansdorff, 2007). Often, these environments with high or low pH have also high concentrations of toxic heavy metals, in which metalophiles, organisms adapted to these mineral species, are very abundant. Halophiles are organisms that require high salinity for growth, in concentration of 200-5900 mM (Edbeib et al., 2016) and are abundant in natural or artificial salt lakes like solar salterns, underground deposits of rock salt as well as salted food products (Gerday & Glansdorff, 2007). Less studied are xerophiles, microorganisms that survive in extremely dry environments (water activity <0.75) (Connon et al., 2007) and permanent exposure of damaging solar radiations (220–320 nm wavelengths UV) typical of elevated altitudes environments and deserts (Gabani et al., 2014). It is worth mentioning that extremophiles are usually defined by one extreme condition, nevertheless, many natural environments possess two or more extreme conditions, making many of these organisms poly-extremophiles. For example, many hot springs are acid or alkaline at the same time, and usually rich in metal content; the deep ocean is generally cold, oligotrophic (very low nutrient content), and exposed to high pressure; and several hypersaline lakes are very alkaline (Canganella & Wiegel, 2011).

Table 2 Summary table of extremophilic microorganisms and their classification based on the environmental condition they live in. Example of organisms are also presented (Hegde & Kaltenegger, 2012).



Environmental parameter	*Class*	*Defining growth condition*	*Environment/Source*	*Remotely detectable observable*	*Example organisms*
High temperature	Hyperthermophile	>80°C	Submarine hydrothermal vents	Water	*Pyrolobus fumarii*, Strain 121
	Thermophile	60–80°C	Hot spring		*Synechococcus lividus*, *Sulfolobus* sp.
Low temperature	Psychrophile	<15°C	Ice, snow	Ice, snow	*Psychrobacter*, *Methanogenium* spp.
High pH	Alkaliphile	pH >9	Soda lakes	Salt	*Bacillus firmus* OF4, *Haloanaerobium alcaliphilum*
Low pH	Acidophile	pH <5 (typically much less)	Acid mine drainage, volcanic springs	Acid mine drainage	*Picrophilus oshimae/torridus*, *Stygiolobus azoricus*
High pressure	Piezophile	High pressure	Deep ocean, *e.g.*, Mariana Trench	Water	*M. kandleri*, *Pyrococcus* sp., *Colwellia* sp.
Radiation	—	Tolerates high levels of radiation	Sunlight, *e.g.*, high UV radiation	Sand, rocks	*Deinococcus radiodurans*, *Thermococcus gammatolerans*
Salinity	Halophile	2–5 *M* NaCl	Salt lakes, salt mines	Salt	Halobacteriaceae, *Dunaliella salina*, *Halanaerobacter* sp.
Desiccation	Xerophile	Anhydrobiotic	Desert, rock surfaces	Sand, rocks	*Artemia salina*, *Deinococcus* sp., lichens, *Methanosarcina barkeri*
Rock-dwelling	Endolith	Resident in rock	Upper subsurface to deep subterranean	Rocks	Lichens, cyanobacteria, *Desulfovibrio cavernae*

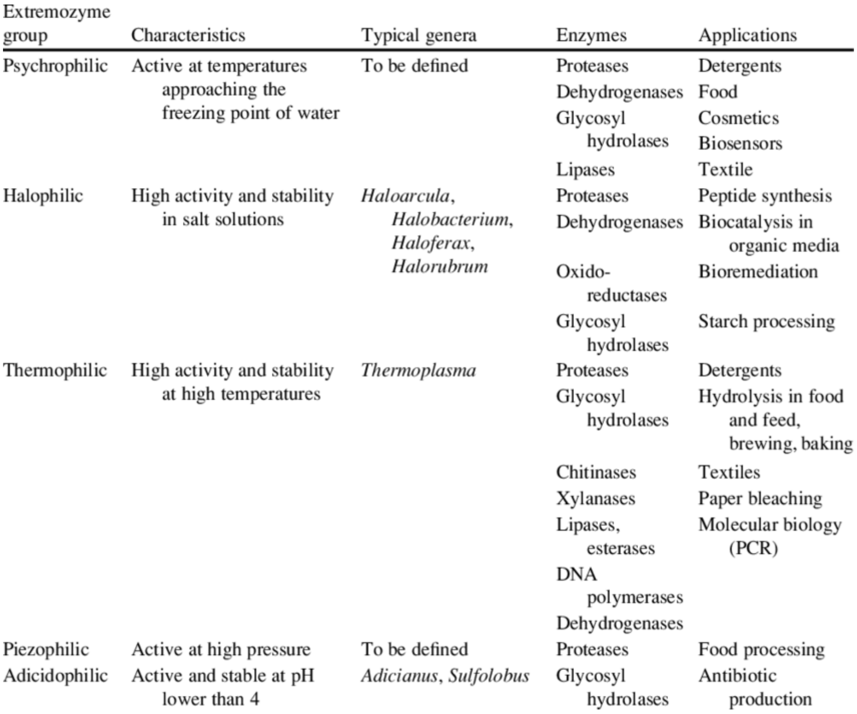
Although taxonomically diverse and widely distributed throughout the phylogenetic tree of life, most extremophiles, either belonging to the same genera or to completely different branches, have evolved biochemical properties optimized for certain extreme conditions, therefore containing enzymes that are perfectly active and functional under the very same circumstances (extremozymes) (Gerday & Glansdorff, 2007). This is the case of different thermophilic, psychrophilic and extreme halophilic enzymes, for instance, for which members may be widespread in the tree of life but their common selective pressure acted on the same structural properties. Nonetheless, specific mechanisms of protein stabilization under extreme conditions may differ depending on the protein family as well as the microbial phylogeny.

The study of extremophiles is a rather difficult field for a series of reasons. First of all, reaching extreme environments is often very challenging and potentially dangerous. Secondly, most of the extremophiles are still part of the microbial dark matter that has not been isolated yet or even discovered at all, due to the complexity of reproducing their ecological niches in laboratories or for lacking appropriate cultivation techniques. Finally, if isolation of a new extremophile is indeed managed, conditions for its optimal growth often go beyond the capabilities of conventional fermentation systems and can lead to considerable expenses (Rampelotto, 2013).

Some of these problems have been overcome, however, during the past decades with the development of new culture-independent techniques such as metagenomics analysis, with improved gene mining tools, as well as proteomics and transcriptomics techniques that base their reliability on an ever increasing pool of sequence-based databases (Lorenz & Eck, 2005). This has allowed to investigate intriguing questions on the nature of extremophiles and to provide industries with an unprecedented chance to bring biomolecules into industrial application. Furthermore, the expression of extremozymes encoding genes in mesophilic hosts (e.g. *Escherichia coli*, *Bacillus subtilis*) avoids problems arising from growing extremophiles and also may provide sufficient quantity of enzymes for practical uses (Hough & Danson, 1999). Nevertheless, the identification of specific enzymes from these modern methods is limited by the current bioinformatic tools, thus a detailed knowledge on the physiology of organisms in culture is still essential to complement genomic or cloning practices and cannot be fully replaced by any other approach.

Many biocatalysts have already been isolated from all sorts of extremophiles living in a great variety of extreme environments (Van den Burg, 2003) and their applications already cover a variety of fields (Adams et al., 1995) (Table 3). Notable is the case of β-galactosidase isolated from the cold-adapted *Kluyveromyces* for the degradation of lactose that result in a lactose hydrolysis up to 70-80% under 24 hours incubation at 5-10°C with the prevention of contaminations, and with higher yields (Cavaille & Combes, 1995). Even more, Taq DNA polymerase isolated from *Thermus aquaticus*, has allowed one of the most dramatic advance in molecular biology with the development of the polymerase chain reaction or PCR (Ishino & Ishino, 2014).

Table 3 Examples of extremozymes isolated from different extreme conditions and their current application in industries (Bonete & Martines-Espinosa, 2011).



Extremozyme group	Characteristics	Typical genera	Enzymes	Applications
Psychrophilic	Active at temperatures approaching the freezing point of water	To be defined	Proteases	Detergents
			Dehydrogenases	Food
			Glycosyl hydrolases	Cosmetics
				Biosensors
			Lipases	Textile
Halophilic	High activity and stability in salt solutions	*Haloarcula*, *Halobacterium*, *Haloferax*, *Halorubrum*	Proteases	Peptide synthesis
			Dehydrogenases	Biocatalysis in organic media
			Oxido-reductases	Bioremediation
			Glycosyl hydrolases	Starch processing
Thermophilic	High activity and stability at high temperatures	*Thermoplasma*	Proteases	Detergents
			Glycosyl hydrolases	Hydrolysis in food and feed, brewing, baking
			Chitinases	Textiles
			Xylanases	Paper bleaching
			Lipases, esterases	Molecular biology (PCR)
			DNA polymerases	
			Dehydrogenases	
Piezophilic	Active at high pressure	To be defined	Proteases	Food processing
Adicidophilic	Active and stable at pH lower than 4	*Adicianus*, *Sulfolobus*	Glycosyl hydrolases	Antibiotic production

Particular attention has been given to thermostable biocatalysts, that is, enzymes that are not denatured by high temperatures (Adams et al., 1995). The source of such enzymes are thermophiles or hyperthermophiles, organisms that can live and grow at optimal temperature between 50°C and 79°C, or above 80°C, respectively (Gerday & Glansdorff, 2007; Stetter, 1996). Currently, the upper temperature limit of life is 122°C and it is held by the Archaea *Methanopyrus kandleri* (Takai et al., 2008). Within the domain of Bacteria, *Aquifex pyrophilus* have the record high growth temperature of 95°C (Burggraf, 1992). For reason of simplification, the term thermophile will be used in this thesis generally to include all microorganisms with T_{opt} > 50°C. When necessary, hyperthermophiles will be distinguished from thermophile.

What are the specific adaptations that thermophilic organisms evolved in order to thrive and grow in the high temperature of their habitat? First, DNA must be prevented from melting. This is achieved by increasing cellular solute levels (e.g. potassium, compatible solutes), synthesis of DNA-binding proteins and, only in hyperthermophilics, the encoding of a unique protein called reverse DNA gyrase, a special DNA topoisomerase, making positively supercoiled DNA. Ribosomal RNA of thermophiles possess higher content in GC nucleic acid that, with their triple hydrogen bond, confer more stability to the helix. Another important adaptation regards cellular membranes stability. In Bacteria, thermostable cytoplasmatic membrane has higher content of long-chain and saturated fatty acids and a lower content of unsaturated fatty acids whereas in Archaea it has increased amount of monobranched fatty alcohol-containing diether lipids (Gerday & Glansdorff, 2007; Koga, 2012; Siliakus et al., 2017). Regarding the properties of their enzymes, the stability of these proteins is based on increased levels of amino acids that promote alpha-helical secondary structures, deletion/shortening of surface loops and immobilization of terminal ends. Proteins have also higher hydrophobic core, increased polar/charged interactions i.e. hydrogen bonds, salt bridges, around the active site and more ionic interactions on the surface (Gerday & Glansdorff, 2007; Reed et al., 2013; Sterner & Liebl, 2001). Furthermore, it has been suggested that surface ion-pair networks and solvent-filled hydrophilic cavities in the core of the protein, provide a degree of resilience and resistance to thermal denaturation (Aguilar et al. 1997).

Thanks to these biochemical adaptations, the more stable and active thermo-enzymes have found a variety of applications in industrial biotechnology where the temperature of reactions is often kept high for several of reasons (Elleuche et al., 2015). As already mentioned, the solubility of many reactants, in particular polymeric substrates, is significantly improved at elevated temperatures. Ordinary proteins denature, exposing the whole polypeptide chain to solute and to catalytic enzymes which active site reaches better their substrate. Moreover, the risk of contamination is reduced by the impossibility of the majority of organisms to survive in such conditions (Littlechild, 2015; Van den Burg, 2003).

It results natural to imagine the countless applications of extremozymes in biomass conversion when elevated temperature can be used to facilitate the degradation of polymers and complex molecules that would otherwise remain insoluble and inaccessible to attack by hydrolytic enzymes. A clear example comes from conversion of starch into more valuable products such as dextrins, glucose, fructose, maltose and other sugars, which requires high temperature to liquefy the material and make it accessible to enzymatic hydrolysis (Gupta et al., 2013). Cellulose, lignin and chitin, but also extremely hard-to-degrade animal proteins such as bones and other hard tissues, are also highly resistant polymers that, if treated at elevated temperatures, can be fully hydrolysed and converted by a synergistic action of different thermoenzymes (Gerday & Glansdorff, 2007; Niehaus et al., 1999; Suzuki et al., 2006).

However, one polymeric substrate that is difficult to bio-degrade by industries is keratin. Keratin is an animal protein present in feathers, hair, skin, wool and horns and is one of the most abundant polymer on Earth together with cellulose and chitin (Gerday & Glansdorff, 2007). The presence of intra-molecular binding of cysteine disulphides and inter- and intra- molecular binding of polar (i.e. hydrogen and ionic bonds) and nonpolar residues (i.e. hydrophobic interactions), makes keratin an extremely stable and resistant polypeptide (Figure 3, A and B) (Parry et al., 1977; Shavandi et al., 2017). These proteins are insoluble in water and resistant to weak acid or alkali solutes. It is also resistant to common proteolytic enzymes such as pepsin or trypsin (Lee et al. 2015). Different types of keratin exist in nature and are grouped based on their secondary structure as well as sulphur content (Lange et al., 2016; Shavandi et al., 2017).

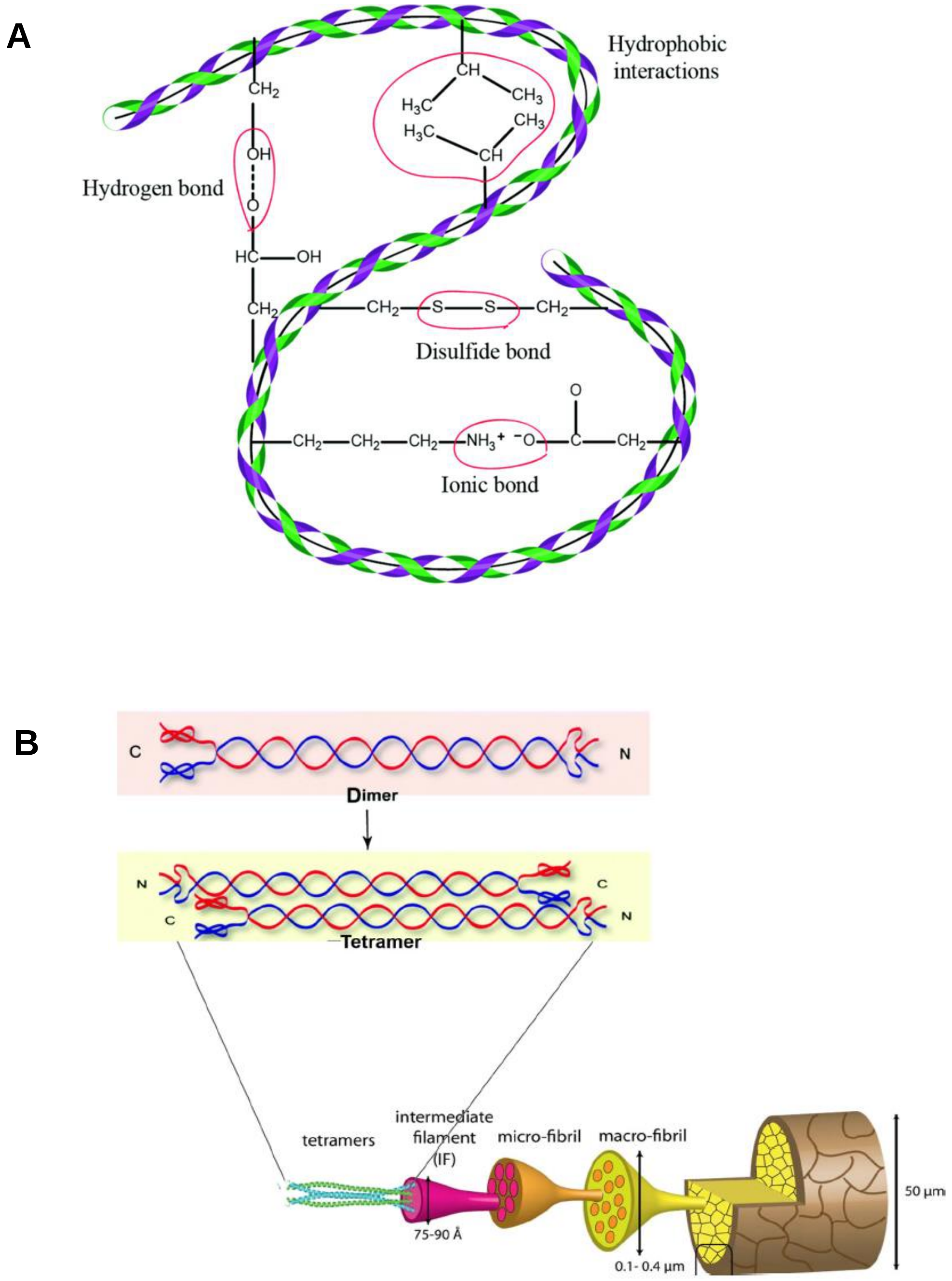


Figure 3 Cartoon showing the structure of keratin polypeptides and how these are combined forming keratin fibers such as hair. A) Two keratin polypeptides form a dimeric coiled coil that interlink by molecular bonds created by important residues typically present in the protein (Shavandi et al., 2017); B) Cartoon of how dimeric coiled coil filaments interconnect to become thicker and thicker keratin fibers (Lange et al., 2016).

The main source of keratin waste is from the production of feather by poultry farming, but also from the production of wool in the fabric industry. It has been estimated that more than 2 million tonnes of wool and 20000 tons of feathers are produced annually worldwide, rising important environmental concern (Friedricht, 1996; FAO, 2013). Keratin-based products are generally hydrolysed by mechanical or chemical treatments to obtain feedstock, fertilizers, glues or foils (Lee, 2015; Williams, 1990). However, their degradation is only partial and most of the essential amino acids they could provide (serine, cysteine and proline) are wasted (Papadopoulos, 1989). These methods, a part from being little efficient, are also very expensive. The ultimate resolution to discard the huge amount of keratin-based products is to bury it in landfills or incineration.

Because of environmental considerations, the use of thermostable proteolytic enzymes in the production of amino acids and peptides from a polymeric substrate such as keratin is becoming attractive for biotechnological applications.

Previous studies showed that feather degradation involves the combination of a complex mixture of enzymes, some specific to catalyse the disulphide bonds (disulphide reductase), some that further hydrolyse the polypeptide chain (keratinases) (Figure 4).

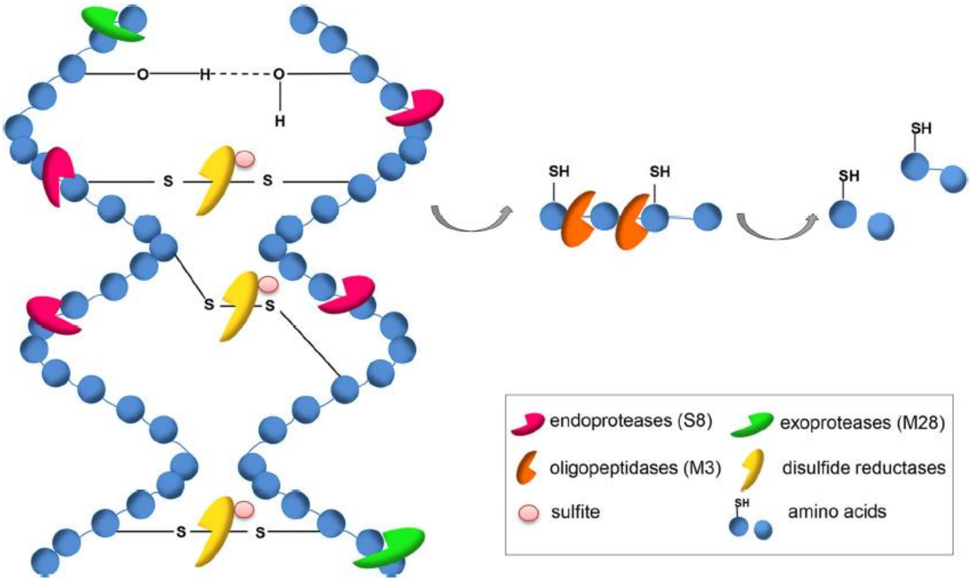


Figure 4 Cartoon showing the dimeric coiled coil composed of two keratin polypeptides where disulphide bonds are cleaved by disulphide reductases whereas the polypeptide chains are degraded by a mixture of endo- and exo- peptidases (Lange et al., 2016).

Keratinases are endo- and exo- proteases that have the peculiarity to act also on keratin substrate (Böckle & Müller, 1997; Lange et al., 2016; Nam et al., 2002; Ramnani & Gupta, 2007). Many of these keratin-degrading proteases produced by Archaea and Bacteria belong to the serine type (Ramnani & Gupta, 2007), in particular to the subtilisin-like serine proteases, also referred to as subtilases (Siezen & Leunissen, 2008). Substilases is a large family of extracellular endo- and exo- peptidases which occur in Archaea, Bacteria, fungi, yeast and higher eukaryotes (Siezen et al., 1991). The mature forms were found to contain a catalytic domain with highly conserved catalytic His, Asp and Ser residues and signal and/or activation-peptides at one of the end of the protein (Shinde & Inouye, 2000). Although mesophilic enzymes play central roles in the degradation of feathers in nature, thermostable subtilases have also been detected and raised interest in the bio-industries. These enzymes can be found in a number of thermophilic bacteria belonging to the genera *Aquifex*, *Thermus*, *Fervidobacterium* and *Thermotoga* (Gerday & Glansdorff, 2007). In particular, in the *Fervidobacterium* genus, anaerobic species have been shown to degrade native feathers both via direct cellular adhesion to the substrate and by secretion of extracellular proteases of the subtilisin family (Cai et al., 2007; Huber et al., 1990; Kang et al., 2019; Suzuki et al., 2006). Recently, a novel thermophilic strain from the genus *Fervidobacterium* was isolated from a hot spring in Tajikistan (Javier-López, 2018). This strain, *Fervidobacterium pennivorans* strain T, belongs to the family Fervidobacteriaceae within the Thermotogae phylum of Bacteria. Thermotogae is one of the deepest branching group within the Bacteria line of descent and possesses a wide range of extremophilic organisms that can catabolise a great variety of substrates (Rosenberg et al. 2014). In the Fervidobacteriaceae family, there are anaerobic and thermophilic bacteria, found in elevated temperature environments all over the globe (Figure 5).

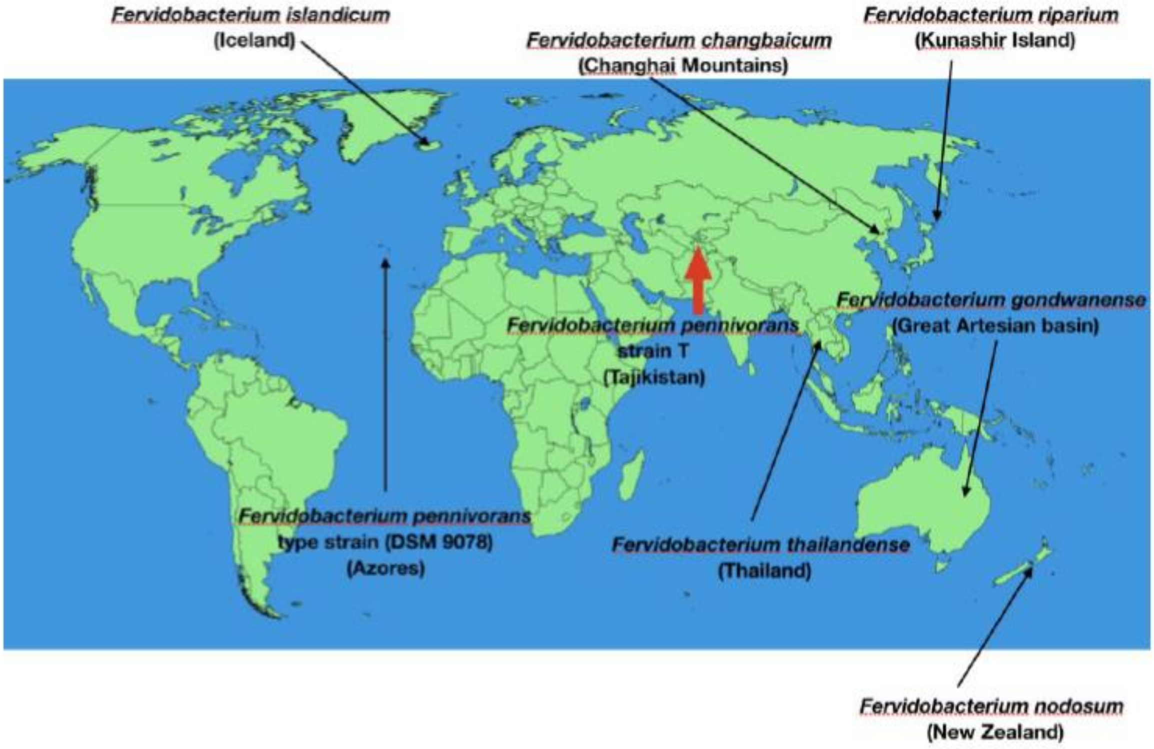

Fervidobacterium islandicum
(Iceland)
Fervidobacterium changbaicum
(Changhai Mountains)
Fervidobacterium riparium
(Kunashir Island)
Fervidobacterium gondwanense
(Great Artesian basin)
Fervidobacterium pennivorans
strain T
(Tajikistan)
Fervidobacterium pennivorans
type strain (DSM 9078)
(Azores)
Fervidobacterium thailandense
(Thailand)
Fervidobacterium nodosum
(New Zealand)

Figure 5 Global distribution of *Fervidobacterium* species currently characterized. Red arrow refer to the location from where *Fervidobacterium pennivorans* strain T was isolated (Javier-López, 2018).

Members of the *Fervidobacterium* genus are gram negative, rod shaped, mostly obligate anaerobes, organotrophic, thermophiles that possess a characteristic outer sheath-like membranous structure called toga, a common trait for all organisms in the Thermotogae phylum (Figure 6) (Frock et al., 2010). Species in this genus have also been well described regarding their physiology (Andrews et al. 1996; Cai et al. 2007; Kanoksilapatham et al. 2016; Nam et al. 2002; Patel et al. 1985; Podosokorskaya et al. 2011; Friedricht et al. 1996).

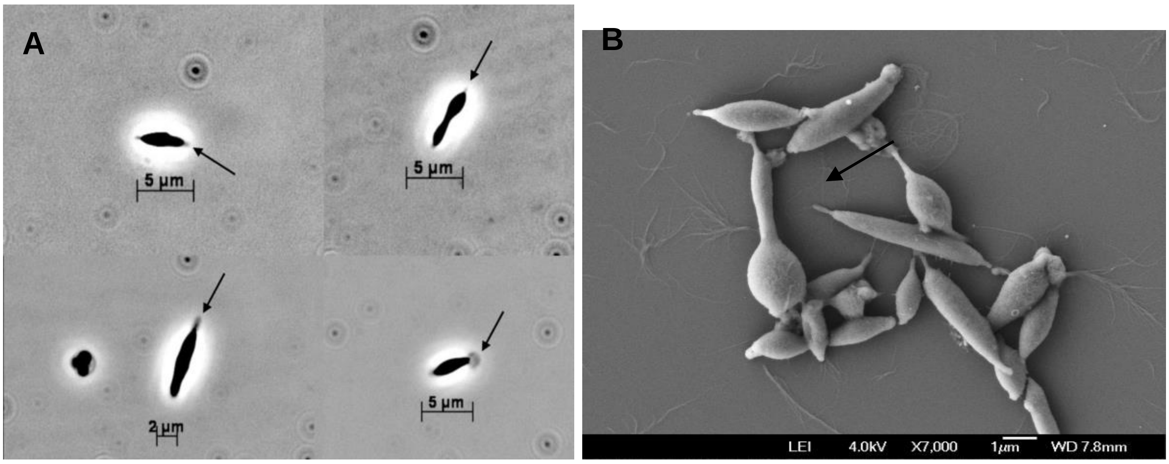


Figure 6 Phase-contrast microscope (A) and scanning electron microscope (B) images of *Fervidobacterium pennivorans* strain T (Javier-López, 2018). The toga has been highlighted.

The physiology of *Fervidobacterium pennivorans* strain T was studied and found to grow anaerobically at optimal temperature of 63°C, tolerate up to 40g/L NaCl concentration and prefer a neutral pH of value around 6.5. It can utilize different kinds of carbon sources, both proteins and carbohydrates (Javier-López, 2018). Examples of sugars sources are pentoses, hexoses, disaccharides, glucans, xylans or sugar alcohols. Among these, glucose, maltose and fructose are the preferred substrates. As protein-derived sources, tryptone, peptone, casein and casamino acids have been described as possible substrates.

Furthermore, *F. pennivorans* strain T have been shown to have the rare ability to degrade feathers, a trait shared only by few other members of the same genus: *F. pennivorans* DSM9078 (type strain) (Friedricht & Antranikian, 1996), *F. islandicum* AW-1 (Nam et al., 2002) and *F. thailandense* (Kanoksilapatham et al., 2016). The phylogenetic relationships of strain T to these two organisms as well as to other members of the same genus is shown in Figure 7.

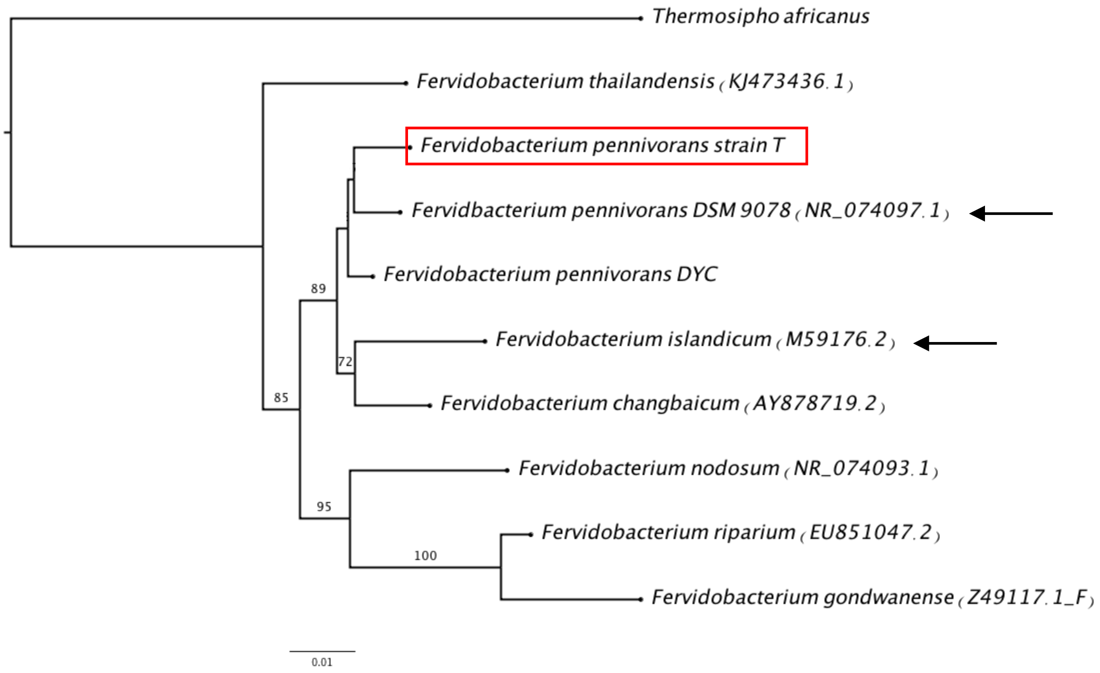

Thermosipho africanus
Fervidobacterium thailandensis (KJ473436.1)
Fervidobacterium pennivorans strain T
Fervidbacterium pennivorans DSM 9078 (NR_074097.1)
Fervidobacterium pennivorans DYC
Fervidobacterium islandicum (M59176.2)
Fervidobacterium changbaicum (AY878719.2)
Fervidobacterium nodosum (NR_074093.1)
Fervidobacterium riparium (EU851047.2)
Fervidobacterium gondwanense (Z49117.1_F)
89
72
85
95
100
0.01

Figure 7 Neighbor-Joining phylogenetic tree of members of the genus *Fervidobacterium* based on 16s rRNA gene sequences (Javier-López, 2018). *F. pennivorans* strain T is highlighted in red. *F. pennivorans* DSM9078 (type strain) and *F. islandicum* AW-1 are marked by black arrows. Bootstrap values higher than 70 are also reported.

F. pennivorans strain T can have important biotechnological potential, although its complex metabolic pathway and catalytic power to degrade feathers remain largely undescribed. Thus, the first aim of this study was to describe the growth pattern of the organism in order to apply a multi-omics approach: genomics, transcriptomics and proteomics, to identify the keratin-degrading enzymes. Particular attention was given to enzymes secreted in the extracellular environment whose combined function leads to feather breakdown. The second aim of this study was to further optimize the expression and testing of the activity of a protease previously identified by Javier-Lopez (2018).

Aims of the project

The overall aim was to enhance the understanding of the keratin-degradation process of *F. pennivorans* strain T, with focus on degradation of feathers.

Sub-goals:

1. Determine the growth pattern of *F. pennivorans* strain T in batch cultures and further determine its physiological properties.

2. Complete genome sequence analysis and gene mining of putative proteolytic enzymes.

3. Up-scaling of growth of the microorganism using a 3 L bioreactor to perform transcriptomics analysis to evaluate expressed genes during feather degradation.

4. Secretomics analyses of cell batches grown on different substrates by proteomics analysis.

5. Bioinformatics characterization of a selected putative keratin degrading enzyme and recombinant expression of its gene in *E. coli*.

6. Biochemical analysis of the expressed protease.

Materials and Methods

Fervidobacterium pennivorans strain T used in this study was previously isolated from a terrestrial hot-spring in Tajikistan in a general anoxic mineral medium (MMF) supplemented with peptone and yeast extract (Javier-López, 2018). The strain was also well characterized in the same work and the knowledge retrieved was used throughout the entire extent of this study.

For all the experiments and laboratory procedures, a fresh 30 ml pre-culture was always used as starting inoculum. The culture was obtained by injecting 1 ml inoculum from a mother culture incubated at room temperature for up to two weeks into new 30 ml anoxic MMF, supplemented with 0.2% peptone, 0.2% yeast extract and incubated at 65°C over night. Before starting new experimental procedures, cellular viability of the pre-culture was checked under a phase-contrast microscope (Nikon Eclipse E400) using oil immersion lens (100X).

1. Physiological analysis

A visual summary of the work flow and main protocols described in this first section is offered in Figure 8.

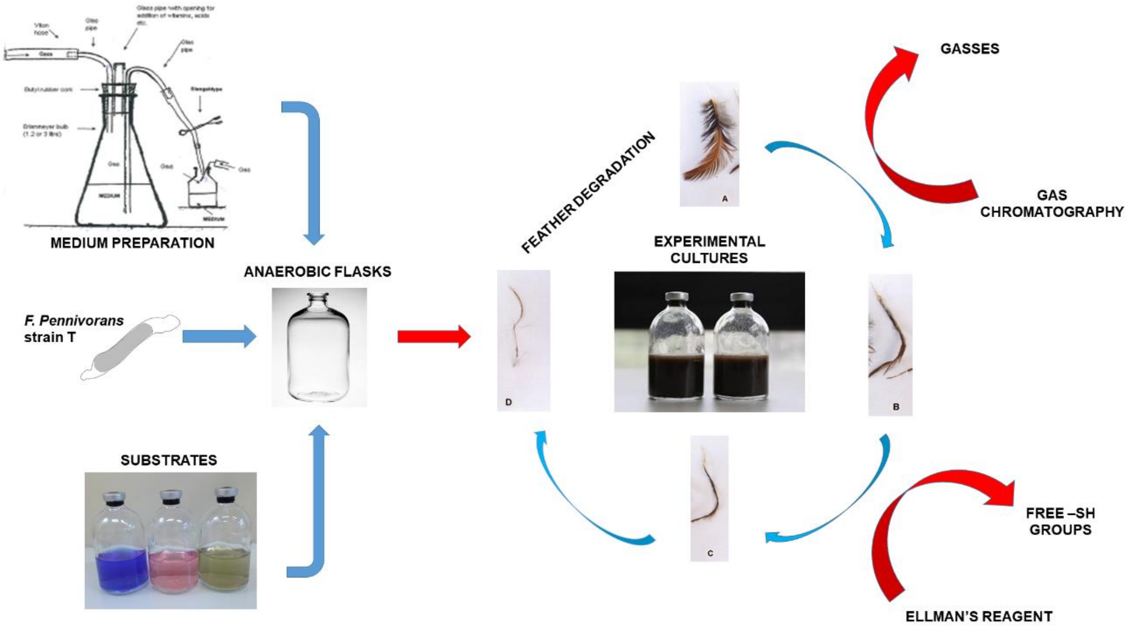


Figure 8 Anoxic medium for F. pennivorans strain T was prepared and was inoculated with fresh pre-cultures and needed substrates in order to assess feather degradation by the microorganisms.

Medium preparation

Fresh cultures of *F. pennivorans* strain T were grown in MMF in anoxic serum bottles at 65°C. MMF was prepared by dissolving all the inorganic components (Table 4,A and B), and 0.2% resazurin as redox indicator, into 1 litre of water in an Erlenmeyer flask. The flask was sealed with a rubber cork (Figure 9,A), autoclaved at 121°C for 20 minutes and, during cooling on ice, it was flushed with sterile nitrogen using the Hungate technique (Macy et al., 1972) to make it anoxic (Figure 9,B). After cooling the solution to approx. 50°C, it was supplemented with 10 ml l^{-1} anoxic vitamin stock solution (Table 5) and cysteine (reducing agent) at a final concentration of 0.05%. The pH was adjusted to 6.7 and the solution aliquoted into small serum flasks using Hungate technique (Figure 9,C). The serum flasks were finally sealed with rubber stoppers and aluminium crimps and kept on the shelf until needed.

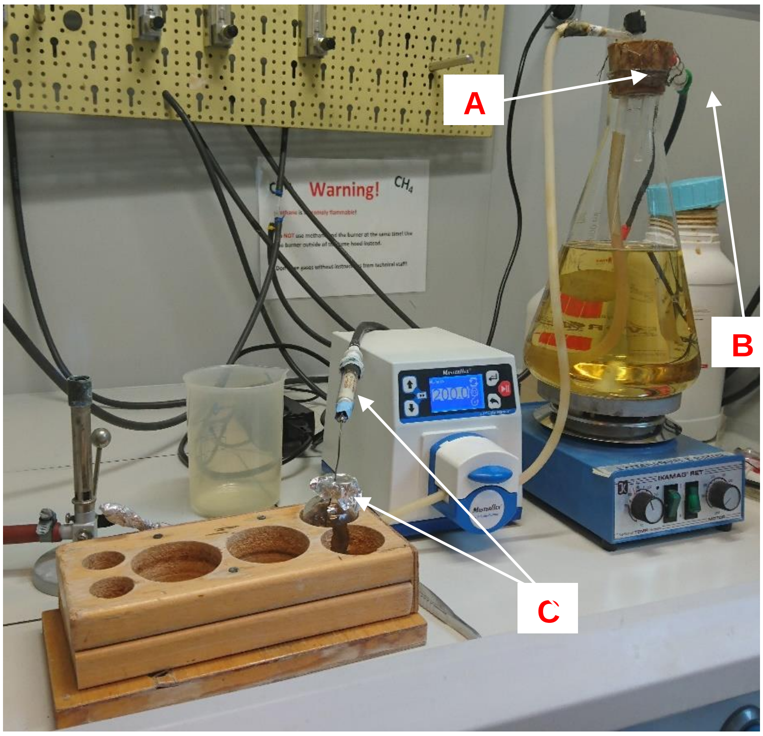
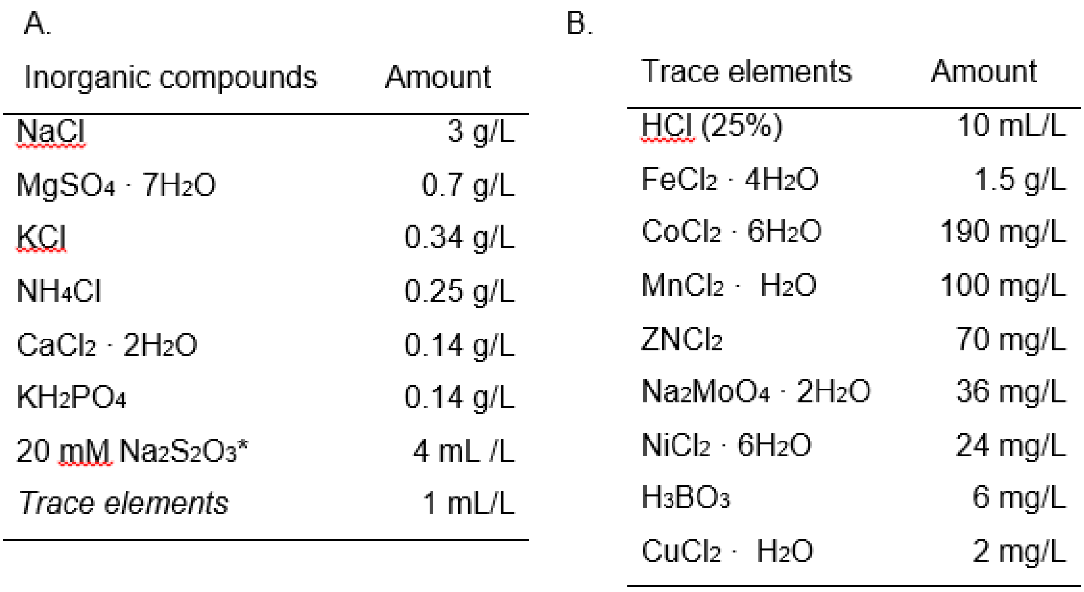


Figure 9 Preparation of anoxic MMF following the Hungate technique (Macy et al., 1972). A, solution with MMF was sealed with a rubber cork and autoclaved; B, while cooling down, MMF was flushed with sterile nitrogen; C, MMF was transferred by a pump into small flasks, flushed with sterile nitrogen, and sealed.

Table 4 Inorganic components of MMF medium with their relative amount. A) The minerals were dissolved into one litre of water together with 0.2% resazurin prior to sterilization by autoclaving. B) Composition of the 1X trace element mix used in the preparation of MMF.



A.

Inorganic compounds	Amount
NaCl	3 g/L
$MgSO_4 \cdot 7H_2O$	0.7 g/L
KCl	0.34 g/L
NH_4Cl	0.25 g/L
$CaCl_2 \cdot 2H_2O$	0.14 g/L
KH_2PO_4	0.14 g/L
20 mM $Na_2S_2O_3$*	4 mL /L
Trace elements	1 mL/L

B.

Trace elements	Amount
HCl (25%)	10 mL/L
$FeCl_2 \cdot 4H_2O$	1.5 g/L
$CoCl_2 \cdot 6H_2O$	190 mg/L
$MnCl_2 \cdot H_2O$	100 mg/L
$ZNCl_2$	70 mg/L
$Na_2MoO_4 \cdot 2H_2O$	36 mg/L
$NiCl_2 \cdot 6H_2O$	24 mg/L
H_3BO_3	6 mg/L
$CuCl_2 \cdot H_2O$	2 mg/L

Table 5 Anoxic vitamin stock solution used in the preparation of the MMF medium. The concentrations reported are for 1X vitamin stock solution. The mix was kept at 4°C until use.

Vitamins	Amount
4-Aminobenzoic acid	8 mg/L
D(+) Biotin	2 mg/L
Nicotinic acid	20 mg/L
Ca-D(+) pantothenate	10 mg/L
Pyridoxamine · 2HCl	30 mg/L
Thiamin dichloride	20 mg/L
Vitamin B12	10 mg/L

Anoxic substrate preparation

MMF was supplemented with organic nutrient sources using anoxic stock solutions. For soluble substrates (i.e. glucose, peptone and casaminoacids), liquid anoxic stocks were prepared using the Hungate technique. Briefly, to prepare e.g. 10% glucose solution, 8 g of glucose was added to a small anoxic serum flask. In the meantime, 100 ml anoxic water was prepared by boiling it, flushing it for 10 minutes with sterile nitrogen and poured into a small anoxic serum flask. The anoxic water was then transferred into the flask containing glucose using a syringe and the solution was autoclaved for 20 minutes at 121°C.

Insoluble substrates (i.e. chicken feathers or keratin azure (Sigma-Aldrich, c.n: K8500) were prepared in a similar way, although chicken feathers were first washed with a solution of 1:1 methanol:ethanol mixture to remove any unwanted material from the feathers, such as dirt or other organic material. Then, the feathers were washed in water and air-dried (method revised from Friedricht et al. 1996). The solid substrates were introduced in an anoxic serum flask and sterilized by autoclaving. Anoxic medium was then added using syringes to desired volume.

Feather degradation

To visually assess the degradation of native chicken feathers by *F. pennivorans* strain T, batches with different toughness of feathers (wings or chest feathers) were prepared. Each batch was prepared with a feather in 30 ml MMF supplemented with 0.1% yeast extract and 0.2% peptone and inoculated with 1 ml of an active pre-culture. The cultures were incubated at 65°C and monitored over time for feather degradation. A negative control batch, without inoculum, was also incubated at 65°C.

Gas chromatography

Gas production was observed in growing cultures and its composition was measured using a gas chromatography system (HP 6890 Series) connected with Shin Carbon Packed Column (60/80 mesh, 1m x 3.2 mm) and equipped with a thermal conductivity detector (GC-8A, Shimadzu, Kyoto, Japan). The injection port, oven, column and detector were used at temperatures of 120, 40, 40 and 100°C, respectively. High purity argon was used at 14 ml per min as carrier gas. The instrument was calibrated with 0.5 ml gas mixture (10% hydrogen, 90% nitrogen, 10% carbon dioxide; based on mol%). Samples of 1 ml gas were measured from cultures growing anaerobically in MMF (0.1% yeast extract) supplemented with 0.5% glucose, 0.5% peptone, 0.5% casamino acids, ~0.5 g keratin azure and ~0.5 g chicken feather after 3 days incubation at 65°C. All measurements were run in biological triplicates.

Culture batch assay

During the degradation of breast chicken feathers by *F. pennivorans* strain T, the sulfhydryl group concentration released in the batches was monitored using Ellman's reaction assay (Ellman, 1959; Kang et al., 2019). Three cultures in 30 ml MMF (0.1% yeast extract, 0.5% glucose, 0.5 g chicken breast feather) were inoculated with 1 ml preculture and grown at 65°C. Twice a day until the complete degradation of feathers, 400 µl culture was extracted and cells were removed by centrifugation at 13000 x g for 7 minutes. From the supernatant, 300 µl sample was mixed with 100 µl reaction buffer (1M KPO_4, pH 8), 600 µl ultra-filtered water and 20 µl DTNB solution (Ellman's powder until saturation into 1 ml reaction buffer). After 15 minutes incubation at room temperature, the absorbance was measured at 412 nm. The instrument was blanked with the same mix but with MMF without inoculum as a sample.

2. Genomics

A visual summary of the work flow and main protocols described in this second section is offered in Figure 10.

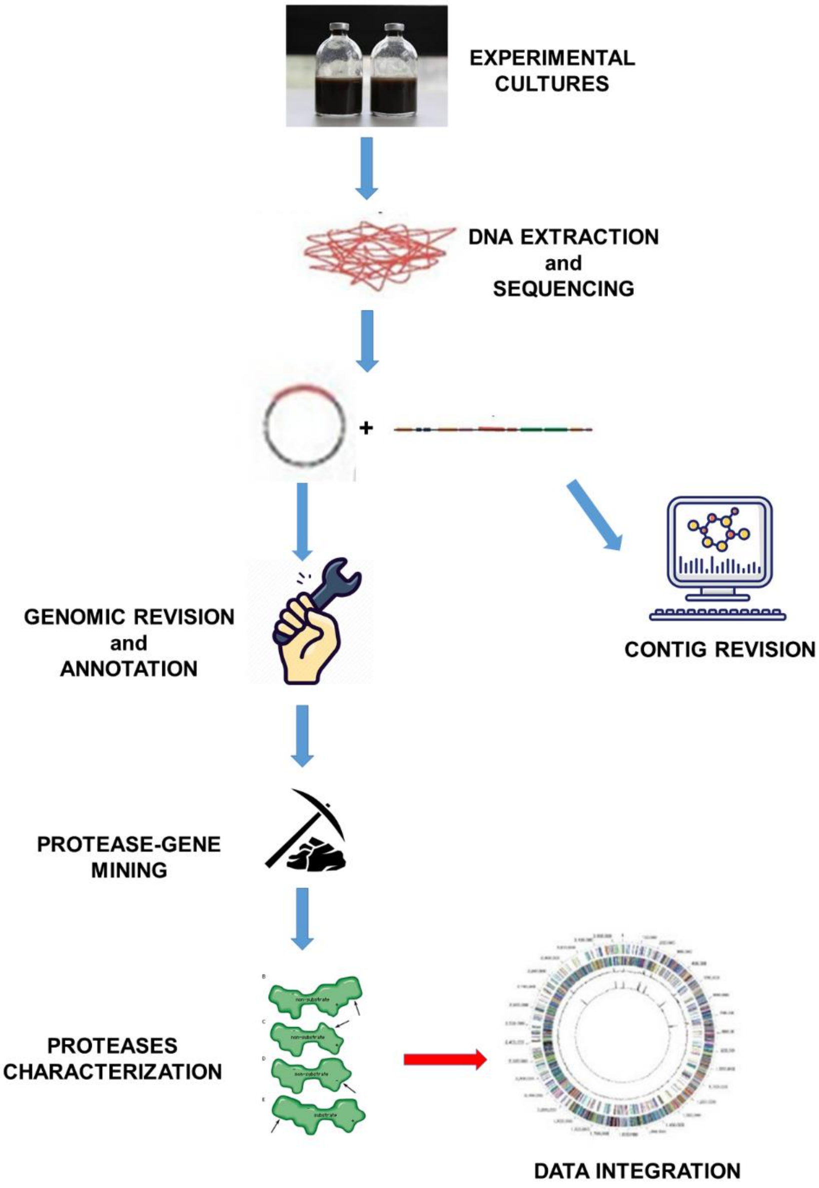

EXPERIMENTAL CULTURES
DNA EXTRACTION and SEQUENCING
+
CONTIG REVISION
GENOMIC REVISION and ANNOTATION
PROTEASE-GENE MINING
PROTEASES CHARACTERIZATION
DATA INTEGRATION

Figure 10 High molecular weight genomic DNA from *F. pennivorans* strain T was extracted from growing cells and sent for complete genome sequencing. Then, the assembled genome was manually inspected and edited to obtain a high quality complete genome sequence. Next, full annotation and gene mining for proteolytic enzymes was performed. Finally, the data obtained were integrated and visually shown.

DNA isolation and sequencing

Complete genome sequencing was performed using PacBio long read technology platform by Eurofins Genomics (https://www.eurofinsgenomics.eu) in Konstanz, Germany. Briefly, genomic DNA was isolated from 100 ml cultures grown for 36 hours in MMFYP (0.1 % yeast extract, 0.5% peptone) using bacterial genomic DNA extraction kit (GenElute^{TM}, Sigma-Aldrich) according to the manufacture's guideline. Cultures were centrifuged at 7000 x g for 10 minutes at 4°C, the supernatant was discarded and the cell pellet kept on ice until DNA extraction. Following extraction, the DNA concentration and purity were determined using NanoDrop™ (One/OneC Microvolume UV-Vis Spectrophotometer, Thermofisher Scientific) whereas its integrity was checked by agarose gel electrophoresis against a DNA ladder (GeneRuler DNA Ladder Mix, ThermoFisher; c.n. SM0333). The 0.8 % agarose gel was run for 45 minutes at 5 V/cm, stained with GelRed (Biotum) and visualized under UV light. The DNA was frozen at -20°C and shipped to the genomic sequence provider in dry ice.

Genomic revision

De novo assembly of the *F. pennivorans* strain T genomic raw sequence read was performed by Eurofins Genomics using HGAP (Hierarchical Genome Assembly Process) (Chin et al., 2013). Contig_1 and contig_2 sequences were separated in two FASTA files and used to perform bioinformatics analysis.

Genomic annotation of raw contig_1 sequence in RAST server was used as a draft in order to reorder the genomic sequence by its chromosomal replication initiator protein gene (dnaA), forerun by a sequence of hundred or so nucleotides that is part of the gene ORF. To choose the length and the starting nucleotide of this precedent sequence to dnaA gene, complete genomes of *F. pennivorans* DSM9078 and *F. islandicum* AW-1 were used as references and inspected in Artemis platform (https://www.sanger.ac.uk/science/tools/artemis) (Rutheford et al., 2000). Then, the corresponding sequence followed by the dnaA gene was searched in the FASTA file of contig_1 and the genome manually edited to make it start with it. The newly ordered single contig was saved as FASTA file, uploaded to and annotated in RAST server, according to the protocol described after.

To inspect the quality of contig_1, the sequence was aligned and compared against *F. pennivorans* DSM9078 and *F. islandicum* AW-1 complete genomes in a multiple sequence alignment. All Genbank sequences were submitted in Mauve software 2.4.0 (GNU GPL) (Darling et al., 2004) using the "align with progressiveMauve" function (http://darlinglab.org/mauve/user-guide/introduction.html;

http://darlinglab.org/mauve/user-guide/aligning.html). Min LCB values were left as default.

Mauve 2.4.0 (GNU GPL) is a genomic alignment software that incorporates results of large-scale evolutionary events such as conserved genomic regions, rearrangements and inversions, with traditional multiple sequence alignments, which detect local changes instead, of nucleotide substitutions (Darling et al., 2004). A unique property to Mauve compared to other multiple alignment systems is that it considers the genomes under study as composed of sets of sequences (blocks) that might be located in different regions in the two genomes under comparison and that might also be conserved in a different order, that is, that the genomes are not collinear. These blocks are marked by the program by the peculiarity of being locally collinear blocks (LCBs), in other words, it identifies conserved segments shared by the genomes under study that appear to be internally free from genomic rearrangement(s). Regions that are in the reverse-complement orientation relative to the first sequence are also shown and appear inverted in the viewer. Once the boundaries of rearrangement have been determined, Mauve represents the logical connection between entire homologous collinear blocks with a single line.

It is important to distinguish between LCBs shared in different genomes that are true genome rearrangement and not random match. To give the confidence, a min LCB weight is used and is defined as the minimum number of matching nucleotides identified in a specific LCB. Mauve's default value of LCB weight is 3 times the minimum match size (considered to be too low) and in general the higher the LCB weight, the higher the confidence that LCBs are actual rearrangements. The procedure to determine a reasonable value for the Min LCB Weight usually involves constructing an initial alignment with the default value and then using the LCB weight slider in the Mauve GUI to find a weight that eliminates all spurious rearrangements. The sequences can then be realigned using the manually determined weight value (https://ecoliwiki.org/colipedia/index.php/Mauve).

After the alignment has been performed, Mauve shows an interactive display layout that offers the possibility to browse and visualize in details specific genes or interesting parts of the genomes under study that might present any differences (http://darlinglab.org/mauve/user-guide/viewer.html).

Furthermore, the software is able to reorder contigs of a draft genome ("move contigs" function) when a reference genome is given (http://darlinglab.org/mauve/user-guide/reordering.html). The outcome is a visual map of the alignment where the contigs of the draft genome are highlighted and reordered based on the reference genome. This action is useful not just to reorder a draft genome, but also to assess the quality of the contigs it is composed of. In fact, if some of the contigs are the result of artefacts or contaminants, they may not align. Thus, after careful evaluation, e.g. Blastn, searches or annotation analysis, these foreign sequences can be deleted from the FASTA sequence output.

To check the possibility that contig_2 was an artefact, that is, a sequence containing variations introduced by non-biological processes during sequencing, a series of bioinformatics analysis were performed.

First, contig_2 was blasted against the nucleotide collection (nr/nt) in NCBI using the Blastn algorithm (Madden 2002; https://blast.ncbi.nlm.nih.gov/Blast.cgi), with expected threshold set to 1e-3. This is an important screening step because, very often, short contigs may not just be artefacts themselves, but sequences belonging to contaminants in the sample. If the latter, the inquired sequence is found belonging to other organisms by a database search and thus may be deleted from the original FASTA file. After it was confirmed that contig_2 did indeed belong to *F. pennivorans* strain T, both contig_1 and contig_2 were aligned against each other using the Blastn algorithm in NCBI. All the parameters were left as default. To obtain a detailed alignment visualization and comparison between contig_1 and contig_2, the two Genbank sequences were submitted into Mauve software 2.4.0 (GNU GPL) (Darling et al., 2004) and the contigs reordered. Finally, RAST server (Aziz et al., 2008) (https://rast.nmpdr.org/) was used to functionally annotate the genes in contig_1 and contig_2 and compare the gene identities between sequences. Both contig_1 and contig_2 were separately uploaded to the server (see next paragraphs for details) and compared using the "genome comparison" function based on sequences. Contig_2 was selected as reference genome and the table outcome reported. As an overall interpretation of the results obtained from the Blastn analyses, Mauve alignment and RAST gene comparison, contig_2 sequence was not considered in any further studies.

Genome annotation

Fully-automated functional annotation of genes of *F. pennivorans* strain T complete genome was conducted using the RAST server database (Aziz et al., 2008) (https://rast.nmpdr.org/).

RAST server (Rapid Annotation using Subsystem Technology; version 2.0) seeks to rapidly produce high-quality assessments of gene functions and an initial metabolic reconstruction for archaeal and bacterial genomes. The entire RAST annotation process is based on subsystems. A subsystem is a gene library collection that is manually curated by experts and is based on what is known from the literature. The proteins encoded by the genes from this collection are used to construct a database containing protein families (FIGfams). These two datasets (expert assertions) are connected and if a gene is found in the subsystem, its encoded protein will automatically be classified with the corresponding annotation from the FIGfams. From the expert assertions, bioinformatics tools project structured gene collections (non-subsystem) that are used to further enlarge the recognition of genes and proteins in new genomes. Thus, there are two classes of genes that RAST produces while annotating a genome: subsystem-based assertions and nonsubsystem-based assertions. For example, if a subsystem coverage of a newly annotated genome is composed of 49% of in-subsystem and 51% of not-in-subsystem annotations, it means that 49% of the genes have been directly recognized with those present in the manually curated library collection, the subsystem, whereas 51% have been recognized thanks to bioinformatical extrapolations. Either way, the RAST server attempts to achieve accuracy, consistency and completeness thanks to an efficient pipeline that involves a step by step gene annotation.

First, a complete or nearly complete (>97%) prokaryotic genome is uploaded in the form of a complete genome or a set of contigs in FASTA or Genbank format. In the uploading process, it is recommended to specify the taxonomy identifier (NCBI taxonomical number) belonging to the organism’s genome as this is used as a handle for analysing it.

Once the job is started, RAST will identify rRNA and tRNA encoding genes first, to then not neglect any protein-encoding genes that significantly overlaps any of these regions (“automatically fix error” function). At this point, the server will try to create a gene pool containing *putative* genes, that is, any genomic parts that resemble actual protein-encoding genes. Once an initial set of *putative* genes has been established, a bunch of universal sequences from the subsystem are searched for in the new genome. These sequences have the property that they are nearly always present in prokaryotes and includes, for example, tRNA synthetases and ribosomial proteins. The outcome of this primary scan is dual: first, the newly found small set of genes from the new genome will become *determined* genes and second, these sequences can be used to obtain genomes that are the closest phylogenetic neighbours to the new one. Once the neighbouring genomes have been detected, they are used by RAST to create a set of genes that is likely to be present in the new genome. Whenever a gene is found, it is moved from the *putative* gene pool into the *determined* gene pool. The *determined* gene pool now obtained is finally used as a training set to identify the protein-encoding genes and estimate the correct starting gene sequences in the new genome. After this major step, all the *putative* genes left unclassified are searched against the entire subsystem, FIGfams and a non-redundant protein database. As a further step, it is also possible to blast large genomic sequences, where no genes were detected, against these databases.

In this study, *F. pennivorans* strain T complete genome was uploaded in RAST server in FASTA format. *Fervidobacterium pennivorans* (NCBI:txid 93466) was used as taxonomical reference and the genetic code translation table 11 (Bacterial and Archaeal) was selected. As RAST annotation scheme, “classic RAST” was chosen whereas “RAST” was selected as gene caller. These parameters make the server run with the standard automatic pipeline as described before, which automatically resolve genes overlapping RNAs regions (automatic error fix function) and blast long gaps for missing genes (backfill gaps). The last updated version of the FIGfams database (Release70) was selected and the work submitted.
The analyses performed for contig_1 and contig_2 revisions followed the protocol just described, as well.

Genome characterization and gene mining

When the annotation of the *F. pennivorans* strain T genome was complete, SEED viewer in the RAST server was used to retrieve general information about the genome (e.g. taxonomy, size, number of coding sequences, RNAs) and to perform a gene mining research for enzymes involved in protein degradation.

From the SEED viewer main page, *Subsystem* Category Distribution of genes were obtained from the interactive pie chart given. Although the Category Distribution function *only* shows genes whose identification matched the ones in the Subsystem (RAST curated database), it could still be used for extrapolating important data. In fact, the list of curated genes given in a specific category was used to extrapolate *key words* that were used to extend the gene mining research in the Genome Browser function, were both the curated genes *and* the predicted ones are listed. Thus, first, the subcategory "Protein Degradation" was selected from the Features in the Subsystem tool. Then, words that were recurring in the protein names listed under "role" column of the table were identified: protease, peptidase and proteolytic. That is, the combination of these three words together gave all the listed proteins in the "Protein Degradation" subcategory. At this point, the Genome Browser tool was opened and the keywords were typed in, one at the time, in the "Function" column of the browser tool. All the proteins listed under a specific keyword were noted.

Although the procedure just described allowed to gather the majority of the putative proteases present in *F. pennivorans* strain T, a bias in gene annotation by the RAST pipeline could have occurred that named genes in different ways than the keywords used. To overcome this problem, complete genomes of *F. pennivorans* DSM9078 and *F. islandicum* AW-1, both downloaded from GenBank in FASTA format (Clark et al., 2016) (https://www.ncbi.nlm.nih.gov/nuccore), were uploaded into the RAST server (see before). The former organism was chosen as being the type strain of the species whereas the latter was selected because it was previously well described as a feather degrading bacterium (Kang et al., 2019; Nam et al., 2002). Then, sequence-based genome comparison was carried out in the RAST server against *F. pennivorans* DSM9078 and *F. islandicum* AW-1 selecting *F. pennivorans* strain T as reference organism. The resulting gene comparison tables were saved (.csv format). At this point, gene mining based on keywords (protease, peptidase and proteolytic) was performed in both *F. pennivorans* DSM9078 and *F. islandicum* AW-1 genomes, and the hits highlighted. Next, the overall genes found in these two organisms were compared with the ones already found in the *F. pennivorans* strain T genome and results integrated. Gene mining comparison among phylogenetically close organisms was used to support RAST annotation, ultimately validating and extending the pool of putative protein-degrading enzymes in the query genome.

A complete pool of putative protein-degrading proteases was finally obtained and their FASTA sequences retrieved from the RAST server.

Peptidase classification

To presumably designate the cellular location of all the proteases gathered, their sequences were submitted into online platforms SignalP 5.0 (Nielsen et al., 2019) (http://www.cbs.dtu.dk/services/SignalP/) and TransMembrane Hidden Markov Model (TMHMM) (Krogh et al., 2001; Sonnhammer & Krogh, 2008) (http://www.cbs.dtu.dk/services/TMHMM/). After combining the results of both platforms, most cellular locations of most proteins were predicted as: intracellular, extracellular, transmembrane and unsolved.

Functional annotation of the enzymes whose cellular location was defined as extracellular, transmembrane or still unsolved, were validated and further characterized using the MEROPS database.

The MEROPS database (Rawlings et al., 2018) (http://www.ebi.ac.uk/merops/) is a manually curated information resource for proteolytic enzymes, their inhibitors and substrates. The database is organized in clusters of homologous sets of peptidase sequences that are organized in families and clans within families. A family contains all related sequences and a clan contains all related tertiary structures. Each category possesses a well-characterized type example (halotype) which all other members of the family or clan must be shown to be related to in a statistically significant manner. MEROPS platform uses NCBI-BLAST+ algorithms for its search and can be operated as a pipeline. It is recommended that a search is first performed against MEROPS-MP to identify that a protein sequence is a peptidase and then against either of the other two libraries (MEROPS-MPRO or –MPEP) to determine whether or not the sequence has a curated report in the collection. If any homologous peptidases to the query are present in the database, they are listed giving the reliability of the match (i.e. e-value). Finally, the detailed description and related information corresponding to a proteolytic enzyme can be retrieved using MEROPS Search tool. MEROPS-MP contains the sequences of peptidase and inhibitor units from all MEROPS family and subfamily type examples and all halotypes; MEROPS-MPRO contains full-length sequences for all the proteins in the MEROPS collection and MEROPS-MPEP contains only the sequences of the peptidase and inhibitor units from all the sequences in the MEROPS collection. Either database selected, analysis is only restricted to the portion of the protein query directly responsible for peptidase or inhibitor activity (unit), which normally corresponds to an active site structural domain, and the retrieved results are reported.

In this study, amino acid sequence of the selected enzymes were submitted to MEROPS-MP to determine whether they were peptidases (i.e. proteolytic enzymes), protease inhibitors or neither. The proteins assessed as peptidases, were submitted to MEROPS-MPRO. If no significant hits were found by MEROPS-MPRO ("unassigned peptidases", i.e. peptidases not present in the database), the proteases were classified according to MEROPS-MP search. Instead, if a submitted query reported significant results in MEROPS-MPRO, the hit with *highest identity percentage* having the *most significant e-value* was chosen and its description, using MEROPS search tool (https://www.ebi.ac.uk/merops/search.shtml), retrieved. When using EMBL-EBI interface for submitting the protein sequences, only the needed database for the research was selected and all other search parameters were left as default.

Genomic and peptidase comparison

BRIG desktop application was run with java 11 in Canonical Ubuntu operating system (version 19.04) and launched opening Ubuntu terminal directly from the program folder using the following command:

$$java\ \ -jar\ ./BRIG.jar$$

First, the location of algorithm BLAST+ (ncbi-blast-2.10.0+-x64-linux) (Camacho et al., 2009) was set for genomic comparison. Then, the Genbank format of *F.*

pennivorans strain T genome was used as reference sequence whereas *F. pennivorans* DSM9078 and *F. islandicum* AW-1, also in Genbank format, were added to the data pool. It is more likely in a larger sequences that an alignment could occur by chance, thus BLAST e-value threshold was changed to 1e-3 typing the following command in the BLAST option section in the BRIG interface:

$$-evalue\ 1e-3$$

Next, rings were created and the corresponding genomic data were selected. For each data ring, the upper and lower identity threshold (%) were set to 90 and 70, respectively. To visualize the locations of the proteolytic enzymes of *F. pennivorans* strain T, the custom feature tool available in BRIG was used and the BRIG manual protocol was followed to create a tab-delimited text file and to set all the parameters (Alikhan, 2011). Once the figure settings were set (i.e. type, size, fonts, colours), the job was submitted and the output image obtained in .jpg format.

DNA-DNA hybridization (DDH) values between *F. pennivorans* DSM9078, *F. islandicum* AW-1 and *F. pennivorans* strain T were calculated submitting the complete genome sequences into the online platform Type Strain Genome Server (TYGS) within the Leibniz Institute DSMZ (https://tygs.dsmz.de/). Values in the "dDDH (d_4, in %)" column were considered. Average Nucleotide Identity (ANI) values between the same organisms were calculated using ANI calculator online platform (http://enve-omics.ce.gatech.edu/ani/) (Goris et al., 2007; Rodriguez-R & Konstantinidis, 2014).

Identity percentages of the 26 protease genes from *F. pennivorans* strain T against the ones in *F. pennivorans* DSM9078 and *F. islandicum* AW-1 were retrieved from the RAST gene comparison described before and their results plotted.

3. Transcriptomics

A visual summary of the work flow and main protocols described in this third section is offered in Figure 11.

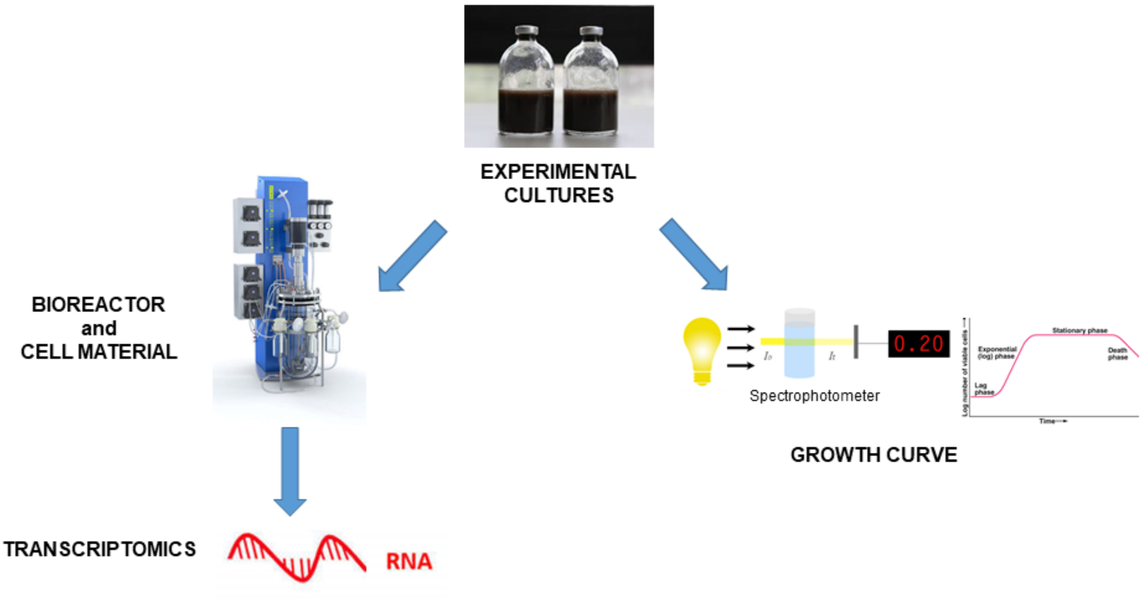


Figure 11 Growth curve of *F. pennivorans* strain T was described to determine the optimal time for harvesting of RNA and secreted proteases. A 3 L bioreactor was run to obtain sufficient cell material from which expressed RNA under feather degradation could be extracted for transcriptomics analysis.

Shake flask growth curve

Cells were grown in 30 ml MMF supplemented with 0.1% yeast extract and 0.5% peptone. One ml of a fresh (< 2 days old), growing (exponential phase) seed culture at 65°C was used as inoculum into new MMF also at 65°C. Then, every two hours, an aliquot of 1 ml was taken by a syringe and centrifuged at 12000 x g for 7 minutes. The cell pellet was resuspended in 1 ml phosphate buffer saline (PBS) (2.5 g l^{-1} Na_2HPO_4, 8 g l^{-1} NaCl, 0.2 g l^{-1} KCl, 0.2 g l^{-1} KH_2PO_4). After absorbance was blanked with PBS, the cell concentration was measured at 600 nm using a spectrophotometer (UV MIN 1240, UV-VIS spectrophotometer, Shimadzu). All the measurements were run in biological triplicates. Generation time (g) for *F. pennivorans* strain T was obtained from OD measurements of the logarithmic curve and calculated from two points of a linear portion of the curve following the equation:

$$g = \frac{ln2}{r}$$

With r representing the growth rate of the organism that takes into consideration the two OD measurements in two time points of the linear portion of the curve according to the following equation:

$$r = \frac{ln^{OD2}/_{OD1}}{t2 - t1}$$

Fermenter

In order to obtain high amount of cellular material from which a good quality and quantity of RNA could be extracted, culturing in a bioreactor was performed. The fermenter (KLF, Bioengineering) was filled with 3 litres of MMF medium supplemented with 0.1% yeast extract, 0.1% glucose and approximately 1.5 g l^{-1} of feathers. After sterilization and anoxidation, a 50 ml starter culture grown in MMF (0.1% yeast extract, 0.1% glucose) was injected. The growth was carried out at 65°C, with a rotor speed of 200 rpm and flushed with 1 vm l^{-1} min^{-1} of sterile nitrogen (volume of sterile nitrogen per volume of medium and minute) for a total of four days, period necessary for the majority of the feathers to be degraded. To compare mRNA normally expressed in the presence of glucose with the mRNA expressed during feather degradation, a sample of 150 ml was taken only after 24 hours of incubation. The sample was collected in falcon tubes using an electric pump, snap cooled in alcohol -80°C for 5 minutes, centrifuged at 7000 x g for 10 minutes at 4°C and the total pellet obtained was weighted and frozen at -80°C. After four days of incubation, 2 litres of culture were collected into 50 ml Falcon tubes following the procedure described before. The supernatant was discarded and the pellet weighted and frozen at -80°C.

The two frozen pellets (before and after feather degradation) were shipped in dry ice to Eurofins genomics for RNA extraction, rRNA depletion and cDNA synthesis and sequencing (RNA Seq analysis).

4. Secretomics

A visual summary of the work flow and main protocols described in this fourth section is offered in Figure 12.

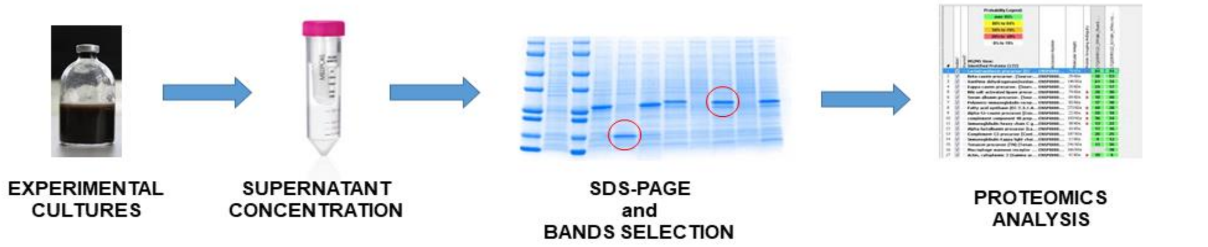


Figure 12 Supernatant of *F. pennivorans* strain T cultures growing on different substrates was concentrated and analysed by SDS-PAGE to identify extracellular proteases secreted by the organism. Relevant bands were sent for LS/MS spectrometry and further visualized by Scaffold viewer (Searle, 2010).

Supernatant concentration

To investigate the different nature and quantity of the proteins secreted in the extracellular environment, cells of *F. pennivorans* strain T were grown in MMF with 0.1% yeast extract and one of the different substrates: 0.5% glucose, 0.1% peptone, 0.5% casamino acids, or 750 g l^{-1} keratin azure. The concentration protocol started after 24 hours incubation for cultures growing in glucose, peptone and casamino acids whereas after 72 hours for cultures growing in keratin azure. For brevity, only the experimental procedure that involved keratin azure as a substrate will be discussed, but the protocols were the same for all the other substrates, as well.

Cells were grown in three anoxic flasks containing 30 ml MMF each (0.1% yeast extract, 0.1% peptone, 750 g l^{-1} keratin azure) and a starting inoculum of 1 ml. After about 72 hours incubation at 65°C, keratin degradation became visual. Next, the cultures were cooled down at room temperature for 15 minutes, collected in cold falcon tubes and centrifuged at 7000 x g for 10 minutes at 4°C. The supernatant was filtered through cold 0.2 µm filters (Whatman, GE healthcare) and collected in a baker on ice. A total volume of 90 ml supernatant was retrieved at this stage and it was distributed into cold 15 ml 10K Amicon® Ultra^{-}15 centrifugal filters (Millipore). Centrifuge cycles of 5000 x g at 4°C, were performed until reaching a 100X concentration factor, recovering a final volume of 1000 µl supernatant. A 30 µl sample was taken for proteomics analysis whereas the remaining concentrated supernatant was frozen at -20°C, after adding glycerol to 10% final concentration.

Any 30 µl sample retrieved after growth with different substrates (glucose, peptone, casaminoacids, and keratin azure) were prepared for sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) analysis by mixing it with 30 µl 2X loading buffer (230 mM Tris-HCl pH 6.8, 25% glycerol, 2% SDS, 0.02% Bromophenol blue, 5% 2-β-Mercaptoetanol), heating it at 95°C for 5 minutes and then loading it into pre-made gels (Mini-PROTEAN® TGX gels, BIO-RAD). Full-range Rainbow^{TM} molecular weight marker (Sigma-Aldrich, GE healthcare) was also added. The gel was run at 200 V for 50 minutes, stained in Coomassie Blue for 1 hour with gentle agitation, destained with destaining solution (50% distilled water, 40% methanol, 10% acetic acid) for 1 hour and washed in distilled water overnight.

Factors taken into consideration in choosing the bends were based on thickness (abundance), run (protein dimension), and quality of the bend (pure protein). The bends were cut out, put in eppendorf and sent on ice to PROBE (Proteomics unit at the University of Bergen, http://www.uib.no/en/rg/probe) for further specific liquid chromatography-mass spectrometry (LC-MS/MS) proteomic analysis by an Orbitrap Elite Hybrid ion Trap-Orbitrap Mass Spectrometer.

Proteomics analysis

In the laboratories of PROBE, each band was disrupted and the proteins precipitated (chloroform/methanol method) before to be denatured in urea solution. Then, after a series of reduction and alkylation, the proteins were digested with porcine trypsin. The small peptides obtained were separated by liquid chromatography, captured by mass spectrometry and resulted in individual experimental MS spectra where each peak theoretically represents a peptide fragment ion. All MS spectra obtained from the samples analysed were then submitted to Proteome Discoverer 2.3 (PD) and compared to the theoretical spectra of protein sequences generated from the complete genome of *F. pennivorans* strain T using Sequest HT and MS Amanda 2.0 search engines. Finally, the data were compiled in a data file that was imported to Scaffold 4.10 (Searle, 2010) and researched with XTandem!. This search engine researches only the detected proteins from PD, but looks for non-tryptic peptides and some other modifications, as well.

The experimental MS spectra obtained from a proteomic analysis are compared against the theoretical spectra generated *in-silico* from a given protein sequence database. Once the experimental data file is uploaded in Scaffold, it is possible to add one or more FASTA sequence databases to which the spectra will be compared against (Manual, 2019). The program converts search engine scores between the spectra obtained from MS spectrometry and the theoretical spectra from the reference genome into probabilities of peptide identification. Different post processing scoring algorithms can be applied e.g. local false discovery rate (LFDR), prefiltered mode, Peptide Prophet Algorithm, ultimately listing the proteins identified in each sample. For each protein found, an identification number will be assigned according to the corresponding protein nomenclature used in the database of reference. However, due to complex procedures involved in sample preparation, a significant portion of acquired spectra may represent chemical or electrical noises. As a consequence, it is not unusual to have peptides identified with high scores but the identifications are due to random matching (false positive rate). On the other hand, it is also not unusual to have “false discoveries” with high scores, although the identifications are due to errors (false discovery rate) (S 1). If not properly controlled, such false positive rate (FPR) and false discovery rate (FDR) may lead to misinterpretation of experimental results. Methods that try to reduce a significant number of both FPR and FDR have been developed. Although unpractical, manual validation remains a common practice, especially for those identifications based on a single peptide. Nonetheless, instruments that use target-decoy database search seems to offer a more reliable solution. In this approach, experimental spectral data are searched against a protein sequence database (the target) and a database comprised of reversed or random amino acid sequences (the decoy). The number of positive identifications from the decoy database is used to estimate FPR and error of FDR in the target database search. By adjusting score thresholds, the number of FPR and FDR can be controlled at desired levels. There are three different score thresholds applicable in Scaffold4, which combined together, will increase or decrease the length of the displayed protein list in the Sample Table: protein threshold, peptide threshold and minimum number of peptides.

The minimum number of peptides option refers to the number of unique peptides that must be found for one protein in order to consider the protein to be identified whereas the protein and peptide thresholds filter the results based on probability (probability that a specific protein is the correct one) or, if the sample were searched using decoys, FDR values. FDR thresholding in Scaffold works by finding the combination of peptide and protein probability thresholds that maximizes the number of proteins identified without exceeding the FDR thresholds and using the selected minimum number of peptides as a lower bound. The values of the selection will be shown in the Dashboard, at the bottom left corner of the program. Understanding these concepts are essential in the visualization of the data as it will affect the number of proteins displayed by the program, thus the results interpretation.

Another interesting function of Scaffold4 allows clustering of data samples in groups, according to the experimental design. For example, samples that represent replicates obtained under the same growing substrate, can be clustered together in one group by editing the BioSample category of each sample and specifying to which it should belong. This is useful whether statistical analysis may be performed among groups or when using the advanced search in the software. A common T-test can be performed among categories using the Quantitative Analysis button, resulting, in the main panel of the Sample section, significant trends (abundance, presence/absence) of specific protein in different samples. Before doing so, it is important to normalize the data among Biosamples to obtain a correct interpretation of the data (done automatically by the software). Once a statistical test has been selected, the user can use the quantitative profile filtering option in advanced search (Sample panel) to highlight which protein belongs to specific filtered categories i.e. proteins found only in the presence of one single substrate. A Venn diagram panel is generated, in the Quantify section of the software, where different subgroups of the filtered proteins can be displayed.

In this study, a target-decoy approach was used to minimize FPR and FDR while criteria for identifications of peptides and proteins in Scaffold were set both at 1.0 % FDR probability if they contained at least two identified peptides. When normalized quantitative value was selected from the view display, the relative abundance of all identified proteins was shown. The most abundant protein within each bend was cross referred to the genomics data and the relevant protease further characterized.

5. Cloning and expression of a putative keratinase gene

A visual summary of the work flow and main protocols described in this fifth section is offered in Figure 13.

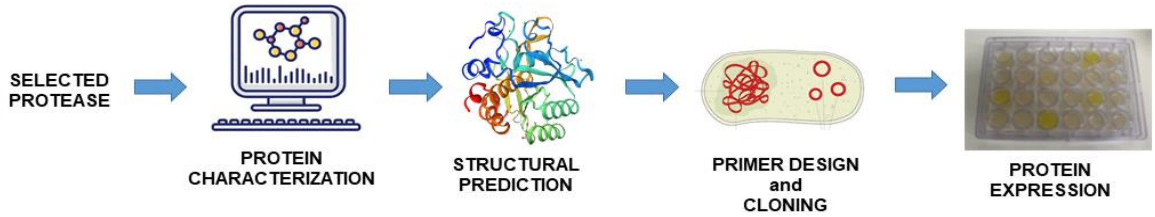


Figure 13 Based on the data and information retrieved from genomics and secretomics, putative serine protease Peg_1025 was selected as candidate enzyme for feather degradation. Its characteristics were described by means of bioinformatics tools and its 3D structure predicted. Primers were design again and cloning in *E. coli* LOBSTR performed both in-vivo and in-silico. Finally, the protease was expressed. Full sequences of the query gene and its encoded protein were retrieved from *F. pennivorans* strain T complete genome annotated in the RAST server.

Protein characterization

Prior to proceeding with the cloning of the serine protease gene (*peg_1025*), more information about its nature, structure, function and catalytic site were retrieved using different bioinformatics tools available in online platforms.

First, basic protein information (e.g. amino acids content, molecular weight, theoretical pI, coefficient of extinction, hydrophobicity) were obtained computing the protein FASTA sequence in the web tool ProtParam (Gasteiger et al., 2005) (https://web.expasy.org/protparam/). Next, the protein sequence was analysed for motifs and domains using scanProSite (Gasteiger et al., 2005) (https://prosite.expasy.org/scanprosite/), pfam (El-Gebali et al., 2019) (https://pfam.xfam.org/), interpro (Hunter et al., 2009) (https://www.ebi.ac.uk/interpro/) and, as already mentioned earlier, SignalIP 5.0 (http://www.cbs.dtu.dk/services/SignalP/) and TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) web tool platforms.

Finally, the information already retrieved by the MEROPS database (Rawlings et al., 2018) described before were integrated with Uniprot (Bateman, 2019) to further unravel the nature of serine protease Peg_1025. From protein family S8 within the MEROPS protein's page, the MEROPS IDs of all homologous peptidases to the query protein which structure was available were cross linked in Uniprot database by the retrieve/ID mapping function. Of the proteases reported as *reviewed* in Uniprot, only the one belonging to the clade Bacteria were selected and FASTA sequences downloaded. Furthermore, some of these proteins had been reported to have a propeptide motif in addition to the catalytic domain and thus these sequences were specifically marked. Then, a multiple sequence alignment was performed between the serine protease Peg_1025 and the reviewed protein sequences obtained from Uniprot using Clustal Omega algorithm (Madeira et al., 2019), in JalView software (Waterhouse et al. 2009). In marking the alignment, colour based on ClustalX algorithm was used with an identity threshold of 40.

Peg_1025 protein sequence was also used to run a local multiple sequence alignment with PSI-BLAST (Madden, 2002) in NCBI within the non-redundant protein sequences (nr) database, a threshold value of 1e-3 and a max target sequence of 1000 hits. The search reached convergence after five interactions (https://www.ncbi.nlm.nih.gov/books/NBK2590/), that is, the last interaction where the most relevant hits were still listed (*F. pennivorans* sequences). All FASTA sequence hits were downloaded and aligned using the same Clustal Omega algorithm described before.

The two alignments, the one with the Uniprot sequences and the other one with the PSI-BLAST sequences, were integrated and taken into consideration together in the identification of conserved motifs and catalytic residues in the query protein. Specifically, for recognition of propeptide motifs in the alignments, previous studies performed by Shinde et al. (1993, 2000) were used. A phylogenetic tree was also calculated based on average distances and BLOSUM62 algorithm, within JalView.

Structural prediction

Based on previous bioinformatics analysis results (i.e. motifs prediction and multiple sequence alignment), the FASTA sequence of serine protease Peg_1025 was divided in two parts composed of a propeptide motif (AA 22 - 127) and a catalytic domain (AA 128 - 439). Propeptide and catalytic domain sequences of serine protease Peg_1025 were individually submitted in SWISS-MODEL (Benkert et al., 2011; Bertoni et al., 2017; Bienert et al., 2016; Guex et al., 2009; Studer et al., 2019; A. Waterhouse et al., 2018) (https://swissmodel.expasy.org/) and their 3D structural model inferred from both fervidolysin (PDB: 1R6V) and subtilisin Ak1 (PDB: 1DBI) templates, when available. Fervidolysin, isolated from *F. pennivorans* DSM9078, and subtilisin Ak1, from *Bacillus stearothermophilus* are well described and characterized proteases (Kluskens et al., 2002; Peek et al., 1993) that resulted the best homologous matches to the query protein.

The structural scores originated from the two templates were compared. In assessing SWISS-MODEL model evaluation, GMQE, QMEAN and its fours terms, local quality plot and model comparison plot were used. For GMQE, higher numbers indicate higher reliability; QMEAN values around zero indicate good agreement between the model structure and the experimental structure; residues of score above 0.6 in the local quality plot are expected to be of high quality and finally, the closer the predicted structure (red star) in the model comparison panel is to the black dots, the more the quality of the predicted structure can be compared to the quality of the experimental structures. The protein prediction with best structural scores was retrieved and visualized using pyMOL Molecular Graphics System (version 1.2r3pre, Schrodinger, LLC). Residues with QMEAN values below threshold compared to the template protein were hidden from the structure. Finally, a superimposed cartoon of the protein structures was also made by fetching both query and template structures in pyMOL, aligning the two amino acids sequences and hiding the residues with low QMEAN scores from the query protein.

Primer design for FX cloning

Primers used to isolate the target gene from *F. pennivorans* strain T genome were already designed in a previous study by Javier-López (2018) (Table 6).

Nonetheless, in this study, primers were designed *de-novo* for learning purposes and only tested *in-silico*. Forward and reverse primers for the open reading frame coding for Peg_1025 were composed of two overall parts: a sequence targeting the gene (target sequence) and a sequence extension on their 5’-end (extension sequence).

The backbone of primers were designed by submitting the gene sequence (devoid of start and stop codons and signal peptide (nucleotides 1 – 63)) into Crystallisation Construct Designer (CCD) web tool (Mooij et al., 2009) (https://ccd.rhpc.nki.nl/). The extension sequences at the 5’ ends in the forward (5’- atatatGCTCTTCtAGTnnn) and reverse (5’-atatatGCTCTTCaTGCnnn) primers were manually added considering the

FX cloning protocol and propagation vector pINITIAL_tet. The target sequence was devoid of its start (ATG) and stop (TAA) codons already present in the expression vectors, in front of the 10xHis-Tag and after the insert, respectively. The signal peptide was omitted according to the expression protocol used. Melting temperature and respective annealing temperature were calculated with online web tool OligoCalc (http://biotools.nubic.northwestern.edu/OligoCalc.html). The primers were tested *in-silico* using the PCR function in Serial Cloner software (SerialBasics) (http://serialbasics.free.fr/Serial_Cloner.html).

Table 6 Forward and reverse primer sequences used in the *in-vitro* cloning protocol, designed from a previous study (Javier-López, 2018).

Forward	5’-atatatGCTCTTCtagtAACTCATTAGAGCCAAGATTTGAACCA
Reverse	5’-atatatGCTCTTCatgcGTATGTCAATGCCTTGTAAGCATCAA

Cloning procedure

The amplified gene of interest (insert) was cloned into expression vectors following a system termed fragment exchange (FX) cloning previously described by Geertsma et al (2011) (Figure 14).

This method is based on type IIS restriction enzymes (in this case *Sap*I) and negative selection markers and involves an initial cloning vector (propagation vector) (Figure 14,B) from which the insert can be sub-cloned into expression vectors (Figure 14,C) designed for the same cloning technique. In this study, the propagation vector chosen was pINIT_tet with tetracycline resistance (Addgene plasmid # 46974) (Geertsma & Dutzler, 2011) (Figure 15) whereas two vectors were chosen for expression, both with kanamycin resistance: plasmids p7xN3H (Addgene plasmid # 47064) (Figure 16,A) and p7xC3H (Addgene plasmid # 47065) (Figure 16,B) (Geertsma & Dutzler, 2011). Both expression plasmids add a tag of 10 histidines to the translated protein, one to the N-terminus of the protein and one to the C-terminus, respectively. The 10xHis-Tag can be later removed by HRV-3C protease digestion leaving a minimal seam of only a single extra amino acid to either side of the protein, Ser to the 5’-N terminal and Ala to the 5’-C terminal.

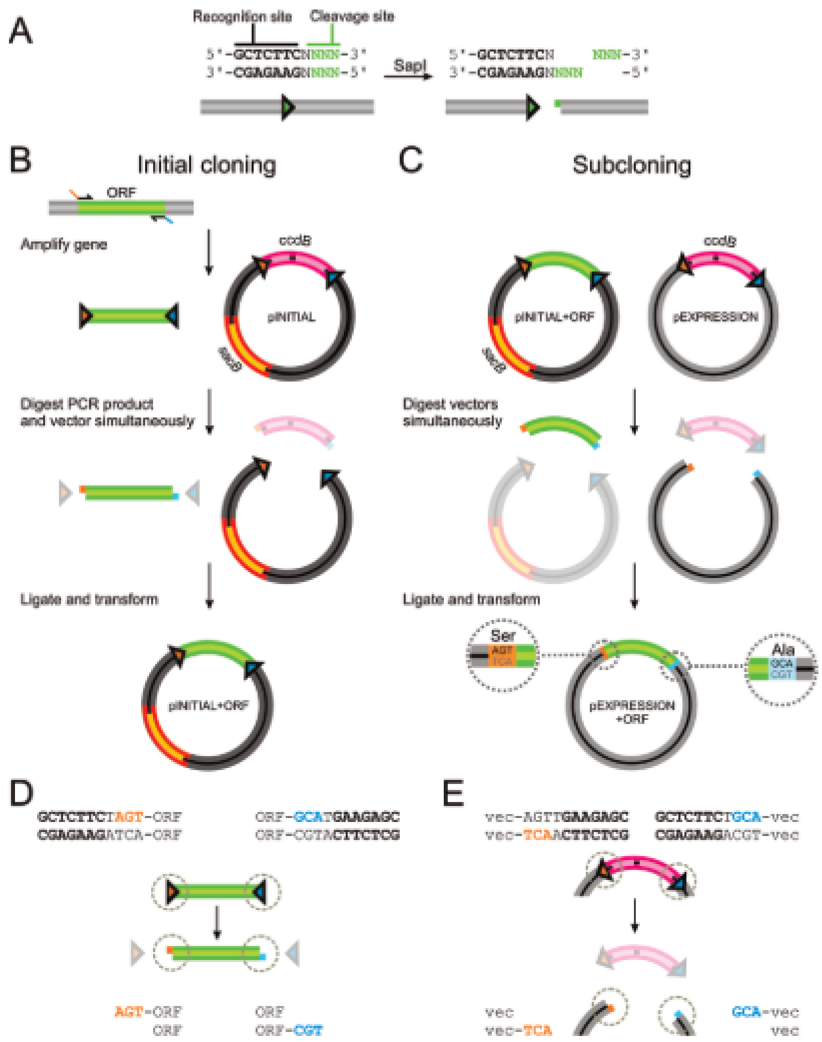


Figure 14 Schematic overview of the FX-cloning method, showing: (A) *Sap*I recognition site (bold letters) with any of the four nucleotides being the cleavage site (green letters). Arrows indicate the direction of the restriction site. (B) Cloning of a PCR product into propagation vector pINITIAL_tet. The genes coding for the counterselection markers ccdB and sacB on pINITIAL are colored in magenta and orange, respectively. (C) Sub-cloning of an ORF into an expression vector. The three nucleotides added to either terminus of the ORF are shown as insets (circle) and are the result of the remaining nucleotides from the cleavage site. (D) Orientation of *Sap*I cleavage sites in the PCR product and pINITIAL and (E) in expression vectors. The single-stranded overhangs generated upon cleavage are shown in orange and magenta.

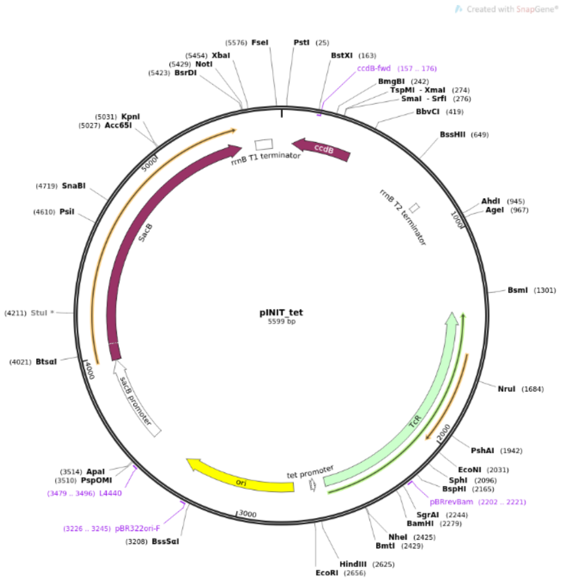


Figure 15 Map of the propagation vector pINIT_tet (Addgene plasmid # 46974), used for the initial cloning of the gene of interest. PCR product of the gene of interest will be inserted in exchange of ccdB counterselective marker gene.

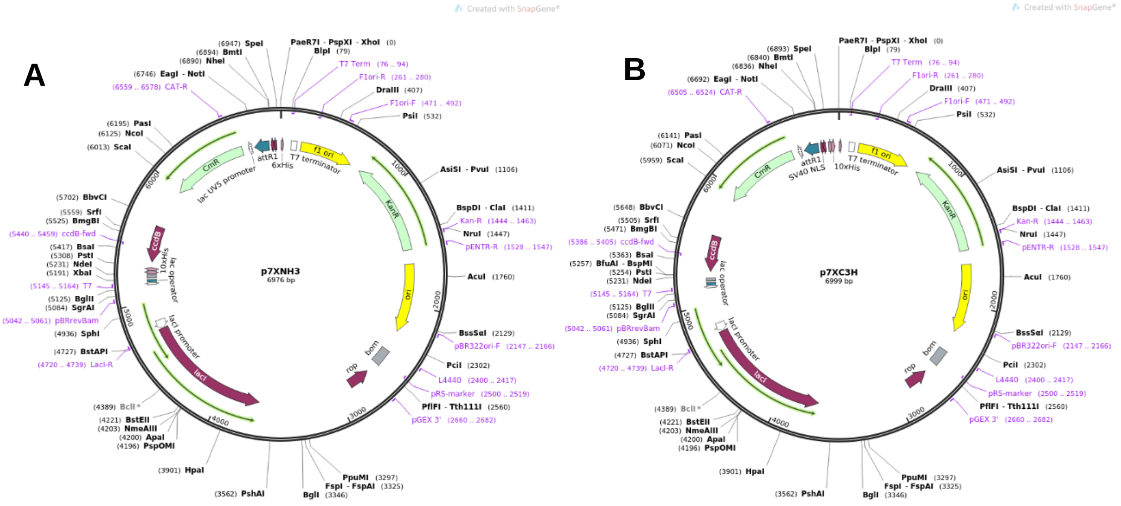


Figure 16 Map and sequences of the expression vectors used in the experiment. p7xN3H (Addgene plasmid # 47064) will add 10x His-tag to the N-terminal of the expressed protein whereas p7xCH3 (Addgene plasmid # 47065) will add 10x His-tag to the C-terminal end. Target gene will be inserted in exchange of the counterselective gene ccdB.

The cloning description that follows was already carried out in the previous study performed by Javier-López (2018). Nonetheless, the whole procedure was performed again in this study, both *in-silico,* with serial cloner software, and *in-vivo*. It will briefly be described but, for a detailed description of the work, it should refer to Javier-López (2018). Gene_1025 was amplified by PCR using Phusion High-Fidelity Polymerase (Table 7) using primers in Table 6. PCR products size and quality were checked in a 1.5% agarose gel electrophoresis run for 45 minutes at 5 V/cm.

Table 7 PCR program used to amplify the gene *peg_1025* from the complete genome of *F. pennivorans* strain T using primers in Table 6.

Step	Temperature (Celsius)	Time (seconds)
Initial denaturation	98	30
30 cycles	98	10
Annealing	52	30
Extension	72	45
Final extension	72	420

In the meantime, *E. coli* strain containing the propagation vector (pINIT_tet) was grown overnight at 37°C in LB medium (tetracycline 10 µg ml-1) (Table 16, A). The plasmid was extracted and 50 ng was mixed with 250 ng of PCR product (final molar ratio 1:5 vector:insert) in 1 µl of 10X buffer adjusted to 9 µl with MQ water. After 1 hour digestion with restriction enzyme *Sap*I (2U), the reaction was stopped and the resulting fragments ligated for 1 hour by the addition of 0.8 µl of T4 DNA Ligase and 1.2 µl of ATP mix (10mM).

Five microliters of this final mixture was used to transform chemically competent *E. coli* cells (One Shot™ TOP10) for 30 minutes in ice and, after heat shock at 45°C for 30 seconds, cells were recovered at 37°C in 250 µl S.O.C. medium (Table 10, C) for 1 hour. Transformed cells were first plated onto LB medium (tetracycline 10 µg ml-1) (1.5% agar) overnight at 37°C, and colonies formed were picked and incubated in 10 ml LB (tetracycline 10 µg ml-1) overnight at 37°C. Finally, the plasmid was isolated and, to verify the insert sequence, a PCR was performed (Table 8, Table 9). The insert length was checked by 1.5 % agarose gel electrophoresis run for 45 minutes at 5 V/cm.

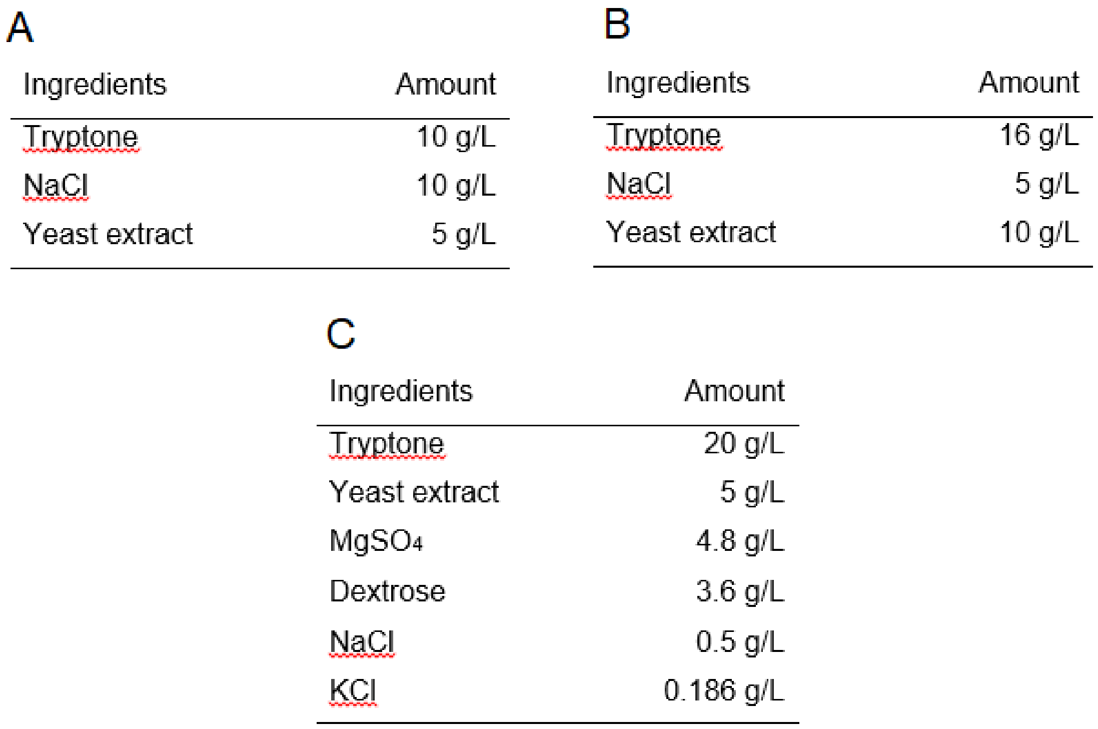
Table 8 Primers used in the PCR conducted to verify the pINIT_tet + gene_1025 construction. The primers were chosen considering the pINIT_tet sequence.

Forward	5’- GAGTAGGACAAATCCGC
Reverse	5’-TGCTTCGCAACGTTCAAATCCGC

Table 9 PCR program used to verify the pINIT_tet + gene_1025 construction.

Step	Temperature (Celsius)	Time (seconds)
Initial denaturation	96	300
25 cycles	96	10
Annealing	46	5
Extension	60	240
Final extension	72	420

Table 10 Composition of media used during the cloning and expression protocols. A, Luria-Bertani (LB) medium; B, 2YT medium; C, S.O.C. medium.



A

Ingredients	Amount
Tryptone	10 g/L
NaCl	10 g/L
Yeast extract	5 g/L

B

Ingredients	Amount
Tryptone	16 g/L
NaCl	5 g/L
Yeast extract	10 g/L

C

Ingredients	Amount
Tryptone	20 g/L
Yeast extract	5 g/L
$MgSO_4$	4.8 g/L
Dextrose	3.6 g/L
NaCl	0.5 g/L
KCl	0.186 g/L

With the propagation vector ready, sub cloning of the insert into expression vectors begun. Two *E. coli* strains containing one expression vectors each, p7xCH3 and p7xN3H, were grown overnight in 10 ml LB medium containing kanamycin (50 µg ml^{-1}). Plasmids were extracted and 50 ng of them were mixed with pINT_tet-insert plasmids to a final molar ratio of 1:4 (expression vector:pINIT_tet plasmid). The same sub-cloning and transformation protocols described before were followed. A derivative strain from chemically competent *E. coli* BL21(DE3) called strain LOBSTR (Low Background Strain) was chosen as expression host. The strain exhibits proteins with reduced affinities to nickel-based resin columns, resulting in a much higher purity of his-tagged target proteins after expression and purification (Andersen et al., 2013).

LOBSTR transformed cells were plated on LB medium (kanamycin 50 µg ml^{-1}) overnight at 37°C first and, then, resulting colonies picked and grown in liquid LB medium (kanamycin 50 µg ml^{-1}). To verify this final transformation procedure, expression vectors were extracted and checked in 0.8% agarose gel electrophoresis run for 45 minutes at 5 V/cm. In addition, a PCR was conducted (Table 12), using specific primers for p7xN3H and p7xC3H (Table 11), to check their presence in the expression vectors.

Table 11 Primers used in the PCR conducted to amplify the expression vectors p7xN3H and p7xC3H.

Forward	5’- TAATACGACTCACTATAGGG
Reverse	5’-GCTAGTTATTGCTCAGCGG

Table 12 PCR program used to verify the presence of p7xN3H and p7xC3H in the expression vectors *E. coli* LOBSTR.

Step	Temperature (Celsius)	Time (seconds)
Initial denaturation	94	60
25 cycles	94	15
Annealing	50	30
Extension	68	60
Final extension	72	420

Two cultures of *E. coli* LOBSTR expression strains, one containing the p7xN3H vector and the other one containing p7xC3H vector, both with gene_1025 inserted, were finally obtained and kept at -80°C. These cultures were used as inocula source for the protein expression protocol.

Protein expression

Precultures of *E. coli* LOBSTR with p7xN3H and p7xC3H vectors were started by scratching a pipette tip onto the frozen cultures and put it into 3 ml 2YT medium (kanamycin 50 µg ml^{-1}) (Table 10, B) and incubated overnight at 37°C. A preculture of *E. coli* LOBSTR without expression vectors was also set and used as negative control for the entire experiment. A control with 3 ml 2YT medium without any inoculum was also set to monitor eventual contaminations throughout the experiment.

Fifty millilitres of 2YT medium (kanamycin 50 µg ml^{-1}) prewarmed at 70°C was inoculated with 1000 µl pre-cultures and incubated at 37°C and 175 rpm agitation. A total of five cultures were set: two cultures growing LOBSTR with p7xN3H, two cultures growing LOBSTR with p7xC3H and one culture with *E. coli* LOBSTR without expression vectors. The absorbance at 600 nm was measured during incubation until a value of 0.3 - 0.4 was reached. At this point, one duplicate from each vector was induced with 200 µl Isopropyl β-D-1-thiogalactopyranoside (IPTG) (0.1 M) to a final concentration of 0.4 mM. After induction, cultures were grown overnight at 20°C, 150 rpm agitation, before they were harvested by centrifugation at 7000 x g for 15 minutes at 4°C.

Harvested cells were resuspended with 5 ml lysis buffer (50 mM Tris-HCl pH 7.5, 50 mM NaCl, 10% glycerol, pH 7.5). Then, the solutions were supplemented with lysozyme (10 mg/ ml) and incubated at 37°C for 30 minutes before being disrupted by 5 cycles of 30 seconds on / 30 seconds off, 50% duty cycles (pulse), 4 output control (amplification) of sonication (Branson Sonifier 250). The cells were kept on ice during this procedure. Cell disruption was checked under phase-contrast microscope before centrifuging the solutions at 7000 x g for 10 minutes at 4°C. The supernatant was heat shocked (HS) at 70°C for 15 minutes and centrifuged at 5000 x g for 20 minutes at 4°C to remove denatured *E. coli* proteins. The supernatant was collected and half of it (2 ml from each culture) was further heat treated at 70°C for 2 hours to activate the expressed protease (HA) (Toogood et al., 2000). Finally, cell lysates containing the expressed protease N-terminal his-tagged (HS and HA samples) and C-terminal his-tagged (HS and HA samples) were concentrated to a final volume of ~500 µl with 15 ml 10K Amicon® Ultra^{-1}5 centrifugal filters (Millipore) and frozen at -20°C until use.

At different stages of the expression, sample aliquots were analysed by SDS-PAGE to monitor expression efficiency. Briefly, 30 µl aliquots from each culture were mixed in 30 µl 2X loading buffer (230 mM Tris-HCl pH 6.8, 25% glycerol, 2% SDS, 0.02% Bromophenol blue, 5% 2-β-Mercaptoetanol), heat-incubated at 95°C for 5 minutes and 20 µl loaded onto pre-made gels (Mini-PROTEAN® TGX gels, BIO-RAD). Colour prestained protein standard (Broad Range, 10-250 kDa, Bio-Labs) was used as molecular weight marker. The gel was run at 200 V for 50 minutes, stained with Coomassie Blue for 1 hour in gentle agitation, distained with distaining solution (50% distilled water, 40% methanol, 10% acetic acid) for 1 hour and washed in distilled water overnight.

Protein concentration

The concentration of total proteins in the cell lysates (N-HA, N-HS, C-HA, C-HS) was measured using 5 µl aliquot following Bradford protein assay (QuickStart^{TM}- BIO-RAD) manual. All the measurements were run in technical triplicates.

6. Enzyme activity evaluation

A visual summary of the work flow and main protocols described in this sixth and final section is offered in Figure 17.

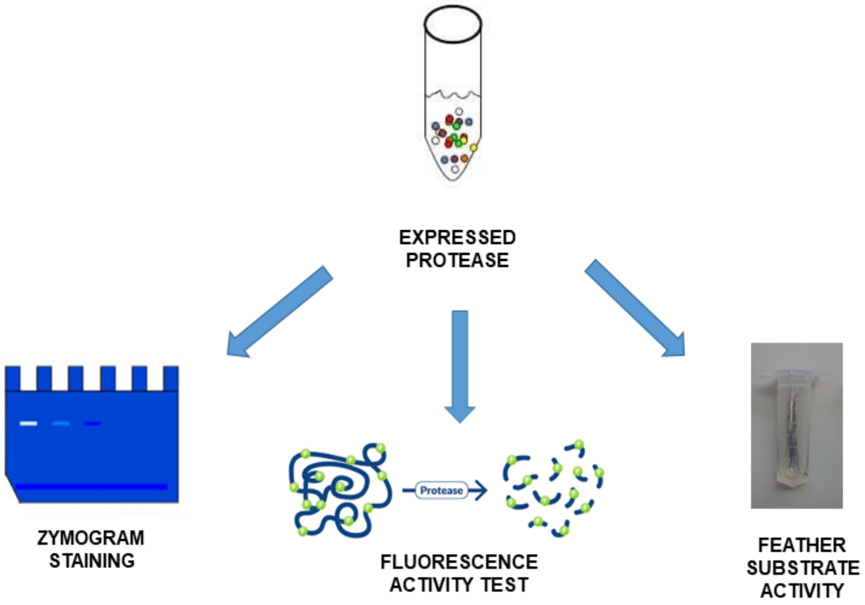


Figure 17 Cell lysates containing newly expressed serine protease *peg_1025* were used to perform three different tests to evaluate the activity of the expressed enzymes: zymogram staining, FITC-casein fluorescence detection kit and feather-substrate enzyme assay. Both heat shocked (-HS) and heat activated (-HA) samples were tested and activity compared.

Zymogram staining

For zymogram analysis, SDS-PAGE gel was made containing 1% skimmed milk. The gel was composed of a resolving gel (3.4 ml MQ, 4.0 ml 30% acrylamide/bis 37:5:1, 2.5 ml 1.5 M Tris-HCl pH 8.8, 0.1 ml 10% SDS, 1.0 ml 10% skimmed milk) and a stacking gel (1.7 ml MQ, 2.0 ml 30% acrylamide/bis 37:5:1, 1.25 ml 1 M Tris-HCl pH 6.8, 0.05 ml 10% SDS), added of 150 µl APS (10%) and 75 µl TEMED for polymerization. Thirty µl cell lysates samples (N-HS, N-HA, C-HS, C-HA) (concentration 0.6 mg ml^{-1}) were mixed with 30 µl 2X loading buffer (62.5 mM Tris-HCl pH 6.8, 25% glycerol, 4% SDS, 0.01% Bromophenol Blue) and loaded in the gel. *E. coli* LOBSTR cell lysate and protease K (2 mg ml^{-1}) were also added as negative and positive control, respectively. Colour prestained protein standard (Broad Range, 10-250 kDa, Bio-Labs) was used as molecular weight marker. After run 50 minutes at 200 V, the gel was washed in renaturing buffer (1X SDS, 2.5% Triton X-100), equilibrated in reaction buffer (50 mM Tris-HCl pH 7.8, 200 mM NaCl, 5 mM $CaCl_2$) for 30 minutes and then incubated at 70°C in reaction buffer with gentle agitation overnight. Finally, the gel was stained in Coomassie blue for 1 hour with gentle agitation, destained with destaining solution (50% distilled water, 40% methanol, 10% acetic acid) for 1 hour and washed in distilled water overnight.

Proteolytic activity assessment

Protease activity was carried out using FITC-labelled casein substrate following a modified version of fluorescent detection kit assay (Sigma-Aldrich, c.n. PF0100) (Bjerga et al., 2016; Twining, 1984). Twenty microliters of reaction buffer (50 mM Tris-HCl pH 7.8, 200 mM NaCl, 5 mM $CaCl_2$) was mixed with 20 µl FITC-casein and 10 µl cell lysates samples (N-HS, N-HA, C-HS, C-HA) (0.6 mg ml^{-1}) and incubated at 70°C. The same reaction was set using 10 µl protease K (2 mg ml^{-1}) and 10 µl *E. coli* LOBSTR cell lysate as positive and negative control, respectively. An extra reaction was set using 10 µl MQ as negative control and to blank the instrument before the readings. After 1 hour incubation, reactions were stopped by addition of 150 µl 0.6 N trichloroacetic acid solution (TCA) and samples further incubated at 37°C for 30 minutes in incubator before centrifuged at 10000 x g for 10 minutes. Finally, 10 µl supernatant was diluted in 200 µl 500 mM Tris-HCl (pH 8.5) into black plate reader wells. Fluorescence intensity was recorded with excitation at 485 nm and emission at 535 nm. The same set up was repeated in technical triplicates and for 24 hours incubation to increase the sensibility of the assay.

Substrate enzyme assay

To assess the effectiveness of the expressed protease on actual substrate, an enzyme assay was run with chicken feather substrates and the cell lysate samples (N-HA, N-HS, C-HA, C-HS) (0.6 mg ml^{-1}). One small eppendorf tube for each sample containing ~0.05 g chicken breast feathers, 900 µl reaction buffer (50 mM Tris-HCl pH 7.8, 200 mM NaCl, 5 mM $CaCl_2$) and 100 µl cell lysate samples were incubated at 70°C for more than 10 days. Tubes containing of *E. coli* LOBSTR cell lysate and protease K (2 mg ml^{-1}) were also set as negative and positive control, respectively.

Results

1. Physiology analysis

Feather degradation

The thermophilic anaerobe, *F. pennivorans* strain T, was grown in anoxic MMF flasks with different substrates and its ability of degrading native feathers was investigated. First, the physiology of fresh growing cultures was studied. Healthy cells were rod-shaped with a characteristic outer sheath-like structure, the toga, occurring singly, in pairs or as short chains. A spheroid “bleb” extension of the toga was detected at one terminal end of the cells (Figure 18). Cells did not form any spores even after long time incubation. Gas production was observed in cultures growing on different substrates and the gas pressure increased inside the serum flasks with the incubation time. The composition of gases detected was on average of 18.3% hydrogen and 6.7% carbon dioxide, independently by the substrate used, on a 75% background of nitrogen derived from the flushing technique. Sulphide could be smelled only in cultures growing on keratin azure or chicken feathers. The periodic release of gas pressure from the anoxic flasks resulted in increased culture turbidity.

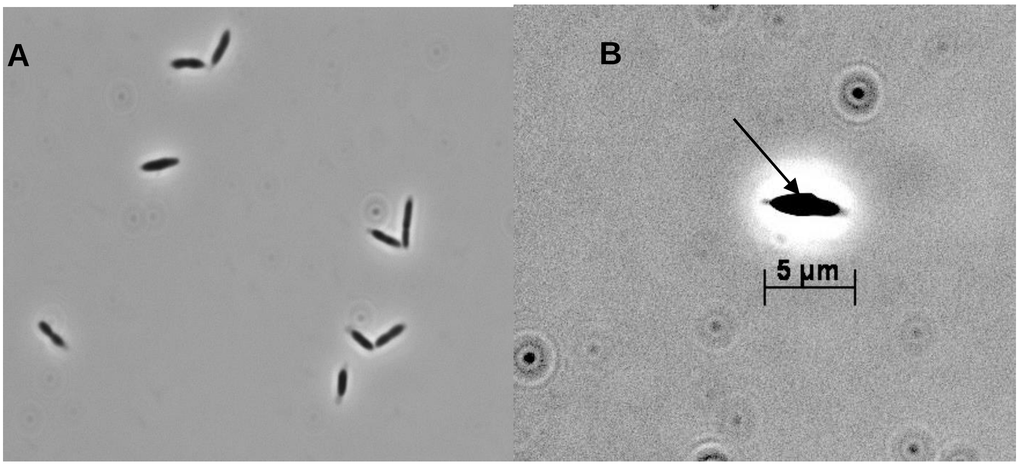


Figure 18 *F. pennivorans* strain T visualized under phase-contrast microscope. A, the organism is characterized by rod-shaped cells occurring singly, in pairs or as short chains; B, black arrow points to the spheroid extension of the toga (“bleb”) found at one terminal end of the cells.

F. pennivorans strain T degraded chicken feathers from breast and wings, with different rates. Chicken breast feathers were soft, light and fluffy whereas wing feathers were stiffened, hard and heavier. Chicken breast feather degradation was almost complete within 3 days as judged by visual inspection of cultures (Figure 19) when cells were grown with MMF supplemented with 0.1% yeast extract and 0.2% peptone whereas wing feathers degradation started only after 7 days of incubation at the same conditions. Full degradation of wing feather was reached within 10 days (Figure 20).

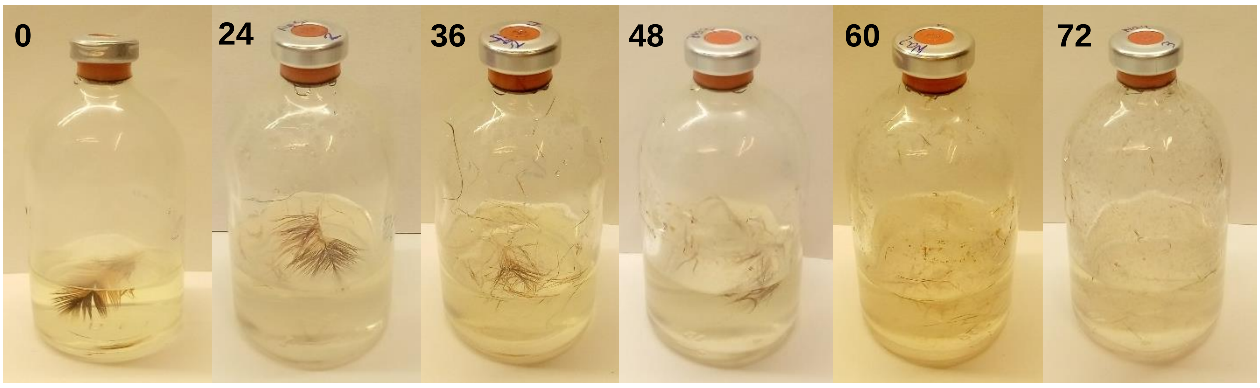


Figure 19 Anaerobic cultures of *F. pennivorans* strain T showing chicken breast feather degradation throughout 3 days incubation at 65°C. Incubation time in hours is also shown.

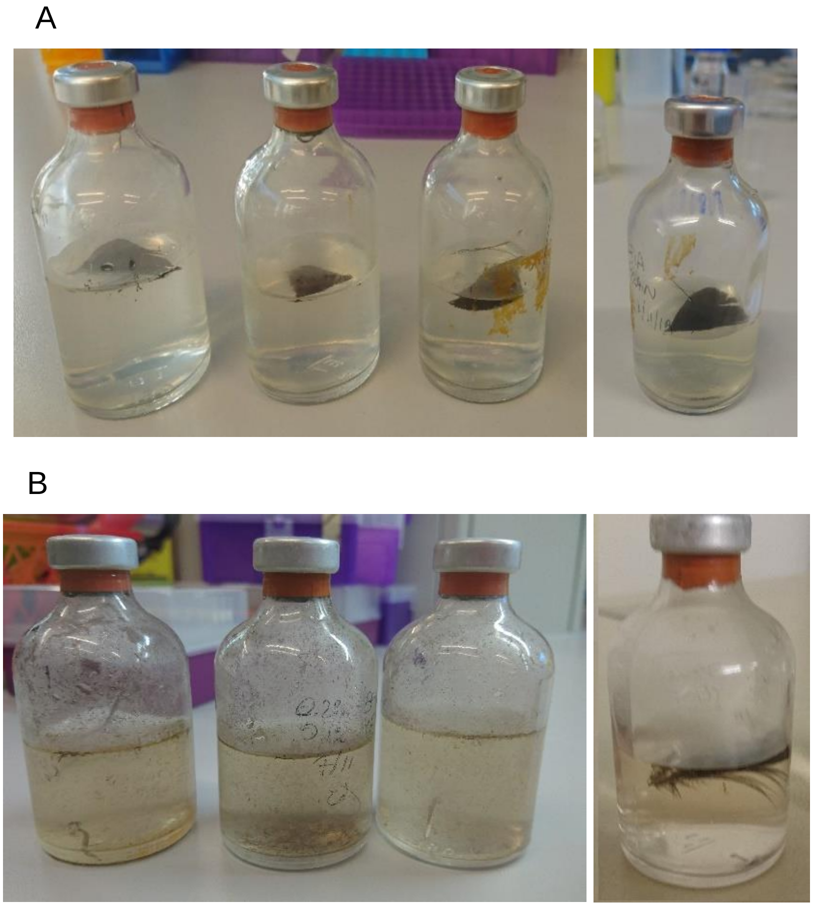


Figure 20 Three replicates of anaerobic cultures of *F. pennivorans* strain T showing chicken wing feather at the beginning (A) and after 10 days (B) incubation at 65°C. A negative control is also shown (right).

Culture batch assay

During the enzyme hydrolysis of feathers by *F. pennivorans* strain T, the sulfhydryl group abundance in the batches was monitored following Ellman’s reagent protocol. The assay did not show any increase in sulfhydryl groups during the course of bacterial growth compared to the negative control, although feathers were clearly being degraded (Figure 19).

2. Complete genome sequence analysis

DNA isolation and sequencing

Genomic DNA was extracted and its quality and size were checked by 0.8% agarose gel electrophoresis. DNA yield was determined by NanoDrop reading. Gel electrophoresis showed a good quality of high molecular weight genomic DNA, with little smearing and abundant material, suggesting highly intact genomic DNA (Figure 21). DNA concentration was measured to 21 µg ml^{-1}.

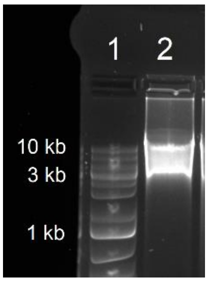


Figure 21 Gel electrophoresis of *F. pennivorans* strain T genomic DNA preparation. Lane 1, DNA ladder (Gene Ruler DNA Ladder MIX, ThermoFisher); lane 2, genomic DNA sample.

Quality and total amount were sufficient for carrying out PacBio sequencing according to the sequencing facility standards (Eurofins Genomics). After de-novo assembly performed by Eurofins Genomics, a long contig (contig_1) composed of 2002515 base pairs, with a coverage of 192 X, and a shorter contig (contig_2) of 12262 base pairs long, with a coverage of 8.6X, were obtained. Details of the sequencing reports are available at the end of the appendix.

Genomic revision

To generate a complete genome sequence of *F. pennivorans* strain T, contig_1 and contig_2 quality was inspected and the sequences manually curated.

Contig_1 was manually reordered and functionally annotated by the RAST server. The complete genome was set to start by the chromosomal replication initiator protein gene (dnaA), as for most bacterial genomes deposited in public databases. In this way, comparison by genome alignments with related strains e.g. *F. pennivorans* DSM9078 and *F. islandicum* AW-1 was easier. The query genome aligned very nicely to the type strain genome, although was slightly shorter (Figure 22). The same pattern was found when strain T genome was aligned with *F. islandicum* AW-1 genome (S 4). Nonetheless, some genomic parts were unique to the query genome and did not have any homologue in the other two relatives. Overall, strain T genome was very similar to the other two, containing all the major conserved locally collinear blocks (LCBs). Thus, contig_1 most probably represented the entire genome of *F. pennivorans* strain T. In particular, three LCBs in strain T genome (Figure 22, yellow, light green and light blue LCBs) resulted inverted compared to the type strain's.

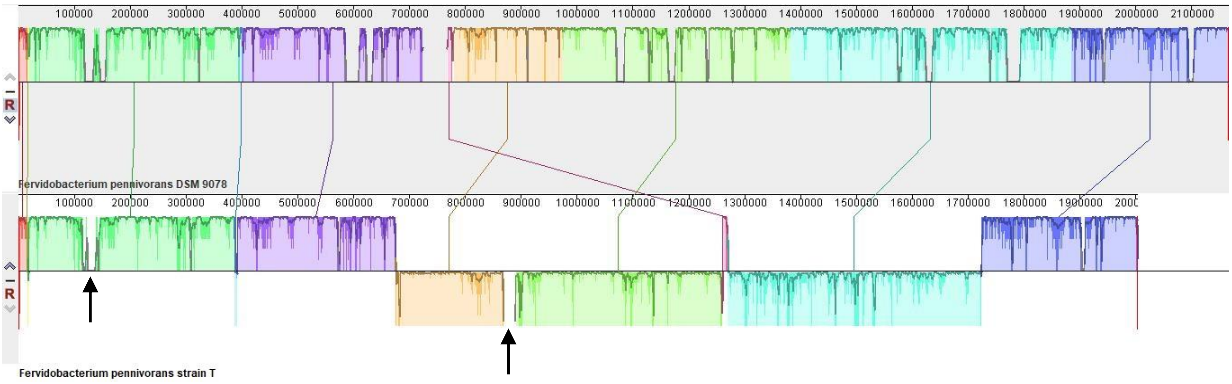


Figure 22 Multiple alignment between *F. pennivorans* DSM9078 (top) and *F. pennivorans* strain T (bottom) complete genomes. Similarly conserved locally collinear blocks (LCBs) are coloured in the same way and connected by lines. Genomic parts unique to strain T are shown by gaps in the query genome and further marked by black arrows. Alignment obtained by MAUVE.

The shorter sequence contig_2 was inquired as an artefact product and its source investigated by Blastn analysis (NCBI), alignment analysis (both within NCBI and Mauve) and genomic annotation tools in RAST server.

Submission of contig_2 sequence within the nucleotide collection (nr/nt) in Blastn reported best hits belonging to *Fervidobacterium pennivorans* strains DYC (Query cover = 100%; E value = 0.0; % identity = 95.86%) and strain DSM9078 (Query cover = 89%; E value = 0.0; % identity = 94.94%). Furthermore, alignment of contig_2 against contig_1 showed a significant match within contig_1 (E value = 0.0; % identity = 99.04%). The details of the alignment were visualized by Mauve software and showed contig_2 to fully align between nucleotides ~272000 and ~286000 in contig_1 (Figure 23). However, contig_2 seemed to be lacking a sequence part, comprised between nucleotides ~279500 and ~281000 of contig_1, corresponding to a CDS for a mobile element protein (*peg_266*) (transposase).

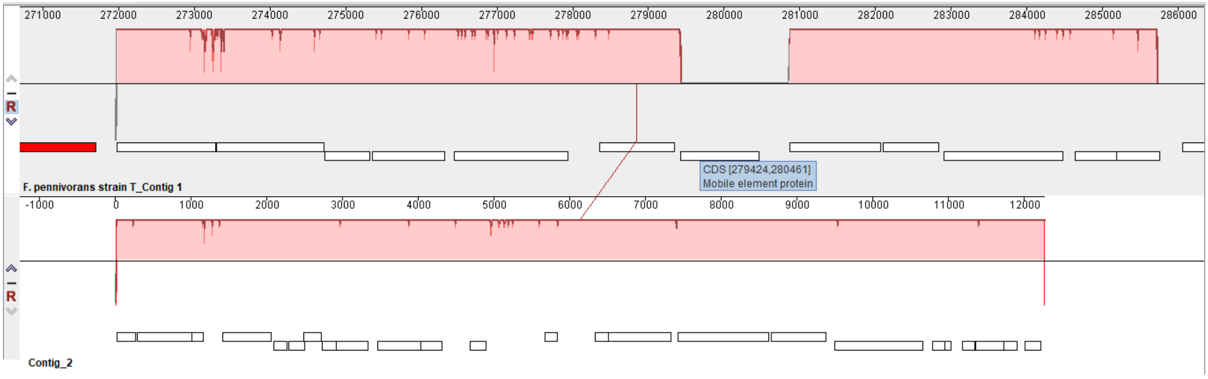


Figure 23 Mauve alignment between contig_1 (top) and contig_2 (bottom). Predicted genes in the two contigs are shown with white bars. Coding sequence for mobile element protein (peg_266) is indicated.

Finally, RAST server annotation revealed that the genes identified in contig_2 were disrupted sequences of original genes in contig_1. The gene coding for a mobile element protein (*peg_266*) was not present in the contig_2 annotation.

As an integrated interpretation of the results, contig_2 sequence was considered being an artefact and deleted from the FASTA file of the complete genome. A single, long contig was obtained and further processed and characterized.

Genomic characterization and gene mining

F. pennivorans strain T complete genome was 2002515 base pair long compared to *F. islandicum* AW-1 and *F. pennivorans* DSM9078 genomes that measured 22373777 and 2166381 base pairs respectively (Table 13). Strain T genome contained a total of 2001 genes, including 1949 protein coding genes and 52 RNAs genes.

Table 13 General characteristics of *F. pennivorans* strain T complete genome retrieved from SEED view in RAST server. *F. pennivorans* DSM9078 and *F. islandicum* AW-1 general information are also reported.

	F. pennivorans strain T	*F. pennivorans* DSM9078	*F. islandicum* AW-1
Domain	Bacteria	Bacteria	Bacteria
Size (base pairs)	2,002,515	2,166,381	2,237,377
GC Content (%)	39.0	38.9	40.7
L50	1	1	1
Number of Subsystems	294	206	199
Number of CDS*	1949	2079	2144
Number of RNAs	52	55	53

*CDS: coding sequences

Of the 1949 annotated coding sequences, 49% percent (937) were identified by gene comparison to the RAST Subsystem (curated database) (Figure 24) and could be further categorized according to the biological function they served. The majority of these genes were involved in essential biological functions such as protein metabolism (202), carbohydrate metabolism (189), amino acids anabolism (119), biosynthesis of cofactors or other secondary metabolites (111) and RNA metabolism (105). Within each category, genes were further divided into more specific features. In this regard, within the protein metabolism category, 29 protein degrading genes were listed and their annotation used to extrapolate the keywords for mining protease encoding genes. Out of 1949 coding sequences (CDS), 472 gene functions were predicted by RAST bioinformatics algorithms whereas 540 protein coding genes were annotated as hypothetical proteins and function remained unknown.

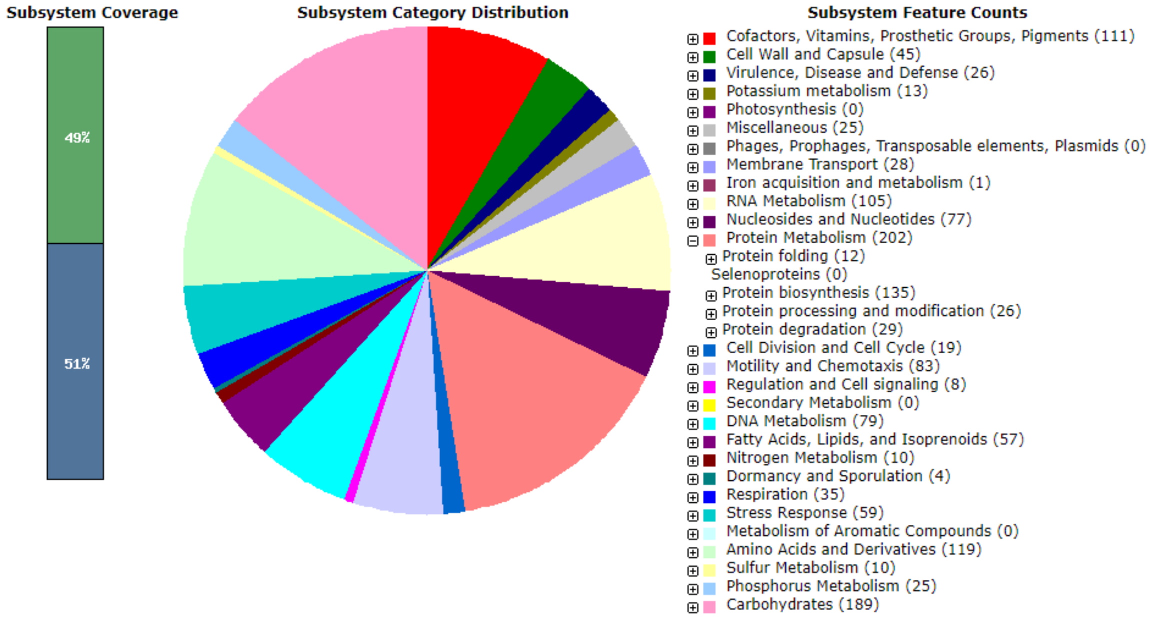


Figure 24 Pie chart showing the distribution of Subsystem-predicted genes based on their biological functions (category) in *F. pennivorans* strain T genome. The number of genes involved in each category is also reported. The category regarded as protein metabolism, involving a total of 202 genes, was opened and shows the detailed gene distribution. On the left, Subsystem coverage is reported, representing the percentage of protein coding genes identified from the RAST Subsystem (curated database) (green) and the ones which function was predicted by bioinformatics algorithms (blue).

Peptidase classification

A total of 58 genes encoding for putative protein-degrading enzymes were found after gene mining and genomic comparison between *F. pennivorans* strain T and its two closely related organisms, *F. pennivorans* DSM9078 and *F. islandicum* AW-1.

The cellular location of the 58 proteases were predicted and shown 29 enzymes that exploited their function intracellularly (S 5), 14 bond to the cellular membrane and 15 were predicted to be secreted in the extracellular environment by the presence of a signal peptide. The 29 enzymes whose function was exploited either extracellularly or on the cellular membrane of the organism, were further validated and characterized by MEROPS database pipeline.

MEROPS and RAST based annotations of the genes were compared. Both databases reported similar information about specific proteins, although the level of annotation by RAST were broader and less specific. For example, *peg_1025* was annotated by RAST as serine protease whereas by MEROPS as subtilisin Ak1. The latter is also a serine protease, but the most available level of information on the protein based on RAST algorithm was the superfamily.

Of the 29 predicted intracellular enzymes, 3 proteins (Peg_1050, Peg_1218, Peg_1608) were not identified as proteases by the MEROPS database and thus were excluded from the report. The others were classified into three main superfamilies, Serine- (12), Metallo- (10) and Aspartic- (4) proteases and further divided into 17 different families (Table 14). No proteases belonging to Cysteine or Threonine protease families were found. Despite of all 26 enzymes being annotated as proteases, not all of them seemed to be involved in catabolism but in more general cellular physiology, instead. For example, some proteases were predicted to be involved in proteins processing (e.g. peg_924, Peg_1261); others in cutting the signal peptide off from proteins directed to the extracellular environment (e.g. Peg_638, Peg_640, Peg_1232, Peg_1654). Other enzymes were involved into cellular division (e.g. Peg_399, Peg_1608).

Table 14 Functional annotation of 26 protease-encoding genes in *F. pennivorans* strain T whose activity is either extracellular (Extr) or bond to the cellular membrane (Trsm) of the organism. Enzymes are listed by their location on the complete genome according to RAST server annotations. The genomic location of these genes is visually indicated in Figure 25.

Locus (fig|93466.12._)	**Cellular location**	**RAST annotation**	**MEROPS annotation**	**MEROPS ID[a]**
peg.362	Trsm	Metalloendopeptidases	Beta-lytic metallopeptidase	M23.013
peg.399	Trsm	Cell division protein FtsH	FtsH peptidase	M41.021
peg.638	Trsm	Lipoprotein signal peptidase	Signal peptidase II	A08.001
peg.640	Trsm	Lipoprotein signal peptidase	Signal peptidase II	A08.001
peg.914	Trsm	Membrane-associated zinc metalloprotease	Site 2 peptidase	M50.A05
peg.924	Extr	HtrA protease/chaperone protein	HtrA peptidase	S01.491
peg.926	Trsm	Prepilin peptidase	Type 4 prepilin peptidase 1	A24.A10
peg.1025	Extr	Serine protease	Subtilisin Ak1	S08.009
peg.1050*	Extr	Multimodular transpeptidase-transglycosylase	Glutamate carboxypeptidase	M20.UPF
peg.1085	Trsm	Membrane protein of glp regulon	Rhomboid YqgP peptidase	S54.014
peg.1218*	Trsm	Activity regulator of membrane protease YbbK	Signal peptide peptidase A	S49.UPC
peg.1232	Trsm	Signal peptidase I	Signal peptidase SipV	S26.006
peg.1261	Trsm	Heat shock protein HtpX	PAB0555-type putative peptidase	M48.010
peg.1364	Trsm	Related to CAAX prenyl protease	Mername-AA052 peptidase	M48.008
peg.1405	Extr	D-alanyl-D-alanine carboxypeptidase	D-Ala-D-Ala carboxypeptidase	S13.002
peg.1439	Trsm	Metal-dependent peptidase	Rhomboid-1	S54.A19
peg.1512	Extr	L-alanoyl-D-glutamate peptidase	Ply118 L-Ala-D-Glu peptidase	M15.020
peg.1516	Extr	L-alanoyl-D-glutamate peptidase	L-alanyl-D-glutamate peptidase	M15.022
peg.1608*	Extr	Peptidoglycan synthetase	D-Ala-D-Ala carboxypeptidase	S11.UPW
peg.1654	Trsm	Lipoprotein signal peptidase	Signal peptidase II	A08.001
peg.1681	Extr	Subtilase-type serine protease	Fervidolysin	S08.021
peg.1693	Extr	Subtilase-type serine protease	Fervidolysin	S08.021
peg.1784	Trsm	Rhomboid family serine protease	RhoII peptidase	S54.027
peg.1828	Extr	Carboxyl-terminal protease	C-terminal processing peptidase-1	S41.004
peg.1868	Extr	Metalloendopeptidases	Endometallopeptidase	M23.009
peg.1949	Extr	D-aminopeptidase DppA	D-aminopeptidase DppA	M55.A01

*Unassigned proteases: proteases not present in MEROPS database;

[a]: Superfamilies (A, aspartic; M, metalloproteases; S, serine); family are indicated by the two numbers after the superfamily (e.g. x08, x23, x55…);

Genomic and peptidases comparison

Genomes of *F. pennivorans* strain T, *F. pennivorans* DSM9078 and *F. islandicum* AW-1 were compared using the BLAST Ring Image Generator (BRIG) platform, also considering the identified proteolytic enzymes found in *F. pennivorans* strain T (Figure 25).

Comparison of these two related microorganism's genomes with *F. pennivorans* strain T showed again clear differences in length and conservation. Some genomic portions of the organisms could not be compared against strain T genome, for example, between 840 - 920 Kbp and 80 – 160 Kbp. That is, the reference genome had gene sequences that were not present in the other genomes and thus could not be blasted by the program, creating gaps in the two outer rings. On the other hand, some genomics portions were highly conserved throughout the three organisms, as it was seen between 80 – 120 Kbp and 640 – 680 Kbp.

When the content of protease-encoding genes from each organism was compered against each other, all 55 genes were found in all genomes. They seemed to be evenly distributed throughout the entire strain T genome, with only few clusters detectable. For example, genes *peg_1512* and *peg_1516*, which exploit similar functions, were consecutive to each other. Some other genes were present in duplicates and still located close to each other, such as genes for signal peptide cleavage (*peg_638* and *peg_640*) or fervidolysin genes (*peg_1681* and *peg_1693*).

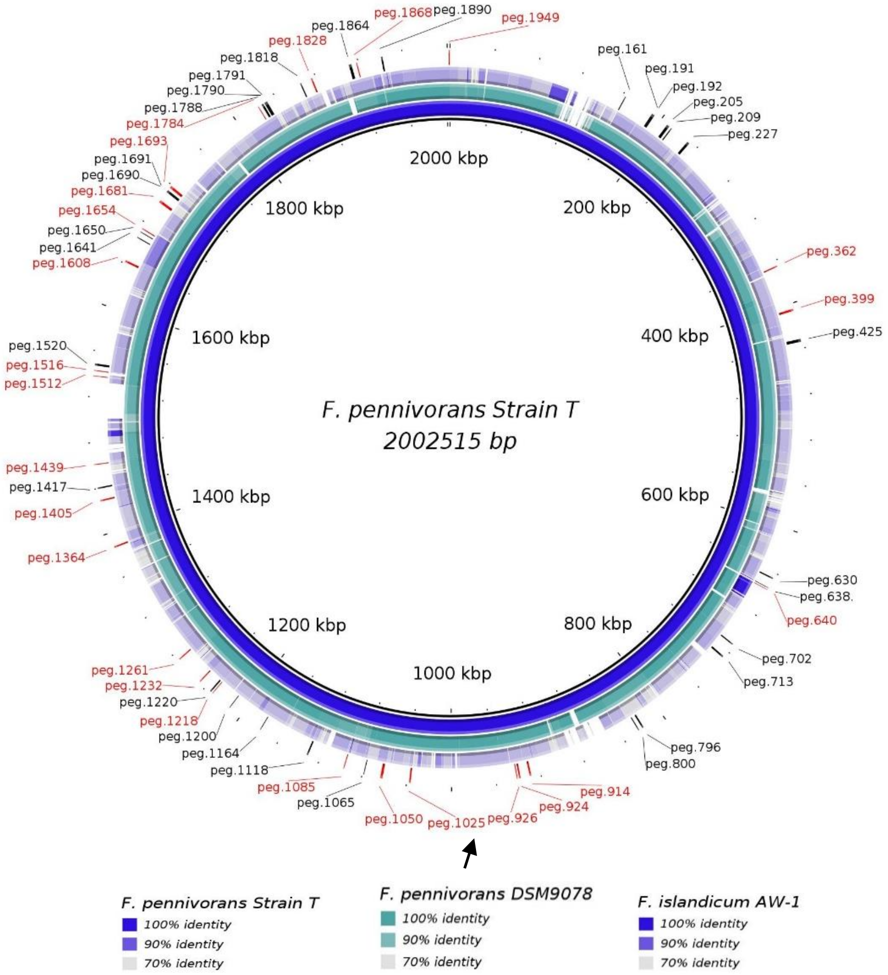


Figure 25 Genomic comparison between *F. pennivorans* strain T, *F. pennivorans* DSM9078 and *F. islandicum* AW-1 and overall protease-encoding gene locations with reference to the strain T genome. Strain T was used as reference genome. Rings (from inside), *F. pennivorans* strain T (dark blue), *F. pennivorans* DSM9078 (green) and *F. islandicum* AW-1 (light blue). The outermost ring shows locations of all 55 protease encoding genes found in the *F. pennivorans* strain T genome. Of these, the 26 enzymes that were further characterized in MEROPS (Table 14) are marked in red. Serine protease *peg_1025* expressed in this study is marked with a black arrow. Gene names and locations refer to *F. pennivorans* strain T genomic location according to RAST server annotation. Genomic length marks measure 40 Kbp each. Figure was constructed using BRIG software.

Genomic similarities between organisms decreased as expected with their phylogenetic distances. Genomic similarities were higher between *F. pennivorans* strain T and *F. pennivorans* DSM9078 than to *F. islandicum* AW-1 (Table 15).

Table 15 Genomic comparison between *F. pennivorans* strain T, *F. pennivorans* DSM9078 and *F. islandicum* AW-1 using both DNA-DNA hybridization (DDH) and Average Nucleotide Identity (ANI). Species threshold in DDH and ANI is defined by at least identity values of 70% and 95% between genomes, respectively.

Organisms	DDH	ANI
F. pennivorans strain T vs *F. pennivorans* DSM9078	76.9 %	97.65 %
F. pennivorans strain T vs *F. islandicum* AW-1	20.4 %	80.90 %
F. pennivorans DSM9078 vs *F. islandicum* AW-1	21.5 %	81.75 %

The same identity patterns were observed when the sequences of the 26 protease genes (Table 14) from *F. pennivorans* strain T were compared against the ones in *F. pennivorans* DSM9078 and *F. islandicum* AW-1 (Figure 26,A and B). In fact, most of these genes were >95% identical between *F. pennivorans* strain T and the type strain, with only few exceptions differing up to of 11% (Figure 26,A). On the other hand, the opposite trend was found when the genes from *F. pennivorans* strain T were compared against *F. islandicum* AW-1 ones, showing most of the identities < 95%, with only two exceptions (Figure 26,B). Serine protease *peg_1025* was 100% identical against the gene in the type strain whereas was 90% similar to the equivalent in *F. islandicum* AW-1.

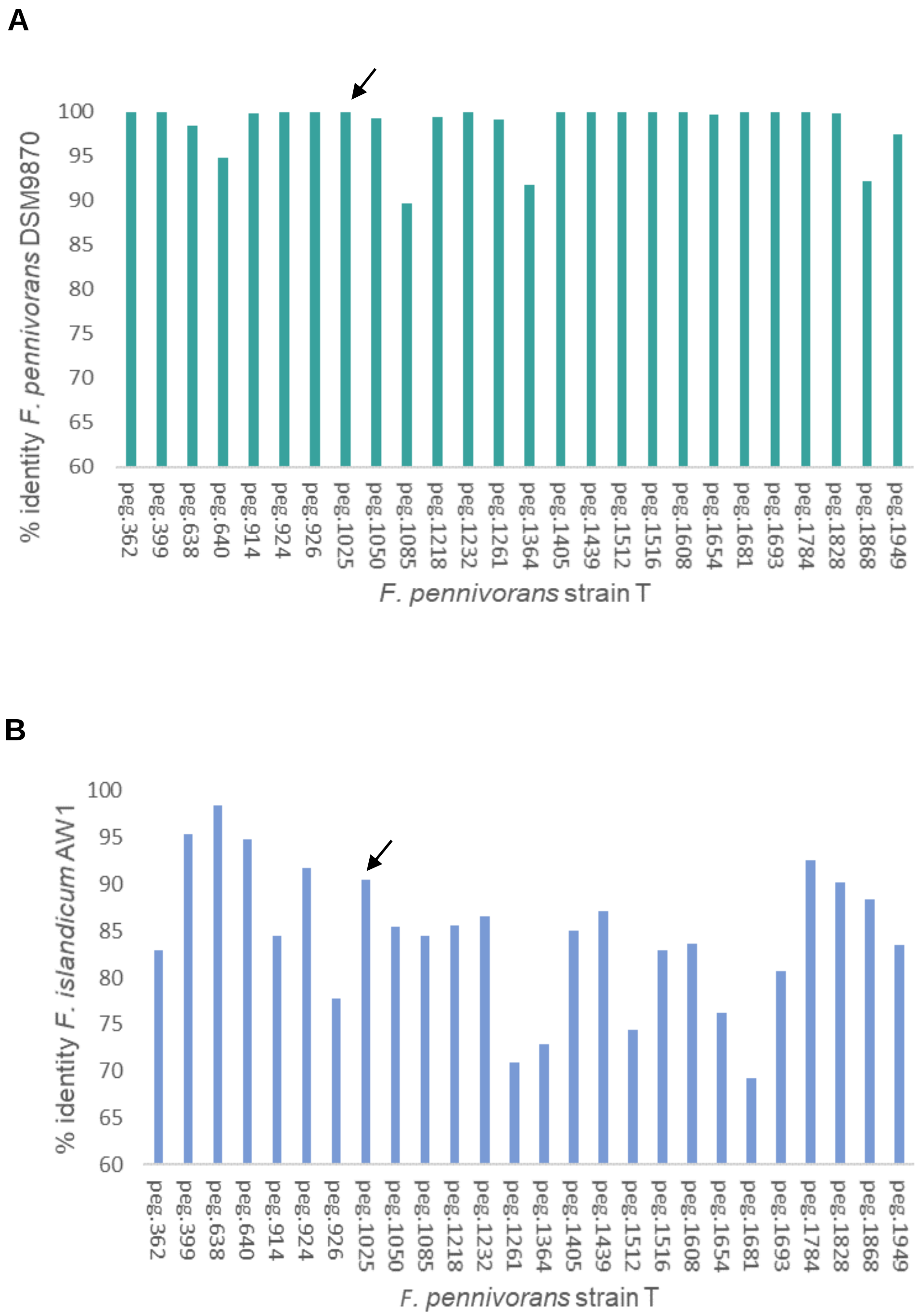


Figure 26 Identity percentage of the 26 protease encoding genes from Table 14. The histogram shows identity percentages of genes in *F. pennivorans* strain T against the equivalents in *F. pennivorans* DSM9078 (A) and in *F. islandicum* AW-1 (B). Serine protease *peg_1025* is pointed by a black arrow.

3. Transcriptomics

Growth curve

F. pennivorans strain T was grown anaerobically at 65°C in MMF supplemented with 0.1% yeast extract and 0.5% peptone and its growth recorded (Figure 27). The growth curve was used to have an idea of when to sample cells and supernatant for the transcriptomics and secretomics analysis, respectively. The growth curve can be divided in three parts sequentially composed of a lag phase, an exponential phase and a stationary phase. The lag phase lasted for about 2 hours. The logarithmic phase spanned 11 hours with an estimated generation time of 150 minutes (Figure 28). After 13 hours, cultures entered a stationary phase. No death phase was observed following incubation for 24 hours.

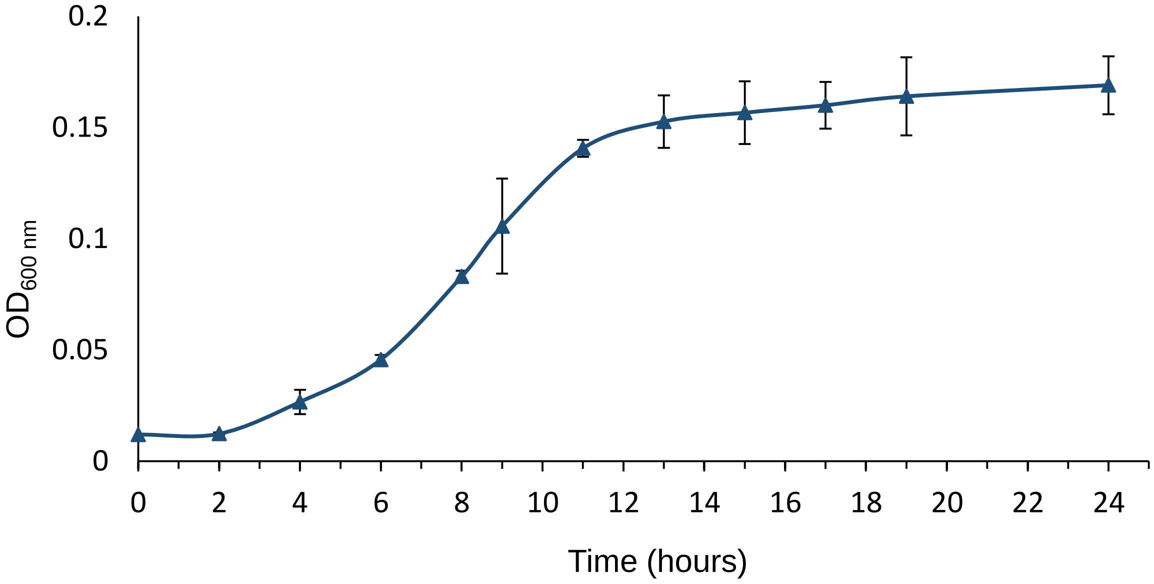


Figure 27 Growth curve of *F. pennivorans* strain T in MMF supplemented with 0.1 % yeast extract and 0.5 % glucose. Three phases of growth can be identified as lag phase (time 0 – 2), exponential phase (time 2 – 13) and stationary phase (from time 13 on). Triangles, with standard deviation for the three replicates, show OD values at different sampling times.

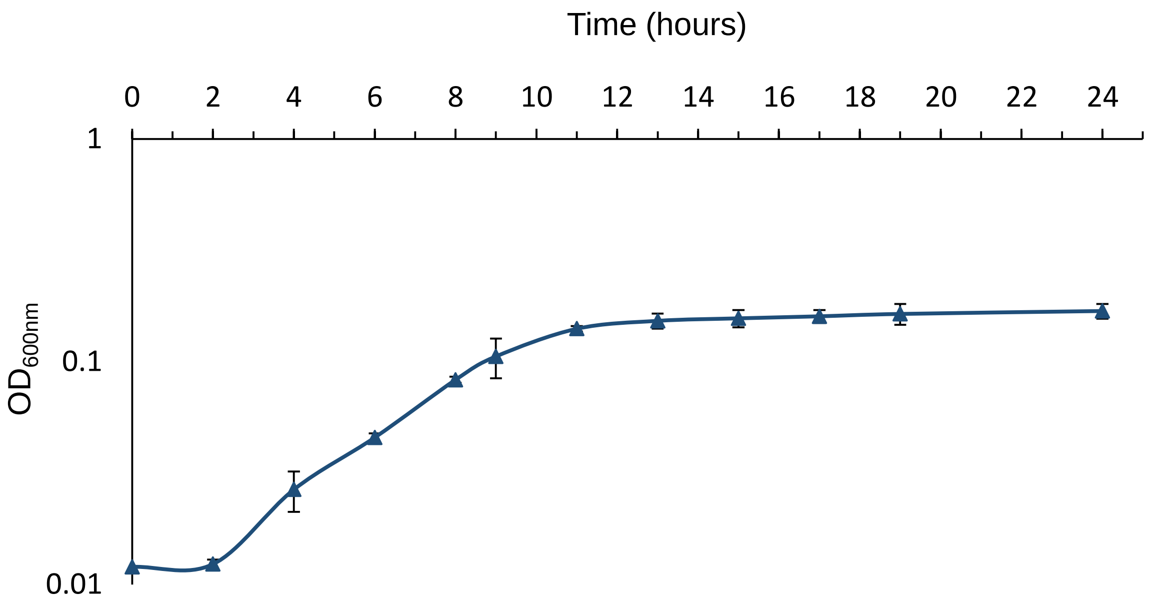


Figure 28 Logarithmic scale of the growth curve for *F. pennivorans* strain T. The linear portion comprised between hours 2 and 9 was used to calculate the generation time of the organism. Triangles, with standard deviation for the three replicates, show logarithmic OD values at different sampling times.

Transcriptomics analysis

Feather degradation by *F. pennivorans* strain T was up-scaled using a bioreactor to further evaluate its potential in industrial applications. Cell material was sampled at the beginning (glucose sample) and at the end of the experiment (feather sample) to compare gene expression under different substrates conditions.

Strain T performed mediocrely in the fermenter growing to a turbid culture after four days incubation. Feather degradation was visually monitored during the whole process but it slowed down by end of incubation, leaving some detritus and small unprocessed material. A cell pellet of 0.6 g (wet weight) was obtained after sampling the culture at the beginning of the experiment (glucose sample) whereas two pellets of 1.2 g (wet weight) each were collected at the end (feather sample). Unfortunately, RNA extraction from the feather samples performed by the sequencing company was repeatedly not successful and no comparison could be performed against the glucose sample.

4. Secretomics

Supernatant concentrate

On the basis of the growth curve of *F. pennivorans* strain T, supernatants of cells growing on glucose, peptone, casamino acids and keratin azure were collected, concentrated by ultrafiltration and analysed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). This gel showed different protein patterns for different growth substrates (Figure 29). No proteins were observed from extracellular medium of the glucose-grown batches while some bands were observed in the media from the other substrate-enriched cultures. Two blurry bands of ~80 and ~100 kDa were detected both in the peptone and casamino acids samples, although of better definition in the latter (Figure 29, lane 3 and 4, respectively). Casamino acid-grown batches seemed to have more intense bands of these two proteins as well as an extra minor band of approx. 50kDa. A similar sized band was also present in the keratin azure sample although in lower amount. Additional bands were seen in the peptone and casamino acids batches but of lower resolutions and band intensity. A smearing was observed at the end of these two samples, likely indicating protein degradation.

The three best defined bands from casamino acids-grown sample and the one corresponding to a size of ~50 kDa from culture grown with keratin azure were cut from the gel and sent for LC-MS/MS proteomics analysis.

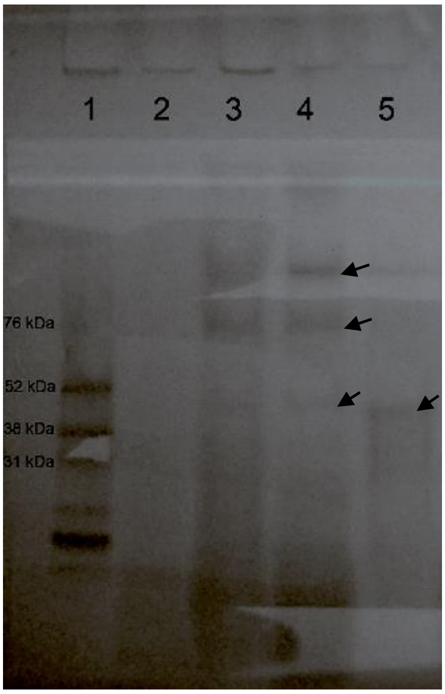
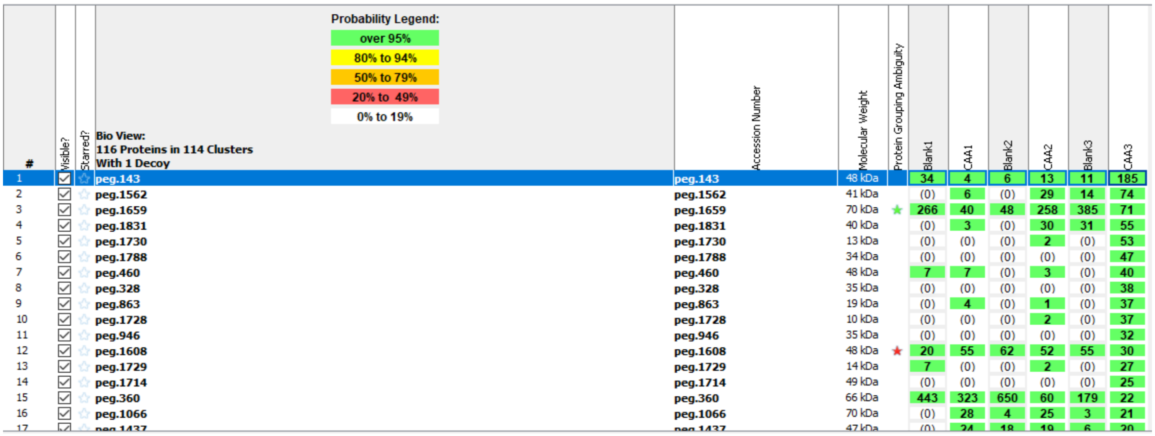


Figure 29 SDS-PAGE of proteins from *F. pennivorans* strain T culture supernatants grown with glucose (line 2), peptone (line 3), casamino acids (line 4) and keratin azure (line 5) as substrates. Full-range Rainbow^{TM} molecular weight marker was added in line 1. The four bands of ~100 kDa, ~80 kDa and ~50 kDa sent for LC-MS/MS proteomics analysis are indicated by black arrows.

Proteomics analysis

Of the four bands sent to PROBE facility for LC-MS/MS proteomics analysis, only the three from the casamino acids-grown sample were successfully resolved. By the use of the Scaffold software, each band was shown to be comprised of a mixture of proteins with different molecular weights (Figure 30). All the molecular weight reported by Scaffold refer to the molecular weight of the original reference protein, thus comprised of signal peptides.



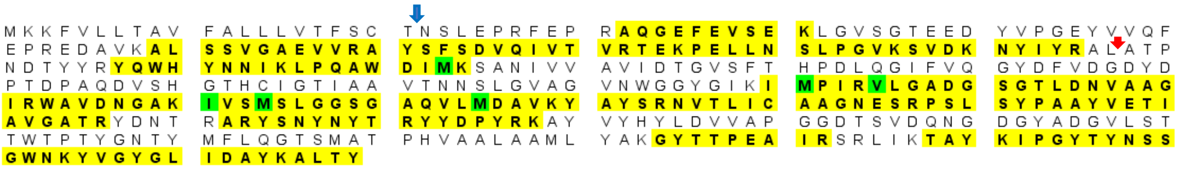
Probability Legend:

- 🟩 over 95%
- 🟨 80% to 94%
- 🟧 50% to 79%
- 🟥 20% to 49%
- □ 0% to 19%

#	Visible?	Starred?	Bio View: 116 Proteins in 114 Clusters With 1 Decoy	Accession Number	Molecular Weight	Protein Grouping Ambiguity	Blank1	CAA1	Blank2	CAA2	Blank3	CAA3
1	☑	☆	peg.143	peg.143	48 kDa		34	4	6	13	11	185
2	☑	☆	peg.1562	peg.1562	41 kDa		(0)	6	(0)	29	14	74
3	☑	☆	peg.1659	peg.1659	70 kDa	★	266	40	48	258	385	71
4	☑	☆	peg.1831	peg.1831	40 kDa		(0)	3	(0)	30	31	55
5	☑	☆	peg.1730	peg.1730	13 kDa		(0)	(0)	(0)	2	(0)	53
6	☑	☆	peg.1788	peg.1788	34 kDa		(0)	(0)	(0)	(0)	(0)	47
7	☑	☆	peg.460	peg.460	48 kDa		7	7	(0)	3	(0)	40
8	☑	☆	peg.328	peg.328	35 kDa		(0)	(0)	(0)	(0)	(0)	38
9	☑	☆	peg.863	peg.863	19 kDa		(0)	4	(0)	1	(0)	37
10	☑	☆	peg.1728	peg.1728	10 kDa		(0)	(0)	(0)	2	(0)	37
11	☑	☆	peg.946	peg.946	35 kDa		(0)	(0)	(0)	(0)	(0)	32
12	☑	☆	peg.1608	peg.1608	48 kDa	★	20	55	62	52	55	30
13	☑	☆	peg.1729	peg.1729	14 kDa		7	(0)	(0)	2	(0)	27
14	☑	☆	peg.1714	peg.1714	49 kDa		(0)	(0)	(0)	(0)	(0)	25
15	☑	☆	peg.360	peg.360	66 kDa		443	323	650	60	179	22
16	☑	☆	peg.1066	peg.1066	70 kDa		(0)	28	4	25	3	21
17	☑	☆	peg.1437	peg.1437	47 kDa		(0)	24	18	19	6	20

Figure 30 Snapshot of Scaffold software showing the quantitative values (in green) of displayed proteins (Accession Number) detected in the bands of ~ 100 kDa (CAA1), ~ 66 kDa (CAA2) and ~ 50 kDa (CAA3) from casamino acids grown supernatant. The values are ordered according to the most abundant protein detected in CAA3, Peg_143 (in text called Peg_1025).

The most abundant protein, accounting 323 number of total sequences (NTS) detected in the band of ~ 100 kDa was of ~ 66 kDa size and was identified as periplasmatic peptide-binding ABC transporter Peg_1720 (as annotated by RAST). A similar peptide-binding ABC transporter protein (Peg_817) was also the most abundant protein in the ~ 80 kDa band (258 NTS), with a molecular weight of 70 kDa. From the ~ 50 kDa band, a protein of 48 kDa was identified as the most abundant protein (185 NTS). This protein had a coverage of 50% from 29 unique peptides with a total of 149 experimental spectra (Figure 31). This protein was identified as Peg_1025 when the protein sequence was blasted within *F. pennivorans* strain T complete genome in RAST, a serine protease. This was probably also present in the keratin-grown culture, and thus might represent a key enzyme for keratin breakdown. No peptides were detected covering the signal peptide sequence (1 – 22 AA). Five unique peptides were found belonging to the propeptide sequence (22 – 127 AA) generating only 7 total experimental spectra.



```
MKKFVLLTAV FALLLVTFSC TNSLEPRFEP RAQGEFEVSE KLGVSGTEED YVPGEYVVQF
EPREDAVKAL SSVGAEVVRA YSFSDVQIVT VRTEKPELLN SLPGVKSVDK NYIYRALATP
NDTYYRYQWH YNNIKLPQAW DIMKSANIVV AVIDTGVSFT HPDLQGIFVQ GYDFVDGDYD
PTDPAQDVSH GTHCIGTIAA VTNNSLGVAG VNWGGYGIKI MPIRVLGADG SGTLDNVAAG
IRWAVDNGAK IVSMSLGGSG AQVLMDAVKY AYSRNVTLIC AAGNESRPSL SYPAAYVETI
AVGATRYDNT RARYSNYNYT RYYDPYRKAY VYHYLDVVAP GGDTSVDQNG DGYADGVLST
TWTPTYGNTY MFLQGTSMAT PHVAALAAML YAKGYTTPEA IRSRLIKTAY KIPGYTYNSS
GWNKYVGYGL IDAYKALTY
```

Figure 31 Spectra coverage (yellow) of serine protease (Peg_1025) sequence visualized by Scaffold software. The spectra were obtained from LC-MS/MS spectrometry of the SDS-PAGE band of molecular size of ~ 50 kDa from the casamino acid-grown sample. The proteomics analysis reported 29 exclusive unique Peg_1025 peptides (highlighted in yellow), covering 50% of the sequence, and a total of 149 experimental spectra. No peptides were detected covering the signal peptide sequence (1 – 22 AA). Five unique peptides were found belonging to the propeptide sequence (22 – 127 AA) generating only 7 total experimental spectra. Signal peptide end and propeptide end are marked by blue and red arrows, respectively (S 2).

On the basis of secretomics analysis as well as genomics results, serine protease Peg_1025 seemed to be involved in the extracellular breakdown of keratin and thus was chosen for carrying out expression experiments. Its sequence gene sequence (S 3) and protein sequence (S 2) were extracted from the complete genome of *F. pennivorans* strain T.

5. Cloning and expression of a putative keratinase gene

Protein bioinformatics

The serine protease with RAST annotation ID Peg_1025 was a 439 amino acids (AA) long protein of molecular weight of 48083.08 Dalton and a theoretical pI (isoelectric point) of 5.6. The most abundant residues were alanine 44 (10%), valine 43 (9.8%), glycine 39 (8.9%) and tyrosine 38 (8.7%). Three cysteines were also present (0.7%). The extinction coefficient had a value of 89745 whereas the grand average of hydropathicity of -0.147 (not hydrophobic).

The protein sequence was analysed for motifs and domains using the scanProSite web tool database. The web tool scanned the protein query against the PROSITE collection of motifs and found one domain (AA 128 – 439) corresponding to a family of serine proteases named Subtilase S8, characterized by a catalytic triad composed of Asp154, His190 and Ser377 (numbers referring to Peg_1025) (Figure 32). The presence of the domain was also confirmed by Interpro and Pfam web tools (results not shown). ScanProSite also identified a signal peptide sequence from residues 1 to 20 with a cleavage site between position 20 and 21 (Figure 32). However, both Interpro and SignalIP 5.0 suggested a cleavage site between position 19 and 20 (probability 0.9979) (Figure 33). As described in Table 14, TMHMM confirmed the protease to be secreted to the extracellular environment by neither having any predicted transmembrane helices (TMH) nor any segments of expected number of hydrophobic AAs in TMHs < 18 (Figure 34). A hydrophobic region was although detected at the beginning of the sequence, referring to the signal peptide motif.

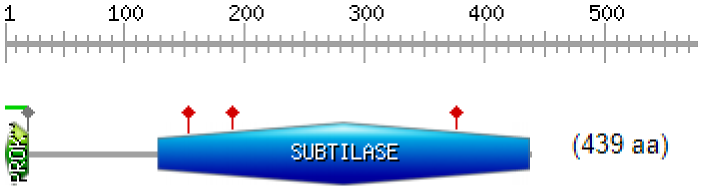


Figure 32 Subtilase S8 domain (AA 128 – 439) and signal peptide motif (AA 1 – 20) predicted in the Peg_1025 serine protease sequence. The three residues that compose the catalytic domain are marked in red symbols (Asp154, His190 and Ser377). A ruler for the sequence length is also reported. Image adapted from PROSITE.

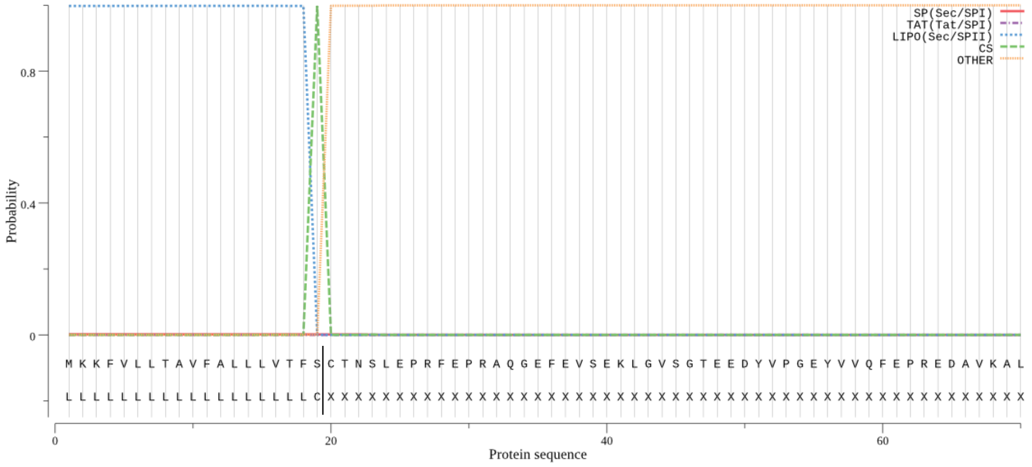


Figure 33 Signal peptide cleavage site based on SignalIP 5.0 prediction. A cleavage site was suggested between residue 19 and 20 (black bar) with a probability of 0.9979.

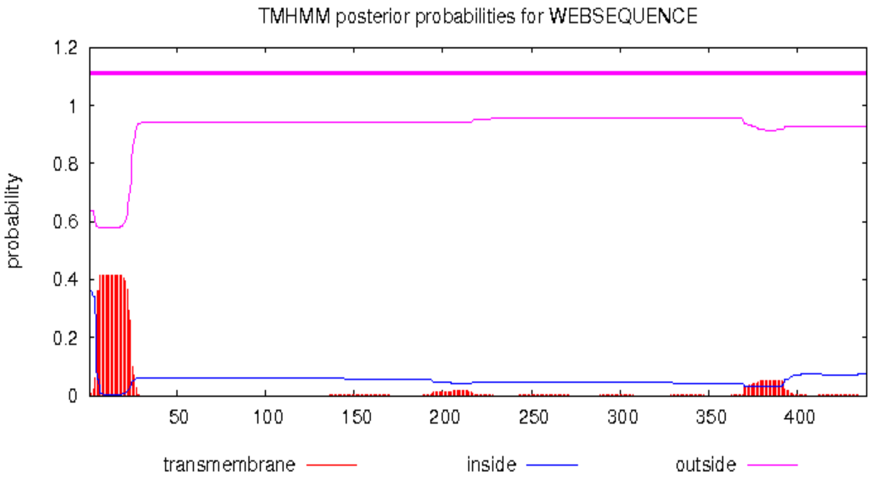


Figure 34 TMHMM cellular location prediction of serine protease Peg_1025. Red bars correspond to hydrophobic parts of the protein, but none of them were over the minimum threshold to be classified as transmembrane. The major hydrophobic region at the beginning of the sequence relate to the signal peptide. Thick pink line (top, between 1 and 1.2) shows the best overall prediction for the protein cellular location. Thin pink line (middle) shows the prediction of cellular location for each residue in the protein.

To better understand general features of the query protein, analysis on its homologous proteins were performed. The most identical homologous protein was, according to Uniprot and MEROPS databases (overall identity = 39%), subtilisin Ak1 (Uniprot ID: Q45670; MEROPS ID: S08.009) from the thermophilic bacterium *Geobacillus stearothermophilus* strain AK1 (Toogood et al., 2000). This peptidase belongs to the protease clan SB, family S8 with serine-type endopeptidase activity (Siezen & Leunissen, 2008). It binds 3 Ca^{2+} ions and 1 Na^{+} ion per subunit, with an optimal pH and temperature dependence of 8.5 and 75°C, respectively. Uniprot reports subtilisin Ak1 composed of three major domains: a signal peptide (position AA 1 – 24), a propeptide (position 25 – 121) and the thermophilic serine proteinase domain (position 122 – 401). It also reports the presence of two cysteines forming a single disulphide bond.

A total of 26 bacterial sequences belonging to family S8 were retrieved from the cross link search between MEROPS and Uniprot (reviewed protease) (S 6) to further extend the characterization of serine protease Peg_102. In the list, a protein belonging to *F. pennivorans* DSM9078 named fervidolysin (Uniprot ID: Q93LQ6; MEROPS ID: S08.021) was also included.

By aligning all the sequences obtained from the PSI-BLAST search and the 26 reviewed sequences from Uniprot, a multiple sequence alignment was built with Clustal Omega in JalView (Figure 35). For practicality and for the scope of this study, only the sequences belonging the Thermotogae phylum retrieved from the PSI-BLAST search were considered. The alignment showed the conservation of the catalytic triad composed of Asp154, His190 and Ser377 (position referring to the Peg_1025 sequence) in all homologous proteases found. The catalytic triad was surrounded by highly conserved regions, as well. Within Thermotogae, residues were well conserved over the whole length of the protein sequences.

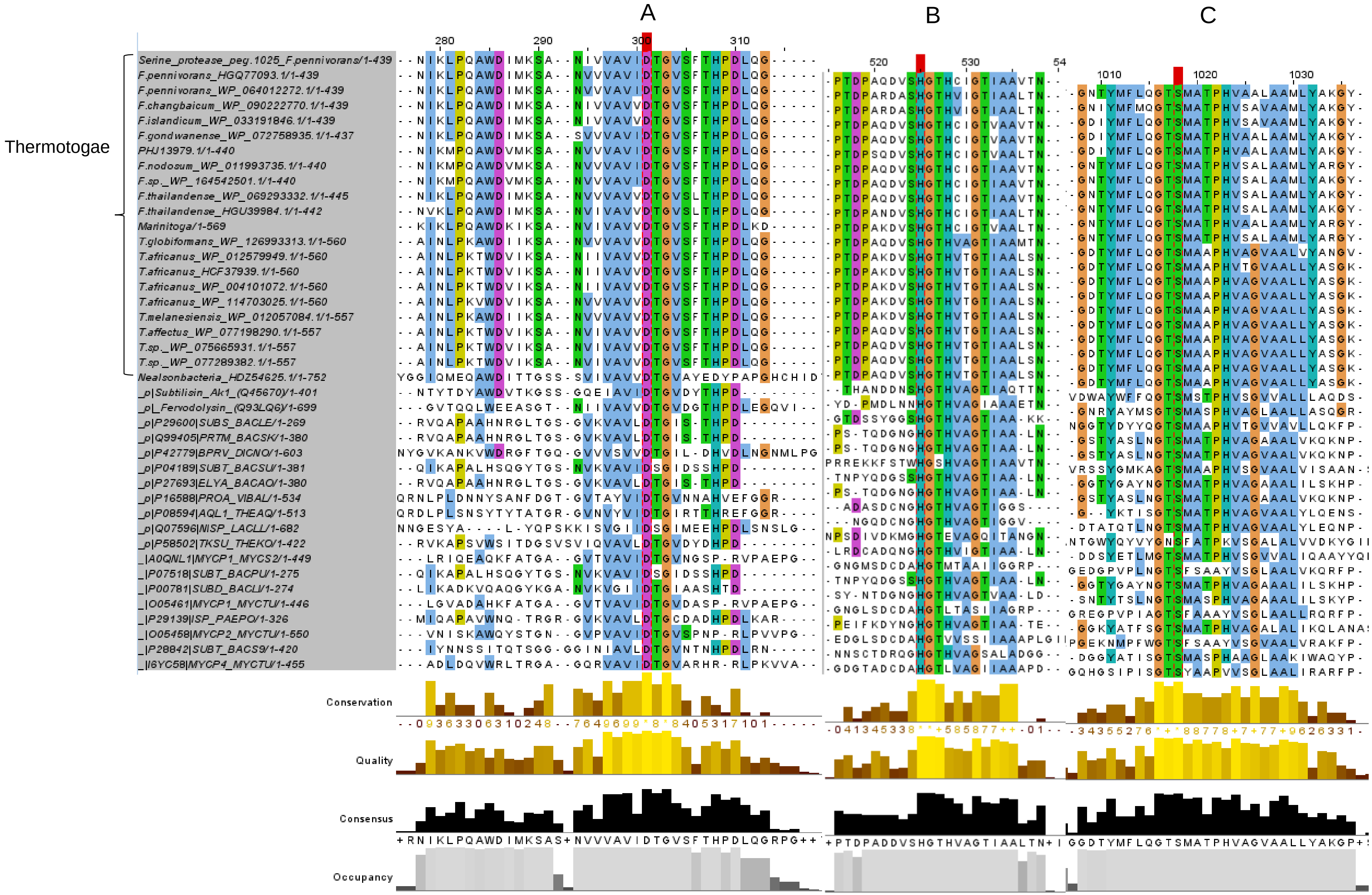

A
B
C
Thermotogae
Conservation
Quality
Consensus
Occupancy

Figure 35 Multiple sequence alignment of the catalytic domain of proteins homologous to serine protease Peg_1025 (fist line) retrieved from PSI-BLAST query and from the curated database Uniprot. From the PSI-BLAST query, only the homologous sequences belonging to the Thermotogae phylum are shown (top 21 sequences). All 26 reviewed proteins from Uniprot are shown (bottom). The catalytic triad composed of Asp154 (A), His190 (B) and Ser377 (C) (position referring to Peg_1025 sequence) are marked in red in the alignment ruler.

Eleven of the 26 reviewed proteases from Uniprot were reported to have, together with the catalytic domain, a propeptide region at the beginning of the sequence (S 6). These sequences were used as references for the identification of a propeptide sequence in serine protease Peg_1025. Multiple sequence alignments of this region showed less overall conservation compared to the catalytic domain region (Figure 36). Nonetheless, two highly conserved motifs could be identified also based on previous studies (Shinde & Inouye, 1993).

The propeptide motif in the reviewed proteins have different length and have an average of 80/90 residues, starting after the signal peptide sequence. For example, in fervidolysin and subtilisin Ak1 the propeptide is identified between residues 22 – 149 in the former and 25 – 121 in the latter. Nonetheless, all sequences aligned nicely at the level of catalytic domain and signal peptide suggesting regions with conserved residues in the sequences in between (S 6). Overall, little conservation was observed between Thermotogae sequences and reviewed sequence from Uniprot. Within the first motif, only residues Y_{111} and V_{113} seemed to be highly conserved throughout all sequences whereas in the second motif only residues V_{169} – L_{192} – P_{196} – V_{198} were conserved (residues numbers referring to the alignment ruler).

By the alignment interpretation, a hypothetical propeptide sequence was defined in serine protease Peg_1025 between residues 22 and 127.

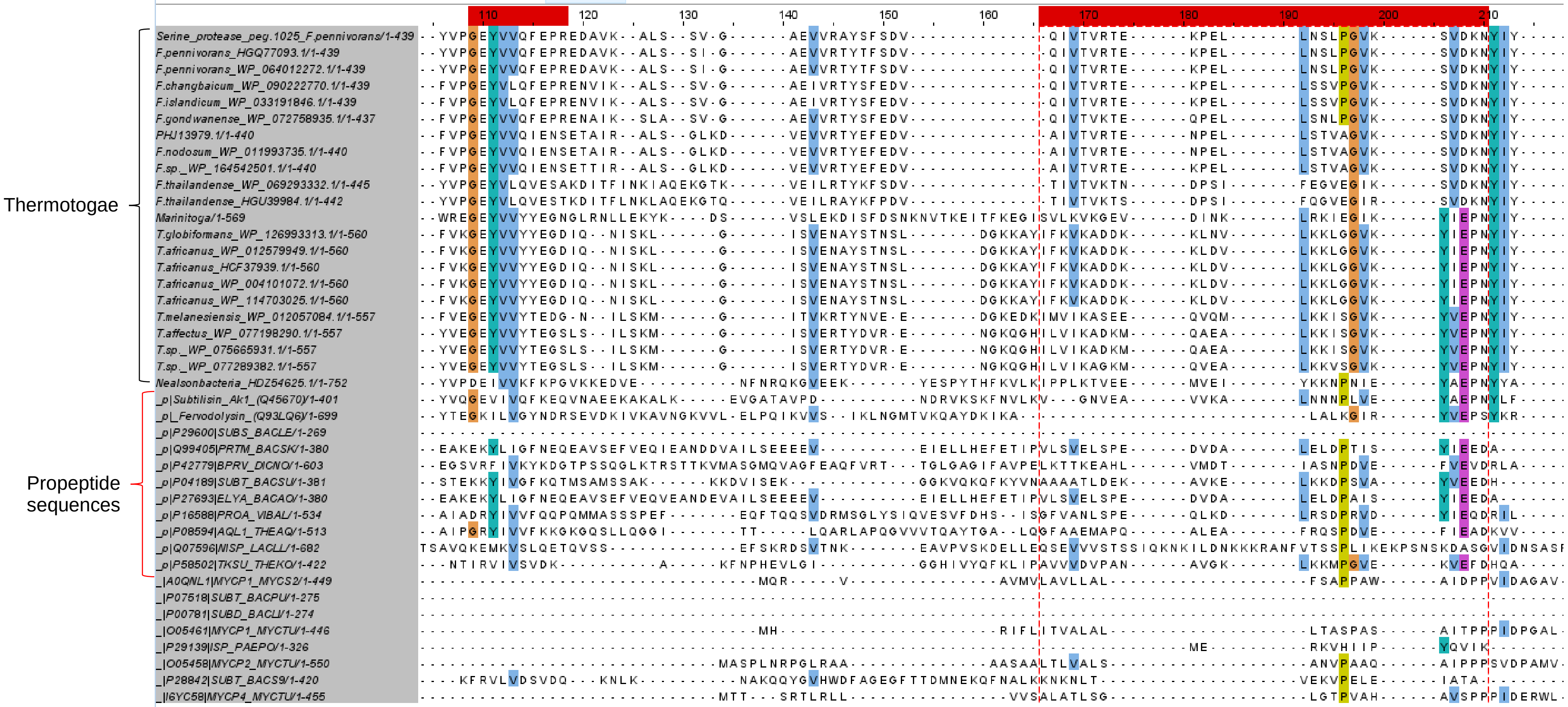

110
120
130
140
150
160
170
180
190
200
210
Thermotogae
Propeptide sequences
Serine_protease_peg.1025_F.pennivorans/1-439
F.pennivorans_HGQ77093.1/1-439
F.pennivorans_WP_064012272.1/1-439
F.changbaicum_WP_090222770.1/1-439
F.islandicum_WP_033191846.1/1-439
F.gondwanense_WP_072758935.1/1-437
PHJ13979.1/1-440
F.nodosum_WP_011993735.1/1-440
F.sp._WP_164542501.1/1-440
F.thailandense_WP_069293332.1/1-445
F.thailandense_HGU39984.1/1-442
Marinitoga/1-569
T.globiformans_WP_126993313.1/1-560
T.africanus_WP_012579949.1/1-560
T.africanus_HCF37939.1/1-560
T.africanus_WP_004101072.1/1-560
T.africanus_WP_114703025.1/1-560
T.melanesiensis_WP_012057084.1/1-557
T.affectus_WP_077198290.1/1-557
T.sp._WP_075665931.1/1-557
T.sp._WP_077289382.1/1-557
Nealsonbacteria_HDZ54625.1/1-752
_p|Subtilisin_Ak1_(Q45670)/1-401
_p|_Fervodolysin_(Q93LQ6)/1-699
_p|P29600|SUBS_BACLE/1-269
_p|Q99405|PRTM_BACSK/1-380
_p|P42779|BPRV_DICNO/1-603
_p|P04189|SUBT_BACSU/1-381
_p|P27693|ELYA_BACAO/1-380
_p|P16588|PROA_VIBAL/1-534
_p|P08594|AQL1_THEAQ/1-513
_p|Q07596|NISP_LACLL/1-682
_p|P58502|TKSU_THEKO/1-422
_|A0QNL1|MYCP1_MYCS2/1-449
_|P07518|SUBT_BACPU/1-275
_|P00781|SUBD_BACLI/1-274
_|O05461|MYCP1_MYCTU/1-446
_|P29139|ISP_PAEPO/1-326
_|O05458|MYCP2_MYCTU/1-550
_|P28842|SUBT_BACS9/1-420
_|I6YC58|MYCP4_MYCTU/1-455

Figure 36 Multiple sequence alignment of the propeptide region of protein homologous to serine protease Peg_1025 (fist line) retrieved from PSI-BLAST query and from curated database Uniprot. From the PSI-BLAST query, only the homologous sequences belonging to the Thermotogae phylum are shown (top 21 sequences). All 26 reviewed proteins from Uniprot are shown (bottom) and the 11 in which a propeptide was reported are bracketed in red. The two motifs which are highly conserved in propeptide sequences are marked in red.

A phylogenetic tree based on average distances obtained from the multiple sequence alignment showed all homologous proteases belonging to Thermotogae phylum clustering together (Figure 37). Serine protease Peg_1025 from *F. pennivorans* strain T and *F. pennivorans* DSM9078 branched independently from *F. islandicum* AW-1, which clustered with *F. changbaicum,* instead. Fervidolysin, branched quite at the origin of the tree and seemed to create an independent node from all the other serine proteases analysed. The sequences from Uniprot reported to have a propeptide did not show any clustering patterns.

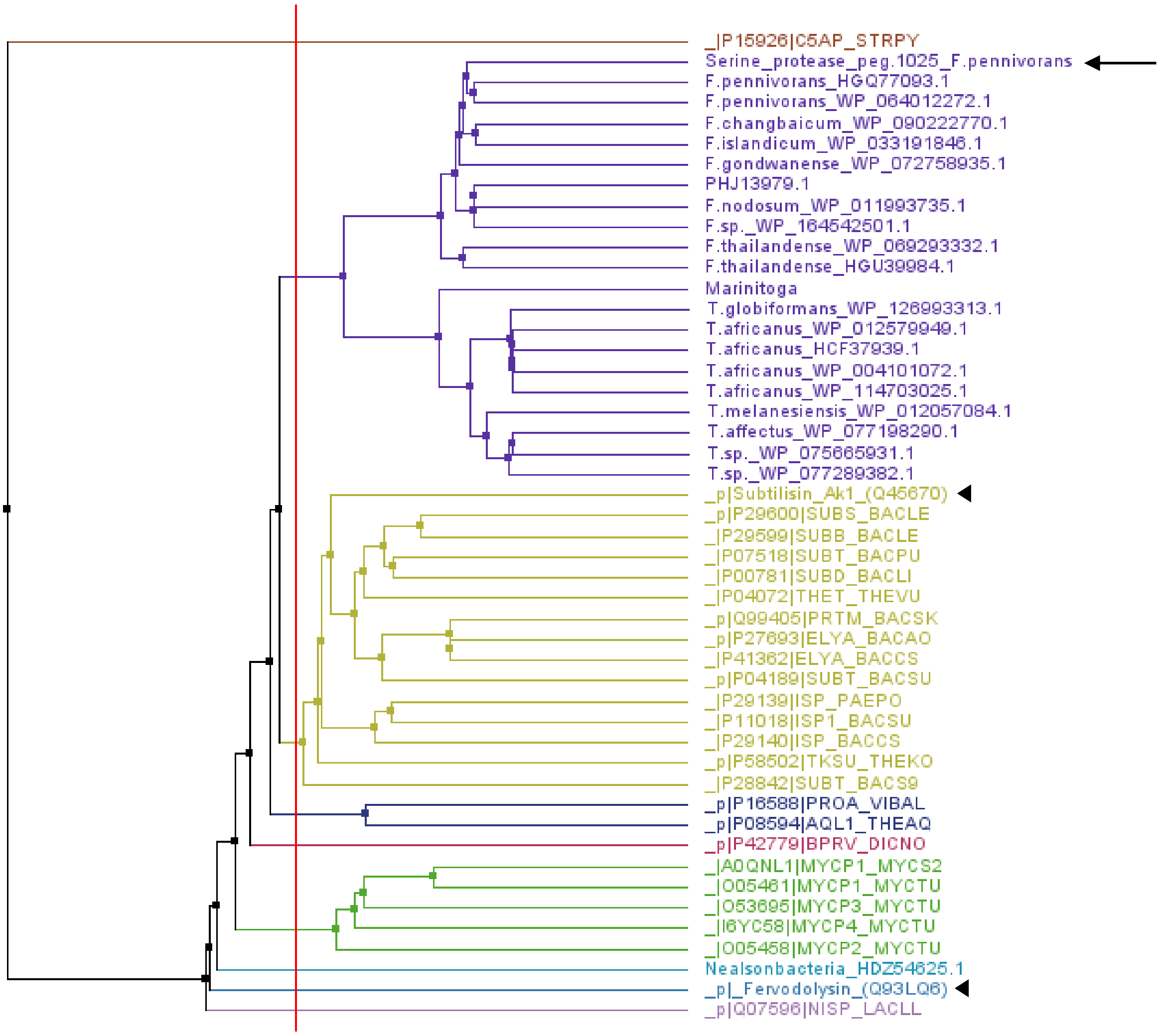


Figure 37 Phylogenetic tree based on average distances of sequences from PSI-BLAST and Uniprot alignment. Protease sequences belonging to Thermotogae phylum clustered together (blue group). Out-group used (top, brown branch) was a protease (Uniprot ID: C5AP) from bacterium *Streptococcus pyogenes*. Serine protease Peg_1025 is marked by black arrow. Subtilisin Ak1 and fervidolysin are also marked by small triangles. Red line across the tree creates the cluster division colours.

Structural prediction

The sequences for the propeptide (AA 22 - 127) and for the catalytic domain (AA 128 - 439) of serine protease Peg_1025 were individually submitted to SWISS-MODEL and their 3D structure inferred from both fervidolysin (PDB: 1R6V) and subtilisin Ak1 (PDB: 1DBI) as template. In assessing SWISS-MODEL model evaluation GMQE, QMEAN and its fours terms, local quality plot and model comparison plot were used. For GMQE, higher numbers indicate higher reliability; QMEAN values around zero indicate good agreement between the model structure and the experimental structure; residues of score above 0.6 in the local quality plot are expected to be of high quality and finally, the closer the predicted structure (red star) in the model comparison panel is to the black dots, the more the quality of the predicted structure can be compared to the quality of the experimental structures.

Both fervidolysin and subtilisin Ak1 could serve as templates in modelling of the catalytic domain of Peg_1025. When subtilisin Ak1 was used as template, Peg_1025 had a sequence identity of 48.19% (Figure 38,A), a Global Model Quality Estimate (GMQE) of 0.68 and a quality mean (QMEAN) of -3.96. Three Ca^{2+} ions ligands were also reported and values for the global quality estimate were close to zero (Figure 38,B). The local quality estimate was in average closer to 1 and the model fitted better with the references in the database (Figure 38,C and D).

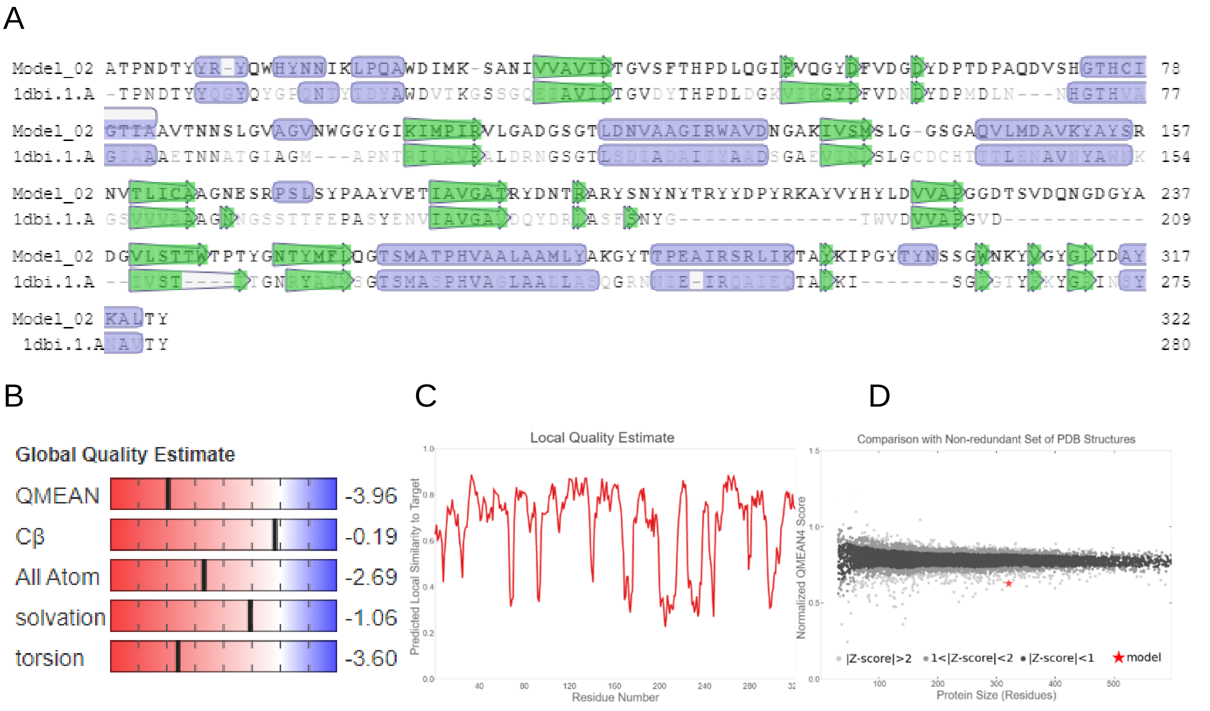


Figure 38 Quality estimates of the structural model for the catalytic domain (AA 97 – 439) of serine protease Peg_1025 obtained using subtilisin Ak1 as template by SWISS-MODEL. A) Structural alignment between serine protease (top line) and subtilisin Ak1 (bottom line); helices are highlighted in blue and beta sheets in green. B) Global quality estimates. C) Local quality estimate. D) Model fitting graph.

On the other hand, the catalytic domain sequence of Peg_1025 had 40.93% identity against fervidolysin template, a GMQE of 0.66 and a QMEAN of -4.27. No ligands were reported. Global quality estimates, local quality estimate and comparison for this model were in general of lower scores then the previous template (S 7).

When the sequence for Peg_1025 propeptide was submitted to SWISS-MODEL, only fervidolysin could be used as template. The modelling scored a sequence identity of 18.84%, a GMQE of 0.44 and a QMEAN of -3.02. Structural alignment identified many shared folding patterns between the sequences (S 8,A), although global quality estimates and modelling comparison were a bit low (S 8,B and D).

As an overall interpretation of modelling quality scores for the structural prediction of Peg_1025 catalytic domain, subtilisin Ak1 model was chosen as best fit (Figure 39). With fervidolysin as template, a model for the propeptide was also predicted (S 9).

In comparing the protein model of Peg_1025 with the template structure of subtilisin Ak-1, some differences were noticed. First, the query protein had some sequence parts that were either very different to the template sequence (five) or completely absent (insertions) (five), resulting in low QMEAN scores (S 10). The structural prediction of these areas with low QMEAN scores was thus not reliable and hidden from the model. Secondly, the two cysteines in the query protein (CYS-194 and CYS-280) resulted far apart from each other (12.6 Å) and unlikely to form a disulphide bond. Finally, some residues at the level of cation binding sites for Ca-1, Ca-3 and Na-1 were not conserved respect to the reference protein subtilisin Ak-1, resulting in lack of predicted cation binding sites. Nonetheless, similarities in major structural conformations could be observed when the superimposed stereo diagram with Peg_1025 and subtilisin Ak-1 was made (Figure 40). The catalytic triad and its motifs were also highly conserved as well as the residues in the cation binding site for Ca-2. As for subtilisin Ak1, structural motifs of Peg_1025 at its N- and C- terminal were different: globular in the former and disordered in the latter. In particular, the C-terminal of both proteins was outside the protein.

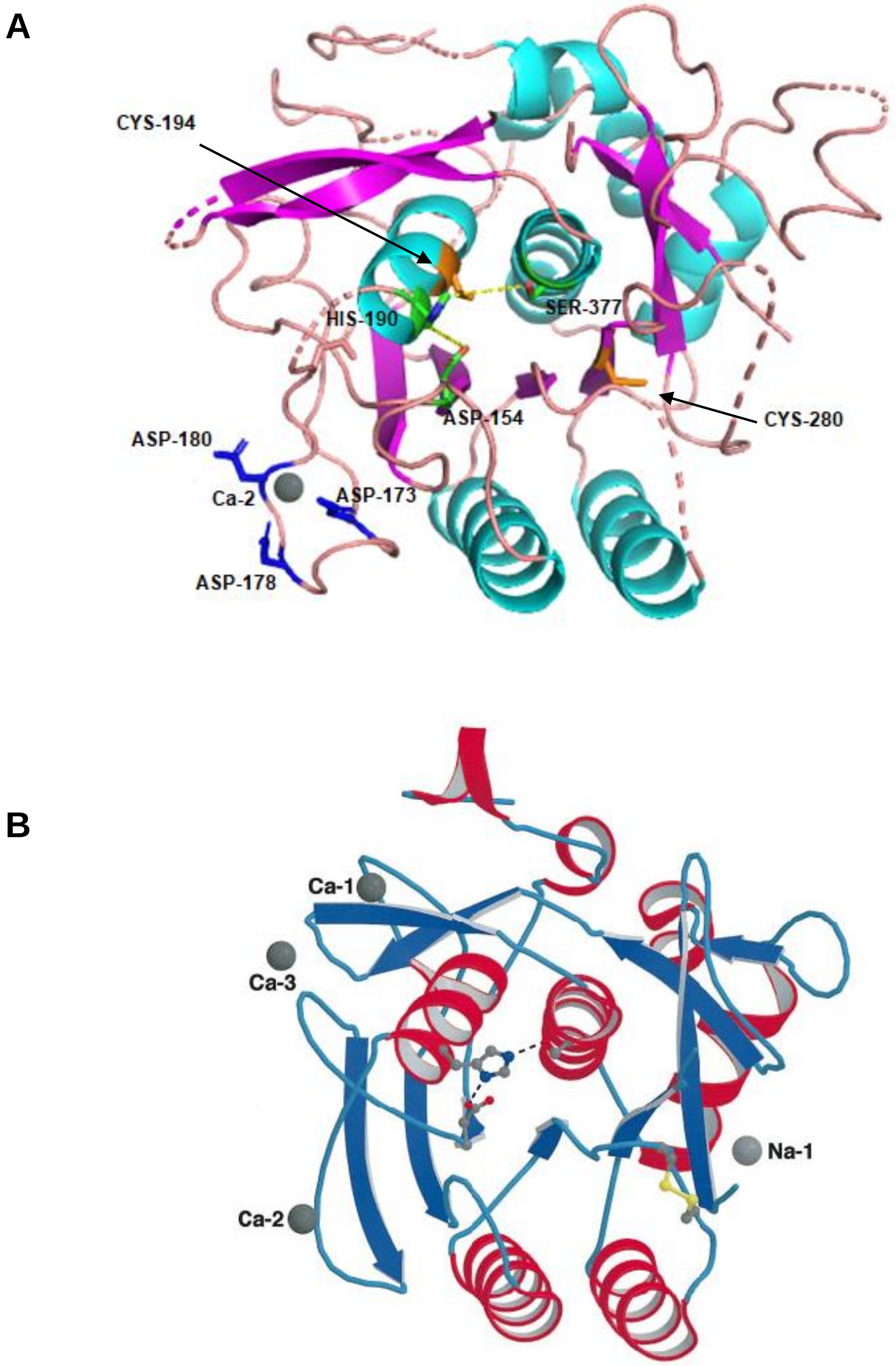


Figure 39 Three-dimensional structural model of serine protease Peg_1025 (A) obtained by submission of the catalytic domain sequence in SWISS-MODEL using subtilisin Ak-1 as template. Catalytic triad Asp154, His190 and Ser377 are highlighted with the carbon colours

(C, green; N, blue; O, red) while the cysteine residues (CYS-194, -280) are highlighted in orange. Calcium binding residues (ASP-173, -178, -180) are coloured in blue. The structure was predicted for the entire catalytic domain (AA 128 – 437) and visualized with pyMOL hiding the residues whose structural prediction scores were low (S 10). Residues numbers correspond to Peg_1025 sequence. B) Structure of subtilisin Ak-1 (PDB: 1DBI) used as template and reported for comparison (Smith et al., 1999).

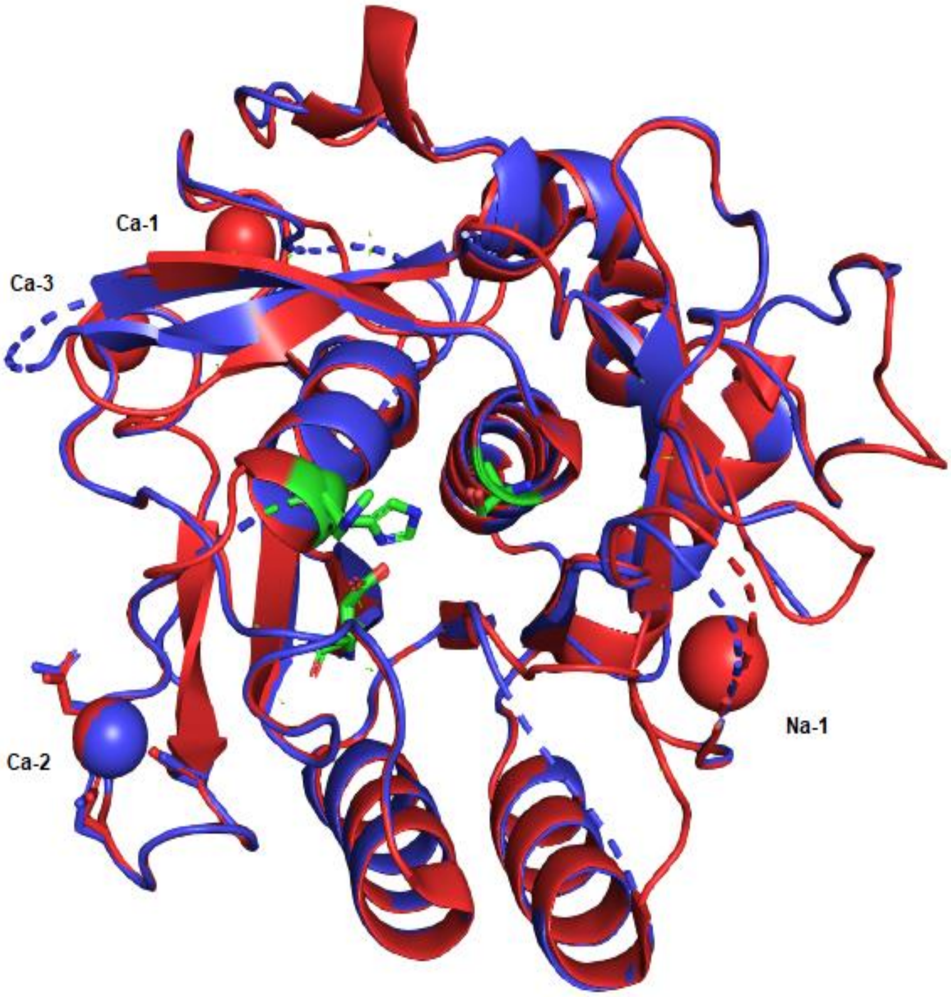


Figure 40 Cartoon diagram showing serine protease Peg_1025 structural model (blues) superimposed with its template, subtilisin Ak-1 (red). Catalytic triad Asp154, His190 and Ser377 are highlighted with the carbon colours (C, green; N, blue; O, red). Cation Ca-2 is bond by both proteins while cations Ca-1, Ca-3 and Na-1 are bond only by subtilisin Ak-1.

Primer design and *in-silico* cloning

Primers designed by Javier-Lopez (2018) were used to target and amplify the *peg_1025* gene without its signal peptide sequence. Nonetheless, for learning purposes, forward and reverse primers for the open reading frame coding for Peg_1025 protease were designed *de-novo* and *in-silico* tested. Primers were divided in two parts: an extension sequence and a target sequence (Table 16) to accommodate cloning in the FX-system. The extension sequence was composed of complementary overhangs to the propagation vector (F: 5'-tagt; R: 5'-atgc); a recognition site for the restriction enzyme *Sap*I (5'-GCTCTTC); a cleavage site generating the overhangs ends (F-5'-tAGT; R-5'-aTGC) and a further extension (5'-atatat). The target sequence was the gene sequence to which primers had to anneal deprived of start and stop codons and the signal peptide. The experimental melting temperature (Tm) obtained from OligoCalc web tool was 53°C for the forward primer and 54°C for the reverse primer.

Table 16 Forward and reverse primer sequences designed in this study and tested in *in-silico* by Serial Cloner software. Primers are divided into two parts: extension sequence and target sequence.

	Extension seq.	Target seq.
Forward	5'-atatatGCTCTTCtagt	AACTCATTAGAGCCAAGATTTGAAC
Reverse	5'-atatatGCTCTTCatgc	GTATGTCAATGCCTTGTAAGCATC

After the cloning protocol, two final cultures of the expression strain *E. coli* LOBSTR containing expression vectors p7xC3H and p7xN3H were obtained and used to carry out expression experiments for serine protease *peg_1025*. The target protease expressed from vector p7xN3H contained an N-terminal 10 x His, PreScission site residues (LEVLFQGP) and an extra serine prior to the Peg_1025 sequence, in the mentioned order (Figure 41). Likewise, an extra alanine, the PreScission site residues (LEVLFQGP) and 10 x His were added to the C-terminal of the protease from the p7xC3H vector (Figure 42). Overall, after the expression, the target protein was expected to have a total molecular weight of 48 kDa.

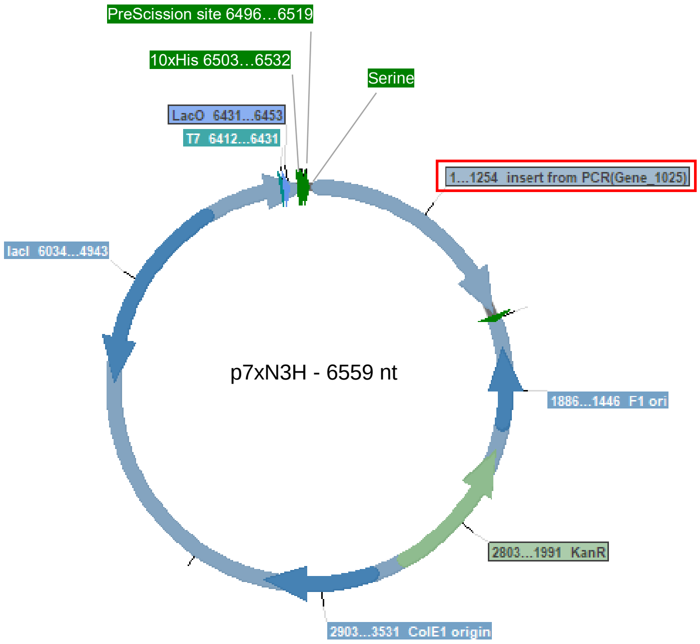


Figure 41 Expression plasmid p7xN3H for serine protease *peg_1025* (red squared) expression. The target protein was expressed with 10 x His, LEVLFQGP residues (PreScission site) and a serine at its N-terminal end. The expression of the insert was regulated by T7 promoter. Image obtained from *in-silico* cloning by Serial Cloner program.

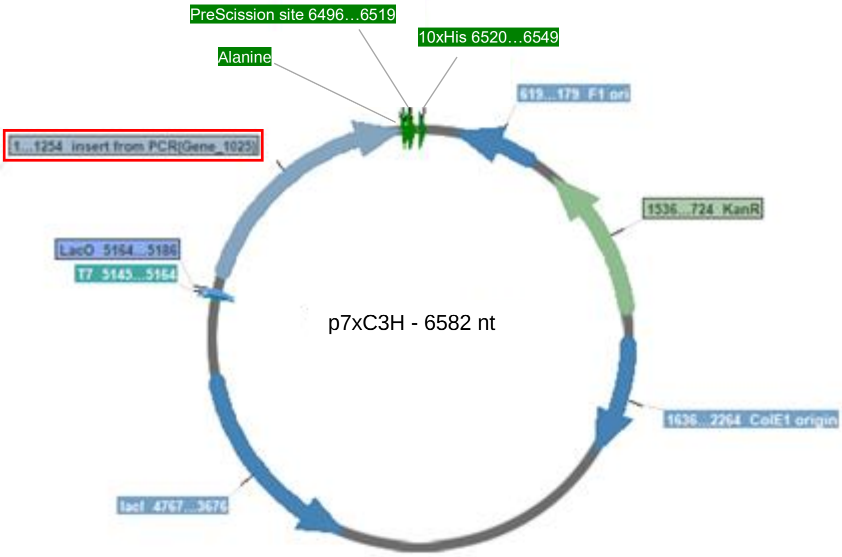


Figure 42 Expression plasmid p7xC3H for serine protease *peg_1025* (red squared) expression. The target protein was expressed with 10 x His, LEVLFQGP residues (PreScission site) and an alanine at its C-terminal end. The expression of the insert was regulated by T7 promoter. Image obtained from *in-silico* cloning by Serial Cloner program.

Protein expression and yield

Serine protease *peg_1025* was successfully expressed by both expression constructs in the *E. coli* LOBSTR host and resulted in soluble protein after cell lysis and lysate treatments. Bands corresponding to the expected size of 48 kDa were identified in all fractions (S 11), also after heat treatments (Figure 43, lanes 3 to 6). A band corresponding to the theoretical mass of the protein was also observed in the control with the uninduced cells (Figure 43, lane 2). No band of 48 kDa was observed in the expression host *E. coli* LOBSTR without the expression vectors (Figure 43, lane 1 vs 3 to 6).

A difference in expression yield was detected between the vector construct p7xC3H and p7xN3H. In fact, after heat shock treatment at 70°C for 20 minutes (samples –HS), more protein material was observed in the lysate fractions of the p7xC3H vector (S 11, lane 6 vs lane 7; lane 8 vs lane 9) together with higher protein concentration: 0.50 mg ml-1 from the p7xN3H vector and 0.62 mg ml-1 from the p7xC3H vector.

The heat activation at 70°C for 2 hours of the lysates (samples –HA) did not produce any visible differences when samples were compared by SDS-PAGE (Figure 43, lane 3 vs lane 5; lane 4 vs lane 6). The full length 48 kDa band seemed unchanged.

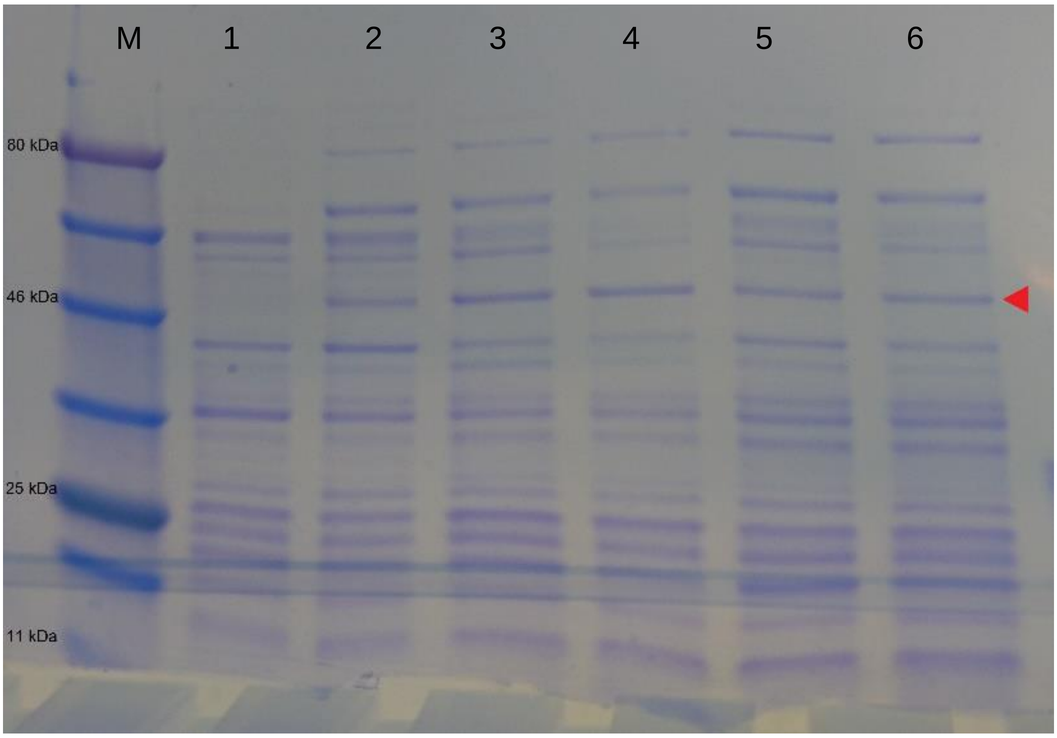


Figure 43 SDS-PAGE showing expression of serine protease Peg_1025 from p7xN3H and p7xC3H in *E. coli* LOBSTR. Band corresponding to the target protein (48 kDa) is marked with red arrow. M, broad rage marker; lane 1, *E. coli* LOBSTR without vector; lane 2, uninduced cells from p7xC3H vector; lane 3 and 4, soluble fraction after heat shock treatment (70°C for 15 minutes), N- and C-tagged, respectively; lane 5 and 6 soluble fraction after heat activation (70°C for 2 hours), N- and C-tagged, respectively.

6. Enzyme activity evaluation

Zymogram staining

To assess the activity of the newly expressed serine protease Peg_1025, the cell lysates from samples N-HS, N-HA, C-HS and C-HA were run in a zymogram staining gel (Figure 44). Unfortunately, it seemed that the samples did not run through the gel and only weakly penetrated the resolving gel. Nonetheless, minor casein-hydrolyzing shades could be detected where the samples were loaded (Figure 44, lanes 3, 4 and 6) which were absent in the negative control (Figure 44, lanes 1 and 2). The hydrolysis was unclear for N-HA sample (Figure 44, lane 5). A clear casein-hydrolyzing shade (smear) was detected for the positive control, proteinase K (Figure 44, lane 7).

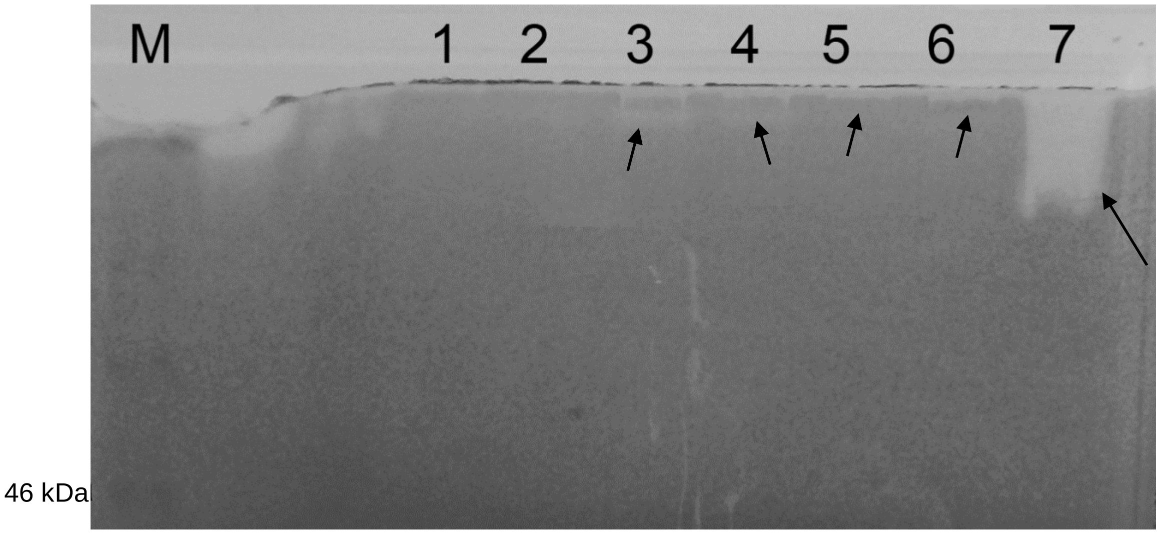


Figure 44 Zymogram staining of caseinolytic activity of the newly expressed serine protease Peg_1025 from cell extract samples. M, broad rage marker; lane 1, water (control); lane 2, *E. coli* LOBSTR without expression vector (control); lane 3 and 4, N-tagged and C-tagged Peg_1025 after heat shock treatment (70°C for 15 minutes), respectively; lane 5 and 6, N-tagged and C-tagged serine protease Peg_1025 after heat activation treatment (70°C for 2 hours), respectively; lane 7, proteinase K. Casein-hydrolyzing shades are indicated by black arrows.

Proteolytic activity assessment

Activity of the newly expressed serine protease Peg_1025 was assessed with fluorescein isothiocyanate (FITC) conjugated casein. When the assay was run for 1 hour incubation, heat activated samples (N-HA and C-HA) seemed more active than the corresponding heat shocked samples (N-HS and C-HS) (Figure 45). In particular, C-HA sample resulted with the highest activity, > 120% compared to the blank. No activity was recorded for N-HS sample.

However, when the assay was run for 24 hours incubation time, no fluorescent could be detected in none of the samples. Solid yellow precipitates were found instead, at the bottom of the small tubes used to run the test. The positive control containing proteinase K resulted active on the substrate.

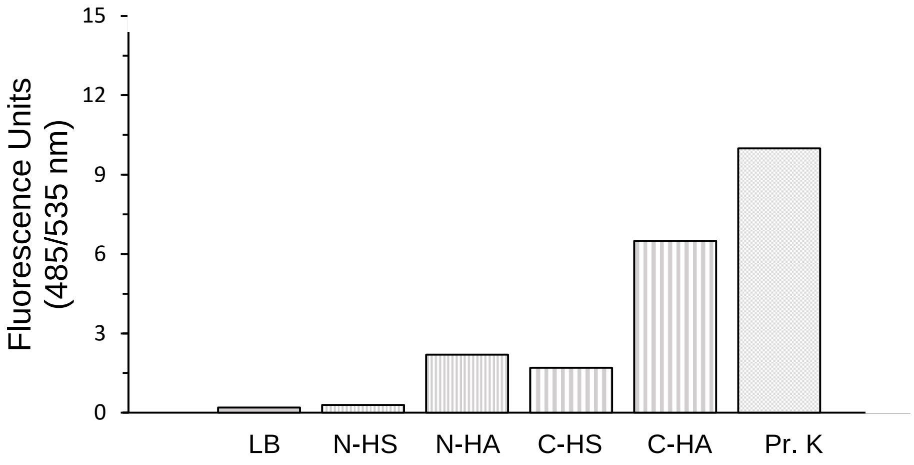


Figure 45 Fluorescent measurements obtained with FITC-casein enzyme test of the cell extracts containing the newly expressed serine protease Peg_1025 after 1 hour incubation with one technical replicate. LB, *E. coli* LOBSTR without expression vectors; N-HS and C-HS, N-tagged and C-tagged serine protease Peg_1025 after heat shock treatment (70°C for 15 minutes), respectively; N-HA, C-HA, N-tagged and C-tagged serine protease Peg_1025 after heat activation treatment (70°C for 2 hours); Pr. K, proteinase K. Standard deviation is also shown.

Substrate enzyme assay

Peg_1025 samples from N-HS, N-HA, C-HS, C-HA samples were added into tubes containing small portions of feathers to test its activity on this substrate. Unfortunately, no evidence for significant feather degradation was observed in any of the tubes (Figure 46).

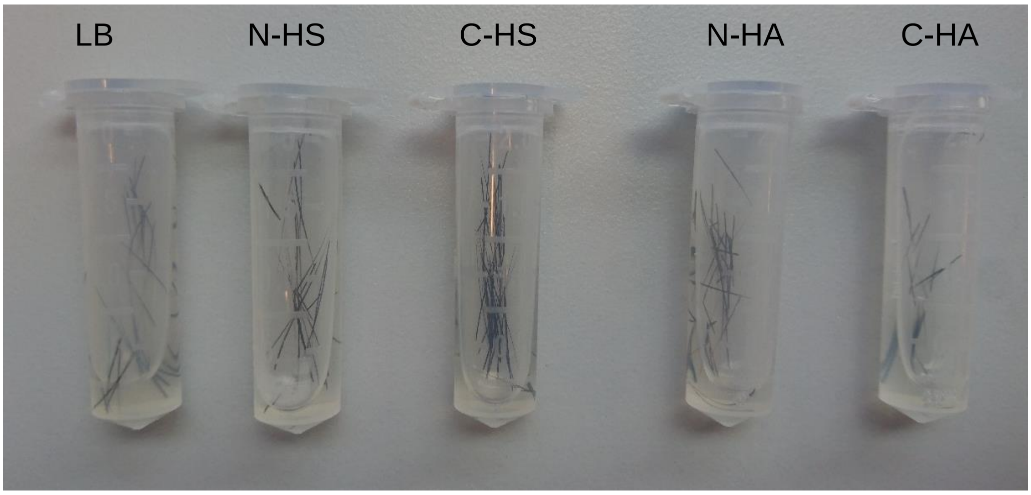


Figure 46 Activity test of the newly expressed serine protease Peg_1025 on feather substrate. LB, *E. coli* LOBSTR without expression vectors. N-HS and C-HS, N-tagged and C-tagged serine protease Peg_1025 after heat shock treatment (70°C for 15 minutes), respectively; N-HA, C-HA, N-tagged and C-tagged serine protease Peg_1025 after heat activation treatment (70°C for 2 hours).

Discussion

Part I. *Fervidobacterium pennivorans* strain T, an extremophilic microorganism

Fervidobacterium pennivorans strain T was isolated from a terrestrial hot-spring in Tajikistan and characterized (Javier-López, 2018). The strain possesses an outer sheath-like envelope, called toga, typical of organisms belonging to the Thermotogae phylum (Rosenberg et al. 2014), but it can be morphologically differentiated by the presence of a terminal spheroid ("bleb") at one end of the cell, a common feature of all members of the *Fervidobacterium* species (Figure 18) (Friedricht & Antranikian, 1996; Frock et al., 2010).

The combination of the toga, a membrane composed of crystalline arrays, long-chained fatty acids and rich in proteins (Bhandari & Gupta, 2014; Huber et al., 1986), with a robust cell wall, essential to withstand the stresses of extreme environments, confers a highly resistant cellular structure that makes extraction of high-quality DNA or RNA challenging. Cells strongly resist mechanical lysis, causing some DNA/RNA extraction kits based on rough physical treatment, such as glass beads (e.g. Qiagen), inefficient. Instead, chemical lysis produces (e.g. Sigma kits) are efficient in breaking the Thermotogae cell wall, and was surely essential in the current study, but extracted material requires more cleaning and handling steps that further deteriorate DNA/RNA, ultimately leading low yield and quality. To compensate for this drawback, a significant amount of starting material (biomass) was necessary, eventually yielding sufficient high-quality genomic DNA that met the Pacbio sequencing requirements.

The complete genome of *F. pennivorans* strain T was 2002515 base pairs long and some manual editing was necessary in order to compare it with other genomes of *Fervidobacterium* species available in Genbank (Table 13).

When the genome was aligned with the genome of *F. pennivorans* DSM9078 and *F. islandicum* AW-1 by MAUVE (Figure 22, S 4), three inverted locally collinear blocks (LCBs) were identified. Dot plot analysis using the RAST server showed the same results (not shown). The GC skew of *F. pennivorans* strain T was visualized by BRIG and indeed seemed to not be coherent with that found in other species (S 12). That is, a clear GC skew pattern was not as obvious as seen in other related bacteria. The GC skew has usually proven to be useful as the indicator of the DNA leading strand, lagging strand, replication origin, and replication terminal (Lobry, 1996; Necşulea & Lobry, 2007). All bacteria contain only one DNA replication origin and, in most of them, the GC skew is usually positive in the leading strand and negative in the lagging strand. It is thus expected to see a switch in GC skew just at the point of DNA replication origin and termination. Preliminary analysis in the query genome detected transposases at the extremities of the inverted regions, suggesting sequence inversions in this organism to be caused by evolutionary events rather than sequence assembly biases. More detailed analysis should be carried out in this regard, for example trying to assemble the raw data of the genome with an additional software and compare the results, or sequencing amplicons targeting the sequence in between LCBs.

Genomes of strain T, *F. pennivorans* DSM9078 and *F. islandicum* AW-1 differed in size, gene content and gene-to-gene similarity as observed by RAST annotation, Mauve alignment and BRIG analysis. Strain T had the shortest genome among these relatives (Table 13, S 4). Shorter genomes may have advantages in the environment as it takes less energy to duplicate and give faster replication times. However, it may also restrain the metabolism capability of an organism, leading to narrower niches and less stress resilience. Indeed, less coding sequences were predicted in *F. pennivorans* strain T when compared to the other two organisms (Table 13), but the nature of the lacking genes remains unknown. Further analysis should be performed in this regard. On the other hand, genes only present in strain T were also detected in Mauve alignment (Figure 22) and caused gaps in the rings of *F. pennivorans* DSM9078 and *F. islandicum* AW-1 when the three genomes were blasted against each other by BRIG (Figure 25). The comparison also showed a higher degree of similarity between genes of strain T to *F. pennivorans* DSM9078 than to *F. islandicum* AW-1. As it will discussed, the same similarity trend was also seen in the protease encoding genes (Figure 26). These patterns were coherent with the DNA-DNA *in-silico* hybridization (DDH) analysis and Average Nucleotide

Identities (ANI) (Table 15). In fact, both DDH and ANI values showed *F. pennivorans* strain T genomic identity closer to *F. pennivorans* DSM9078 than to *F. islandicum* AW-1. That is, genomic and gene similarities between organisms decreases as expected with their phylogenetic distances.

Analysis of the second contig were also performed. The identification of artefact sequences produced by sequencing technologies is usually easy and straightforward if the output sheet provided by the sequencing facility is reviewed. That is, the lower the coverage of a sequence, the more likely that it is the result of technical errors or contaminations. Nonetheless, manual reviewing of extra smaller contigs resulting from assembly of Pacbio raw data is always important and may give interesting insights on the derivation of the sequence. For example, contigs corresponding to contaminants can reveal lack of a pure culture or simply the inefficiency of sterile or aseptic conditions as well as human error factors. On the other hand, contigs belonging to the subject organism may be the result of molecular events such as genomic variability in a population or mutation events. Activation of mobile element such as transposases can lead to duplications and/or subsequence degeneration of nearby genes leading to a pool of cells in a population where part of the genome is distorted. The presence of a mobile element protein (*peg_260*) (Figure 23) in contig_1 and its absence in contig_2 may be interpreted in such way, justifying the presence of many disrupted genes in contig_2 that were nonetheless homologous to the original genes in contig_1.

The strain grew well with different substrates, such as glucose, peptone and casamino acids. However, as a fermentative and strictly anaerobic microorganism, the metabolic capability to use defined carbon sources is limited. Turbid cultures of *F. pennivorans* strain T were only obtained when yeast extract was added as supplemental nutrient, suggesting a need for a mixture of auxiliary nutrients for growth. This behaviour is not unique to this organism and described also for other strains belonging to the *F. pennivorans* species (Bhandari & Gupta, 2014) as well as for many other strict anaerobes.

The growth curve of *F. pennivorans* strain T was determined for the organism when growing with 0.1% yeast extract and 0.5% peptone (Figure 27). Exponentially phase cells are typically in their optimal physiological state, a condition in general preferred for studies of their physiological and metabolic properties (Klumpp et al., 2009; Kumar & Takagi, 1999). Thus, knowing their growth properties was essential for the success of other experiments (i.e. transcriptomics and secretomics). This was actually the first growth curve reported for a *F. pennivorans* species. The time it took new cultures to reach the start of the exponential phase (Figure 27, hour 2) was dependent on different variables: temperature and growth stage of the seed culture, its age and the temperature of the fresh medium. In fact, when stationary phase culture was stored for more than two weeks at room temperature was inoculated into new medium at room temperature, it took up to ~6 10 hours to reach the log phase. When a less than two weeks-old seed culture, stored at room temperature, which incubation was stopped when cells were in their log phase, was inoculated into new medium at 65°C, it would take ~4 hours before reaching the log phase (not shown). Best conditions were found when a fresh seed culture in log phase at 65°C was inoculated into MMF at 65°C. In this case, less than 2 hours was need for cells to enter the log phase (Figure 27). Long lag phase of newly inoculated cultures is a well described phenomenon in microbiology and is usually caused by depletion of various essential nutrients and time required to express new genes enzymes (Klumpp et al., 2009) and to initiate cell growth and DNA replication. Furthermore, in the case of a thermophilic microorganism such *F. pennivorans* strain T, additional time is needed for the medium to reach the optimal grow temperature.

Independently from the duration of the lag phase, the exponential phase of the organism was 10 to 12 hours long. Cells divided every 150 minutes and reached the stationary phase after 13 hours incubation. Thermotogae generation times can differ greatly among members, varying from 0.75 to 4.75 hours (Bhandari & Gupta, 2014), highly depending on the substrate used. *F. islandicum* AW-1 showed similar log phase to strain T when grown in similar conditions (Kang et al., 2019).

Interestingly, stationary phases recorded for *F. pennivorans* strain T never ceased and cultures never enter an apparent death phase, at least within 24 hours. In fact, cultures remained turbid even after long time incubation (over three months). This is furher supported by microscopic observation: cells do not lyse as the cell wall and toga remain intact. The cell wall is therefore likely to prevent cells from bursting even if they cease dividing. Turbidity is still recorded by the spectrophotometer, causing bias in the absorbance measurements. Although, several causes have been formulated to explain why strain T cells start to die, it is normally a combination between depletion of essential nutrients in the culture medium and the accumulation of the organism’s waste products (Madigan et al., 2019). For the former, it is still unclear whether depletion of carbon sources such as yeast extract or peptone is the limiting factor or if it is the depletion of vitamins or elemental minerals. For the latter, the increase of gasses in the headpsace of the anoxic flasks may have a detrimental effect. Elevated gas production containing hydrogen and carbon dioxide were detected and, particularly, hydrogen was described to have inhibitory effects on fervidobacteria (Bhandari & Gupta, 2014). Acetate and ethanol have also been reported as fermentation products (Friedricht & Antranikian, 1996) but were not measured in this study. Sulphide was detected by its particular smell only when the microorganism was grown on feathers or keratin azure. These substrates are particularly rich in sulphur-containing residues (i.e. cysteine and methionine) and when degraded by the microorganism, sulphide is released by the degradation of these amino acids (Rosenberg et al. 2013). The toxic effect of excessive sulphide in the flasks should also not be ignored. In this context, further studies should be performed to fully identify fermentation bio-products released by this organism and reveal its mass balance.

Independently of the causes for cell death, a fundamental question still remains unsolved: how long can *F. pennivorans* strain T cells actually be viable at incubation conditions? Cultures left at room temperature on lab benches could still be used as inocula for new batches even after a long period of time and the longevity of this organism was also reported by previous studies where cultures could be retrieved even after 12 month at 4°C (Friedricht & Antranikian, 1996). Nonetheless, long lasting culture turbidity and unknown cellular longevity caused problems in the evaluation of time sampling for transcriptomics and secretomics experiments of cultures growing on feather substrates.

Members of *Fervidobacterium* genus have been recognized for their ability to catabolize a variety of complex polymer sources, such as chitin (e.g. *F. pennivorans*), cellulose (e.g. *F. islandicum*) and xylan (e.g. *F. riparium*) (Bhandari & Gupta, 2014). In particular, anaerobic degradation of keratin is so far restricted to a few members in this genus: *F. islandicum* AW-1 (Huber et al., 1990), *F. pennivorans* DSM9078 (Friedricht & Antranikian, 1996) and *F. thailandense* (Kanoksilapatham et al., 2016). However, only *F. pennivorans* strain T has been reported to have different degradation times for different types of chicken feathers: breast and wing feathers. The difference in degradation time was significant for these two feather types. Degradation was complete within 72 hours for breast feather (Figure 19) whereas for wing feathers it could take up to 10 days (Figure 20). In fact, feathers are classified according to properties such as toughness, weight and interlinks between sections (barbules) (McDonough, 2013; Thompson, 2014). Chicken breast feathers are soft, light and fluffy whereas wing feathers were stiffened, hard and heavier (more robust).

Beside visual assessment, an increase in free sulfhydryl groups in the cultures was shown to be a more precise method to assess the state of feather degradation (Kang et al., 2019). An increase of free –SH groups is related with broken disulphide bonds and thus to keratin deconstruction. An effort to measure release of sulfhydryl groups in strain T batches was attempted, but unfortunately not successful. Media with different reducing agents have been tested (i.e. cystein and sulphide) as it was thought these chemicals may interact with DTNB Ellman's reagent. Also, different reaction buffers (0.1 M sodium phosphate (pH 8), with and without EDTA) have been tried, as suggested by Ellman's protocols (Ellman, 1959). Finally, the experimental procedure described by Kang et al. (2019) was followed, as they reported successful results with their *F. islandicum* AW-1 strain. Even so, the results obtained in this study remained inconclusive and the causes are still unclear. Other techniques may be considered to qualitatively assess the state of feathers degradation. For example, the more feather degradation was observed in incubated cultures, the more smearing material was detected in SDS-PAGE analysis of their cell-free supernatant. That is, periodic sampling from feather batches can be used to correlate the amount of small peptides in solution, thus relating to the state of feather degradation. Further experiments should be performed in this regard.

F. pennivorans strain T can clearly degrade feathers, but whether is able to use the substrate as source of nutrients is uncertain and yet to be clarified.

Experiments with strain T showed that cultures would only degrade feathers when the initial peptone concentrations were > 0.1%. In the presence of 0.1% yeast extract and 0.5% glucose, dense cultures were obtained, but feather degradation would not occur. This may be explained by the fact that the strain may not possess receptors for intact keratin-like substrates in the absence of peptone, and thus be unable to trigger the regulatory machines that would lead to the expression of enzymes involved in protein degradation or protein metabolism. The regulatory machine may be stimulated only by soluble (or smaller) protein-like substrates, such as peptone components, for which the organisms have adapted receptors. Thus, *only* when protease encoding genes have already been triggered and their corresponding enzymes expressed and secreted in sufficient amounts, keratin may start to be degraded, as well. Smaller keratin peptides may further stimulate protease production.

Secondly, it appeared that, although in dense cultures, cells may actually start to die off during the degradation of feathers. As it was discussed, it is unknown how long cells can remain viable in incubation. Surely, distinct metabolic strategies are exhibited when both soluble and insoluble substrates are present. In fact, soluble nutrients are depleted first, and then feather degradation starts (Kang et al. 2019). Whether cellular viability declines because cells cannot use peptides generated from keratin substrates as nutrient source or as an effect of accumulation of waste products / depletion of other essential nutrients by the time the degradation of feather starts, is unclear. Either way, turbidity of cultures do not increase with longer incubation times although feathers degradation proceeds. It was also reported that cells visualized under phase-contrast microscope did not maintain sharp rod-shapes, suggesting some clues of cellular burst (Javier-López, 2018).

Furthermore, as it will be discussed, from preliminary transcriptomics results and from secretomics experiments, it seemed that the concentration of RNA and proteins retrieved from the pellets and supernatant, respectively, decreased with longer incubation times. That is, the more the feathers were degraded, the less proteins and RNA were detected in the batches. However, degradation of feathers may still continue thanks to the proteases already present in the media. Once the substrate decreases, proteases hydrolyse each other (Toogood et al., 2000), and cannot be further detected in the supernatant. In this sense, further studies should be performed, for example obtaining a growth curve of the organism while growing on feather can give significant insight in the processes in stake, but it would require well planned experimental strategies.

Total RNA extraction performed by Eurofins Genomics from cell pellets grown on feathers were unsuccessful even after several trials. Production of proteases is highly dependent on growth phase and nutrient availability (Klumpp et al., 2009; Kumar & Takagi, 1999) thus the growth curve of the organisms was taken into consideration. However, as already discussed, it was unclear whether *F. pennivorans* strain T cells would be still physiologically active during feather degradation and it was unknown when specific keratin-degrading proteases would be expressed by the organism. Thus, visual degradation of feathers was more importantly taken into consideration when sampling cells from the bioreactor. As it was argued earlier, cells may already be dead at that stage and this may explain why the sequencing company failed repeatedly in the RNA extraction. For further studies, it is suggested to take into consideration shorter incubation times, perhaps when feather degradation just become visible, to preserve as much RNA as possible. The addition of extra essential elements later in the incubation may prevent cellular decline, but it may affect the pH of the medium. Nonetheless, the bioreactor experiment was the first attempt reported of up-scaled feather degradation for industrial applications using a member of the *Fervidobacterium* genus. The results were preliminary and yet to be optimized in light of the problematic just described.

Regarding secretomics, no proteins were detected in the cell-free supernatant of glucose batches by SDS-PAGE gel (Figure 29, line 2). Lack of significant amount of proteins in the supernatant is not surprising in the light of what was hypothesised before. In fact, *F. pennivorans* strain T may secret proteases only when protein-like substrates are detected in the environment. Thus, with glucose, and yeast extract, the protein catabolism is not activated and extracellular proteases are not expressed. Indeed, protein bands were detected in the cell-free supernatants of peptone, casamino acids and keratin azure batches by SDS-PAGE analysis (Figure 29). Bands corresponding to proteins present in the supernatant were either secreted from the bacterium or derived from lysed cells. However, as already discussed, *F.*

pennivorans strain T cells do not seem to lyse easily thanks to the constrain exhibited by the cell wall. The growth curve was also taken into consideration to avoid prolonged incubations and eventual lysis of the cultures. Thus, it can be inferred that all the protein band in the gel were secreted by the organism. It was expected that, in different substrate conditions, different sets of enzymes were secreted by strain T. However, it seemed that the same mixture of proteins was secreted in all three protein-derived substrates. Furthermore, from the results obtained by LC-MS/MS spectrometry of the three major bands visualized in the SDS-PAGE gel, it appeared that only one was identified as protease (serine protease Peg_1025). This protease is clearly secreted in presence of protein-derived substrates and directly involved in their degradation, as also different studies have shown (Kang et al., 2019). It is also one of the proteases detected by gene mining in the *F. pennivorans* strain T complete genome. Although promising, the results from the SDS-PAGE are incomplete and the secretomics analysis should be performed again, also in vision of a feather batch sample. Working with supernatants from extremophilic anaerobic cultures is challenging and the results hard to replicate. As already discussed, supernatants has to be sampled at the right moment during growth, to minimize proteolytic degradation (Toogood et al., 2000) and to avoid large background from keratin substrates that would result in extensive smearing in the SDS-PAGE gels. Particularly, these were all problems faced while working with feather batches, together with others already discussed. Another experimental challenge derives from cultures that do not grow to high density. Thus, higher volumes had to be concentrated to obtain at least some weak bands in the gels. This resulted in more handling steps that may ultimately cause loss of relevant material. To avoid these problems, the whole supernatant concentrate can be sent for LC-MS/MS spectrometry and the results analysed *in-toto* to obtain a more detailed and complete overview of the secretome of this organism.

The results from the secretomics were preliminary and still to be fully explained in light of the multitude of other proteases predicted in this study (Table 14, S 5) and those that are thought to be involved specifically in the degradation of feathers. In fact, it was shown that the biological degradation of keratin is a combination of intracellular proteases, membrane proteases and proteases secreted to the extracellular environment (Böckle & Müller, 1997; Friedricht & Antranikian, 1996; Kang et al., 2019; Ramnani & Gupta, 2007). The details of the mechanism are still unclear, but it seems that a combination of disulphide reductases, which cleave the disulphide bonds, are also needed to allow membrane and extracellular proteases to have access to their specific substrates (Böckle & Müller, 1997; Lange et al., 2016). Overall, the major role in the formation of small peptides is by means of proteases on the outer part of the cellular membrane. In fact, when cell extracts, membrane fractions and expressed proteases were compared, the membrane fraction, that is, lysed cells without detergent treatment, yielded most activity with feather substrate (Friedricht & Antranikian, 1996; Kang et al., 2019).

Even though disulphide reductases are essential for the initiation of feather degradation under aerobic conditions, it can be speculated that under these conditions the reduction of disulphide bonds may happen spontaneously and thus not be essential for keratin degradation. In this study, bioinformatics analysis on the complete genome of *F. pennivorans* strain T yielded a complete overview of proteases whose function is exploited in different cellular locations of the organism: inside the cell, bond to the membrane or in the extracellular environment. By gene mining analysis and genome-to-genome comparisons against *F. pennivorans* DSM9078 and *F. islandicum* AW-1, a total of 55 protease encoding genes were detected in strain T.

As already discussed, these three genomes differed in coding sequence quantity and similarity. Nonetheless, all 55 protease encoding genes were shared by strain T, *F. pennivorans* DSM9078 and *F. islandicum* AW-1 suggesting similar protein catabolism in the three strains, all having been shown to be able to degrade keratin.

The protease encoding genes were evenly distributed throughout the reference genome (Figure 25, Table 14). However, some with similar functions, such as genes *peg_1512* and *peg_1516*, were consequential and closely located to each other, suggesting clustered genes under the same regulatory mechanism. Some other genes were present in duplicates and still located close to each other, such as genes for signal peptide cleavage (*peg_638* and *peg_640*) or fervidolysin genes (*peg_1681* and *peg_1693*). The presence of common protein degrading genes, gene duplicates and the capability of degrading unusual substrates such as keratin, are all examples of different fascinating evolutionary theories yet to be fully understood in these microorganisms. High similarity in protease-encoding genes (Figure 26) is no surprise if the environment from which all these microorganisms have been isolated is taken into consideration: conditions, ecosystems and substrate availability may be similar in many ways.

Gene duplication is the major source of new genes in biology (Freeman & Herron, 2006). Whether new copies of a gene are maintained depends on the microorganism’s need for the encoded protein. For example, fervidolysin, a well characterized keratinase (Kim et al., 2004; Kluskens et al., 2002), may be needed in so high amount for degrading the substrates that both encoding genes may be required by the organism. On the other hand, if just one gene is already sufficient to produce enough protease to meet all the demands, the second gene would not be under selective pressure anymore and may diverge into similar or completely different functions. It may even degenerate into pseudogenes. As it will be discussed in the next section, serine protease *peg_1025* may have actually originated by a gene duplication event from fervidolysin.

The presence of keratinase genes leads also to another interesting evolutionary question: why keratin-degrading genes in the first place? The environments from which these organisms have been isolated, muddy sediments in freshwater hot pools, do not seem to have significant amount of keratin-like substrates from which an evolutionary pressure could arise, at least in recent times. Different theories have been speculated. Thermotogae is a deep-branching group that may had been exposed to keratin substrates early in their evolutionary state and then had retained the genes for unknown reasons. Surely, the term keratinase is used to designate all the proteases that have keratinolytic activity and also have a wide range of other substrate specificity, such as fibrin, elastin, collagen, casein (Bhandari & Gupta, 2014; Gerday & Glansdorff, 2007; Lange et al., 2016). That is, it just happened that these proteases can catabolise keratin, as well. Lastly, the insurgence of keratinase-encoding genes in these organisms may be derived by horizontal gene transfer phenomena, a very common event in fervidobacteria (Cuecas et al., 2017).

Of the 55 protease encoding genes, 26 were further characterized by MEROPS database (Table 14). Fourteen were predicted as cell membrane bond and 12 as extracellular. It was interesting to notice that most of these proteases belonged to the serine- and metallo- protease superfamilies. This is congruent with other studies that reported proteases belonging to these superfamilies are mainly involved in degradation of complex substrates (Kang et al. 2019; Lange et al. 2016; Rosenberg et al. 2014). In experiments with *F. islandicum* AW-1 growing in yeast extract and feathers, 16 of these 55 genes were differentially expressed (Kang et al. 2019). Some of these genes were highly expressed after yeast extract depletion whereas others had constant expression levels regardless of substrate. Despite of the serine protease *peg_1025* homologue in strain AW-1 (RS05775 in Kang's study) was reported to have a constant level of expression in *F. islandicum* AW-1, it was selected as a putative keratinase also in this study. In fact, this was the only protease identified in the SDS-PAGE from batches growing on different substrates and its relevance was further supported by the genomic results reported. The protein was successfully identified in *F. pennivorans* strain T cultures only from the band extracted from batches growing on casamino acids, but the same band was also detected from keratin azure grown batches (Figure 29, line 5) and is likely to be present in feather cultures, as well.

It would have been interesting to further screen the pool of 26 protease encoding genes detected by gene mining analysis with the results of transcriptomics. A wider understanding of the limitations of the methods used in this study may have been reached and further insight on keratin degradation by *F. pennivorans* strain T may have been obtained. Transcriptomics may also have explained why no other proteases were detected in the SDS-PAGE analysis even though 11 additional proteases were predicted to act in the extracellular environment (Table 14). In this perspective, given their high similarity (Figure 26), gene expression levels of these genes from *F. islandicum* AW-1 (Kang et al. 2019) may provide clues to their regulation and roles in *F. pennivorans* strain T.

Nonetheless, a serine protease of intriguing keratin-degrading possibilities was identified, expressed and further described both by bioinformatics tools and by activity tests. The potential of the other enzymes found in this strain is left for future research.

Part II. Biotechnological perspectives

Prior to expression of serine protease *peg_1025*, more information about its nature, structure, function and catalytic site were obtained using different bioinformatics tools available from online platforms. This gave an overall knowledge on the protein and allowed better experimental design. Information about hydrophobicity, cysteine content, ion binding sites, and domains, among others, were used for the preparation of expression and activity test experiments, in the effort to optimize the outcome.

The protease belonged to the protease clan SB, family S8 with serine-type endopeptidase activity, generally called subtilases. Subtilases are characterized by conservation of three catalytic residues, aspartic acid, histidine and serine ((the catalytic triad) (Figure 35) and by the presence of a signal peptide sequence, which is essential for exporting the protein outside the cell (Figure 32, Figure 33) (Siezen & Leunissen, 2008; Siezen et al., 1991). Furthermore, most proteases in this family possess a propeptide, that is, a part of the protein whose function is independent from the main (catalytic) domain of the protein and act both as intramolecular chaperone and inhibitor of the protease (Shinde & Inouye, 2000). Extracellular subtilases are produced and secreted as precursors as a full-length immature protein. Then, the propeptide is autocatalytically cleaved off the mature protein by mediation of the catalytic triad itself and operate as chaperone for the correct folding of the protein. However, although this process is essential for the enzyme to fold into their active conformation, it is not sufficient for activation of the protease. In fact, the propeptide domain binds noncovalently to the catalytic domain after the cleavage, covering its active site and acting as inhibitor until it is released and the protease becomes mature. This interaction is thought to be due to hydrophobic interactions between the two domains (Shinde & Inouye, 1993, 2000). However, it remains unclear whether propeptides have specific residues that act as reactive sites to exploit its function or if just have conserved motifs, instead (Shinde & Inouye, 1993, 2000). Either way, a general trend in hydrophobicity in these propeptide sequences have been observed and specific patterns have been reported (Shinde & Inouye, 1993). On one hand, this complex intramolecular mechanism of cleaving, folding and activation is useful in nature as it prevents unwanted protease activity inside a cell.

Subtilases act on a wide range of substrates, even on themselves (Toogood et al., 2000), and their uncontrolled action can be detrimental. On the other hand, when these enzymes are expressed in hosts for their usage as biocatalysts, their activation is a necessity. Many speculations have been made on how to remove a propeptide or how to activate proteases of this family. For example, the protease is expressed in a functional state intracellularly when the signal peptide sequence is omitted from expression construct (Figure 43), but omission of the propeptide sequence was shown to prevent proper protein folding (Kim et al., 2004). Crude cell extract from the original organism was thought to contain mineral elements or molecular factors that, if added into the expressed protein solution, might help the protease activation (Friedricht & Antranikian, 1996; Kang et al., 2019; Kim et al., 2004; Kluskens et al., 2002). For some other thermophilic proteins, heat treatment was found to be essential for the removal of propeptide and for the activation of the proteins (Peek et al., 1993; Toogood et al., 2000). Finally, other proteins expressed by the organism, might carry out the proteolytic activation of the subtilases (Kim et al., 2004).

In this study, to understand the nature of serine protease Peg_1025 and to unravel the presence of a propeptide domain, its sequence was compared with homologues proteins in a multiple sequence alignment (Figure 35, Figure 36). The alignment showed the conservation of the catalytic triad throughout all the subtilases analysed (Figure 35). It also gave insights about hypothetical propeptide sequence on the protein (Figure 36). In this regard, some degree of sequence conservation was observed (Figure 36) but, as earlier mentioned, the presence of a functional propeptide cannot be clearly stated as it is still unclear whether propeptides in general have specific residues that act as reactive sites and/or if have conserved motifs, instead. The propeptide sequence seemed to be highly conserved when its structural model was predicted using the propeptide sequence from fervidolysin as template (S 8). Also, LC-MS/MS analysis of Peg_1025 extracted from the supernatant of casamino acids grown batches identified very few unique peptides and experimental spectra belonging to the propeptide part of the sequence (22-127 AA) (Figure 31) and thus leading to the interpretation that the secreted protease is mainly found in the active matured form. Finally, heat activated samples (2 hours at 70°C), following the protocol of Toogood et al. (2000), seemed to have higher activity compared to the heat treated ones (15 minutes at 70°C) (Figure 45) indicating a maturation event. In this regard, N-terminal heat-shocked sample (N-HS) did not show any activity compared to the N-HA, maybe because the N-tag interferes with the proper folding and function of the propeptide domain, ultimately leading to inactive proteins.

The primary sequence to which serine protease Peg_1025 had the highest degree of similarity was an endoprotease subtilase from *Geobacillus stearothermophilus* strain AK1 named subtilisin Ak1. Also fervidolysin from *F. pennivorans* DSM9078 showed a high degree of similarity to the query sequence.

The evolutionary relationships between fervidolysin, subtilisin Ak1 and serine protease Peg_1025 were clarified by the phylogenetic tree reconstruction based on the sequences of all homologues found (Figure 37). As discussed before, gene duplication is the most common event for generating new genes. The phylogenetic tree suggested that a duplication event occurred that generated two copies of a fervidolysin-like gene ancestor from one of which all the other proteases reported diverged, forming an independent cluster, including subtilisin Ak1 and serine protease Peg_1025. The phylogenetic tree reported in this study is the first description of the protease belonging to this phylogenetic cluster.

Further clues to the closer relationship of serine protease Peg_1025 to subtilisin Ak1 than to fervidolysin were obtained when their model properties from SWISS-MODEL web tool were considered and their structural prediction scores compared (Figure 38, S 7 and S 8). Both fervidolysin and subtilisin Ak1 are well characterised serine proteases and their defined 3D structures are solved and reported (Kim et al., 2004; Smith et al., 1999). A structural prediction of the matured Peg_1025 protein was obtained, with the catalytic domain based on subtilisin Ak1 (Figure 39) while the propeptide domain (S 9) was based on the fervidolysin structure. In fact, only an inactive form of fervidolysin structure is available, as an unmatured complex of propeptide and catalytic domain (Kim et al., 2004).

Interesting similarities and differences were observed between Peg_1025 and subtilisin Ak1 structures. First, although the majority of the local folds were highly conserved between the proteases (Figure 40), five insertions in Peg_1025 that had no homologue in its template were identified (S 10). Insertions may be involved in further stabilization of the protein or to regulate its activity. They may also act as binding domains for polymeric substrates, but because insertions were short and generally widespread in the sequence, this is unlikely the case.

Another interesting difference was observed regarding the cation binding sites. Cation binding sites for Ca^{2+} and Na^{+} were only found in the subtilisin Ak1 template and not in fervidolysin. However, not all the binding sites were predicted by the modelling of Peg_1025 and only the residues for binding of Ca-2 were conserved (Figure 39). Whether only one Ca^{2+} binding site is actually present in Peg_1025 remains unclear and needs crystal structure determination. Nonetheless, binding of cations on the outer surface of thermophilic proteins is an evolutionary advantage that helps to cope with high temperatures by increasing the overall stability (Gerday & Glansdorff, 2007; Siezen & Leunissen, 2008; Smith et al., 1999; Sterner & Liebl, 2001). Calcium ions, in particular, were found to have important effects on the activity of proteases and in the stability of thermophilic enzymes (Siezen & Leunissen, 2008; Smith et al., 1999; Toogood et al., 2000). If only one Ca-binding site is indeed present in Peg_1025, its stability at high temperatures may be affected and activity lost over time. This may explain the negative results in the casein fluorescence assay with incubation times of 24 hours, where precipitates was observed.

Finally, both subtilisin Ak1 and serine protease Peg_1025 had two cysteines in the polypeptide sequence after maturation. Fervidolysin also has two cysteines, but one is lost with the cleavage of the signal peptide (Kluskens et al., 2002). Subtilisin Ak1 has two cysteines forming a disulphide bridge in the folded active form (Toogood et al., 2000). Serine protease Peg_1025 has three cysteines: one is lost in the signal peptide sequence while the other two are situated in the catalytic domain. However, they reside too far apart from each other in the predicted folded structure and unlikely to form a disulphide bridge (Figure 39, A). Their absence may have important effects both on the stability of the protein and for its substrate specificity. In fact, disulphide bridges are considered another key structural feature of thermophilic enzymes as their strong bond helps maintaining the correct folding and prevent loss of activity due to thermal denaturation (Gerday & Glansdorff, 2007; Reed et al., 2013; Sterner & Liebl, 2001; Toogood et al., 2000). In previous studies on subtilisin

Ak1, it was shown that addition of DTT in the reaction buffer to reduce and break the disulphide bond caused reduced thermal stability of the protein and lower substrate specificity (Toogood et al., 2000). This was observed also in side experiments in this study, where DTT was added to the lyses buffer resulting in the complete precipitation of all proteins after heat treatment. Absence of the disulphide bond was shown to affect the substrate specificity in subtilisin Ak1 leading to more weakly bound catalytic domain-substrate interactions and decreased K_{cat} (Toogood et al. 2000). Lack of a disulphide bond in the active site may also affect the substrate specificity and efficiency of Peg_1025.

Bioinformatics-based protein characterization and structural modelling remain theoretical approaches and experimental methods such as X-ray crystallography should ultimately be used to fully understand the properties of the Peg_1025 protein. Nonetheless, use of predictive bioinformatics tools were fundamental to gather important information on many properties of Peg_1025 and will help to design and guide further studies.

Cloning protocols and their results were reviewed by Javier-López (2018) and will not be further discussed. Instead, a short discussion about primer design and cloning *in-silico* is presented in this study. Primers were designed to amplify gene *peg_1025* in order to follow the FX cloning protocol. Forward and reverse primers for the open reading frame coding for *peg_1025* were designed and both were composed of two parts: one containing a sequence targeting the gene (target sequence) and one containing a sequence extension on their 5’-end (extension) (Table 16). The target sequence annealed to the gene sequence devoid of start (ATG) and stop (TAA) codons as well as the signal peptide (nucleotides 1 – 63). In fact, start and stop codons were already present in the expression vectors, in front of the 10xHis-Tag and after the insert, respectively (Figure 41, Figure 42). The sequence associated with the signal peptide was omitted (nucleotide 1 – 63). Thus, the Forward-primer bound the gene from nucleotide 64 (triplet AAC…) (S 3). As already discussed, the gene sequence associated with the propeptide (nucleotide 64 – 288) was kept (Latiffi et al., 2013; Shinde & Inouye, 1993, 2000; Toogood et al., 2000). The extension sequence for the forward primers was 5’- atatatGCTCTTCtAGTnnn and 5’-atatatGCTCTTCaTGCnnn for the reverse primer (Table 16). The recognition sites

(5’-GCTCTT) for restriction enzyme *Sap*I, are located at the respective 5’ ends of the cleavage site (F-5’-tAGT; R-5’-aTGC), which will result in short single-stranded overhangs at both ends (sticky ends) after *Sap*I digestion of the amplicons. The short single-stranded overhangs are different in the Forward- and Reverse-primers and were designed according to the sticky ends generated in the propagation vector after its *Sap*I digestion. Lowercase letters (5’-atatat) in the extensions can be replaced by any nucleotide and were added to prevent formation of secondary structure elements. More general factors should be taken into consideration when designing primers: the target sequence should not be too long (i.e. between 18-25 nt); the target sequence should contain a minimum of two C/G at the 3’ end; primers should not be complementary to each other; the primers, and in general the whole sequence to be amplified, should not contain the recognition site of the restriction enzyme used (in this case, *Sap*I); finally the melting temperature calculated for both primers should be similar and less than 2°C different between each other.

Once primers were designed, gene *peg_1025* amplification and its cloning into expression vectors p7xC3H and p7xN3H were performed *in-silico* in Serial Cloner software (Figure 41, Figure 42). There are several advantages in running a complex procedure like this *in-silico* before its experimental try-out, for example to check primer design, PCR efficiency and overall cloning outcome. It is also very useful for the researcher to fully understand the procedure and to have detailed insight about the nucleotide sequence that will result and that will ultimately encode the target protein.

Vectors p7xC3H and p7xN3H were used to express serine protease *peg_1025*, with 10 x his tag at its C- terminal and N- terminal, respectively, to test the best tagged protein outcome. Results of this study showed a better expression yield in the p7xC3H vector (S 11), also with higher activity, especially after heat activation (Figure 45). Structural predictions of Peg_1025 at its N- and C- terminal were different: globular in the former and disordered in the latter. In particular, the C-terminal of Peg_1025 was predicted to be in a free state. Extra amino acids added by N-tag to the N-terminal of the propeptide may interfere with both the stability and activity of the protease. Firstly, if the chaperon activity of the propeptide is affected, the proper folding of Peg_1025 will be altered as well. That is, extra residues at the termini can affect protein folding and activity (Geertsma & Dutzler, 2011). Secondly, fouled propeptides by the N-tag may prevent the activation of the protease. As the casein activity assay suggested, the removal of propeptide by heat activation caused higher fluorescence values.

Expression of serine protease *peg_1025* with p7xC3H and p7xN3H vectors yielded soluble proteins which did not affect the growth of the expression host. In fact, *E. coli* LOBSTR, both with and without expression vectors, grew until saturation with the same trend. Interestingly, the gene was expressed in the host, even in the absence of IPTG (Figure 43, lane 2). This result was coherent with Kluskens et al. (2002) for fervidolysin expression.

In the attempt to optimize the expression protocol, several strategies were tested. The composition of the lysis buffer is very important and essential in maintaining the target protein soluble and intact. The lysis buffer contained only Tris-HCl buffer solution, salt and glycerol. It was already discussed the negative effect of DTT in the solubility of the protease when heat treated. Moreover, chemicals that facilitate the lyses of the cells, such as detergents or lysozymes may also be added. In fact, the breakdown of cells was the most critical step of the entire procedure. In this project, cells were lysed by sonication but two major problems were faced. Firstly, the pulses may excessively over heat the samples, even when handed on ice, causing degradation of the target protein. Secondly, the amount of target protein retrieved largely depends on how well cell hosts are broken. This is why microscopic evaluation after sonication is recommended: to check the lysis outcome to avoid handling solutions with little expressed proteins. Incubation of the samples at 37°C with lysozyme facilitated cellular lysis by sonication.

Two major drawbacks from the expression protocols were identified. First, as discussed earlier, N-terminal tags may interfere with the protease folding. It is thus suggested to proceed with the C-terminal tagged construct for further investigations. Secondly, purification of the His-tag proteins by nickel affinity columns was not performed and is recommended for future studies. Heat shock treatment (70°C for 15 minutes) helped in removing most of the *E. coli* proteins, although many others were still detected in solution (Figure 43, S 11). These proteins gave too much background in the SDS-PAGE to assess the efficiency of the heat activation (70°C for 2 hours) on the target protein (Figure 43). If the solution had been purified, three bands of decreasing molecular weight, corresponding to the inactive proteases, active protease and propeptide, should have been detected in the gel (Kluskens et al., 2002; Toogood et al., 2000).

In this study, heat-treated cell lysates were directly used for activity tests. As already discussed, most subtilases have a propeptide that may inhibit protease activity, thus heat activation for two hours was used as a strategy to activate the expressed serine protease Peg_1025, as done in other studies on subtilisin Ak1 (Peek et al., 1993; Toogood et al., 2000). Both heat activated (-HA) and heat shocked (-HS) samples were tested for activity using the FITC-casein fluorescence kit, zymogram staining and feather substrate.

Zymogram staining of the samples was promising, although inadequate. Casein hydrolysing shades were detected on the top layer of the resolving gel when compared to the negative controls (Figure 44). Clearly, the samples did not run through the gel and it was thought that lack of heat denatured of the proteins in loading buffer may have affected their capability to migrate in the gel.

The results from the casein fluorescence detection assay after incubation for one hour showed heat activated samples (N-HA, C-HA) to have higher activity than the heat shocked ones (N-HS, C-HS) (Figure 45). As already discussed, this may be interpreted by the propeptide detachment from the catalytic domain of Peg_1025, resulting in its activation. N-tagged samples performed less well than the C-tagged samples, as the tag may interfere with the propeptide functions. In the attempt of assessing fervidolysin activity, Kluskens et al. (2002) failed in removing the propeptide, leading to negative results. Their study and the N-HS sample show how important it is to identify and acknowledge the presence of these intramolecular chaperones. Nonetheless, these results are only preliminary and more experiments should be performed.

On the other hand, incubation of the samples for 24 hours resulted in yellow precipitates and complete loss of activity. Subtilisin Ak1 was described having a half-life of about 10000 minutes with 5 mM Ca^{2+} in the reaction buffer (Toogood et al., 2000). However, as argued, Peg_1025 has three fewer predicted cation binding sites than subtilisin Ak1, which can limit its stability at high temperatures. Furthermore, it did not seem to possess a disulphide bond, which might further decrease its stability. The enzyme might not be very stable and thus be active only for short period of times, denaturing and precipitating with the FITC-substrate after long time incubations. In this regard, considering to lower both the incubation time to below 24 hours and the temperature less than 70°C, might help prevent degradation of the protease (Toogood et al., 2000) or of the FITC-casein substrate itself (Bjerga et al., 2016; Twining, 1984). It should also not be excluded that Peg_1025 may not recognise casein as an optimal substrate, as studies on fervidolysin suggested (Kluskens et al., 2002). This may also be due to the lack of the disulphide bond creating a looser and less specific active site (Toogood et al. 2000). Experiments with different reaction buffers may provide clues to optimize this assay.

Lastly, for the aims of the project, it was necessary to test serine protease Peg_1025 on feather substrate. However, as already discussed, feather degradation is a combination of a complex mixture of different enzymes which cooperate together for the keratin breakdown. Even though, free amino acid concentration by ninhydrin assay is suggested for further studies as it may detected any occurring partial keratin degradation occurring.

Conclusions

1. The first growth curve of a member of the *F. pennivorans* species was established. It showed a lag phase dependent on both physical and physiological conditions; an exponential phase 11 hours long, with cells having a generation time of about 150 minutes, and, finally, a long stationary phase that opens intriguing questions on the viability of this strain in cultural batches. Gasses produced by the fermentation metabolism of the organism were studied, although the toxicity of hydrogen and sulphide remain unclear.
2. The complete genome sequence of *F. pennivorans* strain T was obtained and analysed, revealing interesting molecular features, e.g. inverted genomic blocks, as well as intriguing similarities and differences compared to its most closely related organisms: *F. pennivorans* DSM9078^T and *F. islandicum* AW-1. ANI value was of 97.65 % versus the former and of 80.90% versus the latter. The strain T genome was slightly shorter (2002515 base pair) and contained less coding sequences, but the same number of predicted protease-encoding genes (55) were found. Eleven proteases were predicted to be secreted by the organism and thus thought to exploit their catalytic power in the extracellular environment.
3. One of the predicted proteases, a serine protease termed Peg_1025, was also detected when the supernatant of actively growing batch cultures was analysed by proteomics. As for the other predicted extracellular proteases, their expression and regulation still remain unknown, also in light of the failed transcriptomics analysis obtained from the fermenter. Nonetheless, the bioreactor experiment was one of the first in its kind, showing a first attempt of up-scaling feather degradation for biotechnological perspectives.
4. Peg_1025, which was chosen for further analysis, was characterized by bioinformatics tools, revealing the conserved catalytic triad typical of members of the subtilase S8 family as well as a signal peptide sequence and a propeptide domain. The identification of the propeptide was essential for the successful outcomes in the structural modelling and in the activity tests. Interesting properties on the enzyme were revealed when its structural model was built using its most similar enzyme characterized as a template: subtilisin

Ak1 of thermophilic bacterium *Geobacillus stearothermophilus* strain AK1. The model showed Peg_1025 to possess several insertions of unknown functions compared to subtilisin Ak1, only one conserved Ca^{2+} binding site of the three originally present in Ak1 as well as lack of a disulphide bond in the active cleft. Important insights on the evolutionary origin of this protease were unravelled by the phylogenetic tree obtained using homologues proteins. The phylogenetic tree reported is also the first description of the protease belonging to this phylogenetic cluster.

5. Expression of *peg_1025* was successfully carried out in *E. coli* and the expressed protease was demonstrated active at 70°C in different enzyme assays using casein as a substrate. The interpretation of the results was the final part of this project and summarised all the information and knowledge so far obtained for this process in extremophilic anaerobic environments.

Future directions

During this project, a member of Thermotogae, *Fervidobacterium pennivorans* strain T, has been assessed for its capability of degrading native chicken feathers. Its complete genome was obtained and protease encoding genes were predicted. A putative keratinase was identified, characterized and expressed. However, the metabolic versatility and the catalytic power of the bacterium as a potential biocatalyst for keratin-laden biomass degradation is yet to be fully explored.

Some advised future research aims are the followings:

- Establish the longevity of strain T cells in different substrates. Especially, determine if the organism is still physiologically active during feather degradation, with particular attention to waste product accumulation and/or limiting substrate factors.
- Identify the genomic parts, and coding sequences, that were absent in *F. pennivorans* DSM9078 and *F. islandicum* AW-1, but were present in strain T as well as the one present in the two relative organisms but absent in the query organism.
- Obtain the total RNA for transcriptomics analysis to further detect active keratin-degrading protease genes during feather degradation process also in light of the ones predicted by gene mining analysis.
- Optimize up-scaled experiments in the bioreactor for eventual bioindustrial application of the organism.
- Perform crystallization and 3 dimensional structural analysis of Peg_1025 by x-ray diffraction.
- Continue with the expression of C-tagged construct for serine protease Peg_1025 and proceed with its purification.
- Further optimize activity tests to exploit the best activity of the protease.
- Determine the catalytic power of the protease using defined chromogenic peptides as in Toogood et al. (2000).
- Express other protease genes that may have keratinolytic potential.

Acknowledgements

I thought that after 100 and more pages of formalities, I could use a couple extra pages to write freely about the Master Internship that I carried out in the laboratories of the Extremophiles and Biotechnology Group at the Department of Biology of the University of Bergen (UiB), Bergen, Norway. Here it is just a brief description of the atmosphere that dragged me in, almost every day, from August 2019 to June 2020.

If you had to randomly walk in the lab at any time of the day, you would meet Akzhigit. Akzhigit is from Kazakhstan and no matter the day, or the time, he would *always* greet you with a big smile, a hand shake and a "como estas". Thank you for making up everyone's day. Then, you would usually meet Dr. Birte Tøpper, the responsible for lab safety and materials, among other things. Birte is originally from Germany and she has the uncomfortable position to have to tell people, with very different cultural backgrounds, to respect the safety rules. Not easy task. Thank you for your patience and your help. At this point, Ruben comes in. Ruben is from Spain and he loves to talk about good food, drinks, sun and quality times. Although some of these things are especially hard to find in Bergen, he is always cheerful and content. He can always make you feel appreciated, always has time for helping or discussing things. He was always there for me, in the good and in the bad times. He was a support and a friend and I owe him a lot. Thank you.

By 10 a.m., the lab would be filled with the "exchange" students. Srisuda and Azan from Thailand, Munnavara from Tajikistan and Natia from Georgia. I have never really understood what Srisuda and Azan worked on, but they always seemed to handle small flasks from where lots of syringes and needles would stuck out from. I thought those were their own version of Voodoo dolls so I always treated them with respect and they always gave me a good laugh… Fortunately. Thank you. Munnavara is usually the name that you can hear people calling alound if they had encounter unexpected problems of any kind. Nonetheless, she would always prepare delicious and very abundant food, making lunches and brakes though. Thank you for your hospitality and good mood. As I mentioned, Natia is from Georgia and she is another particular character. At first sight, she is reserved but yet straightforward on her opinions. Then, you really start to appreciate her comments and her company. She always has unusual incredible adventures to tell and a great vitality ready to boost your day. Thank you. Another fellow who is working hard is Chandini. Chandini, from India, is the person everyone goes to if they need a protocol or scientific advice. She is always busy, but always happy to give you her opinion and expertise on anything. She is very enthusiastic and great to talk to, even for short pauses between lab works. Thank you so much. Just to add a couple of extra personalities to the pool, you would meet Dr. Antonio Garcìa Moyano, from Spain and Dr. Thomas Kruse, from the Netherlands. Thomas is a natural when it comes to details and neat things. You would not meet him often in the lab, but you if you needed his help you would always find him. Thank you. Regarding Antonio, everyone have always considered him one of the cornerstone of the group, and I can only agree with them. He gave me many important advices and his feedbacks were essential in more than one occasion. He always cares about everyone, not just in the lab, but also in the everyday life. Always trying to create an atmosphere where a bunch of friends happened to work together, rather than the opposite. This is rare. Thank you so much.

Nothing of what just said could have happened if at the top of it there had not been a person whose values and ground believes were of curiosity, goodwill and collaboration. The person in subject is Professor Nils-Kåre Birkeland and he is the ground fonder of the group and of everyone’s job. He offered me invaluable guidance throughout my entire staying and always fully supported all my work. Learning from him has been inspiring and fascinating, from many different perspectives. Therefore, my most sincere and deepest gratitude are given to him, for this opportunity, for his help and for his time, especially during the writing process.

Finally, I would like to thank all the people that I have met in this period and that inspired me not just in being a better student, but a better human being; my parents and my family for their support and all my friends, far away or nearby, for their essential and precious existence. I would like also to thank you, who have read the whole thesis or just a part of it: that this project may have encouraged your curiosity towards this fascinating and yet little appreciated universe that is called microbiology.

References

Adams, M. W. W., Perler, F. B., & Kelly, R. M. (1995). Extremozymes : Expanding the Limits of Biocatalysis. *Biotechnology*, *13*, 662–668.

Aguilar, A., Ingemansson, T., & Magnien, E. (1998). Extremophile microorganisms as cell factories: support from the European Union. *Extremophiles*, *2*(3), 367–373. https://doi.org/10.1007/s007920050080

Aguilar, C. F., Sanderson, I., Moracci, M., Ciaramella, M., Nucci, R., Rossi, M., & Pearl, L. H. (1997). Crystal structure of the β-glycosidase from the hyperthermophilic archeon sulfolobus solfataricus: Resilience as a key factor in thermostability. *Journal of Molecular Biology*, *271*(5), 789–802. https://doi.org/10.1006/jmbi.1997.1215

Alikhan, N. (2011). BRIG 0.95 Manual. Retrieved from http://sourceforge.net/projects/brig/

Andersen, K. R., Leksa, N. C., & Schwartz, T. U. (2013). Optimized *E. coli* expression strain LOBSTR eliminates common contaminants from His-tag purification. *Proteins: Structure, Function and Bioinformatics*, *81*(11), 1857–1861. https://doi.org/10.1002/prot.24364

Andrews, K. T., & Patel, K. C. B. (1996). *Fervidobacterium gondwanense* sp. nov., a New Thermophilic Anaerobic Bacterium Isolated from Nonvolcanically Heated Geothermal Waters of the Great Artesian Basin of Australia. *International Journal of Systematic Bacteriology*, *46*(1), 265–269. https://doi.org/10.1099/00207713-46-1-265

Antranikian, G., Vorgias, C. E., & Bertoldo, C. (2005). Extreme Environments as a Resource for Microorganisms and Novel Biocatalysts. In R. Ulber & Y. Le Gal (Eds.), *Marine Biotechnology I* (pp. 219–262). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/b135786

Aziz, R. K., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., & Zagnitko, O. (2008). The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics*, *9*, 1–15.

https://doi.org/10.1186/1471-2164-9-75

Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics (Oxford, England)*, *27*(3), 343–350. https://doi.org/10.1093/bioinformatics/btq662

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific Reports*, *7*(1), 10480. https://doi.org/10.1038/s41598-017-09654-8

Bhandari, V., & Gupta, R. S. (2014). The Phylum Thermotogae BT - The Prokaryotes: Other Major Lineages of Bacteria and The Archaea. In Eugene Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.) (pp. 989–1015). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38954-2_118

Bienert, S., Waterhouse, A., de Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2016). The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, *45*(D1), D313–D319. https://doi.org/10.1093/nar/gkw1132

Bjerga, G. E. K., Arsin, H., Larsen, Ø., Puntervoll, P., & Kleivdal, H. T. (2016). A rapid solubility-optimized screening procedure for recombinant subtilisins in *E. coli. Journal of Biotechnology*, *222*(February 2007), 38–46. https://doi.org/10.1016/j.jbiotec.2016.02.009

Böckle, B., & Müller, R. (1997). Reduction of disulfide bonds by *Streptomyces pactum* during growth on chicken feathers. *Applied and Environmental Microbiology*, *63*(2), 790–792. https://doi.org/10.1128/aem.63.2.790-792.1997

Bonete, M. J., & Martines-Espinosa, R. M. (2011). Enzymes from halophilic archaea: open questions. In A. Ventosa, A. Oren, & Y. Ma (Eds.), *Halophiles and hypersaline environments*. Springer Berlin Heidelberg.

Cai, J., Wang, Y., Liu, D., Zeng, Y., Xue, Y., Ma, Y., & Feng, Y. (2007). *Fervidobacterium changbaicum* sp. nov., a novel thermophilic anaerobic bacterium isolated from a hot spring of the Changbai Mountains, China. *International Journal of Systematic and Evolutionary Microbiology*, *57*(10), 2333–2336. https://doi.org/10.1099/ijs.0.64758-0

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Canganella, F., & Wiegel, J. (2011). Extremophiles: From abyssal to terrestrial ecosystems and possibly beyond. *Naturwissenschaften*, *98*(4), 253–279. https://doi.org/10.1007/s00114-011-0775-2

Cavaille, D., & Combes, D. (1995). Characterization of beta-galactosidase from *Kluyveromyces lactis*. *Biotechnology and Applied Biochemistry*.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569. https://doi.org/10.1038/nmeth.2474

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), D67–D72. https://doi.org/10.1093/nar/gkv1276

Connon, S. A., Lester, E. D., Shafaat, H. S., Obenhuber, D. C., & Ponce, A. (2007). Bacterial diversity in hyperarid atacama desert soils. *Journal of Geophysical Research: Biogeosciences*, *112*(4), 1–9. https://doi.org/10.1029/2006JG000311

Cowan, D. A., Ramond, J. B., Makhalanyane, T. P., & De Maayer, P. (2015). Metagenomics of extreme environments. *Current Opinion in Microbiology*, *25*, 97–102. https://doi.org/10.1016/j.mib.2015.05.005

Cuecas, A., Kanoksilapatham, W., & Gonzalez, J. M. (2017). Evidence of horizontal gene transfer by transposase gene analyses in *Fervidobacterium* species. *PLoS ONE*, *12*(4), 1–21. https://doi.org/10.1371/journal.pone.0173961

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements. *Genome Research*, *14*, 1394–1403. https://doi.org/10.1101/gr.2289704.tion

Edbeib, M. F., Wahab, R. A., & Huyop, F. (2016). Halophiles: biology, adaptation, and their role in decontamination of hypersaline environments. *World Journal of Microbiology and Biotechnology*, *32*(8), 135. https://doi.org/10.1007/s11274-016-2081-9

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Elleuche, S., Schäfers, C., Blank, S., Schröder, C., & Antranikian, G. (2015). Exploration of extremophiles for high temperature biotechnological processes. *Current Opinion in Microbiology*, *25*, 113–119. https://doi.org/10.1016/j.mib.2015.05.011

Elleuche, S., Schröder, C., Sahm, K., & Antranikian, G. (2014). Extremozymes-biocatalysts with unique properties from extremophilic microorganisms. *Current Opinion in Biotechnology*, *29*(1), 116–123. https://doi.org/10.1016/j.copbio.2014.04.003

Ellman, G. L. (1959). Tissue sulfhydryl groups. *Archives of Biochemistry and Biophysics*, *82*(1), 70–77. https://doi.org/10.1016/0003-9861(59)90090-6

Feller, G., & Gerday, C. (2003). Psychrophilic enzymes: Hot topics in cold adaptation. *Nature Reviews Microbiology*, *1*(3), 200–208. https://doi.org/10.1038/nrmicro773

Freeman, S., & Herron, J. (2006). *Evolutionary analysis*. (Upper Saddle River, Ed.) (Fifth edit). NJ: Pearson Prentice Hall.

Friedricht, A. B., & Antranikian, G. (1996). Keratin degradation by *Fervidobacterium pennavorans*, a novel thermophilic anaerobic species of the order thermotogales. *Applied and Environmental Microbiology*, *62*(8), 2875–2882.

Frock, A. D., Notey, J. S., & Kelly, R. M. (2010). The genus Thermotoga: Recent developments. *Environmental Technology*, *31*(10), 1169–1181. https://doi.org/10.1080/09593330.2010.484076

Gabani, P., Prakash, D., & V. Singh, O. (2014). Bio-signature of Ultraviolet-Radiation-Resistant Extremophiles from Elevated Land. *American Journal of Microbiological Research*, *2*(3), 94–104. https://doi.org/10.12691/ajmr-2-3-3

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). The Proteomics Protocols Handbook. *The Proteomics Protocols Handbook*, 571–608. https://doi.org/10.1385/1592598900

Geertsma, E. R., & Dutzler, R. (2011). A versatile and efficient high-throughput cloning tool for structural biology. *Biochemistry*, *50*(15), 3272–3278. https://doi.org/10.1021/bi200178z

Gerday, C., & Glansdorff, N. (2007). *Physiology and Biochemistry of Extremophiles*. *Physiology and Biochemistry of Extremophiles*. ASM Press. https://doi.org/10.1128/9781555815813.ch27

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(1), 81–91. https://doi.org/10.1099/ijs.0.64483-0

Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *ELECTROPHORESIS*, *30*(S1), S162–S173. https://doi.org/10.1002/elps.200900140

Gupta, K., Jana, A. K., Kumar, S., & Maiti, M. (2013). Immobilization of α-amylase and amyloglucosidase onto ion-exchange resin beads and hydrolysis of natural starch at high concentration. *Bioprocess and Biosystems Engineering*, *36*(11), 1715-1724.

Hegde, S., & Kaltenegger, L. (2012). Colors of Extreme Exo-Earth Environments. *Astrobiology*, *13*. https://doi.org/10.1089/ast.2012.0849

Herbert, R. A. (1992). A Perspective on the Biotechnological Potential of extremophiles. *TibTech*, *10*, 395–402. https://doi.org/10.1080/10408410802086783

Hough, D. W., & Danson, M. J. (1999). Extremozymes. *Current Opinion in Chemical Biology*, *3*(1), 39–46. https://doi.org/10.1016/S1367-5931(99)80008-8

Huber, R., Langworthy, T. A., König, H., Thomm, M., Woese, C. R., Sleytr, U. B., & Stetter, K. O. (1986). *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. *Archives of Microbiology*, *144*(4), 324–333. https://doi.org/10.1007/BF00409880

Huber, R., Woese, C. R., Langworthy, T. A., Kristjansson, J. K., & Stetter, K. O. (1990). *Fervidobacterium islandicum* sp. nov., a new extremely thermophilic eubacterium belonging to the “Thermotogales.” *Archives of Microbiology*, *154*(2), 105–111. https://doi.org/10.1007/BF00423318

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., Mcanulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., & Yeats, C. (2009). InterPro: The integrative protein signature database. *Nucleic Acids Research*, *37*(SUPPL. 1), 211–215. https://doi.org/10.1093/nar/gkn785

Ishino, S., & Ishino, Y. (2014). DNA polymerases as useful reagents for biotechnology - The history of developmental research in the field. *Frontiers in Microbiology*, *5*(AUG), 1–8. https://doi.org/10.3389/fmicb.2014.00465

Javier-López, R. (2018). Isolation and characterization of a new keratinolytic *Fervidobacterium pennivorans* strain from a hot spring in Tajikistan. University of Bergen.

Kang, E., Jin, H. S., La, J. W., Sung, J. Y., Park, S. Y., Kim, W. C., & Lee, D. W. (2019). Identification of keratinases from *Fervidobacterium islandicum* AW-1 using dynamic gene expression profiling. *Microbial Biotechnology*. https://doi.org/10.1111/1751-7915.13493

Kanoksilapatham, W., Pasomsup, P., Keawram, P., Cuecas, A., Portillo, M. C., & Gonzalez, J. M. (2016). *Fervidobacterium thailandense* sp. Nov., an extremely thermophilic bacterium isolated from a hot spring. *International Journal of Systematic and Evolutionary Microbiology*, *66*(12), 5023–5027. https://doi.org/10.1099/ijsem.0.001463

Kim, J. S., Kluskens, L. D., De Vos, W. M., Huber, R., & Van Der Oost, J. (2004). Crystal structure of fervidolysin from *Fervidobacterium pennivorans*, a keratinolytic enzyme related to subtilisin. *Journal of Molecular Biology*, *335*(3), 787–797. https://doi.org/10.1016/j.jmb.2003.11.006

Kirk, O., Borchert, T. V., & Fuglsang, C. C. (2002). Industrial enzyme applications. *Current Opinion in Biotechnology*, *13*(4), 345–351. https://doi.org/10.1016/S0958-1669(02)00328-2

Klumpp, S., Zhang, Z., & Hwa, T. (2009). Growth Rate-Dependent Global Effects on Gene Expression in Bacteria. *Cell*, *139*(7), 1366–1375. https://doi.org/10.1016/j.cell.2009.12.001

Kluskens, L. D., Voorhorst, W. G. B., Siezen, R. J., Schwerdtfeger, R. M., Antranikian, G., Van Der Oost, J., & De Vos, W. M. (2002). Molecular characterization of fervidolysin, a subtilisin-like serine protease from the thermophilic bacterium *Fervidobacterium pennivorans*. *Extremophiles*, *6*(3), 185–194. https://doi.org/10.1007/s007920100239

Koga, Y. (2012). Thermal adaptation of the archaeal and bacterial lipid membranes. *Archaea*. https://doi.org/10.1155/2012/789652

Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kumar, C. G., & Takagi, H. (1999). Microbial alkaline proteases: From a bioindustrial viewpoint. *Biotechnology Advances*, *17*(7), 561–594. https://doi.org/10.1016/S0734-9750(99)00027-0

Lange, L., Huang, Y., & Busk, P. K. (2016). Microbial decomposition of keratin in nature—a new hypothesis of industrial relevance. *Applied Microbiology and Biotechnology*, *100*(5), 2083–2096. https://doi.org/10.1007/s00253-015-7262-1

Latiffi, A. A., Salleh, A. B., Rahman, R. N. Z. R. A., Nurbaya Oslan, S., & Basri, M. (2013). Secretory expression of thermostable alkaline protease from *Bacillus stearothermophilus* FI by using native signal peptide and α-factor secretion signal in Pichia pastoris. *Genes and Genetic Systems*, *88*(2), 85–91. https://doi.org/10.1266/ggs.88.85

Lee, S.-J., Lee, D.-W., Park, G.-S., Shin, J.-H., Kim, J.-Y., Park, M.-K., Jeong, H., Lee, Y.-J., Kwak, Y., Lee, S. J., & Kang, H. K. (2015). Genome sequence of a native-feather degrading extremely thermophilic Eubacterium, *Fervidobacterium islandicum* AW-1. *Standards in Genomic Sciences*, *10*(1), 1–9. https://doi.org/10.1186/s40793-015-0063-4

Lee, Y.-J., Ahn, J.-S., Jin, H.-S., Lee, D.-W., Dhanasingh, I., Lee, S. H., & Choi, J. M. (2015). Biochemical and structural characterization of a keratin-degrading M32 carboxypeptidase from *Fervidobacterium islandicum* AW-1. *Biochemical and Biophysical Research Communications*, *468*(4), 927–933. https://doi.org/10.1016/j.bbrc.2015.11.058

Littlechild, J. A. (2015). Enzymes from extreme environments and their industrial applications. *Frontiers in Bioengineering and Biotechnology*, *3*(OCT), 1–9.

https://doi.org/10.3389/fbioe.2015.00161

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, *13*(5), 660–665. https://doi.org/10.1093/oxfordjournals.molbev.a025626

Lorenz, P., & Eck, J. (2005). Metagenomics and industrial applications. *Nature Reviews Microbiology*, *3*(6), 510–516. https://doi.org/10.1038/nrmicro1161

Macy, J. M., Snellen, J. E., & Hungate, R. E. (1972). Use of syringe methods for anaerobiosis. *The American Journal of Clinical Nutrition*, *25*(12), 1318–1323. https://doi.org/10.1093/ajcn/25.12.1318

Madden, T. (2002). Chapter 16 : The BLAST Sequence Analysis Tool. *The NCBI Handbook[Internet]*, (Md), 1–15.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636—W641. https://doi.org/10.1093/nar/gkz268

Madigan, M. T., Bender, Buckley, Sattley, & Stahl. (2019). *Brock Biology of Microorganisms* (Fifteenth). Pearson.

Manual. (2019). *Scaffold User’s Manual* (Release 4.). Proteome Software, Inc.

McDonough, K. (2013). All about feathers. In *All about bird biology*. Ithaca, New york: Corneli Lab of Ornithology.

Mooij, W. T. M., Mitsiki, E., & Perrakis, A. (2009). ProteinCCD: Enabling the design of protein truncation constructs for expression and crystallization experiments. *Nucleic Acids Research*, *37*(SUPPL. 2), 402–405. https://doi.org/10.1093/nar/gkp256

Nam, G. W., Lee, D. W., Lee, H. S., Lee, N. J., Kim, B. C., Choe, E. A., Hwang, J. K., Suhartono, M. T., & Pyun, Y. R. (2002). Native-feather degradation by *Fervidobacterium islandicum* AW-1, a newly isolated keratinase-producing thermophilic anaerobe. *Archives of Microbiology*, *178*(6), 538–547. https://doi.org/10.1007/s00203-002-0489-0

Necşulea, A., & Lobry, J. R. (2007). A new method for assessing the effect of replication on DNA base composition asymmetry. *Molecular Biology and Evolution*, *24*(10), 2169–2179. https://doi.org/10.1093/molbev/msm148

Niehaus, A. F., Bertoldo, C., Kahler, M., & Antranikian, G. (1999). Extremophiles as a source of novel enzymes for industrial application. *Applied Microbial Biotechnology*, *51*, 711–729.

Nielsen, H., Tsirigos, K. D., Brunak, S., & von Heijne, G. (2019). A Brief History of Protein Sorting Prediction. *Protein Journal*, *38*(3), 200–216. https://doi.org/10.1007/s10930-019-09838-3

Orellana, R., Macaya, C., Bravo, G., Dorochesi, F., Cumsille, A., Valencia, R., Rojas, C., & Seeger, M. (2018). Living at the Frontiers of Life: Extremophiles in Chile and Their Potential for Bioremediation. *Frontiers in Microbiology*, *9*(October), 1–25. https://doi.org/10.3389/fmicb.2018.02309

Papadopoulos, M. C. (1989). Effect of processing on high-protein feedstuffs: A review. *Biological Wastes*, *29*(2), 123–138. https://doi.org/10.1016/0269-7483(89)90092-X

Parry, D. A. D., Crewther, W. G., Fraser, R. D. B., & MacRae, T. P. (1977). Structure of α-keratin: Structural implication of the amino acid sequences of the type I and type II chain segments. *Journal of Molecular Biology*, *113*(2), 449–454. https://doi.org/10.1016/0022-2836(77)90153-X

Patel, B. K. C., Morgan, H. W., Daniel, R. M., & Zealand, N. (1985). *Fervidobacterium nodosum* gen. nov. and spec. nov., a new chemoorganotrophic, caldoactive, anaerobic bacterium. *Archives of Microbiology*, *141*, 63–69.

Peek, K., Veitch, D. P., Prescott, M., Daniel, R. M., MacIver, B., & Bergquist, P. L. (1993). Some characteristics of a proteinase from a thermophilic Bacillus sp. expressed in *Escherichia coli*: Comparison with the native enzyme and its processing in *E. coli* and in vitro. *Applied and Environmental Microbiology*, *59*(4), 1168–1175. https://doi.org/10.1128/aem.59.4.1168-1175.1993

Podosokorskaya, O. A., Chernyh, N. A., Merkel, A. Y., Kolganova, T. V., Bonch-Osmolovskaya, E. A., Miroshnichenko, M. L., & Kublanov, I. V. (2011). *Fervidobacterium riparium* sp. nov., a thermophilic anaerobic cellulolytic bacterium isolated from a hot spring. *International Journal of Systematic and Evolutionary Microbiology*, *61*(11), 2697–2701. https://doi.org/10.1099/ijs.0.026070-0

Ramnani, P., & Gupta, R. (2007). Keratinases vis-à-vis conventional proteases and feather degradation. *World Journal of Microbiology and Biotechnology*, *23*(11), 1537–1540. https://doi.org/10.1007/s11274-007-9398-3

Rampelotto, P. H. (2013). Extremophiles and extreme environments. *Life*, *3*(3), 482–485. https://doi.org/10.3390/life3030482

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*, *46*(D1), D624–D632. https://doi.org/10.1093/nar/gkx1134

Reed, C. J., Lewis, H., Trejo, E., Winston, V., & Evilia, C. (2013). Protein adaptations in archaeal extremophiles. *Archaea*, *2013*. https://doi.org/10.1155/2013/373275

Research and Markets. (2020). *Industrial Enzymes Market - Growth, trends, and forecasta (2020 - 2025)*.

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). Bypassing cultivation to identify bacterial species. *Microbe*, *9*(3), 111–118.

Rosenberg, E., Delong, E. F., Lory, S., Stackebrandt, E., & Thompson, F. (2014). *The Prokaryotes_ Major lineages of Bacteria and Archaea*. *The Prokaryotes: Other Major Lineages of Bacteria and The Archaea* (Vol. 9783642389). https://doi.org/10.1007/978-3-642-38954-2_329

Rosenberg, Eugene, DeLong, E. F., Lory, S., Stackebrandt, E., & Thompson, F. (2013). *The Prokaryotes*. *Prokaryotic Biology and Symbiotic Associations* (Vol. 4th Editio). https://doi.org/10.1007/978-3-642-30141-4

Rutheford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics Applications Note*, *16*(10), 944–945.

Schmid, A., Dordick, J. S., Hauer, B., Kiener, A., Wubbolts, M., & Witholt, B. (2001). Industrial biocatalysis today and tomorro. *Nature*, *409*. https://doi.org/10.1016/0300-483x(87)90052-7

Searle, B. C. (2010). Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, *10*(6), 1265–1269. https://doi.org/10.1002/pmic.200900437

Shavandi, A., Silva, T. H., & Bekhit, A. E. A. (2017). Keratin: dissolution, exraction and biomedical application. *Biomaterials Science Application*, *5*, 1699–1735. https://doi.org/10.1039/c7bm00411g

Shinde, U., & Inouye, M. (1993). Intramolecular chaperones and protein folding. *Trends in Biochemical Sciences*, *18*(11), 442–446. https://doi.org/10.1016/0968-0004(93)90146-E

Shinde, U., & Inouye, M. (2000). Intramolecular chaperones: Polypeptide extensions that modulate protein folding. *Seminars in Cell and Developmental Biology*, *11*(1), 35–44. https://doi.org/10.1006/scdb.1999.0349

Siezen, R. J., de Vos, W. M., Leunissen, J. A. M., & Dijkstra, B. W. (1991). Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases. *Protein Engineering, Design and Selection*, *4*(7), 719–737. https://doi.org/10.1093/protein/4.7.719

Siezen, R. J., & Leunissen, J. A. M. (2008). Subtilases: The superfamily of subtilisin-like serine proteases. *Protein Science*, *6*(3), 501–523. https://doi.org/10.1002/pro.5560060301

Siliakus, M. F., van der Oost, J., & Kengen, S. W. M. (2017). Adaptations of archaeal and bacterial membranes to variations in temperature, pH and pressure. *Extremophiles*, *21*(4), 651–670. https://doi.org/10.1007/s00792-017-0939-x

Smith, C. A., Toogood, H. S., Baker, H. M., Daniel, R. M., & Baker, E. N. (1999). Calcium-mediated thermostability in the subtilisin superfamily: The crystal structure of *Bacillus* Ak.1 protease at 1.8 Å resolution. *Journal of Molecular Biology*, *294*(4), 1027–1040. https://doi.org/10.1006/jmbi.1999.3291

Sonnhammer, E. L. L., & Krogh, A. (2008). A hidden Markov model for predicting transmembrane helices in protein sequence. *Sixth Int. Conf. on Intelligent Systems for Molecular Biology*, 8. Retrieved from papers://4b986d00-906f-493f-a74b-71e29d82b719/Paper/p6291

Sterner, R., & Liebl, W. (2001). Thermophilic adaptation of proteins. *Critical Reviews in Biochemistry and Molecular Biology*, *36*(1), 39–106. https://doi.org/10.1080/20014091074174

Stetter, K. O. (1996). Hyperthermophilic procaryotes. *FEMS Microbiology Reviews*, *18*(2–3), 149–158. https://doi.org/10.1016/0168-6445(96)00008-3

Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2019). QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics*, *36*(6), 1765–1771. https://doi.org/10.1093/bioinformatics/btz828

Suzuki, Y., Tsujimoto, Y., Matsui, H., & Watanabe, K. (2006). Decomposition of extremely hard-to-degrade animal proteins by thermophilic bacteria. *Journal of Bioscience and Bioengineering*, *102*(2), 73–81. https://doi.org/10.1263/jbb.102.73

Takai, K., Nakamura, K., Toki, T., Tsunogai, U., Miyazaki, M., Miyazaki, J., Hirayama, H., Nakagawa, S., Nunoura, T., & Horikoshi, K. (2008). Cell proliferation at 122°C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10949–10954. https://doi.org/10.1073/pnas.0712334105

Thompson, M. (2014). Everything You Need To Know About Feathers. Retrieved from https://academy.allaboutbirds.org/feathers-article/

Toogood, H. S., Smith, C. A., Baker, E. N., & Daniel, R. M. (2000). Purification and characterization of Ak.1 protease, a thermostable subtilisin with a disulphide bond in the substrate-binding cleft. *Biochemical Journal*, *350*(1), 321–328. https://doi.org/10.1042/0264-6021:3500321

Twining, S. S. (1984). Fluorescein isothiocyanate-labeled casein assay for proteolytic enzymes. *Analytical Biochemistry*, *143*(1), 30–34. https://doi.org/10.1016/0003-2697(84)90553-0

Van den Burg, B. (2003). Extremophiles as a source for novel enzymes. *Current Opinion in Microbiology*, *6*(3), 213–218. https://doi.org/10.1016/S1369-5274(03)00060-2

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Williams, C. M., Lee, C. G., Garlich, J. D., & Shih, J. (1990). Evaluation of a bacterial feather fermentation product, feather-lysate, as a feed protein. *Poultry Science*, *70*(1), 85–94.

Yayanos, A., Dietz, A., & Van Boxtel, R. (1982). Dependence of reproduction rate on pressure as a hallmark of deep-sea Bacteria. *Applied and Environmental Microbiology*, *44*(6), 1356–1361.

Supplementary materials

S 1 The difference between false positive rate and false discovery rate is in the denominator of the corresponding formulas:

$$FPR = \frac{FP}{FP + TN}$$

$$FDR = \frac{FP}{FP + TP}$$

where FP are false positive counts (incorrect proteins but defined as correct ones) and TP are the true positive counts (truly correct proteins). TN are true negative counts (truly incorrect proteins). In Scaffold, false positive counts are found using the decoy method and the program assumes that there is an equal number of decoy proteins as incorrect proteins from the original search.

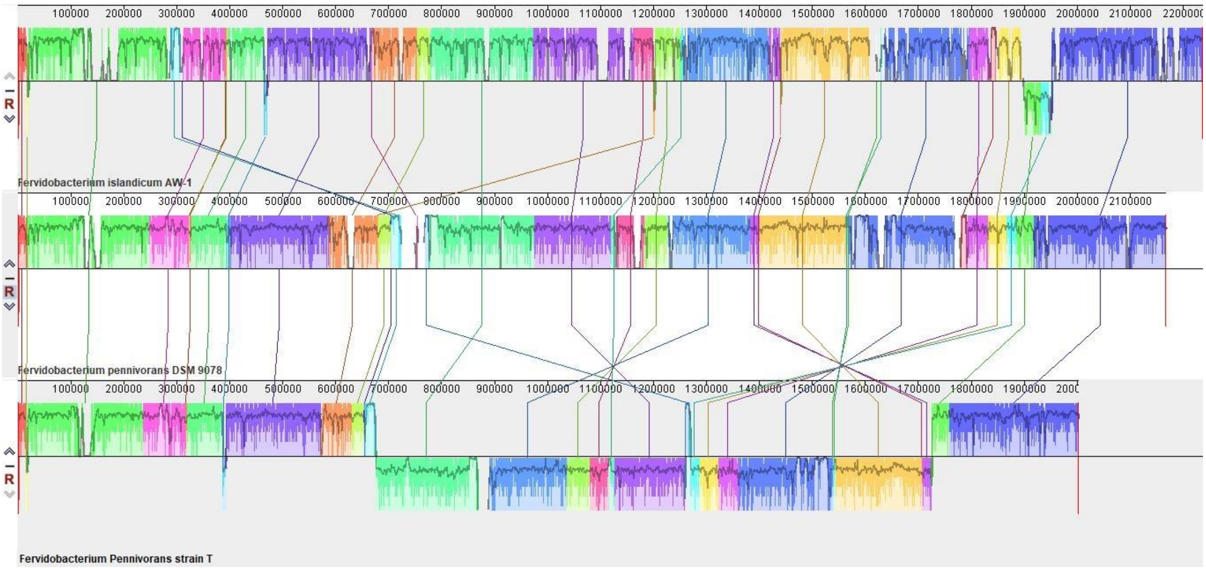
S 2 FASTA sequence of serine protease Peg_1025 isolated from *Fervidobacterium pennivorans* strain T. Signal peptide is highlighted in yellow, hypothetical propeptide is highlighted in green and catalytic triad residues highlighted in red.

>peg.1025 serine protease [*Fervidobacterium pennivorans* strain T]

```
MKKFVLLTAVFALLLVTFSCTNSLEPRFEPRAQGEFEVSEKLGVSGTEEDYVPGEY
VVQFEPREDAVKALSSVGAEVVRAYSFSDVQIVTVRTEKPELLNSLPGVKSVDKNYI
YRALATPNDTYYRYQWHYNNIKLPQAWDIMKSANIVVAVIDTGVSFTHPDLQGIFVQ
GYDFVDGDYDPTDPAQDVSHGTHCIGTIAAVTNNSLGVAGVNWGGYGIKIMPIRVL
GADGSGTLDNVAAGIRWAVDNGAKIVSMSLGGSGAQVLMDAVKYAYSRNVTLICA
AGNESRPSLSYPAAYVETIAVGATRYDNTRARYSNYNYTRYYDPYRKAYVYHYLDV
VAPGGDTSVDQNGDGYADGVLSTTWTPTYGNTYMFLQGTSMATPHVAALAAMLY
AKGYTTPEAIRSRLIKTAYKIPGYTYNSSGWNKYVGYGLIDAYKALTY
```

S 3 Nucleotide sequence of the gene encoding for serine protease Peg_1025 isolated from *Fervidobacterium pennivorans* strain T. Signal peptide is highlighted in yellow and hypothetical propeptide is highlighted in green.

ATG AAG AAG TTT GTT TTA CTT ACA GCA GTT TTC GCA CTT TTG CTG GTA
ACATTCAGCTGTACCAACTCATTAGAGCCAAGATTTGAACCACGCGCACAAGGT
GAGTTTGAGGTTTCAGAGAAACTTGGCGTATCCGGAACAGAAGAAGATTACGTT
CCTGGAGAATATGTTGTCCAGTTCGAGCCAAGAGAAGATGCGGTTAAAGCATTA
TCAAGTGTTGGTGCAGAAGTAGTCAGAGCATATTCATTCAGCGATGTTCAAATC
GTAACAGTAAGAACGGAAAAACCAGAACTTCTTAATTCTCTTCCAGGTGTCAAG
TCAGTTGATAAGAACTACATTTACAGAGCACTCGCAACACCAAACGACACATAC
TACCGATACCAGTGGCACTACAACAATATCAAACTGCCACAGGCATGGGATATC
ATGAAATCTGCTAATATCGTTGTAGCAGTTATTGATACAGGAGTTAGCTTTACAC
ATCCAGACCTGCAAGGCATATTCGTTCAAGGCTATGACTTTGTCGATGGAGATT
ACGATCCGACAGACCCGGCACAGGATGTGAGCCATGGAACACATTGTATAGGA
ACAATAGCCGCTGTTACAAACAACAGCCTTGGTGTTGCCGGAGTTAATTGGGGA
GGATATGGAATAAAGATAATGCCTATCAGGGTTCTTGGCGCAGACGGTTCCGG
AACACTCGATAATGTCGCAGCTGGTATCAGATGGGCAGTTGACAACGGTGCAA
AAATAGTGAGTATGAGTCTTGGTGGTAGCGGTGCACAAGTTCTTATGGATGCCG
TTAAATATGCTTACAGCAGAAATGTAACACTTATCTGCGCAGCAGGAAATGAGA
GTAGACCTTCGCTATCCTATCCAGCAGCATATGTTGAAACGATCGCAGTAGGTG
CAACAAGATACGACAACACACGCGCTCGGTATTCTAACTACAATTACACAAGAT
ACTACGATCCTTACAGAAAAGCGTATGTATACCATTACCTTGACGTTGTTGCTCC
TGGTGGAGATACAAGTGTTGACCAAAACGGTGATGGATACGCAGATGGTGTGC
TCAGCACAACCTGGACACCGACATACGGAAATACATATATGTTCTTGCAAGGTA
CATCGATGGCAACACCACATGTTGCAGCGCTTGCAGCTATGCTTTACGCAAAAG
GTTACACAACACCAGAGGCGATTAGAAGCAGACTTATCAAAACAGCTTATAAGA
TTCCTGGATACACATATAATTCGAGCGGATGGAACAAATACGTTGGCTACGGTT
TAATTGATGCTTACAAGGCATTGACATACTAA



S 4 Multiple alignment between *F. islandicum* AW-1 (top), *F. pennivorans* DSM9078 (middle) and *F. pennivorans* strain T (bottom) complete genomes. Similarly conserved locally collinear blocks (LCBs) are coloured in the same way and connected by lines. Alignemt obtained by MAUVE.

S 5 Twenty nine intracellular proteases identified by gene mining in *Fervidobacterium pennivorans* strain T complete genome.

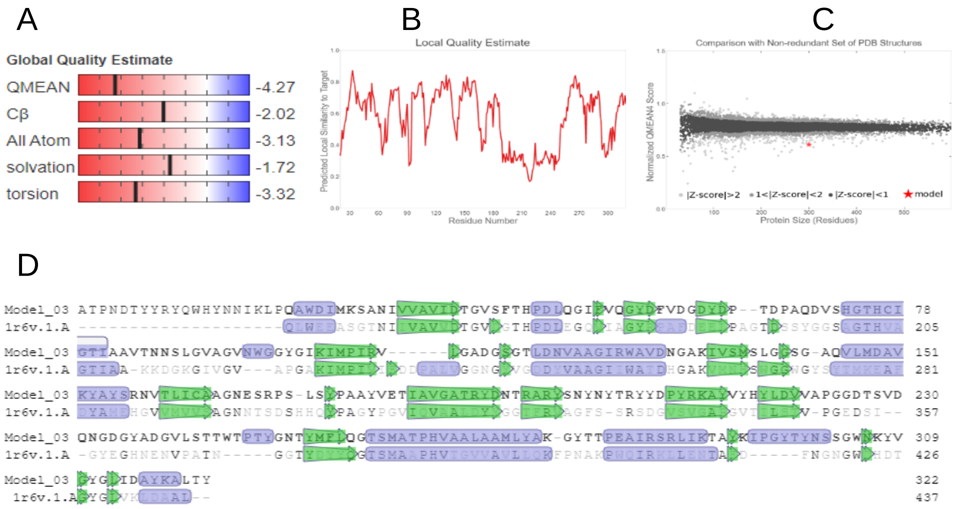
Locus (fig|93466.12._)	Cell. Location	RAST annotation
peg.161	Intracellular	Putative predicted metal-dependent hydrolase
peg.191	Intracellular	Peptidase, M16 family
peg.192	Intracellular	ATP-dependent hsl protease ATP-binding subunit HslU
peg.205	Intracellular	ATP-dependent protease La Type I
peg.209	Intracellular	Signal peptidase-like protein
peg.227	Intracellular	ATP-dependent Clp protease ATP-binding subunit ClpA
peg.425	Intracellular	ATP-dependent Clp protease ATP-binding subunit ClpA
peg.630	Intracellular	Aminopeptidase YpdF
peg.702	Intracellular	Deblocking aminopeptidase
peg.713	Intracellular	Thermostable carboxypeptidase 1
peg.796	Intracellular	Probable M18-family aminopeptidase 1
peg.800	Intracellular	Serine protease
peg.1065	Intracellular	Pyrrolidone-carboxylate peptidase
peg.1118	Intracellular	ATP-dependent Clp protease ATP-binding subunit ClpX
peg.1164	Intracellular	SOS-response repressor and protease LexA
peg.1200	Intracellular	ATP-dependent protease HslV
peg.1220	Intracellular	Aminopeptidase S
peg.1417	Intracellular	Asp-X dipeptidase
peg.1520	Intracellular	Oligoendopeptidase F
peg.1641	Intracellular	Methionine aminopeptidase

peg.1650	Intracellular	ATP-dependent Clp protease proteolytic subunit
peg.1690	Intracellular	Deblocking aminopeptidase
peg.1691	Intracellular	Deblocking aminopeptidase
peg.1788	Intracellular	Oligoendopeptidase F
peg.1790	Intracellular	TldE protein, part of TldE/TldD proteolytic complex
peg.1791	Intracellular	TldD protein, part of TldE/TldD proteolytic complex
peg.1818	Intracellular	Deblocking aminopeptidase
peg.1864	Intracellular	ATP-dependent protease La Type II
peg.1890	Intracellular	Transglutaminase-like enzymes, putative cysteine proteases

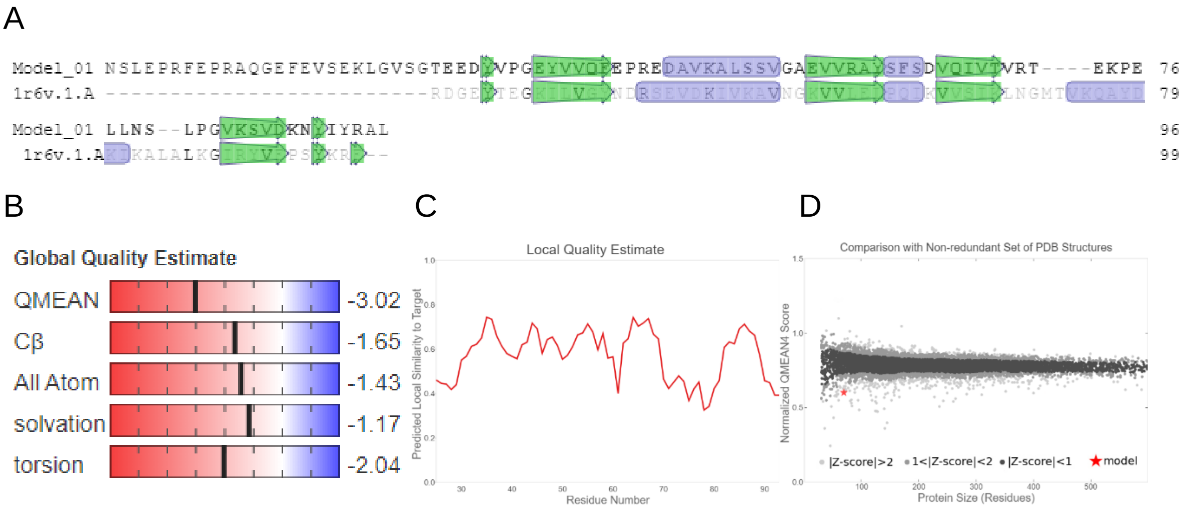
S 6 Twenty six proteins homologous to serine protease Peg_1025 which are reviewed in Uniprot database. All proteins belongs to the family S8. Proteins which a propeptide motif was reported are also marked.

Propeptide	Merops ID	Uniprot ID	Entry name	Protein name	Organism
Yes	S08.009	Q45670	THES_BACSJ	Thermophilic serine proteinase (Ak.1 protease)	*Bacillus sp. (strain AK1)*
Yes	S08.021	Q93LQ6	FLS_FERPE	Fervidolysin (Subtilisin-like serine protease)	*Fervidobacterium pennivorans*
Yes	S08.003	P29600	SUBS_BACLE	Subtilisin Savinase (Alkaline protease)	*Bacillus lentus*
Yes	S08.010	Q99405	PRTM_BACSK	M-protease	*Bacillus clausii (strain KSM-K16)*
Yes	S08.022	P42779	BPRV_DICNO	Extracellular basic protease	*Dichelobacter nodosus (Bacteroides nodosus)*
Yes	S08.036	P04189	SUBT_BACSU	Subtilisin E	*Bacillus subtilis (strain 168)*
Yes	S08.038	P27693	ELYA_BACAO	Alkaline protease	*Bacillus alcalophilus*
Yes	S08.050	P16588	PROA_VIBAL	Alkaline serine exoprotease A	*Vibrio alginolyticus*
Yes	S08.051	P08594	AQL1_THEAQ	Aqualysin-1	*Thermus aquaticus*
Yes	S08.059	Q07596	NISP_LACLL	NisP	*Lactococcus lactis*
Yes	S08.129	P58502	TKSU_THEKO	Tk-subtilisin	*Thermococcus kodakarensis*
	S08.003	P41362	ELYA_BACCS	Alkaline protease	*Bacillus clausii*
	S08.003	P29599	SUBB_BACLE	Subtilisin BL (Alkaline protease)	*Bacillus lentus*
	S08.002	P07518	SUBT_BACPU	Subtilisin (Alkaline mesentericopeptidase)	*Bacillus pumilus*
	S08.007	P04072	THET_THEVU	Thermitase	*Thermoactinomyces vulgaris*
	S08.020	P15926	C5AP_STRPY	C5a peptidase (SCP)	*Streptococcus pyogenes*
	S08.030	P29140	ISP_BACCS	Intracellular alkaline protease	*Bacillus clausii*
	S08.030	P11018	ISP1_BACSU	Major intracellular serine protease (ISP-1)	*Bacillus subtilis (strain 168)*

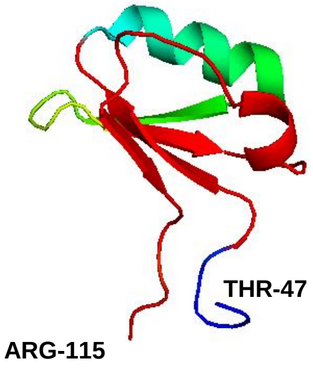
S08.030	P29139	ISP_PAEPO	Intracellular serine protease	*Paenibacillus polymyxa*
S08.037	P00781	SUBD_BACLI	Subtilisin DY	*Bacillus licheniformis*
S08.131	I6YC58	MYCP4_MYCTU	Mycosin-4 (MycP4 protease)	*Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)*
S08.131	O05458	MYCP2_MYCTU	Mycosin-4 (MycP4 protease)	*Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)*
S08.131	O05461	MYCP1_MYCTU	Mycosin-4 (MycP4 protease)	*Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)*
S08.131	O53695	MYCP3_MYCTU	Mycosin-4 (MycP4 protease)	*Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)*
S08.131	A0QNL1	MYCP1_MYCS2	Mycosin-4 (MycP4 protease)	*Mycolicibacterium smegmatis*
S08.140	P28842	SUBT_BACS9	Subtilisin	*Bacillus sp. (strain TA39)*



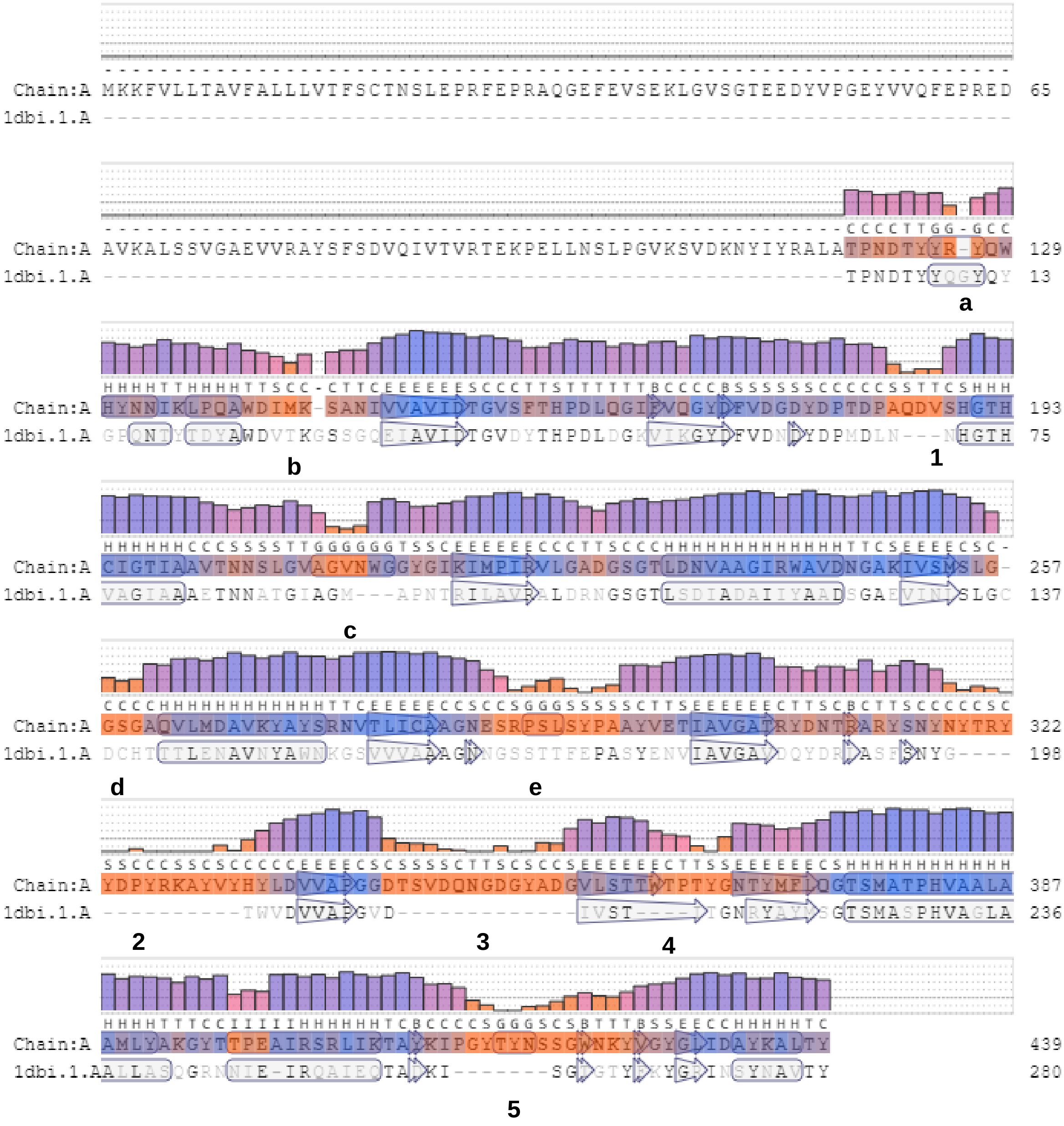
S 7 Overall structure quality estimates between the catalytic domain of Peg_1025 (AA 128 - 439) against fervidolysin as template obtained by SWISS-MODEL.



S 8 Quality estimates of the structural model for the propeptide domain (AA 22 – 96) of serine protease Peg_1025 obtained using fervidolysin as template by SWISS-MODEL. A) Structural alignment between serine protease (top line) and fervidolysin (bottom line); helices are highlighted in blue and beta sheets in green. B) Global quality estimates. C) Local quality estimate. D) Model fitting graph.

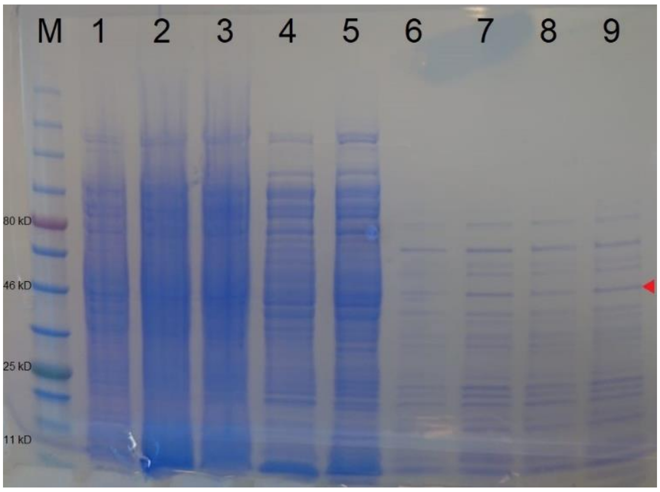


S 9 Three-dimensional model of the propeptide sequence using fervidolysin as a template. Hypothetical propeptide motifs are highlighted in red. The structure was predicted from AA 47 to 115 out of the entire sequence (AA 22 – 127).

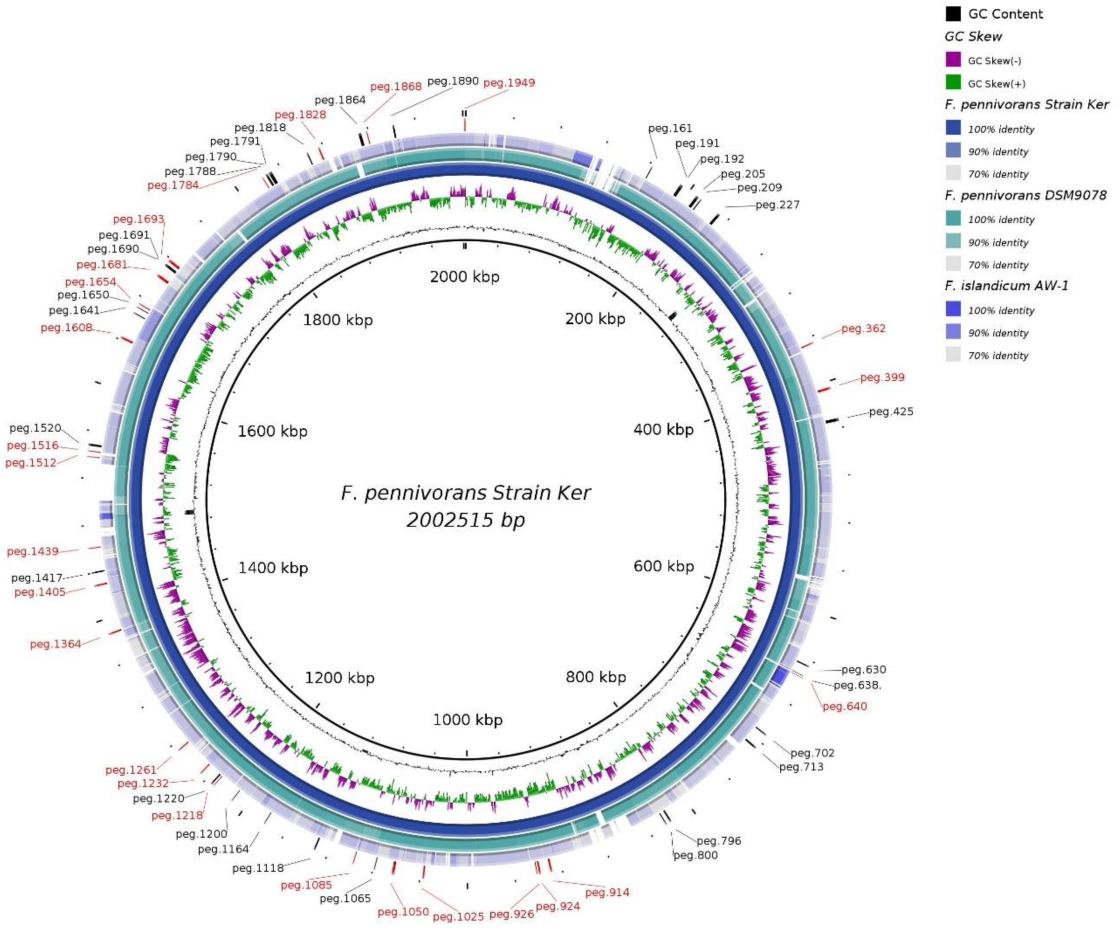


```
         -----------------------------------------------------------------
Chain:A  MKKFVLLTAVFALLLVTFSCTNSLEPRFEPRAQGEFEVSEKLGVSGTEEDYVPGEYVVQFEPRED  65
1dbi.1.A -----------------------------------------------------------------

         -----------------------------------------------------CCCCTTGG-GCC
Chain:A  AVKALSSVGAEVVRAYSFSDVQIVTVRTEKPELLNSLPGVKSVDKNYIYRALATPNDTYYR-YQW  129
1dbi.1.A -----------------------------------------------------TPNDTYYQGYQY  13

         HHHHTTHHHHTTSCC-CTTCEEEEEESCCCTTSTTTTTTBCCCCBSSSSSSCCCCCSSTTCSHHH
Chain:A  HYNNIKLPQAWDIMK-SANIVVAVIDTGVSFTHPDLQGIPVQGYDFVDGDYDPTDPAQDVSHGTH  193
1dbi.1.A GPQNTYTDYAWDVTKGSSGQEIAVIDTGVDYTHPDLDGKVIKGYDFVDNDYDPMDLN---NHGTH  75

         HHHHHHCCCSSSSSTTGGGGGGTSSCEEEEEEECCCTTSCCCHHHHHHHHHHHHHHTTCSEEEECSC-
Chain:A  CIGTIAAVTNNSLGVAGVNWGGYGIKIMPIRVLGADGSGTLDNVAAGIRWAVDNGAKIVSMSLG-  257
1dbi.1.A VAGIAAAETNNATGIAGM---APNTRILAVRALDRNGSGTLSDIADAIIYAADSGAEVINISLGC  137

         CCCCHHHHHHHHHHHHTTCEEEEECCSCCSGGGSSSSSCTTSEEEEEECTTSCBCTTSCCCCCSC
Chain:A  GSGAQVLMDAVKYAYSRNVTLICAAGNESRPSLSYPAAYVETIAVGATRYDNTRARYSNYNYTRY  322
1dbi.1.A DCHTTTLENAVNYAWNKGSVVVAAAGNNGSSTTFEPASYENVIAVGAVDQYDRIASFSNYG----  198

         SSCCCSSCSCCCCCEEEECSCSSSSCTTSCSCCSEEEEEECTTSSEEEEEECSHHHHHHHHHHHH
Chain:A  YDPYRKAYVYHYLDVVAPGGDTSVDQNGDGYADGVLSTTWTPTYGNTYMFLQGTSMATPHVAALA  387
1dbi.1.A ----------TWVDVVAPGVD-------------IVST----TIGNRYAYMSGTSMASPHVAGLA  236

         HHHHTTTCCIIIIIHHHHHHHTCBCCCCSGGGSCSBTTTBSSEECCHHHHHTC
Chain:A  AMLYAKGYTTPEAIRSRLIKTAYKIPGYTYNSSGWNKYVGYGLIDAYKALTY               439
1dbi.1.AALLASQGRNNIE-IRQAIEQTADKI-------SGTGTYPKYGRINSYNAVTY               280
```

S 10 SWISS-MODEL alignment details between serine protease Peg_1025 (top sequence) and subtilisin Ak-1 (bottom sequence). Histograms show the QMEAN values for each residue. Residues with QMEAN values below threshold (red) were hidden from the 3D structure of serine protease Peg_1025, either because very different from the template (black bold letter) or because completely absent, insertions (black bold numbers).



S 11 SDS-PAGE gel showing factions from *E. coli* LOBSTR expression of serine protease Peg_1025 from p7xN3H and p7xC3H vectors. Band corresponding to the target protein (48 kDa) is marked with red arrow. M, Broad rage marker; lane 1, uninduced cells from p7xN3H vector; lane 2 and 3, crude cell extract (N- and C-tagged, respectively); lane 4 and 5, soluble fraction (N- and C-tagged, respectively); lane 6 and 7, soluble fraction after heat shock treatment (70°C for 15 minutes) (N- and C-tagged, respectively); lane 8 and 9 soluble fraction after heat activation (70°C for 2 hours) (N- and C-tagged, respectively).



S 12 GC skew (second ring from inside, green and purple colours) in the complete genome *F. pennivorans* strain T. Figure constructed using BRIG platform. GC skew in this organism may be compared to the GC skew in *F. pennivorans* DSM9078.

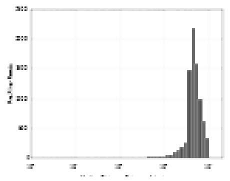
Reports for Job NGLIMSFILTER_81552

SMRT Cells: 1 Movies: 1

Overview

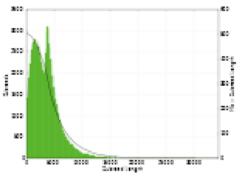
Job Metric	Value
Adapter Dimers (0-10bp)	0.03%
Short Inserts (11-100bp)	0.01%
Number of Bases	510,908,023
Number of Reads	94,920
N50 Read Length	7,998
Mean Read Length	5,382
Mean Read Score	0.84

Adapters



Observed Insert Length Distribution Histogram

Subread Filtering



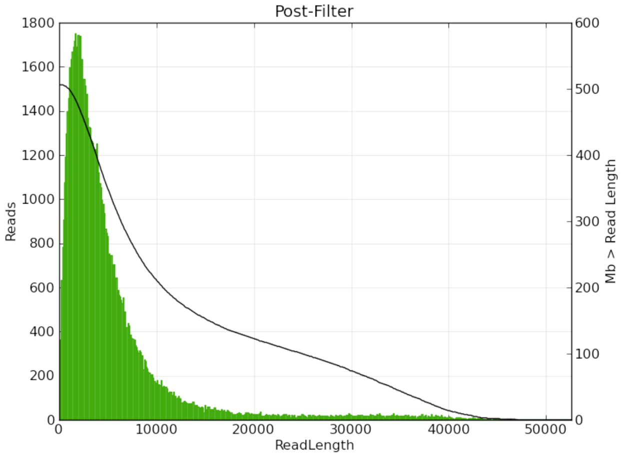
Subread Filtering

Filtering

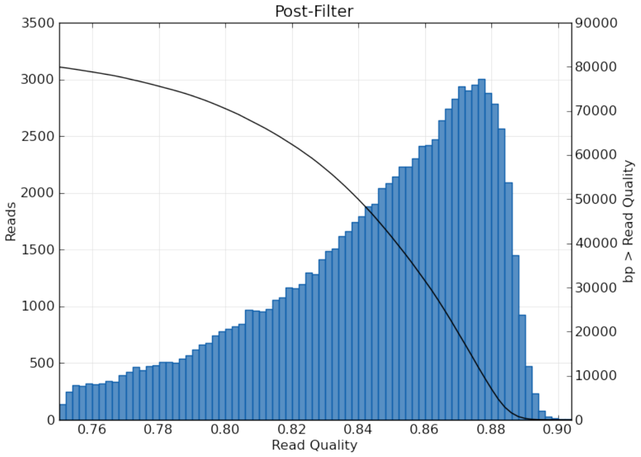
Filtering

Metrics	Pre-Filter	Post-Filter
Polymerase Read Bases	548441893	510908023
Polymerase Reads	150292	94920
Polymerase Read N50	7834	7998
Polymerase Read Length	3649	5382
Polymerase Read Quality	0.564	0.843

Polymerase Read Length

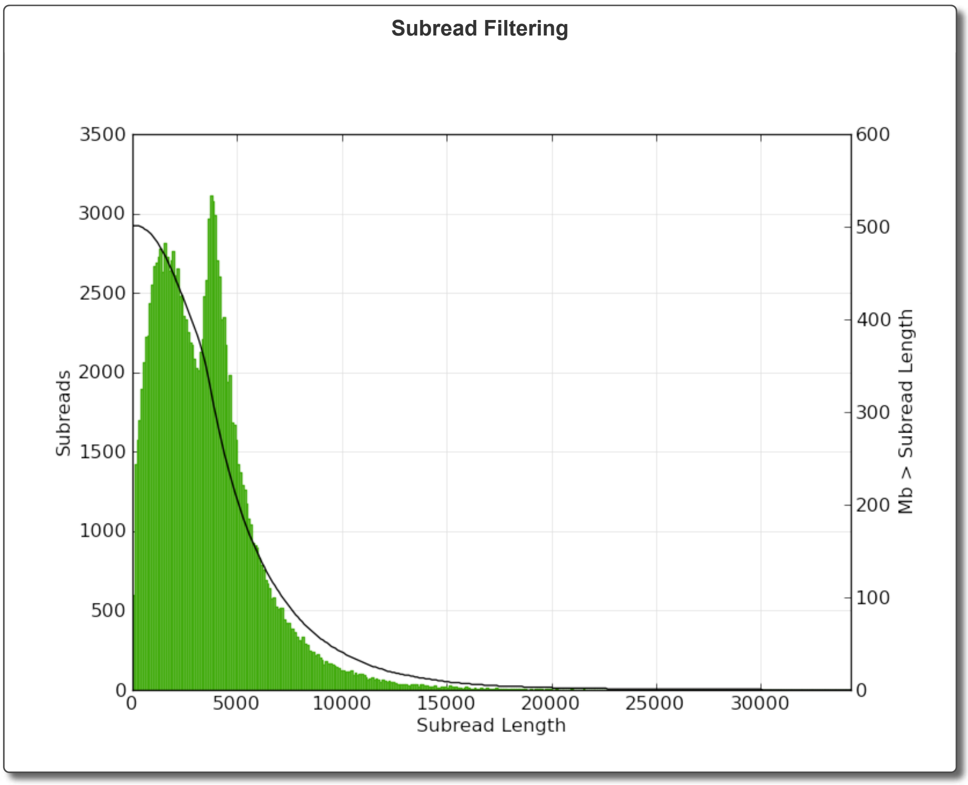


Polymerase Read Quality



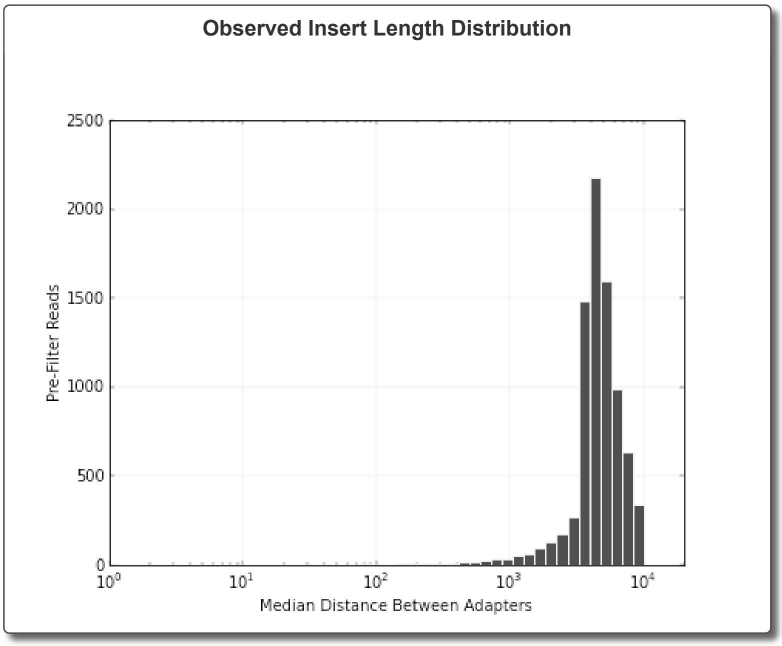
Subread Filtering

Mean Subread length	3,490	N50	4,521
Total Number of Bases	508,395,549	Number of Reads	145,638



Adapters

Adapter Dimers (0-10bp)	0.03%
Short Inserts (11-100bp)	0.01%



Loading

SMRT Cell ID	Productive ZMWs	ZMW Loading For Productivity 0	ZMW Loading For Productivity 1	ZMW Loading For Productivity 2
m200211_080647_42149_c101512282550000001823292705172175	150,292	20.43%	63.16%	16.41%

Generated by SMRT® Portal. Wed Feb 12 09:29:23 CET 2020
For Research Use Only. Not for use in diagnostic procedures.