

Sequence-based enzyme discovery from marine microbial diversity:

Multiple approaches for the heterologous expression of genes in
Escherichia coli

Hasan Arsin

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2021

UNIVERSITY OF BERGEN



Sequence-based enzyme discovery from marine microbial diversity:

Multiple approaches for the heterologous expression of
genes in *Escherichia coli*

Hasan Arsin



Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 09.06.2021

© Copyright Hasan Arsin

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2021

Title: Sequence-based enzyme discovery from marine microbial diversity:

Name: Hasan Arsin

Print: Skipnes Kommunikasjon / University of Bergen

Table of Contents

TABLE OF CONTENTS	II
SCIENTIFIC ENVIRONMENT	IV
ACKNOWLEDGEMENTS	V
ABSTRACT	VII
ABBREVIATIONS	X
LIST OF PUBLICATIONS	XI
1. INTRODUCTION	1
1.1 BACKGROUND.....	1
1.2 METAGENOMICS-BASED ENZYME DISCOVERY	4
1.2.1 <i>Heterologous production of enzymes in Escherichia coli</i>	8
1.2.2 <i>Overcoming codon bias in heterologous expression</i>	10
1.3 VIRUSES OF ARCTIC MARINE BIOMES.....	13
1.4 SERINE PROTEASES.....	17
1.4.1 <i>Subtilisins</i>	20
1.5 NUCLEASES.....	23
1.5.1 <i>T4 Endonuclease V</i>	25
1.5.2 <i>Lambda (λ) exonuclease</i>	25
1.5.3 <i>Exonuclease III</i>	26
1.6 VIRAL ENDOLYSINS	27
2. AIMS OF STUDY	31
3. MATERIALS	32
4. METHODS	33
4.1 HETEROLOGOUS PRODUCTION OF VIRAL NUCLEASE CANDIDATES.....	33
5. RESULTS AND DISCUSSION	36

5.1	BIOINFORMATICAL ANNOTATION AND SELECTION OF TARGET GENES	37
5.2	PHYLOGENY AND TAXONOMY ANALYSES OF PROPHAGE GENES	39
5.3	APPROACHES TOWARDS HETEROLOGOUS PRODUCTION OF SOLUBLE PROTEINS IN <i>E. COLI</i> ...	42
5.3.1	<i>Proteases and solubility tags</i>	42
5.3.2	<i>Codon adjustment strategies for the expression of viral genes</i>	44
5.3.3	<i>Considerations for designing new experiments</i>	46
6.	FUTURE RESEARCH	49
7.	CONCLUSION	51
	REFERENCES	52
	PUBLICATIONS	79
	SUPPLEMENTARY DATA	84
	APPENDIX 1	89

Scientific environment

The work presented in this thesis was carried out at the Department of Biological Sciences at the University of Bergen. The Structural Biology Centre (NorStruct) at the University of Tromsø has also kindly hosted two short research stays for the study of protein structures. This work was part of the projects “NorZymeD: Enzyme development for Norwegian biomass – mining Norwegian biodiversity for seizing Norwegian opportunities in the bio-based economy” (<https://norzymed.nmbu.no/>), NFR-Biotek-2021 (RCN221568) and “Virus-X: Viral Metagenomics for Innovation Value Horizon2020-LEIT-BIO-2015-1” (<http://virus-x.eu/>).

Acknowledgements

A wise wizard once quoted: “Not all who wander are lost”. Although I got to wander longer than many throughout my academic journey, it is thanks to many amazing people around me that I finally get to write these words.

I must begin by thanking my supervisor, Ida Helene Steen. I feel privileged to have been able to work with you and learn from you during my time as a PhD candidate. Your sincere enthusiasm and outlook on science has inspired me to do my best; and your belief in me throughout this process has meant a lot to me. I am truly grateful. My gratitude extends to other amazing members of the team, particularly Runar Stokke and Anita Fedøy, whom I pestered often, and learned a lot from. I would also like to thank Steffen Leth Jørgensen for the opportunity to work closely, albeit briefly, within the Geomicrobiology Lab.

I would like to thank my previous supervisors Hans Torstein Kleivdal, Pål Puntervoll and Gro E.K. Bjerga for the opportunity to begin this journey. I have learned a lot from all of you, and I appreciate the time I had within the Uni research group. In addition, I would like to thank my old colleagues within the group, Øyvind, Yuleima, Antonio, and Ephrem for their valuable company, for patiently sharing their knowledge with me, and also for many fond memories (including amazing cakes).

The friends I got to make along the way were, in fact, real treasures. I am thankful for the company of friends old and recent, and more than a few of them appropriately geeky. Rhian, Ø. Strømland, Lars, I (don't) miss losing to you at MTG. Sven, you're one inspiring guy and I owe you more than a beer. Also, I miss our Friday seminars! Victoria, it has been a joy to work with you and it has been amazing to see you progress; Zavala would be proud. My stalwart partners in shutting up, Petra and Francesca, I'm so thankful for your camaraderie, and I look forward to sharing many adventures in the future. Achim, Andreas, Heidi, I said adventures, so don't forget your d20s. I thank you all, and many others I simply could not list here.

I simply would not be here if not for the loving, ceaseless support from my family here in Bergen, in Finland and Cyprus. Dear Ingvill and Ian, thank you for looking after me and supporting me since my beginnings in Norway. My dear sister, Sıla, may Elune bless your journey. I am proud of you, and I look forward to seeing my name in your thesis. Anneciğim ve babacığım, bu yolun sonuna gelene kadar bana verdiğiniz desteği hiçbir zaman eksik etmediğiniz için size çok teşekkür ederim. Sizi çok seviyorum.

Finally, I must thank the one person who put up with me every single day of this chapter of my life. Victoria, my love; you have been there for me on my best and worst moments, and you have supported me through it all. I could not have been more fortunate to have you. You are amazing and are carving your way in your own doctoral quest. I am so proud of you, and I will do my best to support you. Thank you.

Hasan Arsin, March 2021

Abstract

The vast biodiversity of marine environments is increasingly being recognized as a source for new enzymes with value in both basic research and applications within biotechnology. In this project, the major aim was to identify, produce and perform biochemical and structural characterization on a selection of marine enzymes. These efforts were focused on bacterial proteases (**Paper I, II**) viral nucleases (**Section 4.1, and 5.3.2**), putative endolysins as well as a DNA-polymerase (**Paper III**). A versatile workflow for the cloning and heterologous expression of genes in *Escherichia coli*, focused on subtilisin-like proteases is described in **Paper I**. Fragment exchange cloning was used to insert genes into expression vectors featuring combinations of maltose binding protein (MBP), small ubiquitin-related modifier (SUMO) for improved solubility, and His-tags for protein purification with affinity chromatography. A casein-based assay was also featured to screen for proteolytic activity. All four Bacilli subtilisins tested were successfully produced in soluble and active forms. Constructs featuring N-terminal fusion proteins led to highest soluble yields and activity values for most, but not all genes tested, highlighting the value of including multiple vector configurations. Using this workflow, an intracellular subtilisin protease (ISP) from *Planococcus* sp. AW02J18, was produced and characterized (**Paper II**). Optimal activity was observed at pH 11 and 45 °C, with an active range from pH 7.0 to 11 and no activity above 60 °C. Sequence analyses of the ISP pro-peptide pointed at the presence of a conserved LIPY/F motif, understood to be centrally involved in the autocatalytic maturation of the enzyme via mutational analyses. The 3D structure of *Planococcus* ISP was solved at 1.3 Å resolution with X-ray crystallography, producing the second unique ISP structure to date, and the first with an intact native catalytic triad. The combined mutational study of the LIPY/F motif and the structure of the inhibitory pro-peptide contributed to better understanding of the maturation process of *Planococcus* ISP, and that of ISPs in general.

As a part of the Virus-X project, heterologous production trials were carried out for 42 putative viral nucleases originating from arctic viromes (Aevarsson et al. under

review, see **Appendix I**), with annotated similarity to T4-Endonuclease V, λ -Exonuclease, and Exonuclease III (**Sections 4.1** and **5.3.2**). Candidate genes were synthesized as codon optimized (CO) for expression in *E. coli* and pre-cloned into pET-family vectors encoding a His-tag for IMAC at the N- or C-terminal. Only twelve candidates were produced as soluble proteins, of which six were able to be purified. In turn, the results exemplify the challenges of using *E. coli* as an expression host for environmental viral genes and suggests that CO alone leads to limited success. In **Paper III**, the discovery of a novel prophage is reported from within the genome of *Hypnocyclicus thermotrophus*, a Gram-negative, thermophilic bacterium isolated from the Seven Sisters hydrothermal vent field. Designated *Hypnocyclicus thermotrophus* phage H1 (HTH1), the identified prophage genome was 41.6 kbp long and consisted of 46 protein-coding genes. Analysis and functional annotation of the HTH1 genome with multiple *in silico* approaches suggested closest taxonomic association to the viral family *Siphoviridae*. The lytic cassette of HTH1 showed closest similarity to viruses of Gram-positive bacteria. However, HTH1 was found encoding an N-acetylmuramoyl-L-alanine amidase not observed in other compared phages. Nine genes putatively related to lysis and nucleic acid processing were selected for heterologous production in *E. coli*. Besides CO genes, codon harmonized (CH) variants of each gene were also tested to be able to compare the two approaches' effects on the soluble yield and thermostability of heterologous proteins. Five genes led to soluble protein from their CO variants, of which 4 were also soluble as CH variants. When compared, CO variants achieved higher soluble protein yields, but CH variants led to proteins with higher thermostability, as assessed by differential scanning fluorimetry.

Taken together, this work presents results encompassing key steps of a sequence-based pipeline for the discovery of marine microbial enzymes. In addition, reported findings expand existing knowledge of ISPs and prophages of *Fusobacteria* from hydrothermal vent environments. The cases presented within are connected by a common theme of improving soluble production of heterologous proteins in *E. coli*, via various, compatible means. Coupled with the contemporary bioinformatics tools facilitating the functional annotation of a wider range of viral genes, complementary

implementation of these approaches suggests a promising blueprint for future studies aiming the bioprospecting of marine microbial genetic resources.

Abbreviations

EC: Enzyme Commission, for formal enzyme nomenclature.

ESP: Extracellular subtilisin protease

ISP: Intracellular subtilisin protease

IMAC: Immobilized metal affinity chromatography

MBP: Maltose binding protein

PDB: The Protein Data Bank

SUMO: Small ubiquitin-related modifier

NCBI: National Center for Biotechnology Information

KEGG: Kyoto Encyclopedia of Genes and Genomes

GO: Gene Ontology Resource

CO: Codon optimization

CH: Codon harmonization

List of publications

- I. Bjerga, G. E. K., Arsim, H., Larsen, Ø., Puntervoll, P., and Kleivdal, H. T. (2016). A rapid solubility-optimized screening procedure for recombinant subtilisins in *E. coli*. *J. Biotechnol.* 222, 38–46. doi:10.1016/j.jbiotec.2016.02.009.
- II. Bjerga, G. E. K., Larsen, Ø., Arsim, H., Williamson, A., García-Moyano, A., Leiros, I., et al. (2018). Mutational analysis of the pro-peptide of a marine intracellular subtilisin protease supports its role in inhibition. *Proteins*. 86, 965–977. doi:10.1002/prot.25528.
- III. Arsim, H., Jasilionis, A., Dahle, H., Sandaa, RA., Stokke, R., Nordberg Karlsson, E., and Steen, IH. (2021) “Exploring codon adjustment strategies towards *Escherichia coli*-based production of viral proteins encoded by HTH1, a novel prophage of the marine bacterium *Hypnocyclicus thermotrophus*” (Under review, *Front. Microbiol.*, 03/2021)

1. Introduction

1.1 Background

Enzymes are essential protein catalysts which accelerate and facilitate chemical reactions in and around all living cells. The first documented application of enzymes for the production of desirable goods in human history is thought to stretch back over 7000 years ago to the Mediterranean, where calf stomachs - and the enzymes within, were used for the production of cheese from milk (McClure et al., 2018). Today under the term “biotechnology”, enzymes are increasingly used in a diverse range of industrial, medical and academic applications (Lorenz and Eck, 2005; Ward, 2011; Singh et al., 2016; Chapman et al., 2018). As biological catalysts, enzymes offer a range of advantages over elements traditionally used as catalysts such as palladium, gold, or iridium (Bhaduri and Mukesh, 2014). Their high specificity towards the reactions they catalyse makes them very desirable in the production of fine chemicals and pharmaceutical products (Roy and Abraham, 2006; Sullivan et al., 2009; Sun et al., 2018). They can work in milder conditions that promote substrate or product stability and avoid unwanted side-reactions (Goldberg et al., 2006; Savile et al., 2010; Liu et al., 2017; Patel, 2018). Their biodegradability allows for gentler downstream processing in many applications (listed in more detail for proteases in **Section 1.4**) and lowers their environmental impact (Schmid et al., 2001; Illanes, 2008). In addition, they can be used in food (Raveendran et al., 2018) and cosmetic products when produced in organisms that are recognized as safe (European Food Safety Authority, 2013; Center for Food Safety and Applied Nutrition, 2018). Indeed, the implementation of enzymes for biotechnological applications is not a task without challenges. An ideal enzyme for any application needs to remain stable and display efficient and specific catalytic activity under defined process conditions (Burton et al., 2002; Lorenz and Eck, 2005). Despite their many advantages, enzymes are sensitive proteins. They function optimally under a limited range of physiochemical conditions, often in association with their source organism and its environment (Vieille et al., 1996; Arnold et al., 2001). In order to expand the library of currently available

enzymes, meet the demands of existing biotechnological processes, and stimulate innovation for new methods and products, discovery of new enzymes remains essential.

A common rationale for discovering enzymes with desired properties is to study organisms from relevant natural sources which can be expected to support life under such conditions (Elleuche et al., 2014). Microorganisms in particular, offer broad genetic and physiological diversity, possibility of genetic manipulation, quick generation times and relatively little space requirements for dense growth (Rao et al., 1998; Sumantha et al., 2006). The largest repository of microbial diversity is the marine environments, which represent over 70% of the Earth's surface (Kodzius and Gojobori, 2015). A wide range of environmental niches can be found within the oceans, all housing myriad microorganisms equipped with enzymes molecularly adapted to function under their respective native physiochemical conditions (Elleuche et al., 2014; Brininger et al., 2018). These include cold waters of the Arctic and Antarctic (Marx et al., 2007; de Pascale et al., 2012), as well as some "extreme" environments such as the deep seas (Ferrer et al., 2005; Takai et al., 2008), hypersaline "underwater lakes" (Kim and Dordick, 1997) and hydrothermal vents at the ocean floor with sharp temperature gradients and exposure to unique fluid chemistries (Pedersen et al., 2010; Steen et al., 2016; Schouw et al., 2018; Fredriksen et al., 2019). Marine microorganisms are therefore considered highly valuable sources for the discovery of enzymes and other natural products (Ferrer et al., 2007; Trincone, 2011; De Santi et al., 2016; Indraningrat et al., 2016). Earlier study of marine microorganisms was hampered by their limited cultivability in the laboratory (Staley and Konopka, 1985). However, the emerging methodologies within metagenomics allowed the culture-independent studies of microbial communities (Handelsman, 2004), and became a key approach for the study of marine biodiversity and their enzymes thanks to developments in sequencing technologies (Berenwinkel et al., 2012; Popovic et al., 2015; Goodwin et al., 2016; Ferrer et al., 2019).

In addition to cellular microorganisms, viruses are also promising sources for enzyme discovery. Thanks to their unique way of existence involving the infection and

replication inside of a host cell, they bear a unique selection of enzymes (Lwoff, 1957; Hobbs and Abedon, 2016). Viruses are recognized as the most abundant biological entities in the oceans, estimated to number around 10^{30} - 10^{31} particles (Børsheim et al., 1990; Fuhrman and Suttle, 1993; Youle et al., 2012). With such a large presence, virus-mediated killing of cellular hosts influences microbial communities at large, and by extension affect the nutrient and geochemical cycles within ocean ecosystems (Suttle, 2007). In addition to their near omnipresence, viruses also host a vast breadth of genetic diversity, only a fraction of which is suggested to have been so far explored (Forterre and Prangishvili, 2009; Mokili et al., 2012). A majority of discovered viral genes were noted to share no significant homology with reference genes in existing databases (Paez-Espino et al., 2016; Gregory et al., 2019), leading to the popular “dark matter” term being coined to refer to this unexplored viral sequence space (Youle et al., 2012; Hatfull, 2015; Michalska et al., 2015). Further exploration of this resource is expected to yield a wealth of viral enzymes promising for various applications. Prominent examples include enzymes with lytic activities against prokaryote cells, which may find bactericidal applications within medicine (Hermoso et al., 2007; Plotka et al., 2020); or nucleic acid – modifying enzymes such as DNA polymerases (Dale et al., 1985; Choi, 2012), ligases (Engler and Richardson, 1982; Doherty et al., 1996), or nucleases (Song and Zhang, 2008). These may be utilized to develop diagnostic methods or products within biotechnology, medicine, or academia.

Altogether, the great taxonomic and functional diversity of marine microorganisms presents an immense opportunity to discover new enzymes with desired catalytic properties and novel functions. However, standing challenges in bioinformatic identification, selection, as well as heterologous production of enzyme candidates render the discovery process difficult to streamline. Furthermore, in the shadow of the coronavirus pandemic affecting the world at the time of writing, it is evident that our knowledge of viral pathogens and their workings is far from complete. In this regard, the study of viral enzymes will expand our understanding of viral processes and their enzymatic constituents, potentially contributing to the prevention and/or treatment of similar diseases in the future.

1.2 Metagenomics-based enzyme discovery

In simple terms, “metagenomics” describes the culture-independent study of the collective genomes present within a given microbial community sample. The term was coined by Jo Handelsman and co-workers in their influential study; describing their approach to access the uncultivated genetic diversity of the soil microbiome (Handelsman et al., 1998). Since early assemblies of microbial genomes from lower diversity samples (Tyson et al., 2004), the application of metagenomics approaches has led to ground-breaking discoveries about the evolution of life on Earth (Spang et al., 2015), contributed numerous new lineages to the tree of life (Hug et al., 2016; Parks et al., 2017), allowed prediction of metabolic functions from uncultured microbial lineages (Ravin et al., 2015; Castelle et al., 2018), and provided an unprecedented insight into how microorganisms interact with other species and their environment (Konstantinidis et al., 2009; Takai and Nakamura, 2010; Dahle et al., 2013; Urich et al., 2014; Stokke et al., 2015; Steen et al., 2016; He et al., 2017).

Earlier strategies utilized the generation of extensive clone libraries that were screened for genetic markers via (specific or degenerate) polymerase chain reaction primers, such as the ones targeting the 16S rRNA gene to identify prokaryotic taxa (Schmidt et al., 1991). Thereafter, sub-libraries corresponding to certain groups could be constructed and sequenced to allow detailed study of desired metabolic features (Schleper et al., 1997; Gillespie et al., 2002). While this approach generated valuable new data, it was highly laborious, and the genomic information gathered was often incomplete. In the following years, the development of the so-called “next generation” DNA-sequencing technologies revolutionized the metagenomics approach and propelled it to widespread use (Lynch and Neufeld, 2015; DeCastro et al., 2016; Jünemann et al., 2017; Jo et al., 2020). In contemporary studies, DNA extracted from a sample is often directly subjected to total sequencing, where data processing via bioinformatics tools allows the re-construction of near-complete metagenome assembled genomes (MAGs), their annotation and placement in the tree of life (Hugerth et al., 2015; Parks et al., 2017; Zaremba-Niedzwiedzka et al., 2017).

A multitude of bioinformatical tools and strategies have been developed for the computational analysis, annotation, and further processing of metagenomic datasets. Some examples include the RAST annotation server (Glass et al., 2010), Joint Genome Institute's (JGI) Integrated Microbial Genomes (IMG) genome browsing and annotation platform (Markowitz et al., 2012), JCVI Metagenomics Reports (METAREP) tool for comparative metagenomics (Goll et al., 2010), and more recently the multi-purpose and versatile Elastic MetaGenome Browser (EMGB) (Jünemann et al., 2017) and the open source anvi'o platform (Eren et al., 2021) for the analysis and visualization of omics data.

Metagenomics has also been used as a powerful approach for the discovery of new enzymes, and it has been applied to this context in various ways. A functional screening approach usually involves the generation of metagenomic clone-libraries, their subsequent heterologous expression in a laboratory host (usually *Escherichia coli*) and their functional screening to identify clones with the expected activity. As functional approaches do not rely on any predictions of function based on existing knowledge and therefore (Handelsman, 2004), truly novel enzymes may be discovered. Example studies report the discovery of diverse hydrolytic enzymes, such as proteases (Lämmle et al., 2007; Waschowitz et al., 2009; García-Moyano et al., 2021), esterases and lipases (Ho Jeon et al.; Fu et al., 2013). However, this approach is strongly dependent on the often-challenging production of properly folded and active enzymes to display the expected activity. Furthermore, facilities and specialized methods for high-throughput screening of clones are often necessary to perform such studies at a viable scale (Uchiyama and Miyazaki, 2009; Colin et al., 2015).

In contrast, sequence-based discovery strategies look for genes predicted to code for desired enzymes. In earlier studies, this was performed by screening the metagenome library with targeted PCR primers or hybridization probes (Ferrer et al., 2009). The volume of sequencing data is increasing rapidly with metagenome sequencing studies from diverse ecosystems (Sukul et al., 2017; Gregory et al., 2019), and with the

sequencing of thousands of microbial isolate genomes¹. To be able to process this data and facilitate their efficient mining for desired enzymes, *in silico* tools have been developed (Roumpeka et al., 2017). Once a list of candidate genes is identified, they can be cloned into a suitable vector and heterologously expressed in a host organism of choice (Barone et al., 2014; Liebl et al., 2014). After expression, enzyme candidates are assayed experimentally to confirm their activities and in case of success, proceed to carrying out their functional and/or structural characterization (**Figure 1**). Discoveries via sequence-based approaches include amongst many, proteases (Toplak et al., 2013; de Oliveira et al., 2018), xylanase (Fredriksen et al., 2019), cellulases (Yang et al., 2016) and other glycosyl hydrolases (Wang et al., 2011).

Contemporary sequence-based screens are conducted by assessing the homology of metagenomic sequences against known enzyme-coding sequences stored on public databases such as UniProt (Bateman et al., 2021), NCBI GenBank (Clark et al., 2016) and BRENDA (Placzek et al., 2017). In this context, homology is determined based on excess similarity between sequences (Pearson, 2013), and has been traditionally calculated with the use of search tools such as BLAST (Altschul et al., 1990), SSEARCH (Smith and Waterman, 1981), and FASTA (Pearson and Lipman, 1988)). This comparison can be performed not only between two DNA sequences, but also between protein or translated protein sequences. Sequences with identity values of >40% are considered closely related, whereas statistically significant protein homologs with >20% identity are reportedly observed (Pearson, 2013). Finally, derivative tools such as PSI-BLAST (Altschul et al., 1997) feature sequence comparisons against sequence profiles, which are multiple sequence alignments generated from a group of homologous sequences. Databases such as MEROPS for proteases (Rawlings et al., 2018), CAZy for carbohydrate-active enzymes (Lombard et al., 2014), REBASE for restriction nucleases (Roberts et al., 2015) and CLAE for Lignocellulose-active fungal enzymes (Strasser et al., 2015) provide focused datasets

¹ 178012 Isolate genomes listed on JGI Genomes Online Database (GOLD) at the time of writing (Mukherjee et al., 2021).

for discrete enzyme classes to compare homology to and predict putative function in candidate sequences.

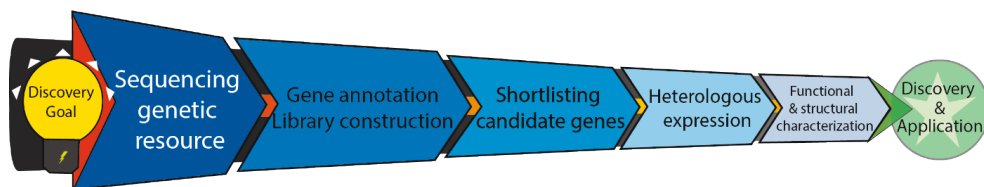


Figure 1: Typical sequence-based enzyme discovery pipeline. The process starts by defining a discovery goal, often a particular group of enzymes. Then, an appropriate resource can be sampled, and its metagenome sequenced. Identified genes are annotated to predict their function, which allows shortlisting of genes with the desired putative activity. Candidate genes are expressed heterologously to produce proteins in amounts needed for functional and structural characterization experiments. Finally, discovered enzymes may provide novel biological insight about their source species or environment and can now be fully assessed for their applications in biotechnology.

Compared to functional screening, sequence-based strategies can be more convenient to perform, as they do not rely on any special equipment or mass-cloning of sequences. Once a shortlist of gene targets fulfilling the desired homology criteria is generated, heterologous production experiments can be carried out at a much smaller scale. Discovery of desired enzymes even with only minor differences from existing examples allow for legitimate patent submissions for new products or applications, similar to the use of mutation or engineering induced changes (Bryan and Pantoliano, 1988; Brode et al., 1994; Fanoë and Mikkelsen, 2007). However, sequence-based approaches are frequently criticised for not being able to detect significantly novel proteins or enzymatic functions (Ferrer et al., 2009; Distaso et al., 2017).

New approaches for the functional annotation and homology prediction of genes are constantly being developed and is expected to further improve the capabilities of sequence-based enzyme discovery efforts. A major development in methodology is the mainstream use of hidden Markov model (HMM) profiles (Söding, 2005; Finn et al., 2011). These profiles hold broader information about multiple aligned homologous sequences and allow more thorough comparisons of homology between

sequences. Current state-of-the-art methods can also represent query sequences as HMMs and are able to carry out profile against profile comparisons. Another example used for the detection of carbohydrate-active enzymes is the “peptide pattern recognition” approach (Busk and Lange, 2013; Agger et al., 2017). Authors have described the approach to rely on the analysis of short, conserved motifs around the enzyme active site allowing for the reliable prediction of enzyme activity even when the overall sequence identity between compared sequences is very different (Busk and Lange, 2013).

Beyond cellular microorganisms, application of metagenomics towards viruses has greatly expanded our understanding of viral genetic diversity (Hatfull, 2015). While the principles of metagenomics remain the same, the study of viral communities need to overcome particular challenges. Viruses vary in their genome architectures, as single or double stranded genomes consisting of RNA or DNA can be found in nature (Grose and Casjens, 2019). Sampling and sequencing of viruses pose technical difficulties and call for specialized methods, as low abundances make it challenging to obtain genetic material of sufficient amount and quality for sequencing via common sampling routines (Sandaa et al., 2018). Furthermore, their inherent genetic diversity greatly complicates functional annotation of viral genes. Developments to improve metagenomic access to the virosphere is expected to facilitate the discovery of viral enzymes with both known and novel physical and catalytic properties. Addressing these specific challenges of viral metagenomics has been the main aim of the Virus-X project, via robust sampling routines, development of specialized bioinformatics tools, heterologous expression approaches, and a thorough characterization pipeline for discovered enzymes (Aevansson et al. under review, see **Appendix I**).

1.2.1 Heterologous production of enzymes in *Escherichia coli*

Whether identified with sequence or function-based screens, enzymes need to be able to be produced in the necessary amounts and purity in order to be further characterized or implemented in any downstream applications. If the enzyme in question comes from a cultivable isolate, it might be possible to obtain enough of it simply by

growing the isolate in desired capacities, also termed native expression (as exemplified in Miyake *et al.*, 2005 and Zhu *et al.*, 2007). However, the desired enzyme may be naturally poorly expressed in the native host. Furthermore, in a case where using the organism of origin is not feasible (i.e. when enzymes are identified from metagenomic sequences or from isolates unable to grow and thrive in the lab) it is often necessary to use another organism to host the expression of the enzyme, using plasmids as a vehicle. This approach, dubbed *recombinant* or *heterologous* expression, has become one of the defining tools in molecular biology, since its demonstration by Cohen and Boyer in 1973 (Cohen *et al.*, 1973).

The most common and perhaps the best-known heterologous expression host is the Gram-negative bacterium *E. coli*. This is due to the wide availability of genetic tools and favourable factors such as higher expression yields, cell growth characteristics, opportunities for intra- or extracellular expression, available post-translational modifications and more (Makrides, 1996). The capabilities of the organism have been extensively reviewed throughout the past decades (Makrides, 1996; Rosano and Ceccarelli, 2014; Rosano *et al.*, 2019). As a heterologous expression host, it offers certain advantages that has given the organism its wide use. While many specialized strains have been developed for various research applications throughout the years, (prominent examples highlighted in Rosano and Ceccarelli, 2014), these advantages remain relevant to many common expression strains such as the BL21 and K-12 families (Daegelen *et al.*, 2009). *Escherichia coli* strains used in protein expression are capable of rapid growth even on simple media such as Luria Broth (LB) (Sezonov *et al.*, 2007), and achieve higher cell densities when growth conditions are further optimized (Shiloach and Fass, 2005). The foreign DNA of interest can be easily and quickly transformed in the cells using chemical competence (Pope and Kent, 1996).

Some of the prominent challenges associated with *E. coli* expression, especially when attempting to express genes from distant genetic origins, include low or no detectable expression, production of insoluble proteins, in misfolded 'inclusion bodies', and the production of soluble but inactive proteins (Rosano and Ceccarelli, 2014). While some of these issues can be alleviated by adjusting the expression parameters (such as

temperature, inducer amount, growth medium and duration etc.), others require design-level implementation of certain elements to overcome some of the more fundamental problems.

The problem of insoluble expression is perhaps the most commonly experienced. It can sometimes be remedied by slowing down the expression by lowering incubation temperatures or by decreasing the amount of inducer applied (Rosano and Ceccarelli, 2014). Solubility enhancing proteins are another well-known method of improving heterologous soluble yields of otherwise challenging proteins in *E. coli*. Currently, more than 20 different proteins are listed for various levels of solubility-enhancing effects (reviewed in detail by Costa et al., 2014; Ki and Pack, 2020). The *E. coli* maltose binding protein (MBP) in particular has been shown to be effective in improving solubility of proteases expression (Kapust and Waugh, 1999; Kwon et al., 2011; Toplak et al., 2013). Some other notable solubility-enhancing fusion proteins include the glutathione S-transferase (GST; EC 2.5.1.18) encoded by *Schistosoma japonicum* (Smith and Johnson, 1988), the small ubiquitin-like modifier (SUMO) (Malakhov et al., 2004) and more recently, the small Fh8 protein extracted from the parasite *Fasciola hepatica* (Costa et al., 2013a).

1.2.2 Overcoming codon bias in heterologous expression

Codon bias is defined as the differing frequencies with which synonymous DNA codons are used for transcription between different organisms, and may vary significantly between different organisms, and even between different proteins within the same organism (Ikemura, 1981; Gouy and Gautier, 1982). This bias indicates the availability of particular tRNAs to the gene expression machinery of each species. Therefore, attempts to express a gene requiring high frequency of “rare” codons in the heterologous host, effectively lacking the tRNA species to perform the task correctly, would lead to low levels of protein of interest produced (Kane, 1995; Gustafsson et al., 2004). In order to overcome the limitations of incompatible codon biases, and improve the effectivity of heterologous protein production, different strategies emerged. Codon adaptation indices (CAI) were generated to facilitate comparison of

codon differences between species (Sharp and Li, 1987; Carbone et al., 2003). To address the problem, one approach is to provide the expression host with the genes to express the ‘rare’ tRNAs. For this purpose, commercially available *E. coli* strains with improved intracellular tRNA pools, such as the BL21-CodonPlus (DE3)-RIL cells commercialised by Agilent (Agilent Technologies, Santa Clara, CA, USA) or the Rosetta (DE3) cells from Merck (Merck, Darmstadt, Germany) can be used.

Codon optimization (CO) is another approach, where the gene of interest is modified to alter the codons to better suit the native codon usage of the expression host, while keeping the target amino acid sequence unchanged. While smaller edits were shown to be successful via site-directed mutagenesis (Kink et al., 1991), total gene synthesis remains an attractive option for the analysis and CO of whole and even multiple genes. When performing CO, perhaps the simplest approach is the “one codon – one amino acid” design aiming to replace all codons for each amino acid with the one most abundant in the expression host, entirely eliminating rare codons. Studies demonstrating increased yields with this strategy have been reported from multiple hosts (reviewed in Gustafsson et al., 2004). However, in oversaturating the cell with an imbalanced tRNA pool, this cruder approach may lead to translation (Kurland and Gallant, 1996) or frameshift (Farabaugh and Björk, 1999) errors, and overproduction of protein to the extent of hindering host cell growth (Gong et al., 2006). In addition, this may cause repetitive mRNA elements, leading to undesirable secondary structure formation and hindering the host cell protein expression machinery (Griswold et al., 2003; Presnyak et al., 2015; Mauger et al., 2019). More refined approaches of CO have been developed in the last decade considering not only CAI values, but also seeking to avoid extreme GC content or undesirable codon pairs (Boycheva et al., 2003), repeating sequences and unfavourable mRNA secondary structures (Griswold et al., 2003; Wu et al., 2004; Goodman et al., 2013). Multiple software solutions are available for researchers to analyse and optimize their sequences with the aforementioned considerations (reviewed in Angov, 2011; Gould et al., 2014); including proprietary optimisation pipelines implemented by many gene-synthesis

companies, such as GenSmart (GenScript, Piscataway, NJ, U.S.A.) (GenScript, 2021) and GeneArt (Thermo Fisher Scientific, Waltham, MA, U.S.A.) (Raab et al., 2010).

Further studies reported that proteins were able to be better expressed with the general lack of rare codons in the expression host (Hale and Thompson, 1998; Chang et al., 2006; Burgess-Brown et al., 2008; Öberg et al., 2011). In other cases, lower yields, formation of truncated proteins or inclusion bodies were observed to be unavoidable, even after CO (Yadava and Ockenhouse, 2003; Farshadpour et al., 2014, **Paper III** and Aevansson et al. under review, see **Appendix I**). As these issues were thought to be associated with incorrect protein folding, studies focusing on this link showed that even synonymous substitution of codons was able to affect the eventual folding quality of the protein of interest (Fedyunin et al., 2012; Liu, 2020; Liu et al., 2021). As total removal of rare codons was recognized a sub-optimal strategy, a new approach, termed “codon harmonization” (CH) was reported (Angov et al., 2008). By attempting to replicate the expression “cadence” of the native host, it was suggested that better folded proteins may be produced in the expression host. This would still be achieved by synonymous substitution of codons, but also considering regions of rare codons in the native host and appropriate substitution towards similarly rare codons in the expression host. Applications of the CH approach have since reported improved protein yields, quality and host cell viability (Hillier et al., 2005; Angov et al., 2011; Wen et al., 2016; Asam et al., 2018; Punde et al., 2019).

Codon bias is also observed in viruses. Particularly in bacteriophages, viral genes were found displaying a high adaptation in codon usage towards that of their host's, but distinct from other unrelated bacteria (Bahir et al., 2009). This occurrence was suggested to be a result of codon-selective pressure inherent within the translational machinery of the host bacteria (Carbone, 2008). Interestingly, a similar finding was reported for viruses infecting humans, but not those infecting other mammals (Bahir et al., 2009). The codon bias similarity was especially pronounced for structural proteins and was noted as potentially related to infectivity of the virus (Lucks et al., 2008; Bahir et al., 2009). Some dsDNA viruses were found to encode tRNAs, and subsequently rely less on codon adaptation to their host (Limor-Waisberg et al., 2011).

This strategy was suggested to be particularly relevant for lytic phages, granting them increased replication effectiveness and virulence, but reported as less common for temperate phages (Bailly-Bechet et al., 2007).

Heterologous expression of many viral genes in *E. coli* has been described in the past four decades (Garapin et al., 1981; Shuman et al., 1988; Braun et al., 1999; Chen et al., 2001). However, low yields and insoluble production of proteins are also reported (Leavitt et al., 1985; Hizi et al., 1988; Li et al., 2005; Lee et al., 2009). From viral metagenomic sequences of ssRNA viruses, Liekniņa and co-workers reported the expression of over 100 coat protein genes, where nearly 40% of tested proteins were observed forming inclusion bodies in initial tests, but this result was able to be improved by lower-temperature expression conditions (Liekniņa et al., 2019). In some other studies, CO approaches was utilized in an attempt to achieve higher protein yields. Improved yields of three structural proteins from Chinese Sacbrood Virus using the one codon – one amino acid application of CO have been reported (Fei et al., 2015). For the expression of chicken anemia virus capsid protein, significant improvements with the use of CO were also observed (Lee et al., 2011). At the time of writing, **Paper III** appears as the sole example of the CH approach applied for the expression of viral genes.

1.3 Viruses of Arctic Marine Biomes

The Arctic ocean houses a broad range of marine biomes, such as surface and deep waters, sea ice, and hydrothermal vents in the ocean floor. Its shallow waters are subject to significant fluctuations in the availability of light and phytoplankton productivity throughout the polar year (Winter et al., 2012; Wilson et al., 2017). Temperature shifts in the water are noted to be less severe, but not less significant, affecting viruses as well as their hosts (De Paepe and Taddei, 2006; Kirchman et al., 2009). As with other marine environments, viruses also play key roles in nutrient and carbon cycles in the Arctic (Stein and MacDonald, 2004; Suttle, 2005; Yamamoto-Kawai et al., 2006), where the viral shunt recycles living cells into dissolved and particulate organic material (Fuhrman, 1999; Wilhelm and Suttle, 1999).

The present understanding of arctic marine viral abundance and diversity has been driven by the large-scale sampling efforts and application of viral metagenomics (Breitbart et al., 2002; Mokili et al., 2012). One of the more recent undertakings was the Tara Oceans Polar Circle (TOPC) expedition, which sampled 41 sites up to 1000 m depth around the Arctic Ocean, generating a metagenomic insight into the diversity of DNA viruses in the region (Gregory et al., 2019). The authors described the Arctic waters as a major hotspot of viral biodiversity. However, among the viral genomes analysed in this study, only 10% of total sequences were able to be taxonomically assigned. These corresponded mainly to the tailed dsDNA bacteriophages of the order *Caudovirales* (*Myoviridae*, *Siphoviridae* and *Podoviridae*) followed by large dsDNA algal viruses in the family *Phycodnaviridae*. In deeper waters, viral assemblages sequenced from 16 sites around the Arctic Ocean, at depths between 10 to 3246 meters were studied (Angly et al., 2006). The authors reported the detection of dsDNA viral hits mainly corresponding to the families *Podoviridae*, followed by *Siphoviridae*. While a total lack of hits for chp1-like ssDNA microphages were reported, a high abundance of prophage-like sequences was observed (Angly et al., 2006).

This finding supported the earlier works of Payet and Suttle, who studied the viruses infecting phytoplankton and bacteria at various locations in the Canadian Arctic via gradient gel electrophoresis fingerprinting of amplicons (Payet and Suttle, 2013, 2014). By targeting genes encoding DNA polymerase B (*polB*) and major capsid protein (*g23*) which act as markers for *Phycodnaviridae* and *Myoviridae*, respectively, *Phycodnaviridae* was observed to be sensitive to changes in hydrological conditions and microbial dynamics, but not T4-like *Myoviridae*. Furthermore, the authors reported that lytic virus infections were associated with periods with high phytoplankton productivity, whereas lysogeny was preferred at periods of lower productivity (Payet and Suttle, 2013, 2014). A similar study used the genes *g23* and *mcp* to target T4-like *Myoviridae* and large dsDNA algal viruses, respectively, and study their seasonal variations north and west of Svalbard. Here, seasonal abundance and diversity of free viral particles was noted as highest during the polar winter and in

deeper water samples, and lowest during the summer months and in surface water samples (Sandaa et al., 2018).

A recent study targeting the viral populations of the deep ocean floor was able to isolate viral DNA and construct viromes from samples taken from 5571 m deep sediments in the Greenland Sea (Corinaldesi et al., 2017). Here, a dominating fraction of viruses that could be taxonomically placed, was noted as tailed dsDNA bacteriophages belonging to the order *Caudovirales*; where *Siphoviridae*, *Myoviridae* and *Podoviridae* were the three most represented taxonomic groups overall. Viruses with ssDNA genomes (mainly of families *Circoviridae*, *Geminiviridae* and *Inoviridae*) as well as retrotranscribing viruses (*Retroviridae*), giant viruses (*Mimiviridae*) were also reported as observed in minor fractions.

Successful isolation of viruses from Arctic Ocean habitats are known, but rare. Four members of *Phycodnaviridae*, infecting the Arctic picophytoplankton *Micromonas polaris* have recently been isolated from Kongsfjorden, Norway (Maat et al., 2017). At Franklin Bay, Canada, a cold-active T5-like siphovirus was isolated from the obligate psychrophilic bacterium *Colwellia psychrerythraea* strain 34H, sampled from a nepheloid layer (Wells and Deming, 2006). One of the colder niches of the Arctic Ocean, sampling of the sea ice led to the isolation of a ssDNA filamentous phage f327 from a *Pseudoalteromonas* strain (Yu et al., 2015). Here, it was also noted the prevalence of similar members of *Inoviridae* in the Arctic sea ice (Yu et al., 2015). Also from sea ice, the isolation of three lytic phages: a Myovirus infecting *Shewanella* and two Siphoviruses infecting *Flavobacterium* and *Colwellia* species have been reported (Borriss et al., 2003).

Viruses of Arctic hydrothermal vents

Hydrothermal vent environments are one of the most remarkable and diverse biomes in the marine biosphere. Hydrothermal vents appear as fissures or tears on the sea floor where geothermally heated seawater is ejected out often forming chimney-like structures (Corliss et al., 1979; Pedersen et al., 2010; Dick, 2019). They feature various harsh physiochemical conditions, including steep gradients of temperature (up

to above 300 °C) and reducing chemistry of vent fluids mixing with the surrounding seawater (Nakamura and Takai, 2014; Steen et al., 2016). The vent plumes may disperse in a wide range, feeding diverse populations of chemolithoautotrophic and heterotrophic prokaryotes (Lam et al., 2004; Takai and Nakamura, 2010, 2011; Steen et al., 2016; Dahle et al., 2018). Although microbiological study of hydrothermal vent environments has received increasing attention in the past decades, studies focusing on viruses have been sparse (Ortmann and Suttle, 2005; Williamson et al., 2008; Lossouarn et al., 2015; Castelán-Sánchez et al., 2019). Even fewer studies examined the viruses of Arctic hydrothermal environments to date.

At the Loki's Castle vent field, a viral study of the deep, surrounding waters and hydrothermal vent plumes was performed (Ray et al., 2012). The authors managed to isolate a near-complete genome of an Enterobacteriophage lambda-like virus via an approach using pulsed field gel electrophoresis (PFGE) (Sandaa et al., 2010) and subsequent sequencing of DNA bands. The *Caudovirales* dsDNA viruses were the most prevalent group in the samples analysed – wherein Siphoviruses were found to be more abundant in the plume samples, and Myoviruses in the surrounding water samples. In-depth analysis of integrase-like genes in the viral sequences indicated a significant lysogenic potential of the viral populations (Ray et al., 2012), in line with previous studies from other hydrothermal sites (Williamson et al., 2008). Above the vent sites at the Arctic Mid-Ocean Ridge (AMOR), a study of T4-like viral communities in the shallow (99-600 m) and deep (1999-3000 m) water column was reported via *g23* gene profiling (Le Moine Bauer et al., 2018). Vent plume induced changes were not reported in the diversity of the viral profiles between water column samples and the vent field plume samples – but instead depth was indicated as the main factor influencing diversity of T4-like *Myoviridae* in these sites (Le Moine Bauer et al., 2018).

The current collection of recovered viromes and viral isolates from arctic marine environments point at the presence of an immense repository of viruses and associated genetic diversity (Paez-Espino et al., 2016). While aforementioned studies have focused on examining the ecology of viral populations, the bioprospecting potential of

this resource has not yet been fully explored, especially not for environments potentially housing extremophiles (Krishnamurthy and Wang, 2017; Dávila-Ramos et al., 2019; Gil et al., 2021). The efforts of the Virus-X project (Aevarsson et al. under review, see **Appendix I**) are therefore of timely importance, aiming the mining of arctic marine viromes towards the discovery of enzymes with biotechnological interest, in addition to expanding existing knowledge on viral ecology.

1.4 Serine proteases

Proteases (or peptidases) (EC 3.4.-.-, (Webb, 1992; McDonald et al., 2009)) are a subset of the hydrolase class of enzymes which catalyse the breaking down of peptide bonds between amino acids, with known members catalogued in the MEROPS peptidase database (<https://www.ebi.ac.uk/merops/>) (Rawlings et al., 2018). Proteases can be classified into two groups according to their pattern of cleavage. Exopeptidases are proteases which cleave peptide bonds at or close to the proteins' exposed amine (N-terminus) or carboxylic (C-terminus) end groups. Conversely, endopeptidases target internal peptide bonds of a protein, able to break down larger proteins into smaller polypeptide chains (Rawlings and Salvesen, 2012). Proteases are further divided into seven main groups based on the mechanistic features they possess, namely: Serine (EC 3.4.21.-), Cysteine (EC 3.4.22.-), Aspartate (EC 3.4.23.-), Metallo- (EC 3.4.24.-), Threonine (EC 3.4.25.-), Glutamate (EC 3.4.23.31), and Asparagine proteases (EC 3.4.23.44) (López-Otín and Bond, 2008; Placzek et al., 2017).

The active sites of the proteases are regarded to be composed of pockets termed “subsites” (Schechter and Berger, 1967). Each subsite binds a corresponding residue on the substrate. These subsites and the amino acid residues around the cleavage site of the substrate are referred to according to a particular set of rules. In this method of labelling, the cleavage site is considered the zero-point, and the substrate residues are labelled outwards from the cleavage site as “····-P₄-P₃-P₂-P₁-P₁'-P₂'-P₃'-P₄'-····” from the N- to the C- termini. Similarly, the corresponding subsites flanking the active site

of the protease outward from the cleavage site are labelled “ \cdots -S₄-S₃-S₂-S₁-S₁'-S₂'-S₃'-S₄'- \cdots ” from the N- to the C- termini (**Figure 2**).

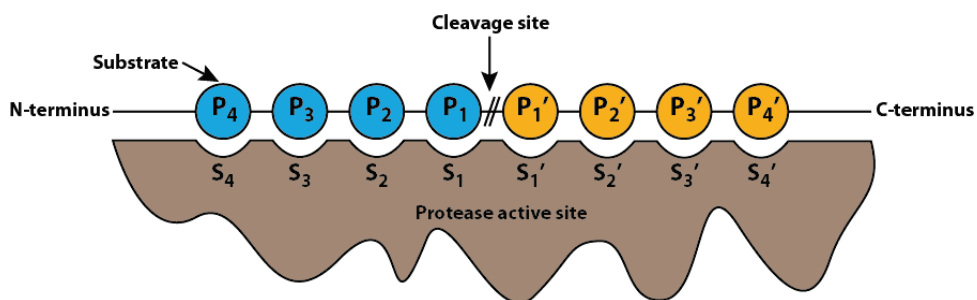


Figure 2: The naming practice of substrate residues and protease subsites around protease active sites. Substrate residues are labelled with the P_x, whereas protease subsites are labelled with the S_x lettering. The double-crossing line indicates the substrate cleavage site between the P₁ and the P₁' position. Only four residues and subsites on each end is displayed for brevity. Adapted from Schechter and Berger, 1967.

Serine proteases comprise the second most abundant group of proteases after metalloproteases (López-Otín and Bond, 2008). Their characteristic incorporation of a nucleophilic serine residue in their active site distinguishes them from other groups of proteases. The serine residue commonly acts in accordance with an aspartate and a histidine residue, forming the common catalytic triad of serine proteases (**Figure 3**) (Ekici et al., 2008). The catalytic serine typically resides in a conserved, glycine-containing peptide in the form of GxSxG (Ward et al., 2009).

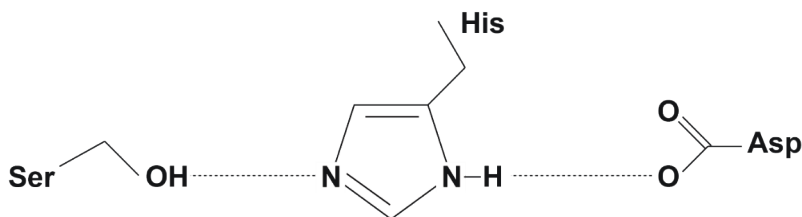


Figure 3: Residues in the active site of subtilisin-like serine proteases. Adapted from Ekici et al., 2008

Serine proteases are commonly found in eukaryotes and prokaryotes alike and are involved in many critical physiological processes. For example, subtilisins commonly act as broad-target digestive enzymes in Gram-positive bacteria (Siezen and Leunissen, 1997; Mitrofanova et al., 2017). Thrombin plays regulatory roles in platelet aggregation, endothelial cell activation and other aspects of vascular biology with the aid of protease-activated receptors (Posma et al., 2016). Granzyme B activates caspases responsible for coordinating apoptosis (Zhou and Salvesen, 1997; Jacquemin et al., 2015). A serine protease activity cascade is also noted to take part in signal transduction influencing the development of *Drosophila* embryos (Veillard et al., 2016). Some extracellular serine proteases secreted by pathogenic microorganisms are found to act as virulence factors (Gaillot et al., 2002; Muszewska et al., 2017; Martínez-García et al., 2018).

In biotechnology, many serine proteases have seen widespread uses in detergent formulations thanks to their broad substrate specificity (Niehaus et al., 2011; Vojcic et al., 2015; Salwan and Sharma, 2019). Other examples include trypsin (EC 3.4.21.4), which is used as a valuable research tool in fields such as molecular biology and proteomics (Gudmundsdóttir and Pálsdóttir, 2005; Toth et al., 2017). Further applications of serine proteases extend to dehairing of leather (Dayanandan et al., 2003; Ward et al., 2009; Zambare and Nilegaonkar, 2017), meat tenderization (Bekhit et al., 2014), in the production of protein hydrolysates as nutritional supplements to food and feed (Aspmo et al., 2005; dos Santos Aguilar and Sato, 2018) in contact lens solutions to remove protein debris from the lens and avoid eye irritation (Rejisha and Murugan, 2020), and more (Rao et al., 1998; Ward et al., 2009).

Serine proteases are typically grouped by their substrate specificities, largely defined by the substrate residue closest to the N-terminal of the cleavage site (P1). Example groups include the trypsin-like (Lys/Arg at P1), chymotrypsin-like (Phe/Tyr/Leu at P1), or elastase-like (Ala/Val at P1) serine proteases. However, some groups deviate from this pattern of categorization, including subtilisins, known for their broader substrate specificity (Siezen and Leunissen, 1997), herpesvirus type serine proteases

with a unique Ser-His-His active site organization (Chen et al., 1996) and more (Ekici et al., 2008).

1.4.1 Subtilisins

Subtilisins (also referred to as subtilases) (EC 3.4.21.62) are a group of serine proteases known for their broad substrate specificity, initially identified from the ubiquitous, Gram-positive soil bacterium *Bacillus subtilis* (Ottesen and Svendsen, 1970; Siezen and Leunissen, 1997). They comprise the S8 protease family in MEROPS, with the family type enzyme denoted as subtilisin Carlsberg from *Bacillus licheniformis* (Evans et al., 2000). Members of the S8B subgroup, exemplified by the *Saccharomyces cerevisiae* kexin (EC 3.4.21.61), have a preference for cleavage after dibasic amino acid residues (Henrich et al., 2005). Additionally, the human proprotein convertase subtilisin/kexin type 9 (PCSK9) has also been an important example for its role in cholesterol regulation and has received great attention in medicine (Lambert et al., 2009; Momtazi-Borojeni et al., 2019). Subtilisins have been studied extensively with a rich collection of structures published on the Protein Data Bank (PDB) (Berman et al., 2000). At the time of writing, 26 structures could be retrieved when searched for subtilisin Carlsberg and 63 for subtilisin BPN', with more being actively added (Wu et al., 2020; Toplak et al., 2021).

Subtilisins commonly display activity at an alkaline pH range, and act on aromatic or hydrophobic residues such as tyrosine, phenylalanine and leucine (Gupta et al., 2002). They are distinguished from other groups of serine proteases by the order (Ser, His, Asp) of their catalytic residues in the amino acid sequence. They also possess a significantly different α/β protein scaffold compared to the β/β scaffold of chymotrypsin-like serine proteases (Siezen and Leunissen, 1997). Two Ca^{2+} binding sites are commonly found in subtilisin structures, adding further stability and in some cases, thermostability to the molecule (Pantoliano et al., 1988; Smith et al., 1999). Like other serine proteases, subtilisins can be commonly inhibited with phenylmethane sulfonylfluoride (PMSF) (Powers et al., 2002). An example structure of subtilisin Carlsberg and its catalytic site can be seen in **Figure 4A and 4B** (Bode et

al., 1987), and in a similar assembly later by Radisky et al. (PDB:1TM5) (Radisky et al., 2004).

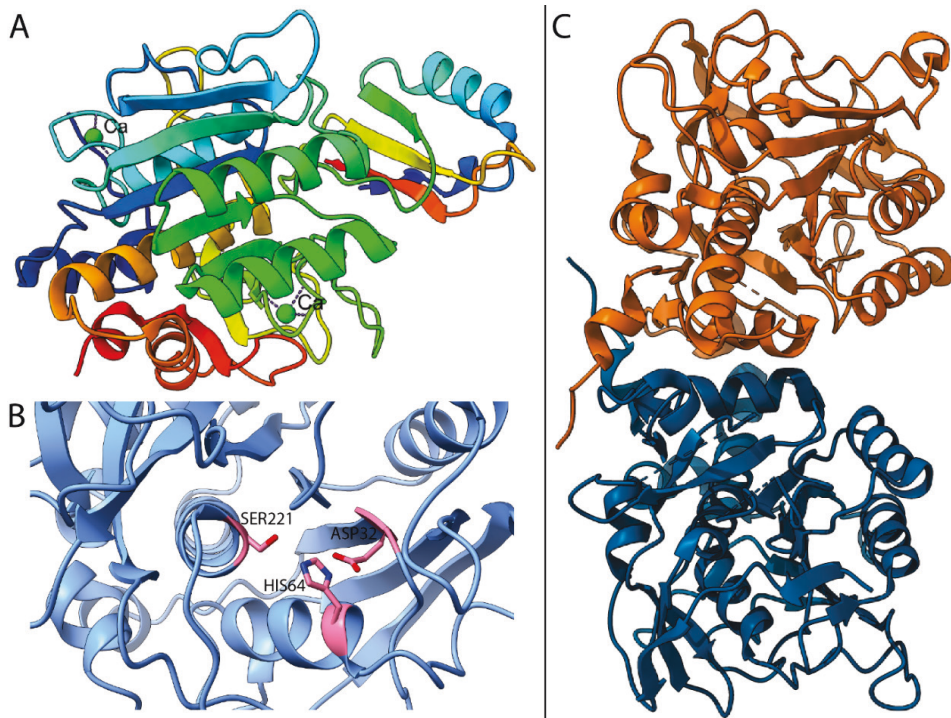


Figure 2: Three-dimensional protein structures obtained using X-ray crystallography. (A) Subtilisin Carlsberg with calcium ions depicted as green spheres, and **(B)** its catalytic triad (PDB:1CSE). **(C)** Homodimeric structure of the *Alkalihalobacillus clausii* ISP (PDB:2WV7). Structures were visualized with UCSF ChimeraX (Pettersen et al., 2021).

Domain architecture of subtilisins

Subtilisins are commonly found in the form of extracellular subtilisin proteases (ESPs). These are produced in the cell in an inactive precursor state called zymogens; and are made up of three main parts (**Figure 5**): An N-terminal leader sequence which directs their secretion outside the cell (Power et al., 1986), a short pro-domain which promotes correct folding and acts as an inhibitor (Zhu et al., 1989; Ohta et al., 1991),

and the catalytic domain (Siezen and Leunissen, 1997). Exceptionally, in some plant subtilisins, an additional C-terminal fibronectin (Fn)-III-like domain was observed, noted to be necessary for activity in some enzymes (Cedzich et al., 2009; Schaller et al., 2018). After secretion, the pro-domain is removed by autoproteolysis in order to achieve maturation and unlock catalytic activity (Power et al., 1986). Although some have been reported to be dimeric (Seki et al., 1994; Silva Lopez and De Simone, 2004), mature ESPs are usually found as monomers (Rawlings and Salvesen, 2013).

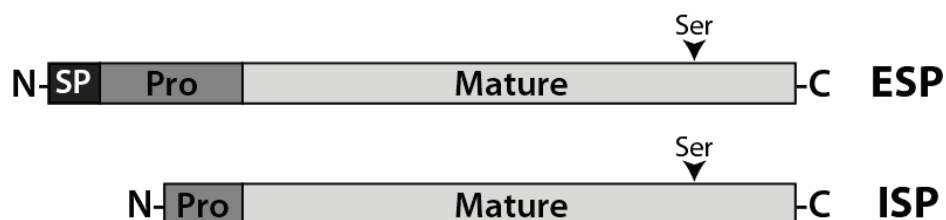


Figure 3: Comparative illustration of the domain structures of extracellular subtilisin protease (ESPs) and intracellular subtilisin proteases (ISPs). Features shown are the ESP signal peptide (SP), pro-domain and the ISP N-terminal pro-peptide (Pro), attached to the Mature protease domains, with respect to N- and C-termini.

Different from the well-studied ESPs, a lesser-known fraction of subtilisins termed intracellular subtilisin proteases (ISPs) exist. They are understood to be a major intracellular agent in *Bacilli* species, responsible for most of the casein and collagenolytic activity observed from within the cells (Orrego et al., 1973; Burnett et al., 1986). As opposed to ESPs, ISPs are found to be homodimeric, exemplified by the three-dimensional structure of *Acidobacillus clausii* ISP depicted in **Figure 4C** (Vévodová et al., 2010).

While ISPs share a 40-50 % amino acid sequence identity when compared to full-length ESP sequences (Vévodová et al., 2010), they display some noteworthy differences in their structure. Most notably, instead of the common ESP pro-domain, they feature a short (16-25 amino acids) N-terminal pro-peptide with no significant sequence homology to the ESP pro-domain (**Figure 5**). This pro-peptide covers the active site of the enzyme blocking the binding of the substrate while present. A

notable LIPY/F sequence motif conserved within known ISPs (but not observed in ESPs) is found within the pro-peptide. This motif is described to play a key role in this steric control of enzymatic activity by introducing a structural shift that prevents the direct binding of substrate to the active site of the enzyme. The pro-peptide is removed during maturation of the enzyme, allowing proteolytic activity to take place (Gamble et al., 2011).

1.5 Nucleases

Nucleases are enzymes which catalyse the cleavage of phosphodiester bonds in nucleic acids. As a subgroup of hydrolases, they are denoted with the enzyme commission number ranging from EC 3.1.11.X to 3.1.31.X. Nucleases are ubiquitous and can be found in all types of organisms. Inside the cell, they are involved with the repair, replication, and recombination of DNA (Marti and Fleck, 2004; Mimitou and Symington, 2009; Ganai and Johansson, 2016), and also in the degradation of exogenous or unnecessary DNA (Pollard et al., 2001). Outside the cell, nucleases may degrade foreign nucleic acids to use as nutrition (Beliaeva et al., 1976; Pinchuk et al., 2008), or act defensively against potentially toxic nucleic acids (Li et al., 2001; Pingoud and Jeltsch, 2001; Baulcombe, 2004; Hsia et al., 2005). Conversely, in pathogenic organisms, similar enzymes act as the aggressors, functioning as virulence factors and working to overcome the target's immune measures (Ma et al., 2017; Dang et al., 2018). In viruses, besides virulence (Kindler et al., 2017), they are also associated with gene recombination (Gammon and Evans, 2009) and the proper generation of virions after infection (Goldstein and Weller, 1998; Li and Rohrmann, 2000). In addition to discrete nucleases, many nucleic acid polymerases are also found to display secondary nucleolytic activities, associated with proofreading functions (Klett et al., 1968; Setlow et al., 1972; Lyamichev et al., 1993; Ganai and Johansson, 2016).

Due to their diversity, the classification of nucleases can be a complex matter. However, some basic properties are most commonly used to refer to many nucleases. Substrate specificity can be an important distinction, divided between targeting DNA

or RNA molecules. They are often considered in terms of their cleavage pattern as well, where exonucleases are noted to cleave singular nucleotides from the ends of the substrate, and endonucleases target phosphodiester bonds within the substrate, releasing mono- or oligonucleotide products, respectively (Yang, 2011). Furthermore, they can also be grouped by their preference for single stranded (ss) or double stranded (ds) targets, or their requirement for metal ions for their activity (Shen et al., 1997; Yang, 2011; Beese and Steitz, 2020). It must be noted that many exceptional nucleases have been reported to deviate from above classifications – such as those with non-specific activity targeting both DNA and RNA (Laskowski Sr, 1982; Rangarajan and Shankar, 2001; Hsia et al., 2005), or those with both endo- and exonuclease activities via a structural targeting preference of cleavage site, such as the Flap endonuclease 1 family (Lyamichev et al., 1993; Harrington and Lieber, 1994) and the Mre11 (Paull and Gellert, 1998) nucleases (3' to 5', and 5' to 3', respectively).

Nucleic acid modification via cleavage is an essential process, and therefore nucleases find applications within a diverse range of fields. Earlier in their study, nucleases saw use as tools for gene mapping (Southern, 1975; Holsinger and Jansen, 1993), which were applicable for DNA comparison methods in forensic analyses (Sajantila and Budowle, 1991; Balazs, 1992). They were also instrumental in the finer study of DNA methylation and accessibility (Bird and Southern, 1978; Grummt and Gross, 1980). However, some of the best-known and used examples of nucleases are restriction endonucleases. Since the reporting of *EcoRI* (EC 3.1.24.4) from *E. coli* (Yoshimori et al., 1972), hundreds more restriction enzymes were identified from diverse organisms, driving cloning and heterologous expression of genes in molecular biology. Over 670 enzymes with diverse cleavage site preferences are commercially available today (New England Biolabs, 2021).

More recently, an endonuclease associated with the clustered regularly interspaced short palindromic repeats (CRISPR) viral immunity system of prokaryotes has been described (Barrangou et al., 2007). The landmark study describing application of the RNA guided, targeted activity of the CRISPR associated protein (Cas) 9 nuclease (Jinek et al., 2012) has made possible the precise, customizable editing of genes and

genomes, earning its authors the Nobel Prize in chemistry in 2020 (Ledford and Callaway, 2020).

Nucleases with homology to three well-described archetypes were produced as part of this work (Aevansson et al. under review, see **Appendix I**). Description of the distinct features of these enzymes are provided below.

1.5.1 T4 Endonuclease V

T4 Endonuclease V (EC 3.1.25.1) is a thoroughly studied enzyme, identified from the lytic T4 Myovirus infecting *E. coli* (Ackermann and Krisch, 1997). It is a 16 kDa protein with a domain architecture comprising a pyrimidine dimer DNA glycosylase domain (Pfam: PF03013) at the N-terminal. T4 endonuclease V is associated with UV-damage repair of DNA molecules, by cleaving pyrimidine dimers at damaged sites (Tanaka et al., 1975). This is performed via a two-pronged activity: a DNA glycosylase activity cuts the glycosylic bond at the 5' pyrimidine of the dimer, generating an apyrimidic site, which is then removed by the endonucleolytic activity of the enzyme (Gordon and Haseltines, 1980; Dodson and Lloyd, 1989). Since its heterologous production (Higgins and Lloyd, 1987), and the elucidation of its structure (Vassilyev et al., 1995; Golan et al., 2006), T4 Endonuclease V became an enzyme of particular interest in dermatology. Referred to as T4N5 or Dimericine, the enzyme has undergone clinical trials and was eventually approved as a means to treat UV damage in human skin and help prevent skin cancer (Cafardi and Elmets, 2008; Zahid and Brownell, 2008).

1.5.2 Lambda (λ) exonuclease

Originating from the Siphovirus phage λ infecting *E. coli* K12 (Lederberg and Lederberg, 1953), the λ exonuclease (EC 3.1.11.3) is 26 kDa and contains a YqaJ-like viral recombinase domain (Pfam: PF09588) towards the N-terminal. The enzyme carries out exonucleolytic cleavage on preferably phosphorylated dsDNA, in the 5'- to the 3'- direction, generating a ssDNA strand and releasing nucleoside 5'-phosphate molecules; although non-phosphorylated dsDNA can also be targeted at a much

slower rate (Little, 1967). This activity requires alkaline conditions, and Mg^{+2} ions as a co-factor, associated with two metal-binding sites within the protein (Little et al., 1967; Carter and Radding, 1971). The structure of the enzyme is described as a toroid formed by three sub-units, able to accommodate and act upon DNA molecules through the central cavity (Kovall and Matthews, 1997), with a central fold reported to be conserved among other nucleases, including type II restriction endonucleases (Kovall and Matthews, 1998). Also referred to as “ λ recombinase”, λ exonuclease is associated with the single strand annealing homologous DNA recombination system of dsDNA break repair in phage λ with the aid of Red β annealase (Radding and Shreffler, 1966; Weller and Sawitzke, 2014). Besides its application in molecular biology, λ exonuclease has also been approved as an investigational drug for the treatment of spinal muscular atrophy under the trade name Spinraza and nusinersen (Haché et al., 2016; Chiriboga, 2017).

1.5.3 Exonuclease III

Exonuclease III (Exo III) (EC: 3.1.11.2) is a multifunctional enzyme identified from *E. coli* K12 (Richardson and Kornberg, 1964; Richardson et al., 1964). Besides its 3'-5'-exonuclease activity, Exo III exhibits phosphatase, ribonuclease H, and endonuclease activities, acting on apurinic or apyrimidinic DNA substrates, likely all from a single active site (Weiss, 1981; Mol et al., 1995). Exonuclease activity is dependent on Mg^{+2} ions (expected to be two per molecule), and it degrades dsDNA (but not ssDNA) in a 3' to 5' direction, releasing 5' phosphomononucleotides as products (Richardson et al., 1964; Mol et al., 1995). The enzyme does not need a blunt dsDNA end for cleavage initiation as it can also act on nicked dsDNA sites and on circular substrates (Thomas And and Olivera, 1978). The inclusion of a short 3'-overhang can be used to prevent cleavage of dsDNA by Exo III (Ding et al., 2019). The 31 kDa enzyme has an N-terminal nuclease domain (Pfam: PF03372), that is shared among other magnesium dependent nucleases and phosphatases (Mol et al., 1995; Dlakić, 2000). Thanks to its broad range of activities, Exo III is noted to be a key enzyme for DNA repair within *E. coli* (Mol et al., 1995; Lovett, 2011). Today, Exo III is commercially available, commonly used in molecular biology for its

exonuclease activity, with particular applications in generating ssDNA templates towards dideoxy sequencing, and intermediates for site-directed mutagenesis (Henikoff, 1984; Vandeyar et al., 1988; Lovett, 2011).

1.6 Viral Endolysins

Endolysins (lysins, peptidoglycan hydrolases) refers to a group of enzymes facilitating the degradation of the bacterial cell wall after viral replication, enabling the release of viral progeny. In dsDNA bacteriophages, endolysins are specialized hydrolytic enzymes targeting one of the 4 main bonds of the peptidoglycan (PG) (also called murein) layer of the bacterial cell envelope (BCE) (Fernandes and São-José, 2018). This layer is present in nearly all bacteria atop the cell membrane (CM), and consists of glycan chains, anchored together by a matrix of cross-linking peptides (Vollmer et al., 2008).

The makeup of the BCE varies between different types of bacteria. In simpler Gram-positive bacteria, BCE consists of the CM and a thicker PG layer which defines the cell wall (CW). Gram-negative bacteria, however, feature a thinner PG layer, but it is delimited on the outside by the lipid bilayer outer membrane (OM) (Silhavy et al., 2010). The space between the PG layer and the OM defines the dense, aqueous periplasm layer, rich in proteins (**Figure 6**) (Mullineaux et al., 2006). Furthermore, the PG in both groups may also contain other proteins that bind to the layer, such as the covalently attached teichoic acids in Gram-positive species (Brown et al., 2013).

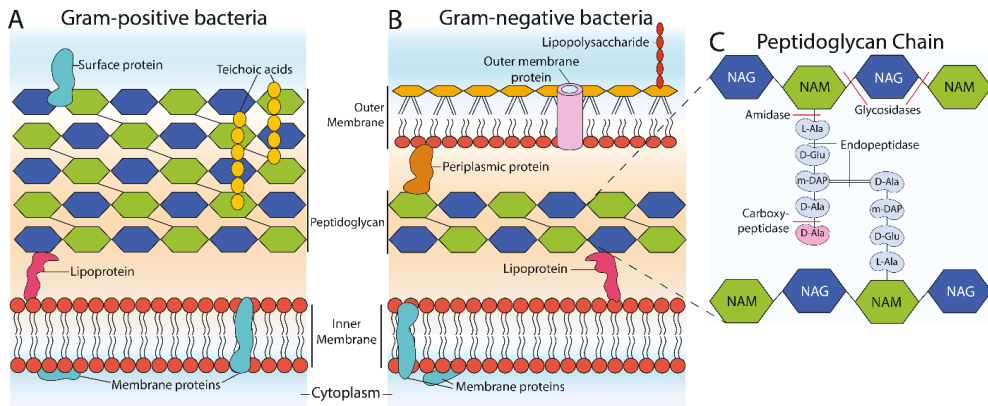


Figure 4: Cell wall architectures of (A) Gram-positive and (B) Gram-negative bacteria. (C) Illustration representing the repeating N-acetylglucosamine-N-acetylmuramic acid chains (NAG – NAM, in blue and green, respectively) cross-linked to form the bacterial PG structure, and various endolysins targeting specific bonds. Adapted from Alcorlo et al., 2017 and Fernandes and São-José, 2018.

In closer detail, the glycan strands of the PG are made up of alternating *N*-acetylglucosamine (NAG) and *N*-acetylmuramic acid (NAM) residues, linked via $\beta(1-4)$ bonds. A short peptide “stem” is also attached to each NAM residue, in place of the D-lactoyl group. These peptides are reported to often consist of the residues L-Ala-g-D-Glu-m-DAP-D-Ala-D-Ala (m-DAP for 2,6-diaminopimelic acid) in Gram-negative bacteria, and L-Ala-g-D-Glu-L-Lys-D-Ala-D-Ala in most Gram-positive bacteria (Vollmer et al., 2008). The cross-linking of the glycan chains is achieved by bonds between amino acid residues at positions 3 (commonly L-Lys or m-DAP) and 4 (D-Ala), via a peptide bond in Gram-negative bacteria, or via interpeptide bridges typically in Gram-positive bacteria (Vollmer et al., 2008). While this common structure defines the basis of the layer, the PG layer is an editable structure, where modifications such as N-deacylation, O-acylation, amidation or amino acid substitutions are understood to take place as a response to changing conditions in and outside of the cell (Vollmer, 2008; Cava and de Pedro, 2014).

Endolysins are often divided into three major groups, based on the bond they target within the PG structure. Namely, these are peptidases, amidases, and glycosidases (Fernandes and São-José, 2018). Peptidases break the peptide bonds between amino acid residues and can include endopeptidases which cleave any of the internal peptide bonds, or carboxypeptidases which detach C-terminal amino acids. Amidases cleave the amide bond between the first amino acid residue (usually L-Ala) of the peptide stem and the NAM (EC 3.5.1.28). Glycosidases act on one of the two glycosidic bonds present between glycan units and can be grouped as N-acetyl- β -D-glucosaminidases (glucosaminidases, EC 3.2.1.30), N-acetyl- β -D-muramidases (muramidases or lysozymes EC 3.2.1.92) and lytic transglycosylases (EC 3.2.1.-) (Alcorlo et al., 2017).

Some viral endolysins bear affinity to different sections of the PG by encoding one or more secondary carbohydrate binding domains. A ubiquitous example is the short Lysin Motif (LysM) domain, which binds to the NAG units in the glycan chain of the PG (Buist et al., 2008; Mesnage et al., 2014). However, as these enzymes are expressed without signal peptides to direct their translocation past the CM, they lack the means to reach the PG without aid. Instead, access to the PG layer is enabled by another class of proteins: the holins. At mid-late phases of phage replication these transmembrane proteins bind to and “form holes” within the CM and compromise its integrity (Wang et al., 2000; Catalão et al., 2013). This allows endolysins to act on the PG and also leads to cell death by causing the collapse of the proton motive force, which drives many essential cell functions (Rice and Bayles, 2008; Cahill and Young, 2019). The viral lysis acting against different bacteria may involve additional associated proteins, and their mechanism may therefore differ from the underlying foundation explained above. The current opinion on these mechanisms and all associated proteins is reviewed in great detail by Fernandes and São-José (Fernandes and São-José, 2018), and also by Cahill and Young (Cahill and Young, 2019).

The application of phages for the control and elimination of bacterial pathogens has been one of the earliest thoughts of viral applications of biotechnology (Ho, 2001). With the era of (meta)genomics and improved access to viral gene resources,

therapeutic use of phages and phage-derived proteins once again garnered high research interest (Dixon, 2004; Matsuzaki et al., 2005). Endolysins' natural bactericidal effects are therefore considered an obvious extension for the concept of phage therapy. Numerous studies have already demonstrated the effects of the potential application of endolysins (Nelson et al., 2001; Briers et al., 2014; Islam et al., 2019). As an alternative to antibiotics treatments, these enzymes could offer higher specificity, and no known toxicity and or triggering of bacterial resistance against their effects (Schmelcher et al., 2012; Rodríguez-Rubio et al., 2013). These proteins could act on antibiotic resistant pathogens, overcoming one of the major medical challenges of the era of antibiotics (Gupta and Prasad, 2011; Zhang et al., 2013; Plotka et al., 2019). In addition, further applications of endolysins have been reported towards protein purification (Joshi and Jain, 2017), crystallography (Boura et al., 2017), and food preservation (van Nassau et al., 2017). For the discovery of many such enzymes, viral genomes appear as an exceptionally rich source (Fernández-Ruiz et al., 2018; Santos et al., 2018), and highlight the possibilities present in the study of viral sequence space.

2. Aims of study

The main goal of this project was to identify, produce and characterize the functional and structural qualities of various enzymes of marine microorganisms, with activities potentially relevant for biotechnology applications.

Additional sub-goals were noted as:

- To assemble a versatile process for the efficient heterologous production of candidate enzymes in *E. coli*, tailored to promote protein solubility.
- To achieve soluble heterologous production of viral proteins in *E. coli*, and investigate the application of the emerging “codon harmonization” approach on proteins produced
- To achieve the identification, production, and characterization of multiple enzymes from psychrophilic and thermophilic microorganisms.

3. Materials

All relevant materials used in each study are described in detail in their respective papers. The marine genetic resources used in this study are sourced as described below:

Planococcus sp. AW02J18 (1379956) featured in **Paper II**, was isolated from a biota sample 135 m below the surface, in the coastal area of Lofoten (68.5025473N°, 015.0046585E°) in 2009. Identified through 16S rRNA gene sequence analysis, it was provided by the bacterial collection at the University of Tromsø (De Santi et al., 2016).

The viral nucleases studied herein were extracted from viromes generated within the Virus-X project (Aevarsson et al. under review, see **Appendix I**). Sampling was carried out on transects from Spitsbergen to the Arctic ocean, Spitsbergen to the Fram Strait and Southeast to the Jan Mayen Fracture Zone at depths of 1500m and - 3594m (Sandaa et al., 2018, Aevarsson et al. under review, see **Appendix I**).

Hypnocyclicus thermotrophus featured in **Paper III** was isolated from a microbial mat situated in a hydrothermal vent field at the Northern Kolbeinsey Ridge, 166 km west of Jan Mayen in the Greenland Sea (Roalkvam et al., 2015).

4. Methods

Detailed materials and methods associated with each part of the study is presented in their respective papers (**Papers I-III**). However, experimental details about the production of viral nuclease candidates are provided below.

4.1 Heterologous production of viral nuclease candidates

The viral nuclease genes presented within this thesis (**Supplementary Table 1**) were identified and selected by collaborators within the Virus-x consortium using a multi-faceted approach (summarized in **Figure 4** of Aevansson et al. under review, see **Appendix I**). This integrative approach featured sequence homology comparisons as well as more advanced HMM-HMM profile similarity searches against multiple databases. These included sequence (Pfam (Finn et al., 2016), GenBank (Clark et al., 2016)) functional (KEGG (Ogata et al., 1999), GO (Ashburner et al., 2000)) and structural (PDB (Berman et al., 2000)) knowledgebases. These resources were co-implementation with the help of contemporary tools such as HH-suite3 (Steinegger et al., 2019), 3DM systems (Kuipers et al., 2010) and the EMGB annotation browser (Jünemann et al., 2017).

All expression constructs were transformed into *E. coli* BL21-Gold(DE3)pLysS (Merck, Darmstadt, Germany) cells using the heat-shock protocol provided by the manufacturer, using 30 ng of plasmid per 15 μ L of bacteria suspension. Single colonies were picked from Lysogeny Broth (LB)-agar plates containing 100 μ g/mL ampicillin after plating and overnight growth at 37 °C, and 4 ml pre-cultures in LB broth with 100 μ g/mL ampicillin and 50 μ g/mL chloramphenicol were subsequently inoculated and incubated overnight at 37 °C with 220 rpm shaking. The next day, 4 ml expression cultures in LB media containing 100 μ g/mL ampicillin were inoculated with 100 μ l of each pre-culture and were grown at 37 °C and 220 rpm until an optical density at 600 nm of 0.5-0.6 was reached. The temperature of the incubator was then reduced to 30 °C and allowed to equilibrate for 30 minutes. Expression was induced with the addition of 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG), and

cultures were allowed to further grow for 4 hours. Following expression, cells were harvested by centrifugation at 5000 x g at 4 °C for 10 min. Collected cells were re-suspended in 1 ml of lysis buffer containing 50 mM HEPES pH 7.5, 300 mM NaCl, 10% Glycerol and were lysed using ultrasonication performed at 4 °C using 4 x 10 second bursts at 10 second intervals, with 27% amplitude. An aliquot representing the total protein fraction was taken and stored at 4 °C from each crude lysate before clarification of lysates by centrifugation at 12000 x g at 4 °C for 3 min. After clarification, aliquots were taken from all samples representing the soluble protein fraction and stored at 4 °C.

Aliquots taken from lysed cell pellets, representing the total protein and soluble protein fractions were run on a gradient (8-16%) SDS-PAGE gel (**Supplementary Figure 1**) (GenScript, Piscataway, NJ, U.S.A.) to assess expression levels. Precision Plus Dual Color (Bio-Rad, Hercules, CA, U.S.A.) protein ladder was used for protein molecular mass determination. Equivalent volumes of lysate (16 µl) were mixed with sample buffer (4 µl) and loaded onto the electrophoresis. The gel was run at 200 V, and subsequently stained using InstantBlue (Expedeon, Cambridge, UK) using staining protocol provided by the manufacturer. After staining was complete, unbound dye was washed off the gel using distilled water on a benchtop shaker to reveal protein bands.

All IMAC protein purification experiments were carried out on an ÄKTA start (Cytiva, Uppsala, Sweden) system equipped with 5 ml HisTrap FF Ni-NTA (Cytiva, Uppsala, Sweden) columns, set up in a cold room at 4 °C. Clear lysate samples were prepared in equilibration buffer (Buffer A) which contained 20 mM HEPES pH 7.5 and 500 mM NaCl; and were loaded with a speed of 5 ml/min into the column pre-equilibrated with Buffer A. The elution buffer (Buffer B) was prepared to contain 20 mM HEPES pH 7.5, 500 mM NaCl and 500 mM imidazole. After washing unbound proteins away with 25 ml Buffer A, bound proteins were eluted stepwise with 10 ml of 100% Buffer B, and fractions were collected for analysis with SDS-PAGE (**Supplementary Figure 2**). Purified protein fractions were pooled, and proteins were shifted into a new buffer containing 20 mM HEPES pH 7.5, 300mM NaCl, 2mM

TCEP (tris(2-carboxyethyl)phosphine) and 30% Glycerol (v/v) with Amicon Ultra - 15 centrifugal filters (Merck Millipore, Cork, Ireland) to be stored at -20 °C.

5. Results and discussion

The main results of each study carried out in this thesis, are presented within their respective papers. The cloning, heterologous expression, and biochemical as well as structural characterization of bacterial serine proteases, are presented in **Paper I** and **II**. The results pertaining to the discovery of the *H. thermothrophus* prophage, its analysis and heterologous expression of its genes are presented and discussed in **Paper III**. Results of the heterologous expression trials of viral nuclease candidates, performed in the context of Virus-X are summarized in **Section 5.3.2**, with relevant tables and figures provided under supplementary data.

Within this thesis, a wide group of enzymes were targeted for discovery, and subsequent heterologous expression trials. These included bacterial serine proteases, as well as endolysins, nucleases and other nucleic acid-processing enzymes of viral origin. Proteases were chosen as a target group of enzymes within the enzyme discovery drive of the NorZymeD project, which aimed for their application in the valorization of by-products of food production. As one of the major sectors of the Norwegian food industry, side-cuts from fish farming was seen as an available resource which could be degraded by proteases to produce peptides with nutritional and economical value (Aspmo et al., 2005; Chalamaiah et al., 2012; NorZymeD, 2021). In the case of the viral enzymes, the viromes made accessible through the Virus-X project presented a near pristine genetic resource for bioprospecting. Therefore, “low-hanging” enzymes, expected to be particularly abundant and diverse within viral genomes were targeted. Endolysins, nucleases, DNA polymerases and similar nucleic-acid processing enzymes are not only of profound ecological interest, but also have the potential to provide valuable tools and applications within academia, medicine, and other fields (Freitag-Pohl et al., 2019; Dorawa et al., 2020; Plotka et al., 2020, Aevansson et al. under review, see **Appendix I**).

5.1 Bioinformatical annotation and selection of target genes

The functional annotation of genes in metagenome libraries still largely depends on our knowledge of previously described proteins and genes, and many genes remain poorly annotated (Fernández-Arrojo et al., 2010; Steinegger et al., 2019). The annotation of metagenomic viral sequences is particularly challenging, due to their inherent diversity and the lack of known precedence for many (Krishnamurthy and Wang, 2017). The lack of an elegant solution to this challenge consequently makes it difficult to mine genetic resources from both terrestrial and marine environments (Howe et al., 2014; Gregory et al., 2019). For an enzyme discovery study, the class of enzymes being sought after, and their native host are key factors defining the complexity of the annotation step and will inform the optimal combination of tools to achieve best results.

Focused databases for particular enzyme classes or activities are valuable assets for the identification of new enzyme targets, as they provide an easily consultable, collected resource for relevant enzymes. For the selection of protease candidates in **Paper I** and **II**, the extensive MEROPS protease database was utilized (Rawlings et al., 2012, 2014, 2018). This process yielded relevant subtilisins for the verification of the expression procedure reported in **Paper I** and the discovery of the *Planococcus* ISP in **Paper II**. Expected proteolytic activity was demonstrated on both accounts, and the MEROPS classification provided helpful contextual insight on other features of proteases such as co-factors, inhibitors as well as structural information. Determination of similar protease targets could be performed without MEROPS today, as resources are available that can examine sequences and search for indicators of desired activity – such as Pfam (Finn et al., 2016) or Interpro (Blum et al., 2021) domains, particular nucleotide, or amino acid sequence identity with a benchmark enzyme, or with the utilization of 3DM super-alignments (Kuipers et al., 2010). Sequence alignment searches against MEROPS and similar databases was an efficient way to detect relevant sequences, but modern tools, such as HH-Suite3, can perform fast, and broader analyses of genes with the computing power available for use today (Steinegger et al., 2019).

For a thorough identification and annotation genes of viral origin within the genome of *H. thermotrophus* in **Paper III** multiple approaches were needed. The PHAge Search Tool Enhanced Release – PHASTER (<https://phaster.ca/>) tool was used for its targeted approach in detecting viral genes not readily picked up by the prokaryotic annotation pipeline of NCBI (Arndt et al., 2016; Li et al., 2021). PHASTER predicts phage genes via homology searches against a custom database, comprising sequences from NCBI phages, a dedicated prophage database (Srividhya et al., 2006), and bacterial genomes previously assessed via the tool (Bleriot et al., 2020; Plotka et al., 2020). Although the referred prophage database no longer appears to be supported (Srividhya et al., 2006), it's understood that the PHASTER database includes the contained sequence information necessary for the analysis.

Compared to other tools for prophage detection, such as PhiSpy (Akhter et al., 2012) and Prophinder (Lima-Mendez et al., 2008) PHASTER offers better usability, and prediction power; although the newer PhageWeb tool is reported as an efficient alternative (Sousa et al., 2018). Different from other tools, PhageWeb also provides additional analyses such as G+C content, presence of tRNAs and also reports hits against public protein databases such as UniProt (Bateman et al., 2021), Pfam (Finn et al., 2016) and Interpro (Blum et al., 2021) which can help inform the functional profile of the prophage genome, and its relationship with the host organism (Sousa et al., 2018).

In **Paper III** the PHASTER annotations were complemented with other tools in order to produce thorough annotation of genes, and best inform the study. Besides the *H. thermotrophus* gene annotations presented on GenBank, and those produced by PHASTER, the EggNog-mapper (<http://eggNog-mapper.embl.de/>) tool was tested, which specializes in reporting orthology relationships between genes and their functional annotations (Huerta-Cepas et al., 2017, 2019). In addition the HHsearch (Söding, 2005) feature of the HH-suite3 web server (<https://toolkit.tuebingen.mpg.de>) was used, which is able to represent and compare both the query and the target proteins as HMM profiles, offering higher resolution comparisons of homology against public sequence and structure databases such as Pfam (Finn et al., 2016),

UniProt (Bateman et al., 2021) and PDB (Zimmermann et al., 2018; Steinegger et al., 2019). While PHASTER was able to detect the presence of viral genes, it was not able to provide thorough annotation of many genes that was not part of the viral backbone. EggnoG-mapper was able to provide better annotations on a few select genes but did not significantly improve the resolution of the viral annotations, likely due to the inherent lack of orthologs of viral genes. Compared to the other approaches, the HHsearch tool was able to provide the most thorough annotation of HTH1 genes, significantly reducing the number of hypothetical genes in the annotation, and also highlighting previously unremarked putative activities of some genes (**Supplementary Tables 1 and 2, Paper III**).

For the inspection of single genes and their putative domain structures, the HMMER web server (<https://www.ebi.ac.uk/Tools/hmmer/>) has been utilized, which can display additional annotations from Pfam (Finn et al., 2016) and Interpro (Blum et al., 2021) for protein domains as well as SignalP (Petersen et al., 2011; Almagro Armenteros et al., 2019) for the presence of signal peptides and transmembrane regions (Finn et al., 2011; Potter et al., 2018). This analysis was used to inform the heterologous expression experimental design, as signal peptides, transmembrane regions, and differences in domain structures may call for truncations, or different placement of His-tags and expression vector systems (Kwon et al., 2011, **Paper I**).

5.2 Phylogeny and taxonomy analyses of prophage genes

Paper III reports on the discovery of a prophage gene cluster within the genome of *H. thermotrophus*, a free-living, Gram-negative, thermophilic bacterium sourced from a vent field in the Arctic Mid-Ocean Ridge (AMOR) (Roalkvam et al., 2015). Analysis of the bacterial genome using PHASTER (Arndt et al., 2016) led to the detection of prophage-associated gene regions. Closer inspection of their functional annotations using multiple approaches, suggested the presence of a 41.6 kbp long prophage region, containing 46 protein coding genes; referred to as *Hypnocyclicus thermotrophus* phage H1 (HTH1). In this study, HTH1 was found to be taxonomically associated with the viral family *Siphoviridae*, via multiple bioinformatical analyses.

Traditional viral taxonomy classifies viruses by their morphology, and their nucleic acid structure (Ackermann, 2007; King et al., 2011). Considering the great diversity of the virosphere, the International Committee on Taxonomy of Viruses (ICTV) communicated support towards sequence-based assignments of taxonomy over phage genomes, with thorough bioinformatical analyses (Simmonds et al., 2017). Soon after, the committee also expanded the existing scheme of viral classification (Wildy, 1971; Francki et al., 1991) to a wider hierarchy in hopes of accommodating this rising exploration of viral genetic diversity (Gorbalenya et al., 2020).

As the lytic induction, isolation, and imaging of HTH1 was not possible within the scope of **Paper III**, multiple bioinformatical approaches were employed to thoroughly assess the taxonomy of HTH1 based on the present sequence information. At the genome-level, all prophage genes were analysed using the MEGAN software (Huson et al., 2016) in order to perform lowest common ancestor (LCA) affiliation analysis. The use of this approach has been previously reported for the taxonomy of marine viruses (Roux et al., 2016), and also specifically in extreme hydrothermal vent environments (Castelán-Sánchez et al., 2019). A comparison of the phage head-neck-tail module genes to existing viral clusters via the Virfam tool (<http://biodev.cea.fr/virfam/>) was also utilized, which leverages the ACLAME database of mobile genetic elements (Leplae et al., 2010; Lopes et al., 2014).

Additional phylogeny analyses on HTH1 at a gene level was performed to gain further insight on its relationship with other viruses. Unlike prokaryotes, where the sequencing of 16S rRNA have become the norm for examining phylogeny (Lane et al., 1985; Pace et al., 1986) viral sequences were long considered to lack similar genes until more genomes were available for study (Rohwer and Edwards, 2002). More recently, a set of orthologous gene groups were reported as “signature genes” by Kristensen and co-workers for the detection of some viral taxa (Kristensen et al., 2013). Herein, holin was identified as one such gene for members of T1-like Siphoviruses; and HTH1 was found to contain a gene annotated as such. As the nucleotide sequence diversity of the gene was found too high for comparisons against gene repositories, the amino acid sequence instead was used to assess protein-level

phylogeny of HTH1 holin homologs. Genome-based phylogeny approaches have also been previously used to compare closely related viruses (Rokas et al., 2003; Olsen et al., 2020). However, as a larger dataset of viruses with multiple homologous genes would be required to carry out a thorough analysis of HTH1, this analysis was not included in **Paper III**. Alignment-free phylogenomic approaches have also been developed (Zhang et al., 2017), which could potentially be applied to analyse HTH1 in a future study.

An analysis was performed using sequences extracted from the viral subset of NCBI nr, queried for similar proteins to HTH1 holin via protein BLAST (blastp) (Altschul et al., 1990). Hosting a rich database of well-described viral isolate genomes, hits against this database were chosen in an effort to improve confidence in the taxonomic assignment of HTH1 (**Figure 1 of Paper III**). This search was further extended to the IMG/VR (Paez-Espino et al., 2016) and Ocean Gene Atlas (Villar et al., 2018) databases to look for similar phages in environmental metagenomes (**Supplementary Figure 2 of Paper III**).

The combined results of different analysis approaches provided generally congruous insight into the taxonomy of HTH1, pointing at closest association with *Siphoviridae* bacteriophages infecting the phylum *Firmicutes*. Interestingly, none of the tools used indicated close similarity to known viruses of *Fusobacteria* or other Gram-negative bacteria, indicating a novel lysogenic interaction seen in *H. thermotrophus*. Without access to morphological data, application of multiple approaches was necessary to suggest a confident sequence-based classification of the prophage. However, it remains interesting to pursue a morphological study of HTH1 in the future. Furthermore, network-based, rather than tree-based representations of viral phylogeny have recently been highlighted (Dion et al., 2020), and phylogeny analysis of a wider viral sequence space around HTH1 may yield further insight into its relation to other viruses.

5.3 Approaches towards heterologous production of soluble proteins in *E. coli*

Throughout this work, *E. coli* has been the expression host used for the screening and characterization of proteases (**Paper I** and **II**), as well as for exploring viral genes both from arctic viromes (**Sections 4.1, 5.3.2**, and Aevansson et al. under review, see **Appendix I**) and the HTH1 prophage (**Paper III**).

5.3.1 Proteases and solubility tags

Proteases are considered challenging proteins to heterologously produce, as they are prone to inclusion body formation, and undesired proteolytic activity by foreign enzymes under production may lead to rapid cell toxicity and greatly reduced yields (Komai et al., 1997; D'alessio et al., 1999; Tang et al., 2004; Li and Li, 2009; Dutta et al., 2010; Pushpam et al., 2011). For the expression of protease candidates in **Papers I** and **II**, experimental design choices were made in an effort to overcome these challenges.

In **Paper I**, the goal was to assemble a streamlined and efficient procedure for the successful expression and activity screening of proteases. The FX cloning approach was used to insert candidate sequences into the cloning vector pINITIAL (Geertsma and Dutzler, 2011), employing multiple selection genes for high cloning efficiency. Candidate genes were then sub-cloned into six expression vectors based on the arabinose inducible pBAD vectors (Geertsma and Dutzler, 2011). All six vectors employed a His-tag at either the N- or C-terminus of candidate genes for the efficient purification of proteins downstream via IMAC. In addition, four vectors also encoded the proteins MBP (Kapust and Waugh, 1999) or SUMO (Malakhov et al., 2004) at the N-terminus to enhance soluble yields of heterologous proteins. Validation experiments for the screening procedure was carried out by demonstrating the production of soluble and active subtilisins in **Paper I**. An expected solubility-enhancing trend was observed from many but not all of the constructs featuring MBP and SUMO tags. Hence, the value of using multiple constructs to test each candidate was underlined to potentially increase the chances of obtaining soluble and active enzymes.

The *Planococcus* ISP featured in **Paper II** was discovered by using the procedure detailed in **Paper I** to screen shortlisted target genes for proteases, sourced from sequence-based mining of marine sequences. The enzyme was then expressed in soluble form in all six expression vectors tested, successfully purified, and further characterized both for its function and structure. Although from a reportedly cold-adapted host, the ISP was found to function optimally at pH 11 and at 45 °C with an active range from pH 7.0 to 11, with no activity observed above 60 °C. Furthermore, the crystal structure of the mature enzyme was determined by X-ray crystallography to a 1.3 Å resolution, which became the highest resolution ISP structure reported to date; and the first one with a native catalytic triad (PDB: 6F9M). Together, these findings allowed the examination of the interplay of structure and function for the regulation of proteolytic activity in the *Planococcus* ISP, contributing new insights towards ISPs as a whole. The biotechnological application potential of the enzyme, however, remains to be assessed. A new ISP from *Bacillus velezensis* SW5 has recently been described by Yang and co-workers, which report significant proteolytic activity of the enzyme against fibrin (Yang et al., 2020). Testing the *Planococcus* ISP on a similar panel of substrates would be a valuable component for such an analysis.

The procedure presented in **Paper I** lies parallel to a similar pipeline developed for protease expression (Kwon et al., 2011) where rapid and efficient cloning via Gateway Cloning (Katzen, 2007) and the use of multiple expression vectors with purification and solubility tags were featured. FX- and Gateway Cloning are comparable approaches for moving target genes into vectors, providing rapid and confident cloning with the use of directional ligation and multiple selection factors for positive clones - including negative selection by the *ccdB* gene in incorrect clones (Bernard et al., 1994; Scholz et al., 2013). However, as Gateway is a proprietary method, FX-cloning remains the more affordable option without sacrificing cloning performance. Tight control of expression is exceptionally important when expressing proteases. The functionality of this approach while using expression vectors featuring the T7 promoter, combined with the choice of the *E. coli* BL21(DE3)pLysS strain in order to

ensure a tightly controllable expression with IPTG induction has been demonstrated (Kwon et al., 2011).

The SUMO tag is a smaller protein and was reported to outperform MBP and other well-established proteins such as glutathione S-transferase (GST), thioredoxin (Trx) and the transcription termination anti-termination factor (NusA) in promoting soluble expression (Marblestone et al., 2006). This was found to be the case in the results of **Paper I (Figure 4)**, where N-terminal SUMO fusion constructs of subtilisin variants yielded the highest proteolytic activity against fluorescently tagged casein. To available-knowledge, this study remains the only example in which the SUMO tag was used for the successful soluble production of subtilisin.

5.3.2 Codon adjustment strategies for the expression of viral genes

As a part of this thesis, production trials were carried out for 42 viral enzymes with putative nuclease activities, chosen by Virus-X partners (**Supplementary Table 1**). Expression constructs were designed based on protein domain analyses, placing the His-tag at the N- or C-terminus, in plasmids pET-3a and pET-21a, respectively (Novagen Inc., 2006). Synthesis and CO of genes was performed through the GenSmart pipeline at the synthesis stage, offered by GenScript (GenScript, 2021). Genes were subsequently obtained synthetically in chosen plasmids and tested for expression. The *E. coli* BL21-Gold(DE3)pLysS strain was chosen for expression in order to hinder potential cytostatic effects which may be caused by leaky gene expression. Among the 42 nucleases tested, only 12 were detected in soluble forms (**Supplementary Figure 1**). Six of the soluble candidates were also able to be purified using IMAC on single experiments (**Supplementary Figure 2**).

Codon optimization has been demonstrated to produce a varying benefit in previous studies targeting expression of viral proteins in *E. coli*. While successful soluble production of viral structural proteins has been reported with the help of CO (Lee et al., 2011), its benefits do not appear guaranteed. For the soluble expression of VLP1 protein from mouse polyomavirus various approaches were tested, where best yields were reportedly achieved without CO, but with the use of *E. coli* Rosetta(DE3)pLysS

cells, which express rare tRNAs (Chuan et al., 2008). Here, use of a GST-fusion construct was also tested, but was reported not to promote significantly higher yields of soluble proteins (Chuan et al., 2008).

Existing information regarding the heterologous expression of viral metagenomic sequences is scarce. In one such study, when 110 ssRNA virus coat proteins from metagenomic sequences were produced in *E. coli*, an approximately 40% protein solubility rate was observed without CO or other experimental optimizations (Liekniņa et al., 2019). This success rate is comparable to the one reported by the Virus-X consortium in Aevansson et al. (under review, see **Appendix I**), and slightly higher than the ratio of soluble heterologous nucleases in this work. Herein, CO was used as a convenient approach to potentially improve the soluble production of viral nucleases at the synthesis stage. Due to constraints of time, requirements of industrial partners and the number of nuclease candidates tested, it was not feasible within the time frame of this project to finely tune expression conditions for each nuclease family or implement fusion tag solutions (similarly to **Paper I**). Fusion tags have been utilized for select targets within the Virus-X project (Aevansson et al. under review, see **Appendix I**), and with further optimization of expression parameters, it is likely that more nucleases can be produced in soluble form.

In **Paper III** a set of 9 prophage genes from HTH1 putatively related to lysis and nucleic acid processing activities were selected for heterologous expression trials in *E. coli*, with an experimental design largely shared with the nuclease candidates. Here CH variants for each chosen gene were also implemented to compare the two approaches for their benefits in expressing viral genes. As harmonization of genes is not similarly available from gene synthesis providers, the design of CH candidates for the experiments presented in **Paper III** were carried out manually. The Codon Harmonization Tool (Claassens et al., 2017), based on the original harmonization algorithm (Angov et al., 2008) was utilized for this step. All nine candidates tested were able to be produced recombinantly, however only five proteins were found soluble from the CO variants. Four were also observed as soluble from their CH variants. On SDS-PAGE images, CO gene variants of candidates were found to yield

higher amounts of soluble protein compared to their CH versions. The four proteins produced in soluble form from both variants, N-acetylmuramoyl-L-alanine amidase, a putative rRNA biogenesis protein, a putative DNase, and a hypothetical protein; were subsequently purified using IMAC. Finally, the thermostability of both protein variants were analysed using differential scanning fluorometry where all CH variants was observed to display higher thermostability than their CO counterparts.

In the case of HTH1 proteins, comparison of the two codon adjustment approaches showed that CO led to the production of higher soluble protein yields, yet CH was able to promote higher thermostability in target proteins. It can be argued that this increased stability indicated a higher folding quality, in line with the expected effect of the harmonization approach reported previously (Angov et al., 2011; Wen et al., 2020). The results suggest that CH could be a particularly preferable approach when structural analyses of the proteins are sought after the production step. It may also be speculated that the improved folding quality would positively affect the activity of the produced enzymes, although further work on functional characterization is necessary to conclusively assess the subject. Furthermore, **Paper III** represents the first application of CH for the expression of viral genes to the best of current knowledge and may provide insight for similar studies in the future.

5.3.3 Considerations for designing new experiments

In this thesis, multiple approaches for generating soluble protein expression in *E. coli*, on previously known genes from publicly available databases (**Paper I**), genes from isolated genomes (**Paper II** and **III**), or viral metagenomes (**Section 4.1**, **5.3.2**, and Aevansson et al. under review, see **Appendix I**). It is clear that it would have been beneficial to be able test, and compare the different tools featured herein (solubilization tags, CO, CH) systematically on each set of genes investigated. Doing so would allow the determination of the combination of optimal approaches for each scenario, and likely greatly improve the odds of obtaining soluble proteins in desired yields. Moreover, the toolbox for soluble expression is constantly expanding, and new features and opportunities should be considered. For example, novel solubility-

enhancing fusion proteins, and their applications are reported. Some examples include the 31 kDa ATP-independent folding chaperone Spheroplast Protein Y (Spy) (Quan et al., 2011; Ruan et al., 2020), the small 8 kDa protein Fh8 sourced from *Fasciola hepatica*, and the tiny 1 kDa H-tag constituted by the first 11 N-terminal amino acids of Fh8 (Costa et al., 2013a). Fh8 was also reported to be able to function as a purification tag for hydrophobic interaction chromatography, facilitating results on par with the traditional His-tag and IMAC strategy (Costa et al., 2013b). While most fusion proteins are reported as aiding soluble heterologous expression in *E. coli* in some context, small fusion tags may be considered favourable over larger ones, as they exert less stress over the cell, and reportedly do not lead to significant loss of solubility after proteolytic tag removal (Costa et al., 2013a, 2014). Ultimately as with many other protein-related challenges, no single “best” fusion tag can be highlighted to ensure the soluble production of a given protein of interest, and therefore testing a selection of most relevant tags remains the most educated strategy.

Considered together, the results obtained in this thesis also show that *E. coli* remains a steadfastly relevant host organism for heterologous expression, even for genes from viral metagenomes. However, other valuable host organisms can also be implemented to complement *E. coli*, and benefit from their unique strengths. Some of the most prominent alternatives include the Gram-positive bacterium *B. subtilis* which also has a wide genetic toolkit available and allows high yields from extra-cellular expression of proteins. In this way, cytotoxic proteins could be expressed outside the cell with high yields. (Vavrová et al., 2010; Biver et al., 2013). Baker’s yeast *Saccharomyces cerevisiae* is a similarly well-studied eukaryotic host, that may be preferable when post-translational modification of produced proteins is required (Damon et al., 2011). For the production of thermophilic proteins, the Gram-negative bacterium *Thermus thermophilus* may be beneficial, compatible with cultivation and screening of enzymes in higher temperatures (Cava et al., 2009). Switching between different hosts can be cumbersome, but shuttle vector systems are available to facilitate the use of multiple expression systems simultaneously (Troeschel et al., 2012; Nakata, 2017), and some work has already been done to expand the procedure presented in **Paper I** to

include various *Bacillus* species alongside *E. coli* (Larsen and Bjerga, 2018). Even more alternative organisms exist and has recently been reviewed by Lewin et al. (Lewin et al., 2017). Ultimately, choice(s) of host organism should be made with the goals and the circumstances of a given study in mind.

The results obtained in this thesis also point to that codon adjustment approaches still hold value to fine tune the codon landscape of target genes and promote higher yields and/or better folding quality (**Paper III**). While these strategies may be unfeasible if the chosen genes are to be cloned manually into desired expression vectors; both CO and CH may be implemented without difficulty if gene synthesis options are available. For studies with greater scopes (such as the Virus-X project), it may be possible to automate the harmonization of target sequences to enable large-scale implementation of the strategy. If the study is focused on a smaller number of genes for heterologous expression, more effort-intensive approaches may be relevant to maximize the chances of successfully producing target proteins. Some examples include refolding of insoluble inclusion bodies (D'alessio et al., 1999; Dutta et al., 2010; Ramón et al., 2014), protein engineering to potentially modify key properties of target proteins (Cherry et al., 1999; Taguchi et al., 1999; Fernandes et al., 2015), and fine tuning of culture conditions (Fouchet et al., 1994; Shiloach and Fass, 2005; Gutiérrez-González et al., 2019).

6. Future research

In this doctoral study, a large part of work has been carried out in the framework of two large research projects. While this setting provided unique opportunities and chance to work with amazing groups of inspiring researchers, at times, it also made it difficult to implement certain ideas due to considerations of time, continuity and compatibility with other research groups, and the overarching scope of the respective projects. Nevertheless, this work contributed to improving access to marine genetic resources for the discovery of enzymes, in a fast-developing era of high-throughput sequencing (van Dijk et al., 2018), omics technologies (Eren et al., 2021), and a global viral pandemic (Pfefferbaum and North, 2020).

Through the course of this doctoral study, a large set of viral genes have been successfully expressed as soluble enzymes in *E. coli* providing a basis for their subsequent in-depth biochemical and structural characterization. Like for the proteases presented in **Paper I** and **II**, enzyme activity assays should be implemented to assess the activity of all proteins produced as soluble.

The viral nucleases have been tested for production in a limited experimental design, and more of them may be possible to produce and purify via solubility tags or other methods discussed above. This is also true for the genes from HTH1, from which the biotechnologically interesting DNA polymerase was found insoluble and could likely benefit from further optimization to achieve soluble production. Activity assays considered for the nucleases, may have to include an array of related assays to check for multiple DNA polymerase, endo- or exonuclease activities, and strand specificities (Fernández-García et al., 2017; del Prado et al., 2019).

Among the purified proteins, the *Planococcus* ISP was functionally and structurally characterized. However, to determine its suitability for biotechnological applications, its activity needs to be further assessed against application relevant substrates, such as fibrin (Yang et al., 2020) and fish proteins (Aspmo et al., 2005). Following their characterization, this is also imperative for the viral nucleases. Purified HTH1

enzymes are currently undergoing structural analyses via protein crystallization (pers. comm. M. Håkansson and S. Al-Karadaghi, SARomics Biostructures), but their functional characterization, including an expected moderate thermophilic activity profile, remains to be assessed.

After characterization, it may be possible to modify key properties of promising enzymes via protein engineering in an effort to push them past the threshold towards applications in biotechnology (Vojcic et al., 2015; Rigoldi et al., 2018). For a study focused on this subject, 3DM systems (Kuipers et al., 2010) could provide a thorough starting knowledgebase to guide engineering efforts and help navigate the patent landscape of similar enzyme applications.

Finally, a potential morphological study, covering the lytic induction, isolation, and electron microscope imaging of HTH1 would be of great interest in strengthening the sequence-based study of prophage taxonomy presented in **Paper III**, and reveal further insights into *H. thermotrophus* and the viral interactions of hydrothermal vent environments

In future endeavours targeting the sequence-based mining of marine genetic resources, the strengths of the different strategies demonstrated herein should be combined in accordance with the requirements of the study. Example features could include the deep sequence analysis capabilities used towards the viral candidates, the versatility and efficiency of the pipeline developed for the expression and activity screening of proteases (tailored to screen for the desired enzymatic activity), and the larger-scale implementation of relevant solubility and quality enhancing measures for the production of target enzymes. Putting together one such “complete” approach would contribute greatly to improving the success rate in marine bioprospecting by bringing more enzyme candidates to a state where they can be functionally and structurally characterized.

7. Conclusion

The work presented in this thesis altogether emphasizes marine microbial diversity as an important resource for enzyme discovery, both towards expanding our understanding of marine microbial life, and also for application of enzymes within various biotechnological fields. The aims set for this doctoral work has largely been met, as pertinent experimental methods and bioinformatical tools were bridged together in an effort to overcome challenges of sequence-based enzyme discovery. This combined competence allowed the efficient screening, discovery, heterologous production, and characterization of a varying portfolio of enzyme classes including proteases, nucleases and endolysins. The multi-faceted application of modern sequence analysis tools allowed not only the analysis of arctic marine viral metagenomes for the identification and heterologous production of novel viral enzyme targets, but also the detection and *in silico* characterization of a novel prophage from the hydrothermal vent bacterium *H. thermotrophus*. The functional and structural characterization of the intracellular subtilisin protease from *Planococcus* sp. AW02J18 expanded our current knowledge about ISPs and their mechanisms of activity regulation. Similarly, the combined genomic and protein-level study of the *H. thermotrophus* phage H1, generated novel insight over the lysogenic phage-host interactions in *Fusobacteria* and hydrothermal vent environments. The efforts towards improving the soluble production of heterologous proteins in *E. coli* have been the common thread that linked the study of different enzyme targets, and as such the approaches demonstrated in this direction stand to benefit a wide range of future studies. Although the use of individual elements such as solubility enhancing fusion proteins, codon adaptation strategies, and parallel testing of gene expression in varied configurations were demonstrated in separate case studies, it is clear that their optimal benefit lies in their complementary application. As no “one approach to express them all” can be named, thorough functional annotation of genes and targeted protein expression workflows constructed towards defined research goals will together allow the deepest access to the vast genetic resources of marine microorganisms.

References

- Ackermann, H. W. (2007). 5500 Phages examined in the electron microscope. *Arch. Virol.* 152, 227–243. doi:10.1007/s00705-006-0849-1.
- Ackermann, H. W., and Krisch, H. M. (1997). A catalogue of T4-type bacteriophages. *Arch. Virol.* 142, 2329–2345. doi:10.1007/s007050050246.
- Agger, J. W., Busk, P. K., Pilgaard, B., Meyer, A. S., and Lange, L. (2017). A new functional classification of glucuronoyl esterases by peptide pattern recognition. *Front. Microbiol.* 8, 309. doi:10.3389/fmicb.2017.00309.
- Akhter, S., Aziz, R. K., and Edwards, R. A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res.* 40. doi:10.1093/nar/gks406.
- Alcorlo, M., Martínez-Caballero, S., Molina, R., and Hermoso, J. A. (2017). Carbohydrate recognition and lysis by bacterial peptidoglycan hydrolases. *Curr. Opin. Struct. Biol.* 44, 87–100. doi:10.1016/j.sbi.2017.01.001.
- Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi:10.1038/s41587-019-0036-z.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389.
- Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol.* 4, 2121–2131. doi:10.1371/journal.pbio.0040368.
- Angov, E. (2011). Codon usage: Nature’s roadmap to expression and folding of proteins. *Biotechnol. J.* 6, 650–659. doi:10.1002/biot.201000332.
- Angov, E., Hillier, C. J., Kincaid, R. L., and Lyon, J. A. (2008). Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One* 3, e2189. doi:10.1371/journal.pone.0002189.
- Angov, E., Legler, P., and Mease, R. (2011). Adjustment of codon usage frequencies by codon harmonization improves protein expression and folding. *Methods Mol. Biol.* 705, 1–13. doi:10.1007/978-1-61737-967-3_1.
- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi:10.1093/nar/gkw387.
- Arnold, F. H., Wintrode, P. L., Miyazaki, K., and Gershenson, A. (2001). How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* 26, 100–106. doi:10.1016/S0968-0004(00)01755-2.
- Asam, C., Roulias, A., Parigiani, M. A., Haab, A., Wallner, M., Wolf, M., et al. (2018). Harmonization of the genetic code effectively enhances the recombinant production of the major birch pollen allergen Bet v 1. *Int. Arch. Allergy Immunol.* 177, 116–122. doi:10.1159/000489707.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556.
- Aspmo, S. I., Horn, S. J., and Eijssink, V. G. H. (2005). Enzymatic hydrolysis of Atlantic cod (*Gadus morhua* L.) viscera. *Process Biochem.* 40, 1957–1966. doi:10.1016/j.procbio.2004.07.011.

- Bahir, I., Fromer, M., Prat, Y., and Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* 5, 311. doi:10.1038/msb.2009.71.
- Bailly-Bechet, M., Vergassola, M., and Rocha, E. (2007). Causes for the intriguing presence of tRNAs in phages. *Genome Res.* 17, 1486–1495. doi:10.1101/gr.6649807.
- Balazs, I. (1992). Forensic applications. *Curr. Opin. Biotechnol.* 3, 18–23. doi:10.1016/0958-1669(92)90120-8.
- Barone, R., De Santi, C., Palma Esposito, F., Tedesco, P., Galati, F., Visone, M., et al. (2014). Marine metagenomics, a valuable tool for enzymes and bioactive compounds discovery. *Front. Mar. Sci.* 1, 38. doi:10.3389/fmars.2014.00038.
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science (80-.)*. 315, 1709–1712. doi:10.1126/science.1138140.
- Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi:10.1093/nar/gkaa1100.
- Baulcombe, D. (2004). RNA silencing in plants. *Nature* 431, 356–363. doi:10.1038/nature02874.
- Beerenwinkel, N., Günthard, H. F., Roth, V., and Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* 3, 329. doi:10.3389/fmicb.2012.00329.
- Beese, L. S., and Steitz, T. A. (2020). “Structural basis for the 3’- 5’ exonuclease activity of escherichia coli dna polymerase i: A two metal ion mechanism,” in *Structural Insights into Gene Expression and Protein Synthesis* (World Scientific Publishing Co.), 245–253. doi:10.1142/9789811215865_0025.
- Bekhit, A. A., Hopkins, D. L., Geesink, G., Bekhit, A. A., and Franks, P. (2014). Exogenous proteases for meat tenderization. *Crit. Rev. Food Sci. Nutr.* 54, 1012–1031. doi:10.1080/10408398.2011.623247.
- Beliaeva, M. I., Kapranova, M. N., and Vitol, M. J. (1976). Nucleic acids utilized as the main source of bacterial nutrition (Russian). *Mikrobiologiya* 45, 420–424. Available at: <https://europepmc.org/article/med/826761> [Accessed February 24, 2021].
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235.
- Bernard, P., Gabarit, P., Bahassi, E. M., and Couturier, M. (1994). Positive-selection vectors using the F plasmid ccdB killer gene. *Gene* 148, 71–74. doi:10.1016/0378-1119(94)90235-6.
- Bhaduri, S., and Mukesh, D. (2014). *Homogeneous Catalysis: Mechanisms and Industrial Applications*. 2nd ed. Wiley Available at: https://books.google.no/books?id=BA1_BAAAQBAJ.
- Bird, A. P., and Southern, E. M. (1978). Use of restriction enzymes to study eukaryotic DNA methylation. I. The methylation pattern in ribosomal DNA from *Xenopus laevis*. *J. Mol. Biol.* 118, 27–47. doi:10.1016/0022-2836(78)90242-5.
- Biver, S., Portetelle, D., and Vandenberg, M. (2013). Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. *Springerplus* 2, 410. doi:10.1186/2193-1801-2-410.
- Bleriot, I., Trastoy, R., Blasco, L., Fernández-Cuenca, F., Ambroa, A., Fernández-García, L., et al. (2020). Genomic analysis of 40 prophages located in the genomes of 16 carbapenemase-producing clinical strains of *Klebsiella pneumoniae*. *Microb. Genomics* 6, 1–18. doi:10.1099/mgen.0.000369.
- Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasamy, S., Mitchell, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi:10.1093/nar/gkaa977.
- Bode, W., Papamokos, E., and Musil, D. (1987). The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech *Hirudo medicinalis*. *Structural*

- analysis, subtilisin structure and interface geometry. *Eur. J. Biochem.* 166, 673–692. doi:10.1111/j.1432-1033.1987.tb13566.x.
- Borriß, M., Helmke, E., Hanschke, R., and Schweder, T. (2003). Isolation and characterization of marine psychrophilic phage-host systems from Arctic sea ice. *Extremophiles* 7, 377–384. doi:10.1007/s00792-003-0334-7.
- Børshøj, K. Y., Bratbak, G., and Heldal, M. (1990). Enumeration and biomass estimation of planktonic bacteria and viruses by transmission electron microscopy. *Appl. Environ. Microbiol.* 56, 352–356. doi:10.1128/aem.56.2.352-356.1990.
- Boura, E., Baumlova, A., Chalupska, D., Dubankova, A., and Klima, M. (2017). Metal ions-binding T4 lysozyme as an intramolecular protein purification tag compatible with X-ray crystallography. *Protein Sci.* 26, 1116–1123. doi:10.1002/pro.3162.
- Boycheva, S., Chkdrov, G., and Ivanov, I. (2003). Codon pairs in the genome of *Escherichia coli*. *Bioinformatics* 19, 987–998. doi:10.1093/bioinformatics/btg082.
- Braun, H., Boller, K., Löwer, J., Bertling, W. M., and Zimmer, A. (1999). Oligonucleotide and plasmid DNA packaging into polyoma VP1 virus-like particles expressed in *Escherichia coli*. *Biotechnol. Appl. Biochem.* 29 (Pt 1), 31–43. doi:10.1111/j.1470-8744.1999.tb01146.x.
- Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14250–14255. doi:10.1073/pnas.202488399.
- Briers, Y., Walmagh, M., Van Puyenbroeck, V., Cornelissen, A., Cenens, W., Aertsen, A., et al. (2014). Engineered endolysin-based “Artilyns” to combat multidrug-resistant gram-negative pathogens. *MBio* 5, 1379–1393. doi:10.1128/mBio.01379-14.
- Brininger, C., Spradlin, S., Cobani, L., and Evilia, C. (2018). The more adaptive to change, the more likely you are to survive: Protein adaptation in extremophiles. *Semin. Cell Dev. Biol.* 84, 158–169. doi:10.1016/j.semcdb.2017.12.016.
- Brode, P. F., Barnett, B. L., and Rubingh, D. N. (1994). Subtilisin bpn’ variants with decreased adsorption and increased hydrolysis. Available at: <https://patents.google.com/patent/WO1995007991A3/en> [Accessed September 16, 2018].
- Brown, S., Santa Maria, J. P., and Walker, S. (2013). Wall teichoic acids of gram-positive bacteria. *Annu. Rev. Microbiol.* 67, 313–336. doi:10.1146/annurev-micro-092412-155620.
- Bryan, P. N., and Pantoliano, M. W. (1988). Subtilisin mutations. Available at: <https://patents.google.com/patent/US5013657> [Accessed September 16, 2018].
- Buist, G., Steen, A., Kok, J., and Kuipers, O. P. (2008). LysM, a widely distributed protein motif for binding to (peptid)glycans. *Mol. Microbiol.* 68, 838–847. doi:10.1111/j.1365-2958.2008.06211.x.
- Burgess-Brown, N. A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., and Gileadi, O. (2008). Codon optimization can improve expression of human genes in *Escherichia coli*: A multi-gene study. *Protein Expr. Purif.* 59, 94–102. doi:10.1016/j.pep.2008.01.008.
- Burnett, T. J., Shankweiler, G. W., and Hageman, J. H. (1986). Activation of intracellular serine proteinase in *Bacillus subtilis* cells during sporulation. *J. Bacteriol.* 165, 139–45. doi:10.1128/jb.165.1.139-145.1986.
- Burton, S. G., Cowan, D. A., and Woodley, J. M. (2002). The search for the ideal biocatalyst. *Nat. Biotechnol.* 20, 37–45. doi:10.1038/nbt0102-37.
- Busk, P. K., and Lange, L. (2013). Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs. *Appl. Environ. Microbiol.* 79, 3380–3391. doi:10.1128/AEM.03803-12.
- Cafardi, J. A., and Elmetts, C. A. (2008). T4 endonuclease V: review and application to dermatology. *Expert Opin. Biol.*

Ther. 8, 829–838. doi:10.1517/14712598.8.6.829.

- Cahill, J., and Young, R. (2019). “Phage lysis: multiple genes for multiple barriers,” in *Advances in Virus Research* (Academic Press Inc.), 33–70. doi:10.1016/bs.aivir.2018.09.003.
- Carbone, A. (2008). Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J. Mol. Evol.* 66, 210–223. doi:10.1007/s00239-008-9068-6.
- Carbone, A., Zinovyev, A., and Kepes, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19, 2005–2015. doi:10.1093/bioinformatics/btg272.
- Carter, D. M., and Radding, C. M. (1971). The role of exonuclease and β protein of phage λ in genetic recombination. *J. Biol. Chem.* 246, 2502–2510. doi:10.1016/s0021-9258(18)62316-6.
- Castelán-Sánchez, H. G., Lopéz-Rosas, I., García-Suastegui, W. A., Peralta, R., Dobson, A. D. W., Batista-García, R. A., et al. (2019). Extremophile deep-sea viral communities from hydrothermal vents: Structural and functional analysis. *Mar. Genomics* 46, 16–28. doi:10.1016/j.margen.2019.03.001.
- Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., and Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* 16, 629–645. doi:10.1038/s41579-018-0076-2.
- Catalão, M. J., Gil, F., Moniz-Pereira, J., São-José, C., and Pimentel, M. (2013). Diversity in bacterial lysis systems: Bacteriophages Show the way. *FEMS Microbiol. Rev.* 37, 554–571. doi:10.1111/1574-6976.12006.
- Cava, F., and de Pedro, M. A. (2014). Peptidoglycan plasticity in bacteria: Emerging variability of the murein sacculus and their associated biological functions. *Curr. Opin. Microbiol.* 18, 46–53. doi:10.1016/j.mib.2014.01.004.
- Cava, F., Hidalgo, A., and Berenguer, J. (2009). *Thermus thermophilus* as biological model. *Extremophiles* 13, 213–231. doi:10.1007/s00792-009-0226-6.
- Cedzich, A., Huttenlocher, F., Kuhn, B. M., Pfannstiel, J., Gabier, L., Stintzi, A., et al. (2009). The protease-associated domain and C-terminal extension are required for zymogen processing, sorting within the secretory pathway, and activity of tomato subtilase 3 (SISBT3). *J. Biol. Chem.* 284, 14068–14078. doi:10.1074/jbc.M900370200.
- Center for Food Safety and Applied Nutrition (2018). Generally Recognized as Safe (GRAS). Available at: <https://www.fda.gov/Food/IngredientsPackagingLabeling/GRAS/ucm2006850.htm> [Accessed May 20, 2018].
- Chalamaiah, M., Dinesh Kumar, B., Hemalatha, R., and Jyothirmayi, T. (2012). Fish protein hydrolysates: Proximate composition, amino acid composition, antioxidant activities and applications: A review. *Food Chem.* 135, 3020–3038. doi:10.1016/j.foodchem.2012.06.100.
- Chang, S.-W., Lee, G.-C., and Shaw, J.-F. (2006). Codon optimization of *Candida rugosa* lip 1 gene for improving expression in *Pichia pastoris* and biochemical characterization of the purified recombinant LIP1 lipase. *J. Agric. Food Chem.* 54, 815–822. doi:10.1021/jf052183k.
- Chapman, J., Ismail, A., Dinu, C., Chapman, J., Ismail, A. E., and Dinu, C. Z. (2018). Industrial applications of enzymes: Recent advances, techniques, and outlooks. *Catalysts* 8, 238. doi:10.3390/catal8060238.
- Chen, P., Tsuge, H., Almassy, R. J., Gribskov, C. L., Katoh, S., Vanderpool, D. L., et al. (1996). Structure of the human cytomegalovirus protease catalytic domain reveals a novel serine protease fold and catalytic triad. *Cell* 86, 835–843. doi:10.1016/S0092-8674(00)80157-9.
- Chen, X. S., Casini, G., Harrison, S. C., and Garcea, R. L. (2001). Papillomavirus capsid protein expression in *Escherichia coli*: Purification and assembly of HPV11 and HPV16 L1. *J. Mol. Biol.* 307, 173–182. doi:10.1006/jmbi.2000.4464.
- Cherry, J. R., Lamsa, M. H., Schneider, P., Vind, J., Svendsen, A., Jones, A., et al. (1999). Directed evolution of a fungal peroxidase. *Nat. Biotechnol.* 17, 379–384. doi:10.1038/7939.

- Chiriboga, C. A. (2017). Nusinersen for the treatment of spinal muscular atrophy. *Expert Rev. Neurother.* 17, 955–962. doi:10.1080/14737175.2017.1364159.
- Choi, K. H. (2012). Viral polymerases. *Adv. Exp. Med. Biol.* 726, 267–304. doi:10.1007/978-1-4614-0980-9_12.
- Chuan, Y. P., Lua, L. H. L., and Middelberg, A. P. J. (2008). High-level expression of soluble viral structural protein in *Escherichia coli*. *J. Biotechnol.* 134, 64–71. doi:10.1016/j.jbiotec.2007.12.004.
- Claassens, N. J., Siliakus, M. F., Spaans, S. K., Creutzburg, S. C. A. A., Nijssse, B., Schaap, P. J., et al. (2017). Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. *PLoS One* 12, e0184355. doi:10.1371/journal.pone.0184355.
- Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Res.* 44, D67–D72. doi:10.1093/nar/gkv1276.
- Cohen, S. N., Chang, A. C., Boyer, H. W., and Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U. S. A.* 70, 3240–4. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/4594039> [Accessed November 12, 2018].
- Colin, P.-Y., Kintsjes, B., Gielen, F., Miton, C. M., Fischer, G., Mohamed, M. F., et al. (2015). Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun.* 6, 10008. doi:10.1038/ncomms10008.
- Corinaldesi, C., Tangherlini, M., and Dell’anno, A. (2017). From virus isolation to metagenome generation for investigating viral diversity in deep-sea sediments. *Sci. Rep.* 7, 1–12. doi:10.1038/s41598-017-08783-4.
- Corliss, J. B., Dymond, J., Gordon, L. I., Edmond, J. M., Von Herzen, R. P., Ballard, R. D., et al. (1979). Submarine thermal springs on the Galápagos Rift. *Science (80-)*. 203, 1073. doi:10.1126/science.203.4385.1073.
- Costa, S., Almeida, A., Castro, A., and Domingues, L. (2014). Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli*: the novel Fh8 system. *Front. Microbiol.* 5, 63. doi:10.3389/fmicb.2014.00063.
- Costa, S. J., Almeida, A., Castro, A., Domingues, L., and Besir, H. (2013a). The novel Fh8 and H fusion partners for soluble protein expression in *Escherichia coli*: a comparison with the traditional gene fusion technology. *Appl. Microbiol. Biotechnol.* 97, 6779–6791. doi:10.1007/s00253-012-4559-1.
- Costa, S. J., Coelho, E., Franco, L., Almeida, A., Castro, A., and Domingues, L. (2013b). The Fh8 tag: a fusion partner for simple and cost-effective protein purification in *Escherichia coli*. *Protein Expr. Purif.* 92, 163–70. doi:10.1016/j.pep.2013.09.013.
- D’Alessio, K. J., McQueney, M. S., Brun, K. A., Orsini, M. J., and Debouck, C. M. (1999). Expression in *Escherichia coli*, refolding, and purification of human procathepsin K, an osteoclast-specific protease. *Protein Expr. Purif.* 15, 213–20. doi:10.1006/prep.1998.1013.
- Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S., and Kim, J. F. (2009). Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* 394, 634–643. doi:https://doi.org/10.1016/j.jmb.2009.09.022.
- Dahle, H., Le Moine Bauer, S., Baumberg, T., Stokke, R., Pedersen, R. B., Thorseth, I. H., et al. (2018). Energy landscapes in hydrothermal chimneys shape distributions of primary producers. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.01570.
- Dahle, H., Roalkvam, I., Thorseth, I. H., Pedersen, R. B., and Steen, I. H. (2013). The versatile in situ gene expression of an *Epsilonproteobacteria*-dominated biofilm from a hydrothermal chimney. *Environ. Microbiol. Rep.* 5, 282–290. doi:10.1111/1758-2229.12016.
- Dale, R. M. K., McClure, B. A., and Houchins, J. P. (1985). A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: Application to sequencing the corn

mitochondrial 18 S rDNA. *Plasmid* 13, 31–40. doi:10.1016/0147-619X(85)90053-8.

- Damon, C., Vallon, L., Zimmermann, S., Haider, M. Z., Galeote, V., Dequin, S., et al. (2011). A novel fungal family of oligopeptide transporters identified by functional metatranscriptomics of soil eukaryotes. *ISME J.* 5, 1871–1880. doi:10.1038/ismej.2011.67.
- Dang, G., Cui, Y., Wang, L., Li, T., Cui, Z., Song, N., et al. (2018). Extracellular sphingomyelinase Rv0888 of *Mycobacterium tuberculosis* contributes to pathological lung injury of *Mycobacterium smegmatis* in mice via inducing formation of neutrophil extracellular traps. *Front. Immunol.* 9, 677. doi:10.3389/fimmu.2018.00677.
- Dávila-Ramos, S., Castelán-Sánchez, H. G., Martínez-ávila, L., Sánchez-Carbente, M. D. R., Peralta, R., Hernández-Mendoza, A., et al. (2019). A review on viral metagenomics in extreme environments. *Front. Microbiol.* 10, 2403. doi:10.3389/fmicb.2019.02403.
- Dayanandan, A., Kanagaraj, J., Sounderraj, L., Govindaraju, R., and Rajkumar, G. S. (2003). Application of an alkaline protease in leather processing: an ecofriendly approach. *J. Clean. Prod.* 11, 533–536. doi:10.1016/S0959-6526(02)00056-2.
- de Oliveira, T. B., Gostinčar, C., Gunde-Cimerman, N., and Rodrigues, A. (2018). Genome mining for peptidases in heat-tolerant and mesophilic fungi and putative adaptations for thermostability. *BMC Genomics* 19, 152. doi:10.1186/s12864-018-4549-5.
- De Paepe, M., and Taddei, F. (2006). Viruses' life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biol.* 4, e193. doi:10.1371/journal.pbio.0040193.
- de Pascale, D., De Santi, C., Fu, J., and Landfald, B. (2012). The microbial diversity of Polar environments is a fertile ground for bioprospecting. *Mar. Genomics* 8, 15–22. doi:10.1016/j.margen.2012.04.004.
- De Santi, C., Altermark, B., de Pascale, D., and Willassen, N.-P. (2016). Bioprospecting around Arctic islands: Marine bacteria as rich source of biocatalysts. *J. Basic Microbiol.* 56, 238–53. doi:10.1002/jobm.201500505.
- DeCastro, M. E., Rodríguez-Belmonte, E., and González-Siso, M. I. (2016). Metagenomics of thermophiles with a focus on discovery of novel thermozymes. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01521.
- del Prado, A., Rodríguez, I., Lázaro, J. M., Moreno-Morcillo, M., de Vega, M., and Salas, M. (2019). New insights into the coordination between the polymerization and 3'-5' exonuclease activities in ϕ 29 DNA polymerase. *Sci. Rep.* 9, 923. doi:10.1038/s41598-018-37513-7.
- Dick, G. J. (2019). The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally. *Nat. Rev. Microbiol.* 17, 271–283. doi:10.1038/s41579-019-0160-2.
- Ding, Y., Li, X., Zhang, X., Li, F., Hou, X., and Wu, P. (2019). Systematic probing of the sequence selectivity of exonuclease III with a photosensitization colorimetric assay. *ACS Omega* 4, 13382–13387. doi:10.1021/acsomega.9b01560.
- Dion, M. B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* 18, 125–138. doi:10.1038/s41579-019-0311-5.
- Distaso, M. A., Tran, H., Ferrer, M., and Golyshin, P. N. (2017). “Metagenomic mining of enzyme diversity,” in *Consequences of Microbial Interactions with Hydrocarbons, Oils, and Lipids: Production of Fuels and Chemicals*, ed. S. Y. Lee (Cham: Springer International Publishing), 1–25. doi:10.1007/978-3-319-31421-1_216-1.
- Dixon, B. (2004). New dawn for phage therapy. *Lancet Infect. Dis.* 4, 186. doi:10.1016/S1473-3099(04)00951-X.
- Dlakić, M. (2000). Functionally unrelated signalling proteins contain a fold similar to Mg²⁺-dependent endonucleases. *Trends Biochem. Sci.* 25, 272–273. doi:10.1016/S0968-0004(00)01582-6.
- Dodson, M. L., and Lloyd, R. S. (1989). Structure-function studies of the T4 endonuclease V repair enzyme. *Mutat. Res. Repair* 218, 49–65. doi:10.1016/0921-8777(89)90011-6.

- Doherty, A. J., Ashford, S. R., Subramanya, H. S., and Wigley, D. B. (1996). Bacteriophage T7 DNA ligase: Overexpression, purification, crystallization, and characterization. *J. Biol. Chem.* 271, 11083–11089. doi:10.1074/jbc.271.19.11083.
- Dorawa, S., Plotka, M., Kaczorowska, A.-K., Fridjonsson, O. H., Hreggvidsson, G. O., Aevarsson, A., et al. (2020). Characterization of DNA polymerase from *Thermus thermophilus* MAT72 phage Tt72. *Proceedings* 50, 38. doi:10.3390/proceedings2020050038.
- dos Santos Aguilar, J. G., and Sato, H. H. (2018). Microbial proteases: Production and application in obtaining protein hydrolysates. *Food Res. Int.* 103, 253–262. doi:10.1016/j.foodres.2017.10.044.
- Dutta, S., Ghosh, R., Dattagupta, J. K., and Biswas, S. (2010). Heterologous expression of a thermostable plant cysteine protease in *Escherichia coli* both in soluble and insoluble forms. *Process Biochem.* 45, 1307–1312. doi:10.1016/j.procbio.2010.04.020.
- Ekici, Ö. D., Paetzel, M., and Dalbey, R. E. (2008). Unconventional serine proteases: Variations on the catalytic Ser/His/Asp triad configuration. *Protein Sci.* 17, 2023–2037. doi:10.1110/ps.035436.108.
- Elleuche, S., Schröder, C., Sahm, K., and Antranikian, G. (2014). Extremozymes — biocatalysts with unique properties from extremophilic microorganisms. *Curr. Opin. Biotechnol.* 29, 116–123. doi:10.1016/j.copbio.2014.04.003.
- Engler, M. J., and Richardson, C. C. (1982). DNA ligases. *Enzymes* 15, 3–29. doi:10.1016/S1874-6047(08)60273-5.
- Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., et al. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* 6, 3–6. doi:10.1038/s41564-020-00834-3.
- European Food Safety Authority (2013). Qualified presumption of safety (QPS). Available at: <https://www.efsa.europa.eu/en/topics/topic/qualified-presumption-safety-qps> [Accessed May 20, 2018].
- Evans, K. L., Crowder, J., and Miller, E. S. (2000). Subtilisins of *Bacillus* spp. hydrolyze keratin and allow growth on feathers. *Can. J. Microbiol.* 46, 1004–1011. doi:10.1139/w00-085.
- Fanoe, T. S., and Mikkelsen, F. F. (2007). Subtilase variants having an improved wash performance on egg stains. A1. Available at: <https://patents.google.com/patent/WO2007122175A1/en> [Accessed September 16, 2018].
- Farabaugh, P. J., and Björk, G. R. (1999). How translational accuracy influences reading frame maintenance. *EMBO J.* 18, 1427–1434. doi:10.1093/emboj/18.6.1427.
- Farshadpour, F., Taherkhani, R., Makvandi, M., Memari, H. R., and Samarbafzadeh, A. R. (2014). Codon-optimized expression and purification of truncated ORF2 protein of hepatitis E virus in *Escherichia coli*. *Jundishapur J. Microbiol.* 7, 11261. doi:10.5812/jjm.11261.
- Fedyunin, I., Lehnhardt, L., Böhmer, N., Kaufmann, P., Zhang, G., and Ignatova, Z. (2012). tRNA concentration fine tunes protein solubility. *FEBS Lett.* 586, 3336–3340. doi:10.1016/j.febslet.2012.07.012.
- Fei, D., Zhang, H., Diao, Q., Jiang, L., Wang, Q., Zhong, Y., et al. (2015). Codon optimization, expression in *Escherichia coli*, and immunogenicity of recombinant Chinese Sacbrood Virus (CSBV) structural proteins VP1, VP2, and VP3. *PLoS One* 10, e0128486. doi:10.1371/journal.pone.0128486.
- Fernandes, P., Aldeborgh, H., Carlucci, L., Walsh, L., Wasserman, J., Zhou, E., et al. (2015). Alteration of substrate specificity of alanine dehydrogenase. *Protein Eng. Des. Sel.* 28, 29–35. doi:10.1093/protein/gzu053.
- Fernandes, S., and São-José, C. (2018). Enzymes and mechanisms employed by tailed bacteriophages to breach the bacterial cell barriers. *Viruses* 10, 396. doi:10.3390/v10080396.
- Fernández-Arrojo, L., Guazzaroni, M. E., López-Cortés, N., Beloqui, A., and Ferrer, M. (2010). Metagenomic era for biocatalyst identification. *Curr. Opin. Biotechnol.* 21, 725–733. doi:10.1016/j.copbio.2010.09.006.
- Fernández-García, J. L., De Ory, A., Brussaard, C. P. D., and De Vega, M. (2017). *Phaeocystis globosa* virus DNA

- Polymerase X: A “Swiss army knife”, multifunctional DNA polymerase-lyase-ligase for base excision repair. *Sci. Rep.* 7. doi:10.1038/s41598-017-07378-3.
- Fernández-Ruiz, I., Coutinho, F. H., and Rodríguez-Valera, F. (2018). Thousands of novel endolysins discovered in encultured phage genomes. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.01033.
- Ferrer, M., Beloqui, A., Timmis, K. N., and Golyshin, P. N. (2009). Metagenomics for mining new genetic resources of microbial communities. *J. Mol. Microbiol. Biotechnol.* 16, 109–23. doi:10.1159/000142898.
- Ferrer, M., Golyshina, O., Beloqui, A., and Golyshin, P. N. (2007). Mining enzymes from extreme environments. *Curr. Opin. Microbiol.* 10, 207–14. doi:10.1016/j.mib.2007.05.004.
- Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Martins dos Santos, V. A. P., Yakimov, M. M., et al. (2005). Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. *Chem. Biol.* 12, 895–904. doi:10.1016/j.chembiol.2005.05.020.
- Ferrer, M., Méndez-García, C., Bargiela, R., Chow, J., Alonso, S., García-Moyano, A., et al. (2019). Decoding the ocean’s microbiological secrets for marine enzyme biodecovery. *FEMS Microbiol. Lett.* 366, 285. doi:10.1093/femsle/fny285.
- Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi:10.1093/nar/gkr367.
- Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi:10.1093/nar/gkv1344.
- Forterre, P., and Prangishvili, D. (2009). The origin of viruses. *Res. Microbiol.* 160, 466–472. doi:10.1016/j.resmic.2009.07.008.
- Fouchet, P., Manin, C., Richard, H., Frelat, G., and Barbotin, J. N. (1994). Flow cytometry studies of recombinant *Escherichia coli* in batch and continuous cultures: DNA and RNA contents; light-scatter parameters. *Appl. Microbiol. Biotechnol.* 41, 584–590. doi:10.1007/BF00178494.
- Francki, R. I. B., Fauquet, C. M., Knudson, D. L., and Brown, F. (1991). Classification and Nomenclature of Viruses: Fifth Report of the International Committee on Taxonomy of Viruses. Virology Division of the International Union of Microbiological Societies., eds. R. I. B. Francki, C. M. Fauquet, D. L. Knudson, and F. Brown Vienna: Springer Vienna doi:10.1007/978-3-7091-9163-7.
- Fredriksen, L., Stokke, R., Jensen, M. S., Westereng, B., Jameson, J. K., Steen, I. H., et al. (2019). Discovery of a thermostable GH10 xylanase with broad substrate specificity from the Arctic Mid-Ocean Ridge vent system. *Appl. Environ. Microbiol.* 85, 1–47. doi:10.1128/AEM.02970-18.
- Freitag-Pohl, S., Jasilionis, A., Håkansson, M., Svensson, L. A., Kovačič, R., Welin, M., et al. (2019). Crystal structures of the *Bacillus subtilis* prophage lytic cassette proteins XepA and YomS. *Acta Crystallogr. Sect. D Struct. Biol.* 75, 1028–1039. doi:10.1107/S2059798319013330.
- Fu, J., Leiros, H. K. S., De Pascale, D., Johnson, K. A., Blencke, H. M., and Landfald, B. (2013). Functional and structural studies of a novel cold-adapted esterase from an Arctic intertidal metagenomic library. *Appl. Microbiol. Biotechnol.* 97, 3965–3978. doi:10.1007/s00253-012-4276-9.
- Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548. doi:10.1038/21119.
- Fuhrman, J. A., and Suttle, C. A. (1993). Viruses in marine planktonic systems. *Oceanography* 6, 51–63. Available at: <http://www.jstor.org/stable/43924641>.
- Gaillot, O., Pellegrini, E., Bregenholt, S., Nair, S., and Berche, P. (2002). The ClpP serine protease is essential for the intracellular parasitism and virulence of *Listeria monocytogenes*. *Mol. Microbiol.* 35, 1286–1294. doi:10.1046/j.1365-2958.2000.01773.x.

- Gamble, M., Künze, G., Dodson, E. J., Wilson, K. S., and Jones, D. D. (2011). Regulation of an intracellular subtilisin protease activity by a short propeptide sequence through an original combined dual mechanism. *Proc. Natl. Acad. Sci. U. S. A.* 108, 3536–41. doi:10.1073/pnas.1014229108.
- Gammon, D. B., and Evans, D. H. (2009). The 3'-to-5' exonuclease activity of *Vaccinia* Virus DNA polymerase is essential and plays a role in promoting virus genetic recombination. *J. Virol.* 83, 4236–4250. doi:10.1128/jvi.02255-08.
- Ganai, R. A., and Johansson, E. (2016). DNA replication—a matter of fidelity. *Mol. Cell* 62, 745–755. doi:10.1016/j.molcel.2016.05.003.
- Garapin, A. C., Colbere - Garapin, F., and Cohen-Solal, M. (1981). Expression of herpes simplex virus type I thymidine kinase gene in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 78, 815–819. doi:10.1073/pnas.78.2.815.
- García-Moyano, A., Diaz, Y., Navarro, J., Almendral, D., Puntervoll, P., Ferrer, M., et al. (2021). Two-step functional screen on multiple proteinaceous substrates reveals temperature-robust proteases with a broad-substrate range. *Appl. Microbiol. Biotechnol.*, 1–15. doi:10.1007/s00253-021-11235-9.
- Geertsma, E. R., and Dutzler, R. (2011). A versatile and efficient high-throughput cloning tool for structural biology. *Biochemistry* 50, 3272–8. doi:10.1021/bi200178z.
- GenScript (2021). GenSmart™ Codon Optimization Tool, GenScript. Available at: <https://www.genscript.com/gensmart-free-gene-codon-optimization.html> [Accessed March 23, 2021].
- Gil, J. F., Mesa, V., Estrada-Ortiz, N., Lopez-Obando, M., Gómez, A., and Plácido, J. (2021). Viruses in extreme environments, current overview, and biotechnological potential. *Viruses* 13, 81. doi:10.3390/v13010081.
- Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., et al. (2002). Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl. Environ. Microbiol.* 68, 4301–4306. doi:10.1128/AEM.68.9.4301-4306.2002.
- Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb. Protoc.* 2010, pdb.prot5368. doi:10.1101/pdb.prot5368.
- Golan, G., Zharkov, D. O., Grollman, A. P., Dodson, M. L., McCullough, A. K., Lloyd, R. S., et al. (2006). Structure of T4 pyrimidine dimer glycosylase in a reduced imine covalent complex with abasic site-containing DNA. *J. Mol. Biol.* 362, 241–258. doi:10.1016/j.jmb.2006.06.059.
- Goldberg, S. L., Nanduri, V. B., Chu, L., Johnston, R. M., and Patel, R. N. (2006). Enantioselective microbial reduction of 6-oxo-8-[4-[4-(2-pyrimidinyl)-1-piperazinyl]butyl]-8-azaspiro[4.5]decane-7,9-dione: Cloning and expression of reductases. *Enzyme Microb. Technol.* 39, 1441–1450. doi:10.1016/j.enzmictec.2006.03.033.
- Goldstein, J. N., and Weller, S. K. (1998). The exonuclease activity of HSV-1 UL12 is required for *in vivo* function.
- Goll, J., Rusch, D. B., Tanenbaum, D. M., Thiagarajan, M., Li, K., Methé, B. A., et al. (2010). METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics* 26, 2631–2632. doi:10.1093/bioinformatics/btq455.
- Gong, M., Gong, F., and Yanofsky, C. (2006). Overexpression of tnaC of *Escherichia coli* inhibits growth by depleting tRNA^{2Pro} availability. *J. Bacteriol.* 188, 1892–1898. doi:10.1128/jb.188.5.1892-1898.2006.
- Goodman, D. B., Church, G. M., and Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science (80-)*. 342, 475–479. doi:10.1126/science.1241934.
- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi:10.1038/nrg.2016.49.
- Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., et al. (2020). The new

- scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* 5, 668–674. doi:10.1038/s41564-020-0709-x.
- Gordon, L. K., and Haseltines, W. A. (1980). Comparison of the Cleavage of Pyrimidine Dimers by the Bacteriophage T4 and *Micrococcus luteus* UV-specific Endonucleases* A two-step model. doi:10.1016/S0021-9258(19)70242-7.
- Gould, N., Hendy, O., and Papamichail, D. (2014). Computational tools and algorithms for designing customized synthetic genes. *Front. Bioeng. Biotechnol.* 2. doi:10.3389/fbioe.2014.00041.
- Gouy, M., and Gautier, C. (1982). Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Res.* 10, 7055–7074. doi:10.1093/nar/10.22.7055.
- Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., et al. (2019). Marine DNA viral macro- and microdiversity from Pole to Pole. *Cell* 177, 1109–1123.e14. doi:10.1016/j.cell.2019.03.040.
- Griswold, K. E., Mahmood, N. A., Iverson, B. L., and Georgiou, G. (2003). Effects of codon usage versus putative 5'-mRNA structure on the expression of *Fusarium solani* cutinase in the *Escherichia coli* cytoplasm. *Protein Expr. Purif.* 27, 134–142. doi:10.1016/S1046-5928(02)00578-8.
- Grose, J. H., and Casjens, S. R. (2019). “Bacteriophage Diversity,” in *Reference Module in Life Sciences* (Elsevier). doi:10.1016/b978-0-12-809633-8.20954-0.
- Grummt, I., and Gross, H. J. (1980). Structural organization of mouse rDNA: Comparison of transcribed and non-transcribed regions. *MGG Mol. Gen. Genet.* 177, 223–229. doi:10.1007/BF00267433.
- Gudmundsdóttir, Á., and Pálsdóttir, H. M. (2005). Atlantic Cod Trypsins: From Basic Research to Practical Applications. *Mar. Biotechnol.* 7, 77–88. doi:10.1007/s10126-004-0061-9.
- Gupta, R., Beg, Q., and Lorenz, P. (2002). Bacterial alkaline proteases: molecular approaches and industrial applications. *Appl. Microbiol. Biotechnol.* 59, 15–32. doi:10.1007/s00253-002-0975-y.
- Gupta, R., and Prasad, Y. (2011). P-27/HP endolysin as antibacterial agent for antibiotic resistant *Staphylococcus aureus* of human infections. *Curr. Microbiol.* 63, 39–45. doi:10.1007/s00284-011-9939-8.
- Gustafsson, C., Govindarajan, S., and Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends Biotechnol.* 22, 346–353. doi:10.1016/j.tibtech.2004.04.006.
- Gutiérrez-González, M., Fariás, C., Tello, S., Pérez-Etcheverry, D., Romero, A., Zúñiga, R., et al. (2019). Optimization of culture conditions for the expression of three different insoluble proteins in *Escherichia coli*. *Sci. Rep.* 9, 1–11. doi:10.1038/s41598-019-53200-7.
- Haché, M., Swoboda, K. J., Sethna, N., Farrow-Gillespie, A., Khandji, A., Xia, S., et al. (2016). Intrathecal Injections in Children with Spinal Muscular Atrophy: Nusinersen Clinical Trial Experience. *J. Child Neurol.* 31, 899–906. doi:10.1177/0883073815627882.
- Hale, R. S., and Thompson, G. (1998). Codon optimization of the gene encoding a domain from human type 1 neurofibromin protein results in a threefold improvement in expression level in *Escherichia coli*. *Protein Expr. Purif.* 12, 185–188. doi:10.1006/prep.1997.0825.
- Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–85. doi:10.1128/MMBR.68.4.669-685.2004.
- Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245-9. doi:10.1016/s1074-5521(98)90108-9.
- Harrington, J. J., and Lieber, M. R. (1994). The characterization of a mammalian DNA structure-specific endonuclease. *EMBO J.* 13, 1235–1246. doi:10.1002/j.1460-2075.1994.tb06373.x.

-
- Hatfull, G. F. (2015). Dark Matter of the Biosphere: the Amazing World of Bacteriophage Diversity. *J. Virol.* 89, 8107–8110. doi:10.1128/jvi.01340-15.
- He, T., Li, H., and Zhang, X. (2017). Deep-sea hydrothermal vent viruses compensate for microbial metabolism in virus-host interactions. *MBio* 8. doi:10.1128/mBio.00893-17.
- Henikoff, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* 28, 351–359.
- Henrich, S., Lindberg, I., Bode, W., and Than, M. E. (2005). Proprotein convertase models based on the crystal structures of furin and kexin: Explanation of their specificity. *J. Mol. Biol.* 345, 211–227. doi:10.1016/j.jmb.2004.10.050.
- Hermoso, J. A., García, J. L., and García, P. (2007). Taking aim on bacterial pathogens: from phage therapy to enzybiotics. *Curr. Opin. Microbiol.* 10, 461–472. doi:10.1016/j.mib.2007.08.002.
- Higgins, K. M., and Lloyd, R. S. (1987). Purification of the T4 endonuclease V. *Mutat. Res. Repair Reports* 183, 117–121. doi:10.1016/0167-8817(87)90053-8.
- Hillier, C. J., Ware, L. A., Barbosa, A., Angov, E., Lyon, J. A., Heppner, D. G., et al. (2005). Process development and analysis of liver-stage antigen 1, a preerythrocyte-stage protein-based vaccine for *Plasmodium falciparum*. *Infect. Immun.* 73, 2109–2115. doi:10.1128/IAI.73.4.2109-2115.2005.
- Hizi, A., McGill, C., and Hughes, S. H. (1988). Expression of soluble, enzymatically active, human immunodeficiency virus reverse transcriptase in *Escherichia coli* and analysis of mutants. *Proc. Natl. Acad. Sci. U. S. A.* 85, 1218–1222. doi:10.1073/pnas.85.4.1218.
- Ho Jeon, J., Kim, J.-T., Jae Kim, Y., Kim, H.-K., Sook Lee, H., Gyun Kang, S., et al. Cloning and characterization of a new cold-active lipase from a deep-sea sediment metagenome. doi:10.1007/s00253-008-1656-2.
- Ho, K. (2001). Bacteriophage therapy for bacterial infections: Rekindling a memory from the pre-antibiotics era. *Perspect. Biol. Med.* 44, 1–16. doi:10.1353/pbm.2001.0006.
- Hobbs, Z., and Abedon, S. T. (2016). Diversity of phage infection types and associated terminology: the problem with “Lytic or lysogenic.” *FEMS Microbiol. Lett.* 363, 47. doi:10.1093/femsle/fnw047.
- Holsinger, K. E., and Jansen, R. K. (1993). Phylogenetic Analysis of Restriction Site Data. *Methods Enzymol.* 224, 439–455. doi:10.1016/0076-6879(93)24034-R.
- Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U. S. A.* 111, 4904–4909. doi:10.1073/pnas.1402564111.
- Hsia, K. C., Li, C. L., and Yuan, H. S. (2005). Structural and functional insight into sugar-nonspecific nucleases in host defense. *Curr. Opin. Struct. Biol.* 15, 126–134. doi:10.1016/j.sbi.2005.01.015.
- Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., et al. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi:10.1093/molbev/msx148.
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi:10.1093/nar/gky1085.
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1, 1–6. doi:10.1038/nmicrobiol.2016.48.
- Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., et al. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol.* 16, 1–18. doi:10.1186/s13059-015-0834-7.

- Huson, D. H., Beier, S., Flade, I., Górská, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Comput. Biol.* 12, e1004957. doi:10.1371/journal.pcbi.1004957.
- Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* 151, 389–409. doi:10.1016/0022-2836(81)90003-6.
- Illanes, A. (2008). *Enzyme Biocatalysis*. , ed. A. Illanes Dordrecht: Springer Netherlands doi:10.1007/978-1-4020-8361-7.
- Indraningrat, A. A. G., Smidt, H., and Sipkema, D. (2016). Bioprospecting Sponge-Associated Microbes for Antimicrobial Compounds. *Mar. Drugs* 14. doi:10.3390/md14050087.
- Islam, M. R., Son, N., Lee, J., Lee, D. W., Sohn, E. J., and Hwang, I. (2019). Production of bacteriophage-encoded endolysin, LysP11, in *Nicotiana benthamiana* and its activity as a potent antimicrobial agent against *Erysipelothrix rhusiopathiae*. *Plant Cell Rep.* 38, 1485–1499. doi:10.1007/s00299-019-02459-1.
- Jacquemin, G., Margiotta, D., Kasahara, A., Bassoy, E. Y., Walch, M., Thiery, J., et al. (2015). Granzyme B-induced mitochondrial ROS are required for apoptosis. *Cell Death Differ.* 22, 862–874. doi:10.1038/cdd.2014.180.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* (80-.). 337, 816–821. doi:10.1126/science.1225829.
- Jo, J., Oh, J., and Park, C. (2020). Microbial community analysis using high-throughput sequencing technology: a beginner's guide for microbiologists. *J. Microbiol.* 58, 176–192. doi:10.1007/s12275-020-9525-5.
- Joshi, H., and Jain, V. (2017). Novel method to rapidly and efficiently lyse *Escherichia coli* for the isolation of recombinant protein. *Anal. Biochem.* 528, 1–6. doi:10.1016/j.ab.2017.04.009.
- Jünemann, S., Kleinbölting, N., Jaenicke, S., Henke, C., Hassa, J., Nelkner, J., et al. (2017). Bioinformatics for NGS-based metagenomics and the application to biogas research. *J. Biotechnol.* 261, 10–23. doi:10.1016/j.jbiotec.2017.08.012.
- Kane, J. F. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* 6, 494–500. doi:10.1016/0958-1669(95)80082-4.
- Kapust, R. B., and Waugh, D. S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 8, 1668–1674. doi:10.1110/ps.8.8.1668.
- Katzen, F. (2007). Gateway[®] recombinational cloning: a biological operating system. *Expert Opin. Drug Discov.* 2, 571–589. doi:10.1517/17460441.2.4.571.
- Ki, M. R., and Pack, S. P. (2020). Fusion tags to enhance heterologous protein expression. *Appl. Microbiol. Biotechnol.* 104, 2411–2425. doi:10.1007/s00253-020-10402-8.
- Kim, J., and Dordick, J. S. (1997). Unusual salt and solvent dependence of a protease from an extreme halophile. *Biotechnol. Bioeng.* 55, 471–479. doi:10.1002/(SICI)1097-0290(19970805)55:3<471::AID-BIT2>3.0.CO;2-9.
- Kindler, E., Gil-Cruz, C., Spanier, J., Li, Y., Wilhelm, J., Rabouw, H. H., et al. (2017). Early endonuclease-mediated evasion of RNA sensing ensures efficient coronavirus replication. *PLOS Pathog.* 13, e1006195. doi:10.1371/journal.ppat.1006195.
- King, A. M. Q., Lefkowitz, E., Adams, M. J., and Carstens, E. B. (2011). *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. Elsevier.
- Kink, J. A., Maley, M. E., Ling, K. Y., Kanabrocki, J. A., and Kung, C. (1991). Efficient Expression of the Paramecium Calmodulin Gene in *Escherichia coli* after Four TAA-to-CAA Changes through a Series of Polymerase Chain

Reactions. *J. Protozool.* 38, 441–447. doi:10.1111/j.1550-7408.1991.tb04814.x.

- Kirchman, D. L., Morán, X. A. G., and Ducklow, H. (2009). Microbial growth in the polar oceans - Role of temperature and potential impact of climate change. *Nat. Rev. Microbiol.* 7, 451–459. doi:10.1038/nrmicro2115.
- Klett, R. P., Cerami, A., and Reich, E. (1968). Exonuclease VI, a new nuclease activity associated with *E. coli* DNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* 60, 943–950. doi:10.1073/pnas.60.3.943.
- Kodzius, R., and Gojobori, T. (2015). Marine metagenomics as a source for bioprospecting. *Mar. Genomics* 24, 21–30. doi:10.1016/j.margen.2015.07.001.
- Komai, T., Ishikawa, Y., Yagi, R., Suzuki-Sunagawa, H., Nishigaki, T., and Handa, H. (1997). Development of HIV-1 protease expression methods using the T7 phage promoter system. *Appl. Microbiol. Biotechnol.* 47, 241–245. doi:10.1007/s002530050920.
- Konstantinidis, K. T., Braff, J., Karl, D. M., and DeLong, E. F. (2009). Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific Subtropical Gyre. *Appl. Environ. Microbiol.* 75, 5345–5355. doi:10.1128/AEM.00473-09.
- Kovall, R. A., and Matthews, B. W. (1998). Structural, functional, and evolutionary relationships between λ -exonuclease and the type II restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 95, 7893–7897. doi:10.1073/pnas.95.14.7893.
- Kovall, R., and Matthews, B. W. (1997). Toroidal structure of λ -exonuclease. *Science (80-)*. 277, 1824–1827. doi:10.1126/science.277.5333.1824.
- Krishnamurthy, S. R., and Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Res.* 239, 136–142. doi:10.1016/j.virusres.2017.02.002.
- Kristensen, D. M., Waller, A. S., Yamada, T., Bork, P., Mushegian, A. R., and Koonin, E. V. (2013). Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* 195, 941–950. doi:10.1128/jb.01801-12.
- Kuipers, R. K., Joosten, H. J., Van Berkel, W. J. H., Leferink, N. G. H., Rooijen, E., Ittmann, E., et al. (2010). 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins Struct. Funct. Bioinforma.* 78, 2101–2113. doi:10.1002/prot.22725.
- Kurland, C., and Gallant, J. (1996). Errors of heterologous protein expression. *Curr. Opin. Biotechnol.* 7, 489–493. doi:10.1016/S0958-1669(96)80050-4.
- Kwon, K., Hasseman, J., Latham, S., Grose, C., Do, Y., Fleischmann, R. D., et al. (2011). Recombinant expression and functional analysis of proteases from *Streptococcus pneumoniae*, *Bacillus anthracis*, and *Yersinia pestis*. *BMC Biochem.* 12, 17. doi:10.1186/1471-2091-12-17.
- Lam, P., Cowen, J. P., and Jones, R. D. (2004). Autotrophic ammonia oxidation in a deep-sea hydrothermal plume. *FEMS Microbiol. Ecol.* 47, 191–206. doi:10.1016/S0168-6496(03)00256-3.
- Lambert, G., Charlton, F., Rye, K. A., and Piper, D. E. (2009). Molecular basis of PCSK9 function. *Atherosclerosis* 203, 1–7. doi:10.1016/j.atherosclerosis.2008.06.010.
- Lämmle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H., et al. (2007). Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J. Biotechnol.* 127, 575–592. doi:10.1016/j.jbiotec.2006.07.036.
- Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U. S. A.* 82, 6955–6959. doi:10.1073/pnas.82.20.6955.
- Larsen, Ø., and Bjerga, G. (2018). Development of versatile vectors for heterologous expression in *Bacillus*. *Microorganisms* 6, 51. doi:10.3390/microorganisms6020051.

- Laskowski Sr, M. (1982). Nucleases: historical perspectives. *Nucleases (Linn, S. Roberts, RJ, Eds.)*, 1–21.
- Le Moine Bauer, S., Stensland, A., Daae, F. L., Sandaa, R. A., Thorseth, I. H., Steen, I. H., et al. (2018). Water masses and depth structure prokaryotic and T4-like viral communities around hydrothermal systems of the Nordic Seas. *Front. Microbiol.* 9, 1002. doi:10.3389/fmicb.2018.01002.
- Leavitt, A. D., Roberts, T. M., and Garcea, R. L. (1985). Polyoma virus major capsid protein, VP1. Purification after high level expression in *Escherichia coli*. *J. Biol. Chem.* 260, 12803–12809. doi:10.1016/s0021-9258(17)38948-2.
- Lederberg, E. M., and Lederberg, J. (1953). Genetic studies of lysogenicity in *Escherichia coli*. *Genetics* 38, 51–64. doi:10.1093/genetics/38.1.51.
- Ledford, H., and Callaway, E. (2020). Pioneers of revolutionary CRISPR gene editing win chemistry Nobel. *Nature* 586, 346–347. doi:10.1038/d41586-020-02765-9.
- Lee, M. S., Hseu, Y. C., Lai, G. H., Chang, W. Te, Chen, H. J., Huang, C. H., et al. (2011). High yield expression in a recombinant *E. coli* of a codon optimized chicken anemia virus capsid protein VP1 useful for vaccine development. *Microb. Cell Fact.* 10, 56. doi:10.1186/1475-2859-10-56.
- Lee, M. S., Lien, Y. Y., Feng, S. H., Huang, R. L., Tsai, M. C., Chang, W. Te, et al. (2009). Production of chicken anemia virus (CAV) VP1 and VP2 protein expressed by recombinant *Escherichia coli*. *Process Biochem.* 44, 390–395. doi:10.1016/j.procbio.2008.11.016.
- Leplae, R., Lima-Mendez, G., and Toussaint, A. (2010). ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.* 38, D57–D61. doi:10.1093/nar/gkp938.
- Lewin, A., Lale, R., and Wentzel, A. (2017). “Expression platforms for functional metagenomics: Emerging technology options beyond *Escherichia coli*,” in *Functional Metagenomics: Tools and Applications* (Springer International Publishing), 13–44. doi:10.1007/978-3-319-61510-3_2.
- Li, A. N., and Li, D. C. (2009). Cloning, expression and characterization of the serine protease gene from *Chaetomium thermophilum*. *J. Appl. Microbiol.* 106, 369–380. doi:10.1111/j.1365-2672.2008.04042.x.
- Li, L., Lin, S., and Yanga, F. (2005). Functional identification of the non-specific nuclease from white spot syndrome virus. *Virology* 337, 399–406. doi:10.1016/j.virol.2005.04.017.
- Li, L., and Rohrmann, G. F. (2000). Characterization of a baculovirus alkaline nuclease †. Available at: <http://jvi.asm.org/> [Accessed February 25, 2021].
- Li, L. Y., Luo, X., and Wang, X. (2001). Endonuclease G is an apoptotic DNase when released from mitochondria. *Nature* 412, 95–99. doi:10.1038/35083620.
- Li, W., O’Neill, K. R., Haft, D. H., Dicuccio, M., Chetvermin, V., Badretin, A., et al. (2021). RefSeq: Expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.* 49, D1020–D1028. doi:10.1093/nar/gkaa1105.
- Liebl, W., Angelov, A., Juergensen, J., Chow, J., Loeschcke, A., Drepper, T., et al. (2014). Alternative hosts for functional (meta)genome analysis. *Appl. Microbiol. Biotechnol.* 98, 8099–8109. doi:10.1007/s00253-014-5961-7.
- Liekņiņa, I., Kalniņš, G., Akopjana, I., Bogans, J., Šišovs, M., Jansons, J., et al. (2019). Production and characterization of novel ssRNA bacteriophage virus-like particles from metagenomic sequencing data. *J. Nanobiotechnology* 17, 61. doi:10.1186/s12951-019-0497-8.
- Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* 24, 863–865. doi:10.1093/bioinformatics/btn043.
- Limor-Waisberg, K., Carmi, A., Scherz, A., Pilpel, Y., and Furman, I. (2011). Specialization versus adaptation: Two strategies employed by cyanophages to enhance their translation efficiencies. *Nucleic Acids Res.* 39, 6016–6028. doi:10.1093/nar/gkr169.

- Little, J. W. (1967). An exonuclease induced by Bacteriophage λ : II. Nature of the enzymatic reaction. *J. Biol. Chem.* 242, 679–686. doi:10.1016/s0021-9258(18)96258-7.
- Little, J. W., Lehman, I. R., and Kaiser, A. D. (1967). An exonuclease induced by Bacteriophage λ : I. Preparation of the crystalline enzyme. *J. Biol. Chem.* 242, 672–678. doi:10.1016/S0021-9258(18)96257-5.
- Liu, W., Li, M., and Yan, Y. (2017). Heterologous expression and characterization of a new lipase from *Pseudomonas fluorescens* Pf0-1 and used for biodiesel production. *Sci. Rep.* 7. doi:10.1038/s41598-017-16036-7.
- Liu, Y. (2020). A code within the genetic code: Codon usage regulates co-translational protein folding. *Cell Commun. Signal.* 18. doi:10.1186/s12964-020-00642-6.
- Liu, Y., Yang, Q., and Zhao, F. (2021). Synonymous but not Silent: the codon usage code for gene expression and protein folding. *Annu. Rev. Biochem.* 90. doi:10.1146/annurev-biochem-071320-112701.
- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi:10.1093/nar/gkt1178.
- Lopes, A., Tavares, P., Petit, M. A., Guérois, R., and Zinn-Justin, S. (2014). Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics* 15. doi:10.1186/1471-2164-15-1027.
- López-Otín, C., and Bond, J. S. (2008). Proteases: multifunctional enzymes in life and disease. *J. Biol. Chem.* 283, 30433–7. doi:10.1074/jbc.R800035200.
- Lorenz, P., and Eck, J. (2005). Metagenomics and industrial applications. *Nat Rev Micro* 3, 510–516. Available at: <http://dax.doi.org/10.1038/nrmicro1161>.
- Lossouarn, J., Dupont, S., Gorlas, A., Mercier, C., Biennu, N., Marguet, E., et al. (2015). An abyssal mobilome: Viruses, plasmids and vesicles from deep-sea hydrothermal vents. *Res. Microbiol.* 166, 742–752. doi:10.1016/j.resmic.2015.04.001.
- Lovett, S. T. (2011). The DNA exonucleases of *Escherichia coli*. *EcoSal Plus* 4. doi:10.1128/ecosalplus.4.4.7.
- Lucks, J. B., Nelson, D. R., Kudla, G. R., and Plotkin, J. B. (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput. Biol.* 4, e1000001. doi:10.1371/journal.pcbi.1000001.
- Lwoff, A. (1957). The concept of virus. *J. Gen. Microbiol.* 17, 239–253. doi:10.1099/00221287-17-2-239.
- Lyamichev, V., Brow, M. A. D., and Dahlberg, J. E. (1993). Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science (80-)*. 260, 778–783. doi:10.1126/science.7683443.
- Lynch, M. D. J., and Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* 13, 217–229. doi:10.1038/nrmicro3400.
- Ma, F., Guo, X., and Fan, H. (2017). Extracellular nucleases of *Streptococcus equi* subsp. *zooepidemicus* degrade neutrophil extracellular traps and impair macrophage activity of the host. *Appl. Environ. Microbiol.* 83. doi:10.1128/AEM.02468-16.
- Maat, D., Biggs, T., Evans, C., van Bleijswijk, J., van der Wel, N., Dutilh, B., et al. (2017). Characterization and temperature dependence of Arctic *Micromonas polaris* viruses. *Viruses* 9, 134. doi:10.3390/v9060134.
- Makrides, S. C. (1996). Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev.* 60, 512–38. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8840785> [Accessed October 9, 2018].
- Malakhov, M. P., Mattern, M. R., Malakhova, O. A., Drinker, M., Weeks, S. D., and Butt, T. R. (2004). SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *J. Struct. Funct. Genomics* 5, 75–86. doi:10.1023/B:JSFG.0000029237.70316.52.
- Marblestone, J. G., Edavettal, S. C., Lim, Y., Lim, P., Zuo, X., and Butt, T. R. (2006). Comparison of SUMO fusion

- technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Sci.* 15, 182–9. doi:10.1110/ps.051812706.
- Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi:10.1093/nar/gkr1044.
- Marti, T. M., and Fleck, O. (2004). DNA repair nucleases. *Cell. Mol. Life Sci.* 61, 336–354. doi:10.1007/s00018-003-3223-4.
- Martínez-García, S., Rodríguez-Martínez, S., Cancino-Díaz, M. E., and Cancino-Díaz, J. C. (2018). Extracellular proteases of *Staphylococcus epidermidis*: roles as virulence factors and their participation in biofilm. *APMIS* 126, 177–185. doi:10.1111/apm.12805.
- Marx, J. C., Collins, T., D'Amico, S., Feller, G., and Gerday, C. (2007). Cold-adapted enzymes from marine Antarctic microorganisms. *Mar. Biotechnol.* 9, 293–304. doi:10.1007/s10126-006-6103-8.
- Matsuzaki, S., Rashel, M., Uchiyama, J., Sakurai, S., Ujihara, T., Kuroda, M., et al. (2005). Bacteriophage therapy: A revitalized therapy against bacterial infectious diseases. *J. Infect. Chemother.* 11, 211–219. doi:10.1007/s10156-005-0408-9.
- Mauger, D. M., Joseph Cabral, B., Presnyak, V., Su, S. V., Reid, D. W., Goodman, B., et al. (2019). mRNA structure regulates protein expression through changes in functional half-life. *Proc. Natl. Acad. Sci. U. S. A.* 116, 24075–24083. doi:10.1073/pnas.1908052116.
- McClure, S. B., Magill, C., Podrug, E., Moore, A. M. T., Harper, T. K., Culleton, B. J., et al. (2018). Fatty acid specific $\delta^{13}\text{C}$ values reveal earliest Mediterranean cheese production 7,200 years ago. *PLoS One* 13, e0202807. doi:10.1371/journal.pone.0202807.
- McDonald, A. G., Boyce, S., and Tipton, K. F. (2009). ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.* 37, D593–D597. doi:10.1093/nar/gkn582.
- Mesnage, S., Dellarole, M., Baxter, N. J., Rouget, J. B., Dimitrov, J. D., Wang, N., et al. (2014). Molecular basis for bacterial peptidoglycan recognition by LysM domains. *Nat. Commun.* 5, 1–11. doi:10.1038/ncomms5269.
- Michalska, K., Steen, A. D., Chhor, G., Endres, M., Webber, A. T., Bird, J., et al. (2015). New aminopeptidase from “microbial dark matter” archaeon. *FASEB J.* 29, 4071–4079. doi:10.1096/fj.15-272906.
- Mimitou, E. P., and Symington, L. S. (2009). Nucleases and helicases take center stage in homologous recombination. *Trends Biochem. Sci.* 34, 264–272. doi:10.1016/j.tibs.2009.01.010.
- Mitrofanova, O., Mardanov, A., Evtugyn, V., Bogomolnaya, L., and Sharipova, M. (2017). Effects of *Bacillus* serine proteases on the bacterial biofilms. *Biomed Res. Int.* 2017. doi:10.1155/2017/8525912.
- Miyake, R., Shigeri, Y., Tatsu, Y., Yumoto, N., Umekawa, M., Tsujimoto, Y., et al. (2005). Two thimet oligopeptidase-like Pz peptidases produced by a collagen-degrading thermophile, *Geobacillus collagenovorans* MO-1. *J. Bacteriol.* 187, 4140–8. doi:10.1128/jb.187.12.4140-4148.2005.
- Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi:10.1016/j.coviro.2011.12.004.
- Mol, C. D., Kuo, C.-F., Thayer, M. M., Cunningham, R. P., and Tainer, J. A. (1995). Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature* 374, 381–386. doi:10.1038/374381a0.
- Momtazi-Borojeni, A. A., Sabouri-Rad, S., Gotto, A. M., Pirro, M., Banach, M., Awan, Z., et al. (2019). PCSK9 and inflammation: a review of experimental and clinical evidence. *Eur. Hear. J. - Cardiovasc. Pharmacother.* 5, 237–245. doi:10.1093/ejcvp/pvz022.
- Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J. C., Lee, J., et al. (2021). Genomes OnLine

- Database (GOLD) v.8: Overview and updates. *Nucleic Acids Res.* 49, D723–D733. doi:10.1093/nar/gkaa983.
- Mullineaux, C. W., Nenner, A., Ray, N., and Robinson, C. (2006). Diffusion of green fluorescent protein in three cell environments in *Escherichia coli*. *J. Bacteriol.* 188, 3442–3448. doi:10.1128/jb.188.10.3442-3448.2006.
- Muszewska, A., Stepniewska-Dziubinska, M. M., Steczkiewicz, K., Pawlowska, J., Dziedzic, A., and Ginalski, K. (2017). Fungal lifestyle reflected in serine protease repertoire. *Sci. Rep.* 7, 1–12. doi:10.1038/s41598-017-09644-w.
- Nakamura, K., and Takai, K. (2014). Theoretical constraints of physical and chemical properties of hydrothermal fluids on variations in chemolithotrophic microbial communities in seafloor hydrothermal systems. *Prog. Earth Planet. Sci.* 1, 1–24. doi:10.1186/2197-4284-1-5.
- Nakata, P. A. (2017). Construction of pDUO: A bicistronic shuttle vector series for dual expression of recombinant proteins. *Plasmid* 89, 16–21. doi:10.1016/j.plasmid.2016.12.001.
- Nelson, D., Loomis, L., and Fischetti, V. A. (2001). Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4107–4112. doi:10.1073/pnas.061038398.
- New England Biolabs (2021). Enzyme Finder. Available at: <https://enzyme finder.neb.com/#!/%23nebheader> [Accessed February 25, 2021].
- Niehaus, F., Gabor, E., Wieland, S., Siegert, P., Maurer, K. H., and Eck, J. (2011). Enzymes for the laundry industries: tapping the vast metagenomic pool of alkaline proteases. *Microb. Biotechnol.* 4, 767–76. doi:10.1111/j.1751-7915.2011.00279.x.
- NorZymeD (2021). NorZymeD. Available at: <https://norzymed.nmbu.no/> [Accessed March 21, 2021].
- Novagen Inc. (2006). “pET System Manual,” in *TB055*, 19–24. Available at: http://kirschner.med.harvard.edu/files/protocols/Novagen_petsystem.pdf [Accessed September 30, 2020].
- Öberg, F., Sjöhamn, J., Conner, M. T., Bill, R. M., and Hedfalk, K. (2011). Improving recombinant eukaryotic membrane protein yields in *Pichia pastoris*: The importance of codon optimization and clone selection. *Mol. Membr. Biol.* 28, 398–411. doi:10.3109/09687688.2011.602219.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34. doi:10.1093/nar/27.1.29.
- Ohta, Y., Hojo, H., Aimoto, S., Kobayashi, T., Zhu, X., Jordan, F., et al. (1991). Pro-peptide as an intermolecular chaperone: renaturation of denatured subtilisin E with a synthetic pro-peptide. *Mol. Microbiol.* 5, 1507–1510. doi:10.1111/j.1365-2958.1991.tb00797.x.
- Olsen, N. S., Forero-Junco, L., Kot, W., and Hansen, L. H. (2020). Exploring the remarkable diversity of culturable *Escherichia coli* phages in the danish wastewater environment. *Viruses* 12, 986. doi:10.3390/v12090986.
- Orrego, C., Kerjan, P., Manca De Nadra, M. C., and Szulmajster, J. (1973). Ribonucleic acid polymerase in a thermosensitive sporulation mutant (ts-4) of *Bacillus subtilis*. *J. Bacteriol.* 116, 636–647. Available at: <https://pdfs.semanticscholar.org/b18c/f398969c46f5cfbc9566503f9c9d9e7d0995.pdf> [Accessed March 12, 2018].
- Ortmann, A. C., and Suttle, C. A. (2005). High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep. Res. Part I Oceanogr. Res. Pap.* 52, 1515–1527. doi:10.1016/j.dsr.2005.04.002.
- Ottesen, M., and Svendsen, I. (1970). The subtilisins. *Methods Enzymol.* 19, 199–215. doi:10.1016/0076-6879(70)19014-8.
- Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1986). “The analysis of natural microbial populations by ribosomal RNA sequences,” in (Springer, Boston, MA), 1–55. doi:10.1007/978-1-4757-0611-6_1.

-
- Paez-Espino, D., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., et al. (2016). IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res.* 45. doi:10.1093/nar/gkw1030.
- Pantoliano, M. W., Whitlow, M., Wood, J. F., Rollence, M. L., Finzel, B. C., Gilliland, G. L., et al. (1988). The engineering of binding affinity at metal ion binding sites for the stabilization of proteins: subtilisin as a test case. *Biochemistry* 27, 8311–7. doi:10.1021/bi00422a004.
- Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542. doi:10.1038/s41564-017-0012-7.
- Patel, R. N. (2018). Biocatalysis for synthesis of pharmaceuticals. *Bioorg. Med. Chem.* 26, 1252–1274. doi:10.1016/j.bmc.2017.05.023.
- Paull, T. T., and Gellert, M. (1998). The 3' to 5' exonuclease activity of Mre11 facilitates repair of DNA double-strand breaks. *Mol. Cell* 1, 969–979. doi:10.1016/S1097-2765(00)80097-0.
- Payet, J. P., and Suttle, C. A. (2013). To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnol. Oceanogr.* 58, 465–474. doi:10.4319/lo.2013.58.2.0465.
- Payet, J., and Suttle, C. (2014). Viral infection of bacteria and phytoplankton in the Arctic Ocean as viewed through the lens of fingerprint analysis. *Aquat. Microb. Ecol.* 72, 47–61. doi:10.3354/ame01684.
- Pearson, W. R. (2013). An introduction to sequence similarity (“Homology”) searching. *Curr. Protoc. Bioinforma.* Chapter 3, Unit3.1. doi:10.1002/0471250953.bi0301s42.
- Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/3162770> [Accessed September 16, 2018].
- Pedersen, R. B., Rapp, H. T., Thorseth, I. H., Lilley, M. D., Barriga, F. J. A. S., Baumberg, T., et al. (2010). Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge. *Nat. Commun.* 1, 1–6. doi:10.1038/ncomms1124.
- Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi:10.1038/nmeth.1701.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., et al. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* 30, 70–82. doi:10.1002/pro.3943.
- Pfefferbaum, B., and North, C. S. (2020). Mental Health and the Covid-19 Pandemic. *N. Engl. J. Med.* 383, 510–512. doi:10.1056/nejmp2008017.
- Pinchuk, G. E., Ammons, C., Culley, D. E., Li, S. M. W., McLean, J. S., Romine, M. F., et al. (2008). Utilization of DNA as a sole source of phosphorus, carbon, and energy by *Shewanella* spp.: Ecological and physiological implications for dissimilatory metal reduction. *Appl. Environ. Microbiol.* 74, 1198–1208. doi:10.1128/AEM.02026-07.
- Pingoud, A., and Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. *Nucleic Acids Res.* 29, 3705–3727. doi:10.1093/nar/29.18.3705.
- Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., et al. (2017). BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* 45, D380–D388. doi:10.1093/nar/gkw952.
- Plotka, M., Kapusta, M., Dorawa, S., Kaczorowska, A. K., and Kaczorowski, T. (2019). Ts2631 endolysin from the extremophilic thermus scotoductus bacteriophage vB_Tsc2631 as an antimicrobial agent against gram-negative multidrug-resistant bacteria. *Viruses* 11. doi:10.3390/v11070657.
- Plotka, M., Szadkowska, M., Håkansson, M., Kovačič, R., Al-Karadaghi, S., Walse, B., et al. (2020). Molecular characterization of a novel lytic enzyme LysC from *Clostridium intestinale* URNW and its antibacterial activity

- mediated by positively charged N-terminal extension. *Int. J. Mol. Sci.* 21, 4894. doi:10.3390/ijms21144894.
- Pollard, H., Toumaniantz, G., Amos, J.-L., Avet-Loiseau, H., Guihard, G., Behr, J.-P., et al. (2001). Ca²⁺-sensitive cytosolic nucleases prevent efficient delivery to the nucleus of injected plasmids. *J. Gene Med.* 3, 153–164. doi:10.1002/jgm.160.
- Pope, B., and Kent, H. M. (1996). High efficiency 5 min transformation of *Escherichia coli*. *Nucleic Acids Res.* 24, 536–537.
- Popovic, A., Tchigvintsev, A., Tran, H., Chernikova, T. N., Golyshina, O. V., Yakimov, M. M., et al. (2015). “Metagenomics as a tool for enzyme discovery: Hydrolytic enzymes from marine-related metagenomes,” in *Prokaryotic Systems Biology, Advances* (Springer, Cham), 1–20. doi:10.1007/978-3-319-23603-2_1.
- Posma, J. J. N., Posthuma, J. J., and Spronk, H. M. H. (2016). Coagulation and non-coagulation effects of thrombin. *J. Thromb. Haemost.* 14, 1908–1916. doi:10.1111/jth.13441.
- Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi:10.1093/nar/gky448.
- Power, S. D., Adams, R. M., and Wells, J. A. (1986). Secretion and autoproteolytic maturation of subtilisin. *Proc. Natl. Acad. Sci. U. S. A.* 83, 3096–100. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/3517850> [Accessed February 14, 2018].
- Powers, J. C., Asgian, J. L., Ekici, Ö. D., and James, K. E. (2002). Irreversible inhibitors of serine, cysteine, and threonine proteases. *Chem. Rev.* 102, 4639–4750. doi:10.1021/cr010182v.
- Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., et al. (2015). Codon optimality is a major determinant of mRNA stability. *Cell* 160, 1111–1124. doi:10.1016/j.cell.2015.02.029.
- Punde, N., Kooken, J., Leary, D., Legler, P. M., and Angov, E. (2019). Codon harmonization reduces amino acid misincorporation in bacterially expressed *P. falciparum* proteins and improves their immunogenicity. *AMB Express* 9. doi:10.1186/s13568-019-0890-6.
- Pushpam, P., Rajesh, T., and Gunasekaran, P. (2011). Identification and characterization of alkaline serine protease from goat skin surface metagenome. *AMB Express* 1, 3. doi:10.1186/2191-0855-1-3.
- Quan, S., Koldewey, P., Tapley, T., Kirsch, N., Ruane, K. M., Pfizenmaier, J., et al. (2011). Genetic selection designed to stabilize proteins uncovers a chaperone called Spy. *Nat. Struct. Mol. Biol.* 18, 262–269. doi:10.1038/nsmb.2016.
- Raab, D., Graf, M., Notka, F., Schödl, T., and Wagner, R. (2010). The GeneOptimizer Algorithm: Using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization. *Syst. Synth. Biol.* 4, 215–225. doi:10.1007/s11693-010-9062-3.
- Radding, C. M., and Shreffler, D. C. (1966). Regulation of λ exonuclease: II. Joint regulation of exonuclease and a new λ antigen. *J. Mol. Biol.* 18, 251–261. doi:10.1016/S0022-2836(66)80244-9.
- Radisky, E. S., Kwan, G., Lu, C. J. K., and Koshland, D. E. (2004). Binding, proteolytic, and crystallographic analyses of mutations at the protease - Inhibitor interface of the subtilisin BPN'/chymotrypsin inhibitor 2 complex. *Biochemistry* 43, 13648–13656. doi:10.1021/bi048797k.
- Ramón, A., Señorale-Pose, M., and Marin, M. (2014). Inclusion bodies: not that bad.... *Front. Microbiol.* 5, 56. doi:10.3389/fmicb.2014.00056.
- Rangarajan, E. S., and Shankar, V. (2001). Sugar non-specific endonucleases. *FEMS Microbiol. Rev.* 25, 583–613. doi:10.1111/j.1574-6976.2001.tb00593.x.
- Rao, M. B., Tanksale, A. M., Ghatge, M. S., and Deshpande, V. V (1998). Molecular and biotechnological aspects of microbial proteases. *Microbiol. Mol. Biol. Rev.* 62, 597–635.

- Raveendran, S., Parameswaran, B., Ummalyama, S. B., Abraham, A., Mathew, A. K., Madhavan, A., et al. (2018). Applications of microbial enzymes in food industry. *Food Technol. Biotechnol.* 56, 16–30. doi:10.17113/ftb.56.01.18.5491.
- Ravin, N. V., Mardanova, A. V., and Skryabin, K. G. (2015). Metagenomics as a tool for the investigation of uncultured microorganisms. *Genetika* 51, 519–28. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/26137633> [Accessed January 13, 2018].
- Rawlings, N. D., Barrett, A. J., and Bateman, A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 40, D343–D350. doi:10.1093/nar/gkr987.
- Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., and 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 Res.* 46, D624–D632. doi:10.1093/nar/gkx1134.
- Rawlings, N. D., and Salvesen, G. (2012). Handbook of proteolytic enzymes. 3rd ed. Academic Available at: <https://www.sciencedirect.com/science/book/9780123822192> [Accessed March 8, 2018].
- Rawlings, N. D., and Salvesen, G. (2013). Handbook of proteolytic enzymes. Academic press.
- Rawlings, N. D., Waller, M., Barrett, A. J., and Bateman, A. (2014). MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 42, 503–509. doi:10.1093/nar/gkt953.
- Ray, J., Dondrup, M., Modha, S., Steen, I. H., Sandaa, R. A., and Clokie, M. (2012). Finding a needle in the virus metagenome haystack - micro-metagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. *PLoS One* 7. doi:10.1371/journal.pone.0034238.
- Rejisha, R. P., and Murugan, M. (2020). Alkaline protease production by halophilic *Bacillus* sp. strain SP II-4 and characterization with special reference to contact lens cleansing. *Mater. Today Proc.* doi:10.1016/j.matpr.2020.08.624.
- Rice, K. C., and Bayles, K. W. (2008). Molecular control of bacterial death and lysis. *Microbiol. Mol. Biol. Rev.* 72, 85–109. doi:10.1128/mubr.00030-07.
- Richardson, C. C., and Kornberg, A. (1964). A deoxyribonucleic acid phosphatase-exonuclease from *Escherichia coli*: I. Purification of the enzyme and characterization of the phosphatase activity. *J. Biol. Chem.* 239, 242–50. doi:10.1016/s0021-9258(18)51775-0.
- Richardson, C. C., Lehman, I. R., and Kornberg, A. (1964). A Deoxyribonucleic Acid Phosphatase-Exonuclease from *Escherichia coli*: II. Characterization of the Exonuclease Activity. *J. Biol. Chem.* 239, 251–258.
- Rigoldi, F., Donini, S., Redaelli, A., Parisini, E., and Gautieri, A. (2018). Review: Engineering of thermostable enzymes for industrial applications. *Cit. APL Bioeng.* 2, 11501. doi:10.1063/1.4997367.
- Roalkvam, I., Bredy, F., Baumberg, T., Pedersen, R. B., and Steen, I. H. (2015). *Hypnocyclicus thermotrophus* gen. Nov., sp. nov. isolated from a microbial mat in a hydrothermal vent field. *Int. J. Syst. Evol. Microbiol.* 65, 4521–4525. doi:10.1099/ijsem.0.000606.
- Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2015). REBASE—a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res.* 43, D298–D299. doi:10.1093/nar/gku1046.
- Rodríguez-Rubio, L., Martínez, B., Rodríguez, A., Donovan, D. M., Götz, F., and García, P. (2013). The phage lytic proteins from the *Staphylococcus aureus* bacteriophage vB_SauS-phiPLA88 display multiple active catalytic domains and do not trigger Staphylococcal resistance. *PLoS One* 8, e64671. doi:10.1371/journal.pone.0064671.
- Rohwer, F., and Edwards, R. (2002). The phage proteomic tree: A genome-based taxonomy for phage. *J. Bacteriol.* 184, 4529–4535. doi:10.1128/jb.184.16.4529-4535.2002.
- Rokas, A., Williams, B. I., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in

molecular phylogenies. *Nature* 425, 798–804. doi:10.1038/nature02053.

- Rosano, G. L., and Ceccarelli, E. A. (2014). Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* 5, 172. doi:10.3389/fmicb.2014.00172.
- Rosano, G. L., Morales, E. S., and Ceccarelli, E. A. (2019). New tools for recombinant protein production in *Escherichia coli*: A 5-year update. *Protein Sci.* 28, 1412–1422. doi:10.1002/pro.3668.
- Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., and Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front. Genet.* 8, 23. doi:10.3389/fgene.2017.00023.
- Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537, 689–693. doi:10.1038/nature19366.
- Roy, J. J., and Abraham, T. E. (2006). Continuous biotransformation of pyrogallol to purpurogallin using cross-linked enzyme crystals of laccase as catalyst in a packed-bed reactor. *J. Chem. Technol. Biotechnol.* 81, 1836–1839. doi:10.1002/jctb.1612.
- Ruan, A., Ren, C., and Quan, S. (2020). Conversion of the molecular chaperone Spy into a novel fusion tag to enhance recombinant protein expression. *J. Biotechnol.* 307, 131–138. doi:10.1016/j.jbiotec.2019.11.006.
- Sajantila, A., and Budowle, B. (1991). Identification of individuals with DNA testing. *Ann. Med.* 23, 637–642. doi:10.3109/07853899109148096.
- Salwan, R., and Sharma, V. (2019). Trends in extracellular serine proteases of bacteria as detergent bioadditive: alternate and environmental friendly tool for detergent industry. *Arch. Microbiol.* 201, 863–877. doi:10.1007/s00203-019-01662-8.
- Sandaa, R.-A., Short, S. M., and Schroeder, D. C. (2010). Fingerprinting aquatic virus communities. *Man. Aquat. Viral Ecol. ASLO*, 9–18.
- Sandaa, R. A., Storesund, J. E., Olesin, E., Paulsen, M. L., Larsen, A., Bratbak, G., et al. (2018). Seasonality drives microbial community structure, shaping both eukaryotic and prokaryotic host–viral relationships in an arctic marine ecosystem. *Viruses* 10. doi:10.3390/v10120715.
- Santos, S. B., Costa, A. R., Carvalho, C., Nóbrega, F. L., and Azeredo, J. (2018). Exploiting bacteriophage proteomes: the hidden biotechnological potential. *Trends Biotechnol.* 36, 966–984. doi:10.1016/j.tibtech.2018.04.006.
- Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., et al. (2010). Biocatalytic asymmetric synthesis of sitagliptin manufacture. *Science (80-)*. 329, 305–310. doi:10.1126/science.1188934.
- Schaller, A., Stintzi, A., Rivas, S., Serrano, I., Chichkova, N. V., Vartapetian, A. B., et al. (2018). From structure to function - a family portrait of plant subtilases. *New Phytol.* 218, 901–915. doi:10.1111/nph.14582.
- Schechter, I., and Berger, A. (1967). On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* 27, 157–162. doi:10.1016/S0006-291X(67)80055-X.
- Schleper, C., Swanson, R. V., Mathur, E. J., and DeLong, E. F. (1997). Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum*. *J. Bacteriol.* 179, 7803–11. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9401041> [Accessed May 4, 2018].
- Schmelcher, M., Donovan, D. M., and Loessner, M. J. (2012). Bacteriophage endolysins as novel antimicrobials. *Future Microbiol.* 7, 1147–1171. doi:10.2217/fmb.12.97.
- Schmid, A., Dordick, J. S., Hauer, B., Kiener, A., Wubbolts, M., and Witholt, B. (2001). Industrial biocatalysis today and tomorrow. *Nature* 409, 258. Available at: <https://doi.org/10.1038/35051736>.
- Schmidt, T. M., DeLong, E. F., and Pace, N. R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173, 4371–4378. doi:10.1128/jb.173.14.4371-4378.1991.

- Scholz, J., Besir, H., Strasser, C., and Suppmann, S. (2013). A new method to customize protein expression vectors for fast, efficient and background free parallel cloning. *BMC Biotechnol.* 13, 12. doi:10.1186/1472-6750-13-12.
- Schouw, A., Vulcano, F., Roalkvam, I., Hocking, W., Reeves, E., Stokke, R., et al. (2018). Genome analysis of *Vallitalea guaymasensis* Strain L81 isolated from a deep-sea hydrothermal vent system. *Microorganisms* 6, 63. doi:10.3390/microorganisms6030063.
- Seki, N., Muta, T., Oda, T., Iwaki, D., Kuma, K., Miyata, T., et al. (1994). Horseshoe crab (1,3)-beta-D-glucan-sensitive coagulation factor G. A serine protease zymogen heterodimer with similarities to beta-glucan-binding proteins. *J. Biol. Chem.* 269, 1370–1374. Available at: <http://www.jbc.org/content/269/2/1370.abstract>.
- Setlow, P., Brutlag, D., and Kornberg, A. (1972). Deoxyribonucleic acid polymerase: two distinct enzymes in one polypeptide. I. A proteolytic fragment containing the polymerase and 3' leads to 5' exonuclease functions. *J. Biol. Chem.* 247, 224–231. doi:10.1016/S0021-9258(19)45779-7.
- Sezonov, G., Joseleau-Petit, D., and D'Ari, R. (2007). *Escherichia coli* physiology in Luria-Bertani broth. *J. Bacteriol.* 189, 8746–8749. doi:10.1128/jb.01368-07.
- Sharp, P. M., and Li, W. H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. doi:10.1093/nar/15.3.1281.
- Shen, B., Nolan, J. P., Sklar, L. A., and Park, M. S. (1997). Functional analysis of point mutations in human flap endonuclease-1 active site. *Nucleic Acids Res.* 25, 3332–3338. doi:10.1093/nar/25.16.3332.
- Shiloach, J., and Fass, R. (2005). Growing *E. coli* to high cell density—A historical perspective on method development. *Biotechnol. Adv.* 23, 345–357. doi:10.1016/j.biotechadv.2005.04.004.
- Shuman, S., Golder, M., and Moss, B. (1988). Characterization of vaccinia virus DNA topoisomerase I expressed in *Escherichia coli*. *J. Biol. Chem.* 263, 16401–16407. doi:10.1016/s0021-9258(18)37607-5.
- Siezen, R. J., and Leunissen, J. A. M. (1997). Subtilases: The superfamily of subtilisin-like serine proteases. *Protein Sci.* 6, 501–523. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2143677/pdf/9070434.pdf> [Accessed May 19, 2018].
- Silhavy, T. J., Kahne, D., and Walker, S. (2010). The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* 2, a000414. doi:10.1101/cshperspect.a000414.
- Silva Lopez, R. E. da, and De Simone, S. G. (2004). A Serine protease from a detergent-soluble extract of *Leishmania (Leishmania) amazonensis*. *Zeitschrift für Naturforsch. C* 59, 590–598. doi:10.1515/znc-2004-7-825.
- Simmonds, P., Adams, M. J., Benk, M., Breitbart, M., Brister, J. R., Carstens, E. B., et al. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* 15, 161–168. doi:10.1038/nrmicro.2016.177.
- Singh, R., Kumar, M., Mittal, A., and Mehta, P. K. (2016). Microbial enzymes: industrial progress in 21st century. *3 Biotech* 6, 174. doi:10.1007/s13205-016-0485-8.
- Smith, C. A., Toogood, H. S., Baker, H. M., Daniel, R. M., and Baker, E. N. (1999). Calcium-mediated thermostability in the subtilisin superfamily: the crystal structure of *Bacillus* Ak.1 protease at 1.8 Å resolution. *J. Mol. Biol.* 294, 1027–1040. doi:10.1006/jmbi.1999.3291.
- Smith, D. B., and Johnson, K. S. (1988). Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* 67, 31–40. doi:10.1016/0378-1119(88)90005-4.
- Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7265238> [Accessed September 16, 2018].
- Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960. doi:10.1093/bioinformatics/bti125.

- Song, Q., and Zhang, X. (2008). Characterization of a novel non-specific nuclease from thermophilic bacteriophage GBSV1. *BMC Biotechnol.* 8, 43. doi:10.1186/1472-6750-8-43.
- Sousa, A. L. de, Maués, D., Lobato, A., Franco, E. F., Pinheiro, K., Araújo, F., et al. (2018). PhageWeb – Web interface for rapid identification and characterization of prophages in bacterial genomes. *Front. Genet.* 9, 644. doi:10.3389/fgene.2018.00644.
- Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98. doi:10.1016/S0022-2836(75)80083-0.
- Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., et al. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521, 173–179. doi:10.1038/nature14447.
- Srividhya, K. V., Rao, G. V., Raghavenderan, L., Mehta, P., Prilusky, J., Manicka, S., et al. (2006). “Database and Comparative Identification of Prophages,” in *Intelligent Control and Automation* (Springer Berlin Heidelberg), 863–868. doi:10.1007/978-3-540-37256-1_110.
- Staley, J. T., and Konopka, A. (1985). Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial Habitats. *Annu. Rev. Microbiol.* 39, 321–346. doi:10.1146/annurev.mi.39.100185.001541.
- Steen, I. H., Dahle, H., Stokke, R., Roalkvam, I., Daae, F.-L., Rapp, H. T., et al. (2016). Novel barite chimneys at the Loki’s Castle vent field shed light on key factors shaping microbial communities and functions in hydrothermal systems. *Front. Microbiol.* 6, 1510. doi:10.3389/fmicb.2015.01510.
- Stein, R., and MacDonald, R. W. (2004). The organic carbon cycle in the Arctic Ocean. , eds. R. Stein and R. W. MacDonald Berlin, Heidelberg: Springer Berlin Heidelberg doi:10.1007/978-3-642-18912-8.
- Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 20, 473. doi:10.1186/s12859-019-3019-7.
- Stokke, R., Dahle, H., Roalkvam, I., Wissuwa, J., Daae, F. L., Tooming-Klunderud, A., et al. (2015). Functional interactions among filamentous *Epsilonproteobacteria* and *Bacteroidetes* in a deep-sea hydrothermal vent biofilm. *Environ. Microbiol.* 17, 4063–4077. doi:10.1111/1462-2920.12970.
- Strasser, K., McDonnell, E., Nyaga, C., Wu, M., Wu, S., Almeida, H., et al. (2015). mycoCLAP, the database for characterized lignocellulose-active proteins of fungal origin: resource and text mining curation support. *Database* 2015, 8. doi:10.1093/database/bav008.
- Sukul, P., Schäkermann, S., Bandow, J. E., Kusnezowa, A., Nowrousian, M., and Leichert, L. I. (2017). Simple discovery of bacterial biocatalysts from environmental samples through functional metaproteomics. *Microbiome* 5, 28. doi:10.1186/s40168-017-0247-9.
- Sullivan, B., Carrera, I., Drouin, M., and Hudlicky, T. (2009). Symmetry-Based Design for the Chemoenzymatic Synthesis of Oseltamivir (Tamiflu) from Ethyl Benzoate. *Angew. Chemie Int. Ed.* 48, 4229–4231. doi:10.1002/anie.200901345.
- Sumantha, A., Larroche, C., and Pandey, A. (2006). Microbiology and industrial biotechnology of food-grade proteases: A perspective. *Food Technol. Biotechnol.* 44, 211–220.
- Sun, H., Zhang, H., Ang, E. L., and Zhao, H. (2018). Biocatalysis for the synthesis of pharmaceuticals and pharmaceutical intermediates. *Bioorg. Med. Chem.* 26, 1275–1284. doi:10.1016/j.bmc.2017.06.043.
- Suttle, C. A. (2005). Viruses in the sea. *Nature* 437, 356–361. doi:10.1038/nature04160.
- Suttle, C. A. (2007). Marine viruses - Major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812. doi:10.1038/nrmicro1750.
- Taguchi, S., Ozaki, A., Nonaka, T., Mitsui, Y., and Momose, H. (1999). A Cold-Adapted Protease Engineered by

Experimental Evolution System. *J. Biochem.* 126, 689–693. Available at: https://www.jstage.jst.go.jp/article/biochemistry1922/126/4/126_4_689/_pdf [Accessed July 23, 2018].

- Takai, K., and Nakamura, K. (2010). “Compositional, physiological and metabolic variability in microbial communities associated with geochemically diverse, deep-sea hydrothermal vent fluids,” in *Geomicrobiology: Molecular and Environmental Perspective* (Springer Netherlands), 251–283. doi:10.1007/978-90-481-9204-5_12.
- Takai, K., and Nakamura, K. (2011). Archaeal diversity and community development in deep-sea hydrothermal vents. *Curr. Opin. Microbiol.* 14, 282–291. doi:10.1016/j.mib.2011.04.013.
- Takai, K., Nakamura, K., Toki, T., Tsunogai, U., Miyazaki, M., Miyazaki, J., et al. (2008). Cell proliferation at 122°C and isotopically heavy CH₄ production by a hyperthermophilic methanogen under high-pressure cultivation. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10949–10954. doi:10.1073/pnas.0712334105.
- Tanaka, K., Sekiguchi, M., and Okada, Y. (1975). Restoration of ultraviolet induced unscheduled DNA synthesis of xeroderma pigmentosum cells by the concomitant treatment with bacteriophage T4 endonuclease V and HVJ (Sendai virus). *Proc. Natl. Acad. Sci. U. S. A.* 72, 4071–4075. doi:10.1073/pnas.72.10.4071.
- Tang, X. M., Shen, W., Lakay, F. M., Shao, W. L., Wang, Z. X., Prior, B. A., et al. (2004). Cloning and over-expression of an alkaline protease from *Bacillus licheniformis*. *Biotechnol. Lett.* 26, 975–979. doi:10.1023/B:BILE.0000030042.91094.38.
- Thomas And, K. R., and Olivera, B. M. (1978). Processivity of DNA exonucleases*. doi:10.1016/S0021-9258(17)38226-1.
- Toplak, A., Teixeira de Oliveira, E. F., Schmidt, M., Rozeboom, H. J., Wijma, H. J., Meekels, L. K. M., et al. (2021). From thiol-subtilisin to Omniligase: Design and structure of a broadly applicable peptide ligase. *Comput. Struct. Biotechnol. J.* 19. doi:10.1016/j.csbj.2021.02.002.
- Toplak, A., Wu, B., Fusetti, F., Quaedflieg, P. J. L. M., and Janssen, D. B. (2013). Proteolysin, a novel highly thermostable and cosolvent-compatible protease from the thermophilic bacterium *Coprothermobacter proteolyticus*. *Appl. Environ. Microbiol.* 79, 5625–32. doi:10.1128/AEM.01479-13.
- Toth, C. A., Kuklenyik, Z., Jones, J. I., Parks, B. A., Gardner, M. S., Schieltz, D. M., et al. (2017). On-column trypsin digestion coupled with LC-MS/MS for quantification of apolipoproteins. *J. Proteomics* 150, 258–267. doi:10.1016/j.jprot.2016.09.011.
- Trincone, A. (2011). Marine biocatalysts: Enzymatic features and applications. *Mar. Drugs* 9, 478–499. doi:10.3390/md9040478.
- Troeschel, S. C., Thies, S., Link, O., Real, C. I., Knops, K., Wilhelm, S., et al. (2012). Novel broad host range shuttle vectors for expression in *Escherichia coli*, *Bacillus subtilis* and *Pseudomonas putida*. *J. Biotechnol.* 161, 71–79. doi:10.1016/j.jbiotec.2012.02.020.
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43. doi:10.1038/nature02340.
- Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* 20, 616–622. doi:10.1016/j.copbio.2009.09.010.
- Urlich, T., Lanzén, A., Stokke, R., Pedersen, R. B., Bayer, C., Thorseth, I. H., et al. (2014). Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environ. Microbiol.* 16, 2699–2710. doi:10.1111/1462-2920.12283.
- van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.* 34, 666–681. doi:10.1016/j.tig.2018.05.008.
- van Nassau, T. J., Lenz, C. A., Scherzinger, A. S., and Vogel, R. F. (2017). Combination of endolysins and high pressure

- to inactivate *Listeria monocytogenes*. *Food Microbiol.* 68, 81–88. doi:10.1016/j.fm.2017.06.005.
- Vandeyar, M. A., Weiner, M. P., Hutton, C. J., and Batt, C. A. (1988). A simple and rapid method for the selection of oligodeoxynucleotide-directed mutants. *Gene* 65, 129–133. doi:10.1016/0378-1119(88)90425-8.
- Vassylyev, D. G., Kashiwagi, T., Mikami, Y., Ariyoshi, M., Iwai, S., Ohtsuka, E., et al. (1995). Atomic model of a pyrimidine dimer excision repair enzyme complexed with a dna substrate: Structural basis for damaged DNA recognition. *Cell* 83, 773–782. doi:10.1016/0092-8674(95)90190-6.
- Vavrová, E., Muchová, K., and Barák, I. (2010). Comparison of different *Bacillus subtilis* expression systems. *Res. Microbiol.* 161, 791–797. doi:10.1016/j.resmic.2010.09.004.
- Veillard, F., Troxler, L., and Reichhart, J. M. (2016). *Drosophila melanogaster* clip-domain serine proteases: Structure, function and regulation. *Biochimie* 122, 255–269. doi:10.1016/j.biochi.2015.10.007.
- Vévodová, J., Gamble, M., Künze, G., Ariza, A., Dodson, E., Jones, D. D., et al. (2010). Crystal structure of an intracellular subtilisin reveals novel structural features unique to this subtilisin family. *Structure* 18, 744–755. doi:10.1016/j.str.2010.03.008.
- Vieille, C., Burdette, D. S., and Zeikus, J. G. (1996). Thermozymes. *Biotechnol. Annu. Rev.* 2, 1–83. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9704095> [Accessed July 22, 2018].
- Villar, E., Vannier, T., Vernet, C., Lescot, M., Cuenca, M., Alexandre, A., et al. (2018). The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Res.* 46, W289–W295. doi:10.1093/nar/gky376.
- Vojcic, L., Pitzler, C., Körfer, G., Jakob, F., Ronny Martinez, Maurer, K.-H., et al. (2015). Advances in protease engineering for laundry detergents. *N. Biotechnol.* 32, 629–634. doi:10.1016/j.nbt.2014.12.010.
- Vollmer, W. (2008). Structural variation in the glycan strands of bacterial peptidoglycan. *FEMS Microbiol. Rev.* 32, 287–306. doi:10.1111/j.1574-6976.2007.00088.x.
- Vollmer, W., Blanot, D., and De Pedro, M. A. (2008). Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.* 32, 149–167. doi:10.1111/j.1574-6976.2007.00094.x.
- Wang, H., Gong, Y., Xie, W., Xiao, W., Wang, J., Zheng, Y., et al. (2011). Identification and characterization of a novel thermostable *gh-57* gene from metagenomic fosmid library of the Juan De Fuca Ridge hydrothermal vent. *Appl. Biochem. Biotechnol.* 164, 1323–1338. doi:10.1007/s12010-011-9215-1.
- Wang, I.-N., Smith, D. L., and Young, R. (2000). Holins: The protein clocks of bacteriophage infections. *Annu. Rev. Microbiol.* 54, 799–825. doi:10.1146/annurev.micro.54.1.799.
- Ward, O. P. (2011). *Comprehensive biotechnology*. Elsevier doi:10.1016/B978-0-08-088504-9.00222-1.
- Ward, O. P., Rao, M. B., and Kulkarni, A. (2009). *Proteases, production*. Elsevier doi:10.1016/B978-012373944-5.00172-3.
- Waschkowitz, T., Rockstroh, S., and Daniel, R. (2009). Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. *Appl. Environ. Microbiol.* 75, 2506–16. doi:10.1128/AEM.02136-08.
- Webb, E. C. (Edwin C. (1992). *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. San Diego, California: Academic Press Available at: <http://www.sbc.sqmul.ac.uk/iubmb/enzyme/> [Accessed April 7, 2018].
- Weiss, B. (1981). Exodeoxyribonucleases of *Escherichia coli*. *Enzymes* 14, 203–231. doi:10.1016/S1874-6047(08)60338-8.

- Weller, S. K., and Sawitzke, J. A. (2014). Recombination promoted by DNA viruses: Phage λ to herpes simplex virus. *Annu. Rev. Microbiol.* 68, 237–258. doi:10.1146/annurev-micro-091313-103424.
- Wells, L., and Deming, J. (2006). Characterization of a cold-active bacteriophage on two psychrophilic marine hosts. *Aquat. Microb. Ecol.* 45, 15–29. doi:10.3354/ame045015.
- Wen, J., Lord, H., Knutson, N., and Wikström, M. (2020). Nano differential scanning fluorimetry for comparability studies of therapeutic proteins. *Anal. Biochem.* 593. doi:10.1016/j.ab.2020.113581.
- Wen, Z., Boddicker, M. A., Kaufhold, R. M., Khandelwal, P., Durr, E., Qiu, P., et al. (2016). Recombinant expression of *Chlamydia trachomatis* major outer membrane protein in *E. coli* outer membrane as a substrate for vaccine research. *BMC Microbiol.* 16. doi:10.1186/s12866-016-0787-3.
- Wildy, P. (1971). “Classification and nomenclature of viruses,” in *1st Report of the International Committee on Nomenclature of Viruses*. (Karger Publishers), 7–23. doi:10.1159/000392076.
- Wilhelm, S. W., and Suttle, C. A. (1999). Viruses and nutrient cycles in the sea. *Bioscience* 49, 781–788. doi:10.2307/1313569.
- Williamson, S. J., Cary, S. C., Williamson, K. E., Helton, R. R., Bench, S. R., Winget, D., et al. (2008). Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. *ISME J.* 2, 1112–1121. doi:10.1038/ismej.2008.73.
- Wilson, B., Müller, O., Nordmann, E. L., Seuthe, L., Bratbak, G., and Øvreås, L. (2017). Changes in marine prokaryote composition with season and depth over an Arctic polar year. *Front. Mar. Sci.* 4, 95. doi:10.3389/fmars.2017.00095.
- Winter, C., Payet, J. P., and Suttle, C. A. (2012). Modeling the winter-to-summer transition of prokaryotic and viral abundance in the Arctic Ocean. *PLoS One* 7. doi:10.1371/journal.pone.0052794.
- Wu, S., Nguyen, T. T. T. N., Moroz, O. V., Turkenburg, J. P., Nielsen, J. E., Wilson, K. S., et al. (2020). Conformational heterogeneity of Savinase from NMR, HDX-MS and X-ray diffraction analysis. *PeerJ* 2020. doi:10.7717/peerj.9408.
- Wu, X., Jörnvall, H., Berndt, K. D., and Oppermann, U. (2004). Codon optimization reveals critical factors for high level expression of two rare codon genes in *Escherichia coli*: RNA stability and secondary structure but not tRNA abundance. *Biochem. Biophys. Res. Commun.* 313, 89–96. doi:10.1016/j.bbrc.2003.11.091.
- Yadava, A., and Ockenhouse, C. F. (2003). Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems. *Infect. Immun.* 71, 4961–4969. doi:10.1128/IAI.71.9.4961-4969.2003.
- Yamamoto-Kawai, M., Carmack, E., and McLaughlin, F. (2006). Nitrogen balance and Arctic throughflow. *Nature* 443, 43. doi:10.1038/443043a.
- Yang, C., Xia, Y., Qu, H., Li, A. D., Liu, R., Wang, Y., et al. (2016). Discovery of new cellulases from the metagenome by a metagenomics-guided strategy. *Biotechnol. Biofuels* 9, 138. doi:10.1186/s13068-016-0557-3.
- Yang, H., Liu, Y., Ning, Y., Wang, C., Zhang, X., Weng, P., et al. (2020). Characterization of an intracellular alkaline serine protease from *Bacillus velezensis* SW5 with fibrinolytic activity. *Curr. Microbiol.* 77, 1610–1621. doi:10.1007/s00284-020-01977-6.
- Yang, W. (2011). Nucleases: Diversity of structure, function and mechanism. *Q. Rev. Biophys.* 44, 1–93. doi:10.1017/S0033583510000181.
- Yoshimori, R., Roulland-Dussoix, D., and Boyer, H. W. (1972). R factor-controlled restriction and modification of deoxyribonucleic acid: restriction mutants. *J. Bacteriol.* 112, 1275–1279. doi:10.1128/jb.112.3.1275-1279.1972.
- Youle, M., Haynes, M., and Rohwer, F. (2012). “Scratching the surface of biology’s dark matter,” in *Viruses: Essential*

Agents of Life (Springer Netherlands), 61–81. doi:10.1007/978-94-007-4899-6_4.

- Yu, Z. C., Chen, X. L., Shen, Q. T., Zhao, D. L., Tang, B. L., Su, H. N., et al. (2015). Filamentous phages prevalent in *Pseudoalteromonas* spp. Confer properties advantageous to host survival in Arctic sea ice. *ISME J.* 9, 871–881. doi:10.1038/ismej.2014.185.
- Zahid, S., and Brownell, I. (2008). Repairing DNA damage in xeroderma pigmentosum: T4N5 lotion and gene therapy. *J. Drugs Dermatol.* 7, 405–408. Available at: <https://europepmc.org/article/med/18459526> [Accessed February 25, 2021].
- Zambare, V. P., and Nilegaonkar, S. S. (2017). Proteases in leather processing. *Ind. Biotechnol. Sustain. Prod. Bioresour. Util.*, 209.
- Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, Di., Juzokaite, L., Vancaester, E., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541, 353–358. doi:10.1038/nature21031.
- Zhang, Q., Jun, S. R., Leuze, M., Ussery, D., and Nookaew, I. (2017). Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. *Sci. Rep.* 7, 1–13. doi:10.1038/srep40712.
- Zhang, W., Mi, Z., Yin, X., Fan, H., An, X., Zhang, Z., et al. (2013). Characterization of *Enterococcus faecalis* phage IME-EF1 and its endolysin. *PLoS One* 8. doi:10.1371/journal.pone.0080435.
- Zhou, Q., and Salvesen, G. S. (1997). Activation of pro-caspase-7 by serine proteases includes a non-canonical specificity. *Biochem. J.* 324, 361–364. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1218439/pdf/9182691.pdf> [Accessed May 18, 2018].
- Zhu, W., Cha, D., Cheng, G., Peng, Q., and Shen, P. (2007). Purification and characterization of a thermostable protease from a newly isolated *Geobacillus* sp. YMTC 1049. *Enzyme Microb. Technol.* 40, 1592–1597. doi:10.1016/j.enzmictec.2006.11.007.
- Zhu, X., Ohta, Y., Jordan, F., and Inouye, M. (1989). Pro-sequence of subtilisin can guide the refolding of denatured subtilisin in an intermolecular process. *Nature* 339, 483–484. doi:10.1038/339483a0.
- Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., et al. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* 430, 2237–2243. doi:10.1016/j.jmb.2017.12.007.

Publications

I



A rapid solubility-optimized screening procedure for recombinant subtilisins in *E. coli*



Gro Elin Kjæreng Bjerga^{a,*}, Hasan Arsin^b, Øivind Larsen^a, Pål Puntervoll^a,
Hans Torstein Kleivdal^a

^a Centre for Applied Biotechnology, Uni Research AS, Thormøhlensgt. 55, N-5008 Bergen, Norway

^b Department of Biology, University of Bergen, Thormøhlensgt. 53 A/B, N-5008 Bergen, Norway

ARTICLE INFO

Article history:

Received 21 October 2015

Received in revised form 1 February 2016

Accepted 3 February 2016

Available online 6 February 2016

Keywords:

Subtilisin

Serine protease

Recombinant expression

FX-cloning

FITC-casein

ABSTRACT

Subtilisins and other serine proteases are extensively used in the detergent, leather and food industry, and frequently under non-physiological conditions. New proteases with improved performance at extreme temperatures and in the presence of chemical additives may have great economical potential. The increasing availability of genetic sequences from different environments makes homology-based screening an attractive strategy for discovery of new proteases. A prerequisite for large-scale screening of protease-encoding sequences is an efficient screening procedure. We have developed and implemented a screening procedure that encompasses cloning of candidate sequences into multiple expression vectors, cytoplasmic expression in *E. coli*, and a casein-based functional screen. The procedure is plate-format compatible and can be completed in only four days, starting from the gene of interest in a suitable cloning vector. The expression vector suite includes six vectors with combinations of maltose-binding protein (MBP) or the small ubiquitin-related modifier (SUMO) for increased solubility, and polyhistidine tags for downstream purification. We used enhanced green fluorescent protein and four *Bacilli* subtilisins to validate the screening procedure and our results show that proteins were expressed, soluble and active. Interestingly, the highest activities were consistently achieved with either MBP or SUMO fusions, thus demonstrating the merit of including solubility tags. In conclusion, the results demonstrate that our approach can be used to efficiently screen for new subtilisins, and suggest that the approach may also be used to screen for proteins with other activities.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Proteases are proteolytic enzymes that have great industrial, therapeutic and academic value because of their ability to degrade proteins and peptides (reviewed in Li et al., 2013). In particular, serine proteases have broad applications in the detergent, leather and food industry, due to their broad substrate specificity and their activity at neutral to alkaline pHs (reviewed in Gupta et al., 2002). The industrial conditions under which these proteases are applied may be non-physiological, and include high temperatures, cocktails of detergents and other chemical additives. Discovery and development of proteases that are applicable to industry have a great economic potential, and both sequence- and function-based dis-

covery can be used (exemplified in Biver et al., 2013; Kwon et al., 2011). These approaches require a robust production line, which should include molecular cloning and recombinant expression, and functional screening procedures tailored to the relevant proteases (Kwon et al., 2011; Sroga and Dordick, 2002).

Subtilisin-like serine proteases are extensively used in industry, mainly in detergents, and according to a HERA-report,¹ the European Union used about 1000 tons of pure subtilisins in 2002. These proteases are well represented among species of *Bacilli*, are active at an alkaline pH range, and show specificity towards aromatic or hydrophobic residues (Groen et al., 1992). Subtilisins are involved in nutritional regulation in their native hosts and are frequently secreted. They are produced as inactive precursor proteins called pre-pro-proteins or zymogens consisting of a leader sequence that direct their export, a pro-sequence and the catalytic domain. The

* Corresponding author.

E-mail addresses: Gro.Bjerga@uni.no (G.E.K. Bjerga), Hasan.Arsin@uib.no (H. Arsin), Oivind.Larsen@uni.no (Ø. Larsen), Pal.Puntervoll@uni.no (P. Puntervoll), Hans.Kleivdal@uni.no (H.T. Kleivdal).

<http://dx.doi.org/10.1016/j.jbiotec.2016.02.009>

0168-1656/© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

¹ Human & Environmental Risk Assessment on ingredients of household cleaning products, Edition 2.0 February 2007.

pro-sequence acts as an inhibitor and as a molecular chaperone to guide correct folding of the active enzyme both *in vivo* and *in vitro* (Ikemura et al., 1987; Ohta et al., 1991; Zhu et al., 1989).

Autoproteolytic maturation poses a challenge for heterologous production, but subtilisin-like proteases have been successfully produced in *E. coli* by periplasmic expression (Ikemura and Inouye, 1988; Ikemura et al., 1987). Increased solubility and yield have been achieved with solubility tags, such as the maltose-binding protein (MBP²) (Bedouelle and Duplay, 1988; di Guan et al., 1988; Kapust and Waugh, 1999; Kapust and Waugh, 1999), fused to the N-terminus of active serine proteases (Kwon et al., 2011; Sakaguchi et al., 2008). For identification and downstream purification of mature subtilisin-like proteases, they are often expressed in fusion to C-terminal affinity tags, such as polyhistidine (his) (Ghasemi et al., 2012; Hu et al., 2013; Sroga and Dordick, 2002).

In this study, we have developed a screening procedure for rapid and efficient identification of functional subtilisins. The procedure uses *E. coli* as host, implements a versatile cloning system tailored to proteases, followed by recombinant expression in the cytosol, and direct activity screening of lysates using fluorescein isothiocyanate (FITC)-conjugated casein as substrate. The cloning system allows expression of candidate proteases with terminal his-tags for purification, optionally in combination with N-terminal fusions to MBP and the small ubiquitin-related modifier (SUMO) for improved solubility. To validate our approach we have tested four homologous *Bacilli* subtilisins, and show that our screening procedure effectively expresses and detects active, recombinant subtilisins.

2. Material and methods

2.1. Construction of a tailored vector suite

A tailored suite of six different vectors were designed and constructed for the recombinant expression of proteases (Fig. 1, Table 1). p1–p3 vectors contain N-terminal decahistidine tags, and p12, p6 and p7 vectors contain C-terminal hexahistidine tags. The preference for the C-terminal hexahistidine to decahistidine was based on experimental data showing better yields of expression and solubility of eGFP with shorter tag (data not shown). The p2 and p3 vectors were generated by introducing genes encoding the MBP and SUMO fusion partners, respectively, between the existing His-tag and HRV 3C protease (3C) protease site of the p1 vector (Table 1) (Geertsma and Dutzler, 2011). The p12 vector was constructed based on p5 (Table 1) (Geertsma and Dutzler, 2011), where the 3C cleavage site and decahistidine tag was replaced with a hexahistidine tag. Vectors p6 and p7 were constructed by introducing the above-mentioned fusion partners between the start codon and the first SapI site, replacing the 3C and decahistidine region.

For downstream applications, N-terminal affinity and solubility tags were made removable by the utility of specific proteases. A 3C site was designed to p2 and p7, identical to the design in p1 to allow cleavage by 3C protease (Cordingley et al., 1989). No linear motif was added to p3 and p6, as the ubiquitin-like protease 1 (Ulp1) specifically recognizes the tertiary structure of SUMO to remove the tag (Mosessova and Lima, 2000).

Synthetic genes (GenScript) optimized for *E. coli* production encoding either the *E. coli* *malE* gene or the *S. cerevisiae* *smt3* gene served as templates for megaprimer PCRs of MBP (aa 27–391, GenBank acc. no.: [A1Z93193](#)) and SUMO (aa 2–97, GenBank acc. no.:

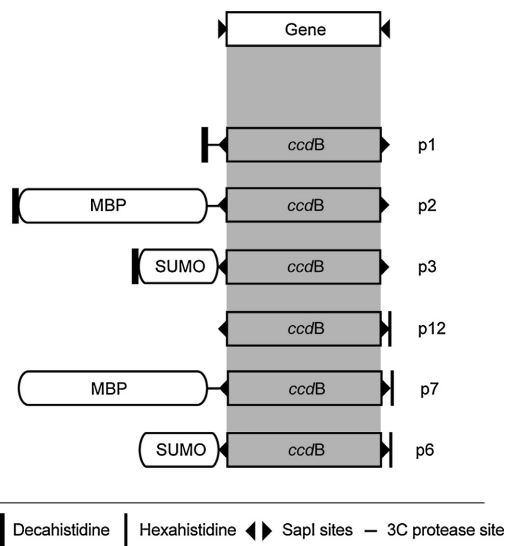


Fig. 1. Schematic representation of the expression vector suite.

Upon restriction-ligation in the FX-cloning regime the gene of interest is replacing the *ccdB* counterselection gene in expression vectors. All vectors, named p1, p2, p3, p12, p7 and p6, contain his-tags, either N-terminally (decahistidine) or C-terminally (hexahistidine). In addition, four of the vectors contain the MBP and SUMO solubility partners as fusions to the N-termini of the protease sequences. Triangles show placement and orientation of the SapI sites used in the FX-cloning procedure. Linear sequence motifs are introduced to allow tag removal by 3C protease in p1, p2 and p7 (black line), whereas the SUMO fusion in p3 and p6 can be removed by Ulp1 protease after tertiary structure recognition.

DAA12341), respectively, by exponential megaprimer PCR cloning (EMP, Ulrich et al., 2012). In brief, the megaprimers of genes encoding MBP and SUMO were amplified using Phusion polymerase (NEB), purified with the QIAquick PCR purification kit (Qiagen) and inserted to vectors by linear (in case of p2 and p3) or exponential plasmid amplification (in case of p6 and p7) and treated as described in the protocols (Ulrich et al., 2012; van den Ent and Löwe, 2006). To remove parental DNA, the PCR products were digested with 10U DpnI (NEB) and transformed into *E. coli* MC1061 cells. Plasmids were isolated using the NucleoSpin plasmid purification kit (Macherey-Nagel). Sanger sequencing was used to confirm correct cloning of all vectors. Primers for RF and EMP cloning were designed using an online tool (Bond and Naus, 2012). Information on primers and vectors used in this study is summarized in Table S1 and Table 1, respectively.

2.2. Molecular cloning of eGFP and subtilisins to vector suite

The pCMV cyto-EGFP-myc plasmid served as template for the amplification of a mutated *gfp* gene (*egfp*) encoding enhanced green fluorescent protein (eGFP, residues 2–239) (Cormack et al., 1996) by Phusion PCR and primers in Table S1. The *egfp* gene was integrated into the pINITIAL cloning vector (Table 1) by digesting the PCR product and the vector with SapI (NEB) and ligating with T4 DNA ligase (NEB) according to the fragment exchange (FX) cloning protocol (Geertsma, 2014). The ligation reaction was transformed into *E. coli* MC1061 cells and clones were selected on LB-agar (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, 1.5% (w/v) agar-agar) supplemented with kanamycin (50 µg/ml, Sigma-Aldrich). Plasmids were isolated as above, and sequencing was used to confirm correct cloning of all pINITIAL constructs. Sub-cloning from pINITIAL to the expression vectors (Fig. 1) were carried out as outlined

² Abbreviations: his, polyhistidine; MBP, maltose binding protein; SUMO, small ubiquitin-related modifier; FX-cloning, fragment exchange cloning; RF, restriction-free cloning; EMP, exponential megaprimer PCR; 3C, HRV 3C protease; Ulp1, ubiquitin-like protease 1; FITC, fluorescein isothiocyanate; HRP, horseradish peroxidase; eGFP, enhanced green fluorescent protein.

Table 1
Vectors used in this study.

Vector name	Fusion partners	His-tag length (aa)	Fusion partner size (kDa)	Promoter	Tag removal ^a	Resistance gene	Vector type	Reference
pINITIAL	–	–	–	–	–	<i>cam</i>	Cloning	Geertsma and Dutzler (2011)
p1 (pBXNH3)	N-his	10	2.6	<i>pBAD</i>	3C	<i>amp</i>	Expression	Geertsma and Dutzler (2011)
p2	N-his-MBP	10	42.7	<i>pBAD</i>	3C	<i>amp</i>	Expression	This study
p3	N-his-SUMO	10	12.6	<i>pBAD</i>	Ulp1	<i>amp</i>	Expression	This study
p5 (pBXC3H)	C-his	10	2.3	<i>pBAD</i>	3C	<i>amp</i>	Expression	Geertsma and Dutzler (2011)
p12	C-his	6	0.8	<i>pBAD</i>	–	<i>amp</i>	Expression	This study
p7	N-MBP, C-his	6	41.4	<i>pBAD</i>	3C	<i>amp</i>	Expression	This study
p6	N-SUMO, C-his	6	11.3	<i>pBAD</i>	Ulp1	<i>amp</i>	Expression	This study
pUC57: <i>smt3</i> (S2-G97)-opt	(<i>smt3</i> template)	–	–	–	–	<i>kan</i>	Cloning	This study
pUC57:His- <i>malE</i> (K27-Q391) 3C-opt	(<i>malE</i> template)	–	–	–	–	<i>kan</i>	Cloning	This study
pUC57.kan, SapI-free	–	–	–	–	–	<i>kan</i>	Cloning	GenScript

^a For p6 and p7 the N-terminal fusion partners, but not the C-terminal his tags, are removable.

above, except that clones were selected on LB agar with ampicillin (100 µg/ml, Sigma-Aldrich).

Codon-optimized *apr* genes (GenScript) encoding *Bacilli* subtilisins (Table 2) flanked by SapI sites served as templates for FX cloning of full-length genes to expression vectors, or served as templates for PCRs in the case of truncated genes. Otherwise, genes were cloned as described above. Empty vectors were generated by replacing the *ccdB* gene with a GSGSGS linker to allow their cloning and expression in *E. coli* MC1061 cells. The GSGSGS-linker was constructed by hybridizing two oligos to one double stranded DNA fragment designed to contain sticky SapI overhangs to allow its integration to the pINITIAL vector. Information on primers used is summarized in Table S1.

2.3. Site-directed mutagenesis

A Phusion PCR using mutagenesis primers (Table S1) was designed to generate a S325A mutation in the *apr* gene of *B. licheniformis* DSM13 in pINITIAL. Parental DNA was removed by DpnI digestion. Mutants were then transformed into *E. coli* MC1061 and selected to LB agar containing 34 µg/ml chloramphenicol (Sigma-Aldrich). The mutant subtilisin was sub-cloned to the expression vector suite as described above.

2.4. Recombinant expression

Recombinant expression was carried out according to a protocol described elsewhere (Vincentelli et al., 2011). The *E. coli* MC1061 strain was utilized for expression due to its inability to metabolize

the inducer, L-arabinose. Precultures in LB media were inoculated directly from positive clones on LB agar with 100 µg/ml ampicillin and grown 16–20 h at 250 rpm at 37 °C. In deep 24-well plates, 4 ml 2YT media (1.6% (w/v) tryptone, 1% (w/v) yeast extract and 0.5% (w/v) NaCl) containing 100 µg/ml ampicillin was inoculated with 100 µl precultures and incubated at 37 °C at 250 rpm for 2 h to reach log phase. Cultures were equilibrated to 20 °C for 30 min before induction of the *pBAD* promoter with 0.1% (v/v) L-arabinose (Sigma-Aldrich) for 16–20 h at 250 rpm at 20 °C. To improve the protocol, cultures can also be autoinduced by replacing lactose with arabinose in a trace-metal free version of the ZYP-5052 media (Studier, 2005). 100 µl culture was used for OD measurements by reading absorbance at 600 nm using the Hidex Sense microplate reader (Kem-En-Tec Nordic). The cells were harvested using an Allegra X-12R benchtop centrifuge (Beckman Coulter) at 4750 rpm for 15 min. Cells were resuspended in 1 ml 8.5 N lysis buffer (50 mM Tris HCl pH 8.5, 50 mM NaCl, 0.25 mg/ml lysozyme, 10% (v/v) glycerol), and incubated at 20 min at room temperature during gentle agitation for lysis. Lysis was completed by ultrasound using five seconds pulse two times at 40–60% amplitude with a CV-18 probe powered by an Ultrasonic Homogenizer 4710 (Cole Parmer). Lysates were cleared by centrifugation at 4750 rpm for 15 min. Cleared lysate samples (representing soluble fraction) were analysed by SDS-PAGE (Laemmli, 1970).

2.5. Immunoblot

Proteins from cleared lysates were analysed by SDS-PAGE, and transferred onto a nitrocellulose membrane (Towbin et al., 1979)

Table 2
Subtilisins used in this study.

Origin (short name)	Enzyme name	GenBank acc. no.	ATCC reference	Reference (protease)	Reference (genome)	Length (aa)	Sequence identity to <i>B. licheniformis</i> DSM13 (%).
<i>Bacillus licheniformis</i> DSM13 (<i>Bli</i>)	Subtilisin Carlsberg AprE	AAU40017	14580	Jacobs et al. (1985)	Rey et al. (2004)	379	100
<i>Bacillus paralicheniformis</i> ATCC 9945a (<i>Bpa</i>)	Subtilisin Carlsberg AprE	AGN35600	9945a	Jacobs et al. (1985)	Rachinger et al. (2013)	379	98.2
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168 (<i>Bsu</i>)	SubtilisinE	CAB12870	23857	Stahl and Ferrari (1984)	Kunst et al. (1997)	381	66.1
<i>Bacillus amyloliquefaciens</i> ATCC 23844 (<i>Bam</i>)	BPN' subtilisin	AAB05345	23844	Vasantha et al. (1984)	–	382	66.1

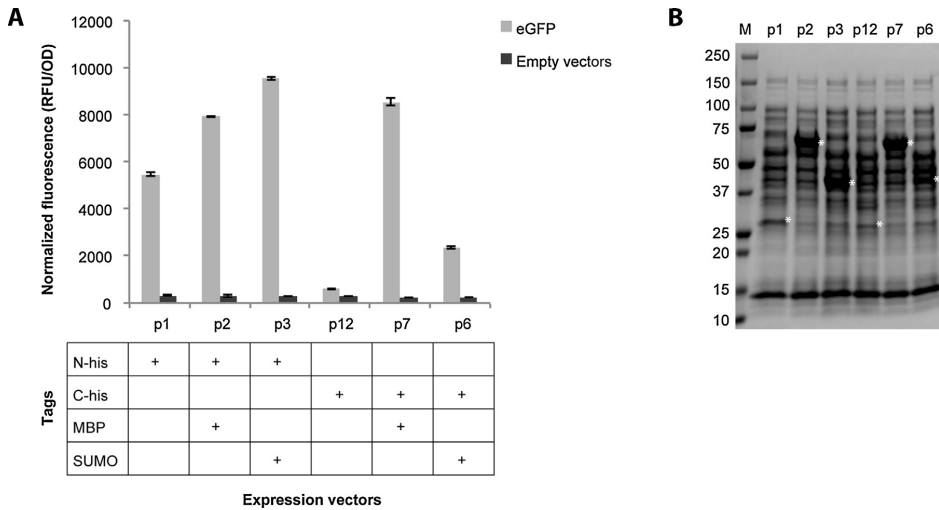


Fig. 2. Recombinant expression of eGFP in the vector suite.

A. Fluorescence from cultures containing eGFP in six different fusion constructs. Fluorescence was normalized to the optical density of the bacterial cultures. Expression from empty vectors was used as a control. The Table below shows the presence of tags in the different constructs. B. Analysis of protein integrity of soluble eGFP fusion proteins by SDS-PAGE. Sizes of molecular weight standard (M) are shown to the right (in kDa). Asterisks indicate the expected mass of the recombinant proteins.

using the Trans-Blot Turbo (BioRad) transfer system. A mouse monoclonal anti-polyhistidine antibody (H1029, Sigma-Aldrich) and a mouse monoclonal anti-MBP antibody (M1321, Sigma-Aldrich) were used to detect recombinant expression of subtilisin mutants. The primary antibodies were detected with a secondary rabbit HRP-linked mouse IgG (NA931 V, GE Healthcare). The HRP-reaction was developed with the Clarity Western ECL Substrate (BioRad), and imaged in the Chemi-Doc gel imager (BioRad).

2.6. eGFP fluorescence measurement

Upon harvest, 100 μ l culture from eGFP expression was used for OD measurements by reading the absorbance at 600 nm using the FluoStar Optima microplate reader (BMG Labtech). Fluorescence from cell cultures was measured at excitation 485 nm and emission 520 nm, and values were normalized to optical density of the bacterial cultures to compensate for growth differences.

2.7. Casein-based protease assay

The protease fluorescent detection kit (PF0100, Sigma-Aldrich) was used for detection of proteolytic activity (Twining, 1984). 5 μ l cleared lysates were used in activity assays with 10 μ l FITC-labelled casein and adjusted to 25 μ l with 20 mM phosphate buffer pH 7.6, and incubated at 60 $^{\circ}$ C for 1 h. Unhydrolyzed proteins were precipitated by 75 μ l 0.6 N Trichloroacetic Acid (TCA) at 37 $^{\circ}$ C for 30 min. Hydrolyzed FITC-peptides were then collected by centrifugation in a Heraeus Pico centrifuge (Thermo Scientific) at 13,000 rpm for 5–10 min. 2 μ l supernatant was added to 200 μ l 0.5 M TrisHCl pH 8.5 in MicroFluor 1 plates (Thermo Scientific) and fluorescence was measured at excitation 485 nm and emission 520 nm using the Hidex Sense microplate reader. As a control, Alcalase $^{\circ}$ 2.4L (P4860, Sigma-Aldrich) was used at a dilution 1:50 000 in lysis buffer.

3. Results and discussion

3.1. Screening procedure design and vector suite validation

Our aim was to develop a quick but robust screening procedure for subtilisin-like proteases. We identified a set of design requirements for such a screening procedure: the entire procedure should be compatible with a plate format for high throughput; expression should be carried out in the cytoplasm of *E. coli* for high yield; the cloning system should facilitate simple, parallel sub-cloning into a set of vectors for exploitation of purification and solubility tags; and finally, a functional screen should be implemented for assessment of activity from recombinant proteases in lysates. A recently developed vector suite, which facilitates sub-cloning based on fragment exchange (FX) of the gene of interest from a cloning vector into multiple expression vectors, was chosen as a starting point (Geertsma and Dutzler, 2011). The approach is highly effective due to the directional cloning caused by the orientation of the type IIS restriction sites (SapI) and the presence of counter selection genes. Based on available vectors (p1 and p5, Table 1), we constructed five new ones, p2, p3, p6, p7 and p12 (Fig. 1, Table 1), designed to express serine proteases in fusion to selected affinity and solubility tags. We chose to include two solubility tags for use in N-terminal fusions, namely MBP (Kapust and Waugh, 1999) and SUMO (Malakhov et al., 2004). In case the tags would interfere with maturation and/or proteolytic activity, specific protease sites were included for tag removal (3C and Ulp1 proteases, respectively). All constructs contained either an N-terminal or C-terminal his-tag to enable detection and protein purification. The C-terminal his-tag containing vectors were constructed to allow purification of mature protease (Ghasemi et al., 2012; Hu et al., 2013; Sroga and Dordick, 2002). Cloning was scaled to a plate-based format that allowed parallel cloning of many constructs simultaneously.

To validate the expression vectors, we chose the gene encoding eGFP for simple fluorescence-based monitoring of recombinant protein production. The *egfp* gene was cloned into the pINITIAL cloning vector, verified by sequencing, and sub-cloned into our

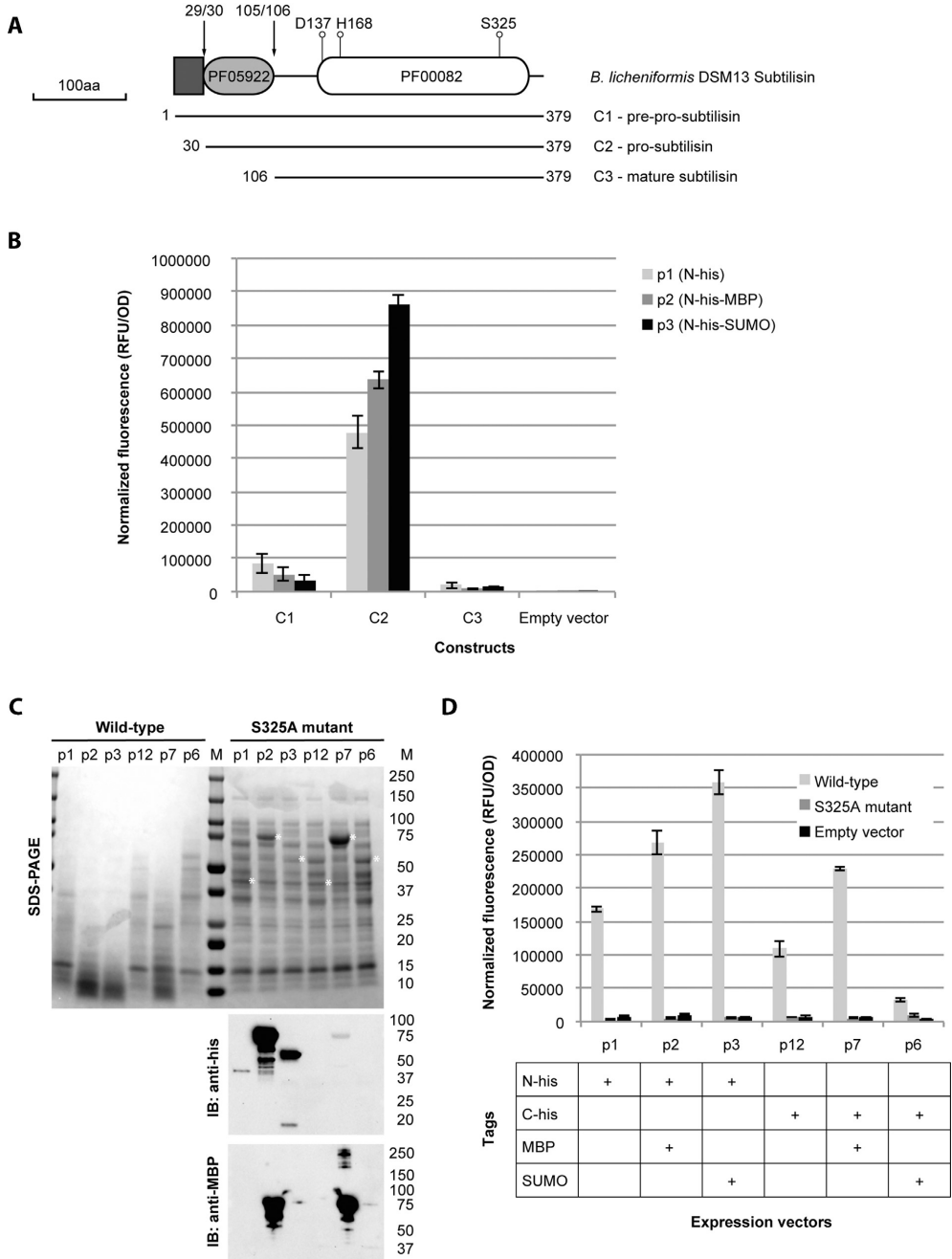


Fig. 3. Activity of recombinant *B. licheniformis* DSM13 pro-subtilisin in *E. coli*.
A. Cartoon of the annotated *B. licheniformis* DSM13 subtilisin with leader sequence (square-shaped box), proteolytic cleavage positions (arrows), catalytic triad (open rings) and Pfam domain annotations (PF-coded oval-shaped box). Below the constructs used in this study are shown: the full-length pre-pro subtilisin sequence (residues 1–379, C1), the pro-subtilisin construct (residues 30–379, C2), and the mature subtilisin (residues 106–379, C3). Cartoon is drawn to scale. **B.** The soluble fraction of extracts containing subtilisin constructs (C1–C3) in three vectors, p1, p2 or p3, were screened for proteolytic activity against FITC-conjugated casein. Fluorescence was normalized to optical density of expression cultures, to account for growth effects. Expression from empty vectors was used as a control. Error bars represent standard deviation between two independent experiments. **C.** Analysis of protein integrity of native pro-subtilisin and the catalytic S325A mutant by SDS-PAGE (upper panel) and immunoblots (IB). Antibodies against the his-tag (middle panel) and the MBP tag (lower panel)

panel of expression vectors. Next, the vectors were used to express the eGFP fusion proteins in *E. coli* MC1061, and the expression cultures were monitored directly for *in vivo* fluorescence (Fig. 2A) (such as applied in Scholz et al., 2013). The N-terminally his-tagged eGFP construct (p1) had a much higher level of fluorescence than the one with a C-terminal his-tag (p12), which fluoresced just above background. Both the N- and C-terminally his-tagged constructs displayed increased fluorescence when fused to either MBP or SUMO (Fig. 2A). Protein expression levels in the soluble fraction correlated well with the *in vivo* fluorescence data (Fig. 2B), suggesting that the differences in fluorescence were due to different amounts of soluble protein. These results validated the integrity of the vectors, and demonstrated that the MBP and SUMO tags increased the solubility of the target protein, in line with previous data (Marblestone et al., 2006).

3.2. Validation of the screening procedure using four *Bacillus subtilis* subtilisins

To validate the suitability of the vector suite for expression of serine proteases, the industrially relevant subtilisin from *Bacillus licheniformis* (trade name: Alcalase 2.4 L, Novozymes™) was chosen as a reference protease. As periplasmic expression of subtilisin in *E. coli* has shown low expression levels (Ikemura and Inouye, 1988; Ikemura et al., 1987), we decided to attempt cytoplasmic expression (e.g. Hu et al., 2013; Maciver et al., 1994; Zhang et al., 2005).

To investigate whether expressing the full-length or other truncated versions of the *apr* gene would affect solubility and activity, we cloned three versions of the gene of the *Bacillus licheniformis* DSM13 *apr* into the p1, p2, and p3 vectors (Fig. 3A). The first construct contained the entire *apr* gene, encoding pre-pro subtilisin (residues 1–379, C1). The second construct contained a truncated gene encoding pro-subtilisin (residues 30–379, C2) without the leader sequence. The third construct contained a truncated gene where the part encoding both the pre and pro sequences was removed (residues 106–379, C3). All gene versions were codon-optimized to improve protein expression. The three *apr* gene variants were cloned into pINITIAL, sub-cloned into the three expression vectors, and expressed in the *E. coli* MC1061 strain.

To measure the activity of the expressed subtilisin fusions, we chose an *in vitro* casein-based protease assay (Twining, 1984). In this assay, the use of fluorescein isothiocyanate (FITC) conjugated casein allows direct screens for protease activity in the cleared extracts, thus eliminating elaborate purification steps. Casein was chosen to facilitate screening for proteases with broad specificity for peptide bonds. The soluble extracts containing subtilisin fusions were tested in the assay, and compared to extracts from strains carrying empty plasmid vectors (Fig. 3B). The negative controls using empty vectors did not exhibit any activity, which suggests that the activity observed in the extracts containing subtilisin fusion was not due to *E. coli* host proteases. Hence, the proteolytic degradation of FITC-casein in those extracts is likely caused by the activity of the expressed native subtilisin fusions. Cultures expressing the C2 version were shown to be the optimal construct (Fig. 3B). The C1 version showed significant growth deficiencies, which may suggest that the subtilisin leader sequence caused misfolding. Despite this, some activity was observed with soluble extracts containing C1 in the FITC-casein assay. In contrast, soluble extracts containing the C3 construct showed no activity. In line with previous data, these

results strongly suggest that putative leader sequences should be removed when screening for new subtilisins using our system, but that the pro sequence should be kept intact (Ikemura et al., 1987; Ohta et al., 1991; Zhu et al., 1989).

It was not obvious that N-terminal fusion tags would promote correct folding of subtilisin, as these might hinder folding assisted by the pro domain. The highest activity was, however, observed with the extracts containing the N-terminal his-SUMO (p2) and his-MBP (p3) fusion constructs of the C2 version of subtilisin (Fig. 3B). These results show that fusion to a large solubility partner at the N-terminal of pro-subtilisin promoted production of active serine proteases. Our data is in line with previous data obtained with MBP (Kwon et al., 2011; Sakaguchi et al., 2008). To our knowledge, this is the first time subtilisin has been successfully expressed as a SUMO fusion protein in *E. coli*.

As a C-terminal his-tag may allow downstream purification and detection after maturation, the C2 subtilisin construct was also cloned to the p6, p7 and p12 vectors. As an additional negative control, we constructed a mutant of the truncated *apr* gene where the codon for the nucleophilic serine S325 in the catalytic triad was mutated to an alanine codon (Fig. 1) (Carter and Wells, 1988). The soluble fractions of sonicated cells were analysed by SDS-PAGE, revealing striking differences in the observed protein patterns of the native subtilisin fusion proteins and their cognate mutant controls (Fig. 3C). The lack of high molecular weight proteins in the native subtilisin-containing fractions suggests that they are expressed and have proteolytic activity. However, there were no visible protein bands corresponding to the native subtilisin fusion proteins, and mass spectrometry (MS) analysis of the samples also failed to identify them. Instead, the MS analysis revealed the presence of known protease resistant *E. coli* membrane and periplasmic proteins (Table S2). In contrast to the native subtilisin fusion proteins, most of the mutant subtilisins, with the exception of the p12 mutant construct, were identified, either directly on the Coomassie-stained SDS-PAGE, or by immunoblotting using anti-his and anti-MBP antibodies (Fig. 3C). Overall, this suggests that the subtilisin fusion proteins are expressed, soluble and active.

All subtilisin fusion proteins were able to digest FITC-casein (Fig. 3D), but again the highest activity was observed with the extracts containing the N-terminal his-SUMO and his-MBP fusion constructs. Interestingly, the conventional subtilisin construct design with a C-terminal his-tag (p12) was the second lowest performing construct in the activity assay. The extent of hydrolysis observed by the SDS-PAGE analysis (Fig. 3C) correlated with the level of proteolytic activity observed in the FITC-casein assay (Fig. 3D).

Furthermore, there is a striking similarity to the eGFP validation experiment (Fig. 2); the effect of the different fusion constructs on the levels of fluorescence in the eGFP experiment is similar to the activities observed with the subtilisin fusion proteins (Fig. 3D). This suggests that the vector suite may have some generic features, and that its utility is not restricted to proteases. To conclude, the described *E. coli* based expression system was used to successfully express soluble *B. licheniformis* DSM13 subtilisin fusion proteins, which were active in the casein-based activity assay.

As a final validation step of our screening procedure, we performed a mini screen of the *B. licheniformis* DSM13 *apr* gene and three other *Bacilli* genes from *B. paralicheniformis* ATCC 9945a, *B. amyloliquefaciens* ATCC 23844 and *B. subtilis* subsp. *subtilis* str. 168 (Table 2). The subtilisins encoded by these genes are thermophilic

were used to identify the recombinant mutant constructs. Sizes of molecular weight standard (M) are shown to the right (in kDa). Asterisks indicate the expected mass of the unprocessed recombinant proteins. The mature subtilisin is 27 kDa. FITC-casein activity of soluble *E. coli* extracts containing fusion constructs of native pro-subtilisin from *B. licheniformis* DSM13 and the catalytic S325A mutant. Hydrolysis was measured as in B, except error bars show standard deviation in two replicates in one representative experiment.

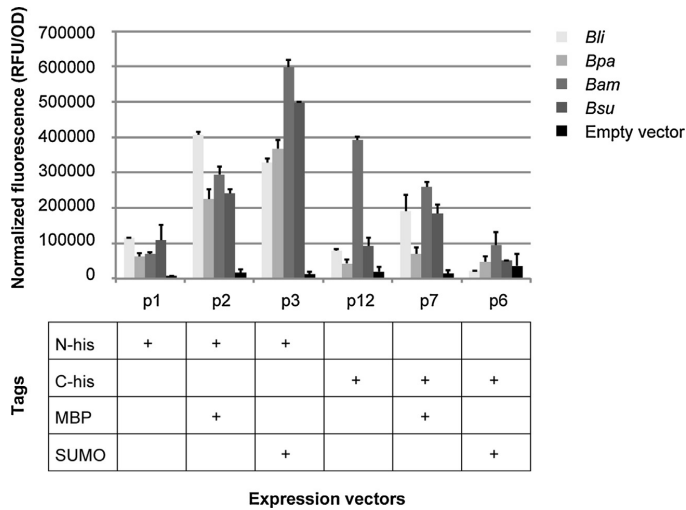


Fig. 4. Proteolytic activity of four *Bacilli* subtilisin homologues.

The soluble fractions of *E. coli* extracts containing fusion constructs of the four homologous subtilisins were screened for proteolytic activity against FITC-conjugated casein. Hydrolysis was measured as in Fig. 3, where error bars show standard deviation in two replicates in one representative experiment. *Bli*, *Bacillus licheniformis* DSM13; *Bpa*, *Bacillus paralicheniformis* ATCC 9945a; *Bsu*, *Bacillus subtilis* subsp. *subtilis* str. 168; *Bam*, *Bacillus amyloliquefaciens* ATCC 23844.

with temperature optimums above 40 °C (Table 2) (examples in Ferreira et al., 2003; Peng et al., 2003; Sellami-Kamoun et al., 2008), and their pairwise sequence identities range from 66 to 98% (Table 2). The three additional *apr* genes, without the leader sequence encoding part, were cloned into the six expression vectors. Each of the four subtilisins was expressed from the six expression plasmids, and soluble extracts were tested for protease activity using the FITC-casein assay (Fig. 4). All four subtilisin targets were positively identified as active proteases in four or five of the constructs, using three times the background signal to determine the threshold. The two constructs that showed the most consistent activity across the different subtilisins were the his-MBP (p2) and his-SUMO fusions (p3). This suggests that one could restrict future screens for new subtilisins to only the p2 and p3 vectors, thus increasing the throughput of the system. However, the high activity of the p12 construct of *B. amyloliquefaciens* ATCC 23844 subtilisin (Fig. 4) demonstrates the merit of including additional vectors in the screen, as the C-terminal his tag allows for downstream purification and characterization. This construct also indicates that the success of recombinant expression from a particular vector may be protein-dependent, and that a range of expression vehicles with different properties may be valuable to identify positive candidates.

4. Conclusion

The development of a screening procedure for the identification of recombinant subtilisins has been described. The overall procedure enables a four-day approach from the sub-cloning of candidate genes in expression vectors to the evaluation of enzymatic activity (Fig. 5). The approach utilizes a rapid cloning method for sub-cloning of genes into six expression vectors and subsequent recombinant expression in *E. coli* MC1061, an ultrasound-based cell lysis and direct activity assessment of soluble fractions with FITC-casein. The screening approach has been validated using four homologous *Bacilli apr* encoded subtilisins of varying degree of sequence identities (ranging from 65 to 98%) and temperature profiles (ranging roughly 40–80 °C). All subtilisins were identi-

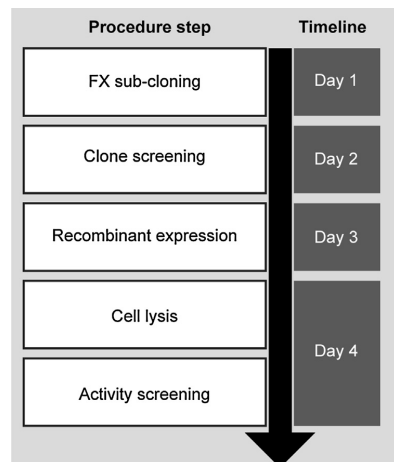


Fig. 5. A flow-chart of the rapid, solubility-optimized screening procedure for recombinant subtilisins in *E. coli*.

The procedure has been developed and tailored to subtilisins, and enables a four-day approach succeeding from sub-cloning of candidate genes in expression vectors to the evaluation of enzymatic activity. The screening-approach utilizes a rapid and versatile FX-cloning regime, subsequent recombinant production in *E. coli* MC1061, an ultrasound-based cell lysis and activity assaying with FITC-casein.

fied as active proteases in at least four of the constructs. Hence, this approach is suitable for application in discovery of novel subtilisin-like proteases or for application on engineered subtilisin-like proteases (Bryan, 2000; Wells and Estell, 1988) to effectively screen production of mutants and their enzymatic activity. We also showed, using eGFP, that the utility of the vector suite is not restricted to proteases. Other applications may, however, require tailoring to the specific type of protein, depending on their properties, such as metal-requirement, temperature and pH range.

Contributions

GEKB, PP, ØL, HTK jointly conceived and designed the study, GEKB has performed the bioinformatical analyses, GEKB and HA carried out the experiments and supervised students, GEKB and PP prepared the manuscript. All authors have discussed the results and read, edited and approved the final manuscript.

Acknowledgements

We would like to thank the students Odin Blomset and Anni Geithus for excellent technical assistance. The pCMV-cyto-EGFP-myc plasmid was kindly provided by Prof. Mathias Ziegler's lab, University of Bergen. This study was financed by the Norwegian Research Council as part of the NorZymeD initiative (project ID: 221568).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at <http://dx.doi.org/10.1016/j.jbiotec.2016.02.009>.

References


- Bedouelle, H., Duplay, P., 1988. Production in *Escherichia coli* and one-step purification of bifunctional hybrid proteins which bind maltose: export of the Klenow polymerase into the periplasmic space. *Eur. J. Biochem.* 171, 541–549.
- Biver, S., Portetelle, D., Vandenberg, M., 2013. Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. *Springerplus* 2, 410. <http://dx.doi.org/10.1186/2193-1801-2-410>.
- Bond, S.R., Nau, C.C., 2012. RF-Cloning.org: an online tool for the design of restriction-free cloning projects. *Nucleic Acids Res.* 40, W209–13. <http://dx.doi.org/10.1093/nar/gks396>.
- Bryan, P.N., 2000. Protein engineering of subtilisin. *Biochim. Biophys. Acta* 1543, 203–222. [http://dx.doi.org/10.1016/S0167-4838\(00\)00235-1](http://dx.doi.org/10.1016/S0167-4838(00)00235-1).
- Carter, P., Wells, J.A., 1988. Dissecting the catalytic triad of a serine protease. *Nature* 332, 564–568. <http://dx.doi.org/10.1038/332564a0>.
- Cordingley, M.G., Register, R.B., Callahan, P.L., Garsky, V.M., Colonna, R.J., 1989. Cleavage of small peptides in vitro by human rhinovirus 14 3C protease expressed in *Escherichia coli*. *J. Virol.* 63, 5037–5045.
- Cormack, B.P., Valdivia, R.H., Falkow, S., 1996. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* 173, 33–38. [http://dx.doi.org/10.1016/0378-1119\(95\)00685-0](http://dx.doi.org/10.1016/0378-1119(95)00685-0).
- di Guan, C., Li, P., Riggs, P.D., Inouye, H., 1988. Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene* 67, 21–30. [http://dx.doi.org/10.1016/0378-1119\(88\)90004-2](http://dx.doi.org/10.1016/0378-1119(88)90004-2).
- Ferreira, L., Ramos, M.A., Dordick, J.S., Gil, M.H., 2003. Influence of different silica derivatives in the immobilization and stabilization of a *Bacillus licheniformis* protease (Subtilisin Carlsberg). *J. Mol. Catal. B Enzym.* 21, 189–199. [http://dx.doi.org/10.1016/S1381-1177\(02\)00223-0](http://dx.doi.org/10.1016/S1381-1177(02)00223-0).
- Geertsma, E.R., Dutzler, R., 2011. A versatile and efficient high-throughput cloning tool for structural biology. *Biochemistry* 50, 3272–3278. <http://dx.doi.org/10.1021/bi200178z>.
- Geertsma, E.R., 2014. FX cloning: a simple and robust high-throughput cloning method for protein expression. In: Valia, S., Lale, R. (Eds.), *DNA Cloning and Assembly Methods, Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 153–164. <http://dx.doi.org/10.1007/978-1-62703-764-8>.
- Ghasemi, Y., Dabbagh, F., Ghasemian, A., 2012. Cloning of a fibrinolytic enzyme (subtilisin) gene from *Bacillus subtilis* in *Escherichia coli*. *Mol. Biotechnol.* 52, 1–7. <http://dx.doi.org/10.1007/s12033-011-9467-6>.
- Groen, H., Meldal, M., Breddam, K., 1992. Extensive comparison of the substrate preferences of two subtilisins as determined with peptide substrates which are based on the principle of intramolecular quenching. *Biochemistry* 31, 6011–6018. <http://dx.doi.org/10.1021/bi00141a008>.
- Gupta, R., Beg, O.K., Lorenz, P., 2002. Bacterial alkaline proteases: molecular approaches and industrial applications. *Appl. Microbiol. Biotechnol.* 59, 15–32. <http://dx.doi.org/10.1007/s00253-002-0975-y>.
- Hu, H., He, J., Yu, B., Zheng, P., Huang, Z., Mao, X., Yu, J., Han, G., Chen, D., 2013. Expression of a keratinase (kerA) gene from *Bacillus licheniformis* in *Escherichia coli* and characterization of the recombinant enzyme. *Biotechnol. Lett.* 35, 239–244. <http://dx.doi.org/10.1007/s10529-012-1064-7>.
- Ikemura, H., Inouye, M., 1988. In vitro processing of pro-subtilisin produced in *Escherichia coli*. *J. Biol. Chem.* 263, 12959–12963.
- Ikemura, H., Takagi, H., Inouye, M., 1987. Requirement of pro-sequence for the production of active subtilisin E in *Escherichia coli*. *J. Biol. Chem.* 262, 7859–7864.
- Jacobs, M., Eliasson, M., Uhlén, M., Flock, J.L., 1985. Cloning: sequencing and expression of subtilisin Carlsberg from *Bacillus licheniformis*. *Nucleic Acids Res.* 13, 8913–8926.
- Kapust, R.B., Waugh, D.S., 1999. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 8, 1668–1674. <http://dx.doi.org/10.1110/ps.8.8.1668>.
- Kunst, F., Ogasawara, N., Moser, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C., Caldwell, V., Capuano, B., Carter, V., Choi, N.M., Codani, S.K., Connerton, J.J., Danchin, I.F., 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256. <http://dx.doi.org/10.1038/36786>.
- Kwon, K., Hasseman, J., Latham, S., Grose, C., Do, Y., Fleischmann, R.D., Pieper, R., Peterson, S.N., 2011. Recombinant expression and functional analysis of proteases from *Streptococcus pneumoniae*, *Bacillus anthracis*, and *Yersinia pestis*. *BMC Biochem.* 12, 17. <http://dx.doi.org/10.1186/1471-2091-12-17>.
- Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227, 680–685. <http://dx.doi.org/10.1038/227680a0>.
- Li, Q., Yi, L., Marek, P., Iverson, B.L., 2013. Commercial proteases: present and future. *FEBS Lett.* 587, 1155–1163. <http://dx.doi.org/10.1016/j.febslet.2012.12.019>.
- Maciver B., Mchale R.H., Saul D.J., Bergquist P.L., 1994. Cloning and sequencing of a serine proteinase gene from a thermophilic *Bacillus* species and its expression in *Escherichia coli*. Cloning and Sequencing of a Serine Proteinase Gene from a Thermophilic *Bacillus* Species and Its Expression in *Escherichia coli* 60.
- Malakhov, M.P., Mattern, M.R., Malakhova, O.A., Drinker, M., Weeks, S.D., Butt, T.R., 2004. SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *J. Struct. Funct. Genomics* 5, 75–86. <http://dx.doi.org/10.1023/B:SFG.0000029237.70316.52>.
- Marblestone J.G., Edavettal S.C., Lim Y., Lim P., Zuo X.U.N., Butt T.R., 2006. Comparison of SUMO fusion technology with traditional gene fusion systems: Enhanced expression and solubility with SUMO 182–189. 10.1101/ps.051812706for.
- Mossesso, E., Lima, C.D., 2000. Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell* 5, 865–876.
- Ohta, Y., Hojo, H., Aimoto, S., Kobayashi, T., Zhu, X., Jordan, F., Inouye, M., 1991. Pro-peptide as an intramolecular chaperone: renaturation of denatured subtilisin E with a synthetic pro-peptide [corrected]. *Mol. Microbiol.* 5, 1507–1510. <http://dx.doi.org/10.1111/j.1365-2958.1991.tb00797.x>.
- Peng, Y., Huang, Q., Zhang, R.H., Zhang, Y.Z., 2003. Purification and characterization of a fibrinolytic enzyme produced by *Bacillus amyloliquefaciens* DC-4 screened from douchi, a traditional Chinese soybean food. *Comp. Biochem. Physiol.—B Biochem. Mol. Biol.* 134, 45–52. [http://dx.doi.org/10.1016/S1096-4959\(02\)00183-5](http://dx.doi.org/10.1016/S1096-4959(02)00183-5).
- Rachinger, M., Volland, S., Meinhardt, F., Daniel, R., Liesegang, H., 2013. First insights into the completely annotated genome sequence of *Bacillus licheniformis* strain 9945A. *Genome Announc.* 1, 525–526. <http://dx.doi.org/10.1128/genomeA.00525-13>.
- Rey, M.W., Ramaiya, P., Nelson, B.A., Brody-Karpin, S.D., Zaretsky, E.J., Tang, M., Lopez de Leon, A., Xiang, H., Gusti, V., Clausen, I.G., Olsen, P.B., Rasmussen, M.D., Andersen, J.T., Jørgensen, P.L., Larsen, T.S., Sorokin, A., Bolotin, A., Lapidus, A., Galleron, N., Ehrlich, S.D., Berka, R.M., 2004. Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biol.* 5, R77. <http://dx.doi.org/10.1186/gb-2004-5-10-r77>.
- Sakaguchi, M., Niimiya, K., Takezawa, M., Toki, T., Sugahara, Y., Kawakita, M., 2008. Construction of an expression system for aqualysin I in *Escherichia coli* that gives a markedly improved yield of the enzyme protein. *Biosci. Biotechnol. Biochem.* 72, 2012–2018. <http://dx.doi.org/10.1271/bbb.80132>.
- Scholz, J., Besir, H., Strasser, C., Suppmann, S., 2013. A new method to customize protein expression vectors for fast, efficient and background free parallel cloning. *BMC Biotechnol.* 13, 12. <http://dx.doi.org/10.1186/1472-6750-13-12>.
- Sellami-Kamoun, A., Haddar, A., Ali, N.E.H., Ghorbel-Frikha, B., Kanoun, S., Nasri, M., 2008. Stability of thermostable alkaline protease from *Bacillus licheniformis* RP1 in commercial solid laundry detergent formulations. *Microbiol. Res.* 163, 299–306. <http://dx.doi.org/10.1016/j.micres.2006.06.001>.
- Sroga, G.E., Dordick, J.S., 2002. A strategy for in vivo screening of subtilisin E reaction specificity in *E. coli* periplasm. *Biotechnol. Bioeng.* 78, 761–769. <http://dx.doi.org/10.1002/jbit.10269>.
- Stahl, M.L., Ferrari, E., 1984. Replacement of the *Bacillus subtilis* subtilisin structural gene with an in vitro-derived deletion mutation. *J. Bacteriol.* 158, 411–418.
- Studier, F.W., 2005. Protein production by Auto-induction in high-density shaking cultures. *Protein Expr. Purif.* 41, 207–234. <http://dx.doi.org/10.1016/j.pep.2005.01.016>.
- Towbin, H., Staehelin, T., Gordon, J., 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U. S. A.* 76, 4350–4354. <http://dx.doi.org/10.1073/pnas.76.9.4350>.
- Twining, S.S., 1984. Fluorescein isothiocyanate-labeled casein assay for proteolytic enzymes. *Anal. Biochem.* 143, 30–34. [http://dx.doi.org/10.1016/0003-2697\(84\)90553-0](http://dx.doi.org/10.1016/0003-2697(84)90553-0).
- Ulrich, A., Andersen, K.R., Schwartz, T.U., 2012. Exponential megapriming PCR (EMP) cloning—seamless DNA insertion into any target plasmid without sequence constraints. *PLoS One* 7, e35360. <http://dx.doi.org/10.1371/journal.pone.0053360>.

- Vasantha, N., Thompson, L.D., Rhodes, C., Banner, C., Nagle, J., Filpula, D., 1984. Genes for alkaline protease and neutral protease from *Bacillus amyloliquefaciens* contain a large open reading frame between the regions coding for signal sequence and mature protein. *J. Bacteriol.* 159, 811–819.
- Vincentelli, R., Cimino, A., Geerlof, A., Kubo, A., Satou, Y., Cambillau, C., 2011. High-throughput protein expression screening and purification in *Escherichia coli*. *Methods* 55, 65–72, <http://dx.doi.org/10.1016/j.ymeth.2011.08.010>.
- Wells, J.A., Estell, D.A., 1988. Subtilisin—an enzyme designed to be engineered. *Trends Biochem. Sci.* 13, 291–297, [http://dx.doi.org/10.1016/0968-0004\(88\)90121-1](http://dx.doi.org/10.1016/0968-0004(88)90121-1).
- Zhang, R.H., Xiao, L., Peng, Y., Wang, H.Y., Bai, F., Zhang, Y., 2005. Gene expression and characteristics of a novel fibrinolytic enzyme (subtilisin DFE) in *Escherichia coli*. *Lett. Appl. Microbiol.* 41, 190–195, <http://dx.doi.org/10.1111/j.1472-765X.2005.01715.x>.
- van den Ent, F., Löwe, J., 2006. RF cloning: a restriction-free method for inserting target genes into plasmids. *J. Biochem. Biophys. Methods* 67, 67–74, <http://dx.doi.org/10.1016/j.jbbm.2005.12.008>.
- Zhu, X.L., Ohta, Y., Jordan, F., Inouye, M., 1989. Pro-sequence of subtilisin can guide the refolding of denatured subtilisin in an intermolecular process. *Nature* 339, 483–484, <http://dx.doi.org/10.1038/339483a0>.

II

RESEARCH ARTICLE

Mutational analysis of the pro-peptide of a marine intracellular subtilisin protease supports its role in inhibition

Gro E. K. Bjerga¹  | Øivind Larsen¹ | Hasan Arsin² | Adele Williamson³ | Antonio García-Moyano¹ | Ingar Leiros³ | Pål Puntervoll¹

¹Uni Research, Center for Applied Biotechnology, Thormøhlens gate 55, Bergen, 5006, Norway

²Department of Biological Sciences, University of Bergen, Thormøhlens gate 53, Bergen, 5006, Norway

³The Norwegian Structural Biology Centre (NorStruct), Department of Chemistry, UiT The Arctic University of Norway, Tromsø, 9037, Norway

Correspondence

Gro Elin Kjæreng Bjerga, Uni Research, Center for Applied Biotechnology, Thormøhlens gate 55, Bergen, 5006, Norway
E-mail: gro.bjerga@uni.no

Abstract

Intracellular subtilisin proteases (ISPs) have important roles in protein processing during the stationary phase in bacteria. Their unregulated protein degrading activity may have adverse effects inside a cell, but little is known about their regulatory mechanism. Until now, ISPs have mostly been described from *Bacillus* species, with structural data from a single homolog. Here, we study a marine ISP originating from a phylogenetically distinct genus, *Planococcus* sp. The enzyme was successfully overexpressed in *E. coli*, and is active in presence of calcium, which is thought to have a role in minor, but essential, structural rearrangements needed for catalytic activity. The ISP operates at alkaline pH and at moderate temperatures, and has a corresponding melting temperature around 60 °C. The high-resolution 3-dimensional structure reported here, represents an ISP with an intact catalytic triad albeit in a configuration with an inhibitory pro-peptide bound. The pro-peptide is removed in other homologs, but the removal of the pro-peptide from the *Planococcus* sp. AW02J18 ISP appears to be different, and possibly involves several steps. A first processing step is described here as the removal of 2 immediate N-terminal residues. Furthermore, the pro-peptide contains a conserved LIPY/F-motif, which was found to be involved in inhibition of the catalytic activity.

KEYWORDS

ISP, LIPY/F-motif, *Planococcus*, protease structure, subtilisin

1 | INTRODUCTION

ISPs have key roles in cell cycle regulation, specifically in protein recycling by processing proteins during transition to the stationary phase.^{1,2} To prevent proteolysis that may be lethal to the cell, the activity of an intracellular protease must be tightly controlled. Although a potential ISP inhibitor protein has been identified,^{3,4} the primary mechanism of regulation is likely intrinsic.^{5,6} In the precursor protein, an N-terminal pro-peptide of typically 16–20 residues binds across the active site and inhibits activity. As shown for a few homologs,^{6,7} the pro-peptide is released by intra-molecular maturation allowing the enzyme to act on exogenous substrates. ISPs are homodimeric,⁶ which contributes to making ISPs a structurally distinct family of subtilases. The catalytic domain of ISPs are homologous to those of other members of the Subtilisin superfamily, such as the

extracellular subtilisin proteases (ESPs), which is a "Peptidase S8" domain in the Pfam classification.⁸ According to the MEROPS peptidase database,⁹ both ISPs and ESPs belong to the S08 family in clan SB.

Within this domain a catalytic triad is arranged as an aspartate, a histidine and a serine (Asp32, His64, and Ser221, respectively, referring to the processed SubtilisinE from *B. subtilis*; Uniprot ID: CAB12870). In brief, the nitrogen-bonded protein (Ne2-H) of His64 is hydrogen bonded to the hydroxyl group proton of Ser221. This interaction causes a charge separation of the hydroxyl, deprotonating the serine oxygen and activating it for nucleophilic attack on the carbonyl of the peptide substrate, which ultimately leads to breakage of the peptide bond.¹⁰ Aside from homology within the catalytic domain, significant architectural differences exist between ISPs and ESPs. The N-termini of ESPs contain short leader sequences of about 20–30

residues for protein secretion,¹¹ followed by a pro-domain of typically 60–80 residues,^{12,13} which is not conserved in sequence, but vital to their folding and function.¹⁴ In an analogous manner to the ISP pro-peptide, the ESP pro-domain is processed intra-molecularly during maturation of the enzyme into an active conformation. The pro-domain has dual roles in acting as an inhibitor,^{15,16} and as a molecular chaperone that guides folding of the active enzyme.^{16–18}

The first ESP structure was solved in 1969,¹⁹ and has since been reported for several homologues^{20,21} and a number of engineered mutants.²² For ISPs, however, structural information is known from a single homologue, the *Bacillus clausii* ISP,^{5,6} with 4 structures reported (PDB IDs: 2WVT, 2WWT, 2X8J, and 2XRM). The 4 structures represent 2 activity states: the inactive state with the inhibitory pro-peptide bound and the active state without the pro-peptide bound. Notably, all *B. clausii* structures are from inactive mutants carrying catalytic Ser250 to Ala mutations.

In ISPs, the leader sequence and pro-domain of ESPs are replaced with a pro-peptide (also termed N-terminal extension). The pro-peptide binds across the active site, with residues Phe4–Leu6 forming a central β -strand of a 3-stranded antiparallel β -sheet.⁶ The pro-peptide also contains a LIPY/F motif, not found in ESPs. In *B. clausii* ISP this motif is involved in inhibiting the active site. Residues within the motif contribute to disruption of the conformation of the catalytic triad by shifting the catalytic Ser and His residues apart.⁵ According to a standardized residue nomenclature for peptide binding to the active site,²³ residues N-terminal to the scissile bond of the peptide substrate are termed P4, P3, P2, and P1, and those C-terminal to the bond are termed P1', P2', P3', and P4', where the scissile bond is between P1 and P1'. The corresponding sites in the enzyme are S4, S3, S2, S1, S1', S2', S3', and S4'. In *B. clausii* ISP, Leu6 and Ile7 correspond to P2 and P1 and are pointing inwards into the hydrophobic pocket at the S2 and S1 sites, respectively. Pro8 holds a unique position at the centre of a small curve, which displaces the peptide bond between Ile7 (P1 site) and Pro8 (P1' site) out of reach of the active site Ser, whereas Tyr9 is occupying the S1' site. The proline-centred curve is unique in *B. clausii* ISP, and contrasts the scissile bond in ESPs, which is positioned to allow autoproteolytic processing. Altogether, the structure suggests that the residues in the pro-peptide are involved in blocking the active site serine.^{5,6}

Both ESPs and *B. clausii* ISP harbor a conserved high affinity metal-binding site occupied by a metal ion that serves a structural role.^{5,6,24,25} The high affinity metal-binding site in ESPs is occupied by calcium,^{24,26} whereas in *B. clausii* ISP it is occupied by sodium.^{5,6} In addition, *B. clausii* ISP has 2 unique binding sites for divalent metal ions, probably occupied by calcium ions, in each monomer: 1 close to the dimer interface and 1 in proximity to the active site. The latter is involved in ordering a loop that contributes to formation of 1 of the binding sites (S1) involved in catalysis. Due to the processing of the pro-peptide and the positioning of calcium, the catalytic triad and substrate binding cleft are significantly rearranged, especially at the S1 binding site.⁵ In a proposed model for ISP regulation,²⁷ it was suggested that once a minor fraction of the pool of ISPs adopts an open conformation, calcium binding takes place and reshapes the S1 binding site, which ultimately releases the pro-peptide within this population and leads to a cascade of activation of other ISPs. The sequence

of events and details of how the maturation precedes, in particular the role of calcium, are not known.

This study reports an ISP from a marine isolate, *Planococcus* sp. AW02J18, which is from a related, but phylogenetically distinct genus to *B. clausii*. Here, we present biochemical data for the recombinant enzyme, showing it is active in presence of calcium, at alkaline pH and moderate temperatures. We furthermore present a high-resolution structure of an ISP with an intact catalytic triad and an inhibitory pro-peptide bound across the active site. The structure supports previous findings and unique features of ISPs, such as its dimeric nature, sodium binding in the high-affinity metal-binding site and active site blocking by the pro-peptide. The processing of the pro-peptide appears however to be different from reported ISPs, possibly involving multiple processing steps. We also present mutagenesis data supporting an inhibitory role of the LIPY/F motif of the pro-peptide.

2 | MATERIALS AND METHODS

2.1 | *In silico* identification of an intracellular subtilisin protease

The ISP sequence was identified from sequence-based mining of a marine bacterial isolate, *Planococcus* sp. AW02J18 (Supporting Information Table S1). This isolate was collected during expeditions in the coastal areas of Lofoten in 2009, and is stored in a bacterial collection at the University of Tromsø. The sampling procedure and collection has been presented elsewhere.²⁸ Genomic material was isolated for Illumina sequencing (MiSeq). Using a sequence-based approach, translated genomic sequences from a marine bacterial collection were mined for subtilisin-like proteases by searching for S08 family homologs against the MEROPS database.⁹ The ISP candidate was identified in this data set, and the sequence has been deposited in GenBank with the accession code MG786190.

2.2 | The LIPY/F sequence conservation

Sequences homologous to *Planococcus* sp. AW02J18 ISP were identified using the UniProt blast search engine (default settings) against the UniRef90 database (UniProt release 2017_10).²⁹ Sequence hit number 156, UniRef90_A0A136C445, was the first sequence to contain 2 motif mutations (LVNE) making the motif unlikely to be functional and was used to define the distance cut-off (expect value 4e-107; 57% sequence identity to *Planococcus* sp. AW02J18 ISP). Hence, the top 155 sequence hits were used to make a multiple sequence alignment (MAFFT, default settings).³⁰ Three sequences were fragments that lacked the LIPY/F motif, and were manually removed (UniRef90: UPI00098840FB, UPI000590D2A7, UPI000689F3EC). The alignment containing the remaining 152 sequences was used to construct a sequence logo (default parameters).³¹

2.3 | Sub-cloning of the *isp* gene to expression vectors

To facilitate enzyme expression we used our previously developed screening procedure for subtilisin-like serine proteases.³² The

Planococcus sp. AW02J18 ISP protein sequence was used as template for gene synthesis (GenScript), and the synthetic *isp* gene was codon-optimized to improve its expression in *E. coli*. The *isp* gene was synthesized with flanking *SapI* sites, and delivered in a customized *SapI*-free pUC57 vector with kanamycin selection marker. The *isp* gene was sub-cloned from the delivery vector to a suite of expression vectors using a fragment exchange (FX) cloning method.³³ Construction of the expression vectors have been described previously.³²

2.4 | Gene truncation and mutagenesis

Truncation constructs and mutants were prepared from the pUC57 template. Primers were designed to contain a *SapI*-cloning site and a 15–20 bp gene-specific region targeting the desired truncation start. Primers in Supporting Information Table S2 were used to amplify the truncated ISP versions by PCR using Phusion polymerase. Gene fragments were purified, and cloned into the pINITIAL cloning vector by FX-cloning.³² Plasmids were sequenced to confirm correct truncations. Point mutations were prepared by site-directed mutagenesis using primers in Supporting Information Table S2. Truncation constructs and mutants were sub-cloned into the p12 expression vector, using FX cloning.

2.5 | Small-scale expression and analysis of protein integrity

Small-scale recombinant expression was carried out according to the protocol described previously³² in 4 mL culture volumes. Following expression, cells were collected and resuspended in 1 mL lysis buffer (50 mM Tris-HCl pH 8.5, 50 mM NaCl, 0.25 mg/mL lysozyme, 10% (v/v) glycerol). Lysis was completed by ultrasonication using two 5-s pulses at 40–60% amplitude with a CV-18 probe powered by an Ultrasonic Homogenizer 4710 (Cole Parmer). Lysates were cleared by centrifugation at 4600 × *g* for 20 min. Cleared lysate samples (representing soluble fraction) were analyzed by SDS-PAGE and immunoblot as described previously.³² As background controls, lysates containing empty vector were used, herein termed GS due to the insertion of triple GS encoding sequence as a replacement of the *ccdB* gene in the expression vector.³²

Semi-quantitative analysis of recombinant protein in cleared extracts was performed in Image Lab 3.0 (BioRad). Target band intensities were extracted from image data of Coomassie-stained SDS-PAGE gels, and normalized to the total protein intensities in the lane excluding the target band intensities to adjust for variable growth rates and protein expression levels.

2.6 | Large-scale expression

E. coli MC1061 cells containing the p1:ISP, p12:ISP, or the p12:ISP-S251A (catalytic mutant) constructs were grown in 1 L terrific broth medium (1.2% tryptone, 2.4% yeast extract, 0.4% glycerol, 17 mM KH₂PO₄, and 72 mM K₂HPO₄) supplemented with ampicillin (100 µg/mL) in 2.5 L Thomson's Ultra Yield™ flasks (Thomson Instrument Company). Protein expression was induced by 0.1% (w/v) *L*-arabinose overnight at 20 °C with 250 rpm shaking. Cells were collected by

centrifugation (JLA-9.1000 rotor, Beckman) at 7500 × *g*, 30 min at 4 °C, and stored at –20 °C.

2.7 | Protein purification

Frozen cell pellets from about 1 L culture were resuspended in 50 mM Tris HCl pH 7.5 at room temperature (RT, roughly around 20 °C), 150 mM NaCl and 0.25 mg/mL lysozyme. After incubation for 30 min at 37 °C and 250 rpm, the cell suspension was cooled on ice before sonication in a final concentration of 500 mM NaCl. Cell debris was removed by centrifugation at 20,000 × *g* for 20 min at 4 °C (JA-25.50 rotor, Beckman). The cleared lysate was loaded onto 2 × 5 mL HisTrap FF crude columns (GE Healthcare) equilibrated with 50 mM Tris-HCl pH 7.5 (at RT), 500 mM NaCl and 10 mM imidazole on the ÄKTA Pure (GE Healthcare) system. Bound proteins were eluted in the same buffer containing 800 mM imidazole. Fractions containing protein were pooled and dialyzed 2 times in Spectra Por® dialysis tubes (Spectrum Laboratories, Inc.) with 6–8 kilo dalton (kDa) MWCO against 1 L 20 mM Tris-HCl pH 7.5 overnight at 4 °C. 1 mM CaCl₂ was added to a 50 µg/mL ISP solution and incubated overnight at RT during slow stirring, yielding what we herein term Asn3-ISP (approx. 35 kDa). Protein solutions were concentrated using Amicon 10 kDa MWCO spin-filter columns (Merck) with buffer exchange to 50 mM Tris HCl pH 7.5, 50 mM NaCl and stored in aliquots at 4 °C at concentrations 80 mg/mL (WT) and 150 mg/mL (mutant). From 1 L expression culture yields of 0.2 g of purified Asn3-ISP (no tags), and 0.4 g of the catalytic mutant (with C-terminal His-tag) were typically achieved. Purity and protein mass estimation was assessed by quantitative analysis in Image Lab 3.0 (BioRad), by extracting the target band intensities from image data of Coomassie-stained SDS-PAGE gels.

The size exclusion chromatography experiments were ran using a pre-calibrated Superdex 200 10/300 GL (GE Healthcare) column on the ÄKTA Explorer (GE Healthcare) system. The system was equilibrated using the loading buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl). 500 µl Asn3-ISP at 0.8 mg/mL concentration was loaded onto the column in loading buffer in the absence or presence of 2 mM CaCl₂ or 1 mM EDTA. The CaCl₂ supplemented sample was prepared immediately before the chromatography experiment to avoid autolysis, and the EDTA-treated samples were prepared overnight in order to allow time for chelation.

Mass spectrometry (MS) analyses were performed at the PROBE facility (University of Bergen, Norway). N-Terminal amino acid sequencing was carried out at Alta Bioscience (University of Birmingham, United Kingdom).

2.8 | Protease activity assays

The protease fluorescent detection kit (Sigma-Aldrich) was used for routine detection of proteolytic activity as previously described.^{32,34} Briefly, 10 µL lysate or 5 µM enzyme was assessed for activity on FITC-casein in 50 mM TrisHCl pH 8.5 (at RT), 50 mM NaCl, in absence or presence of 1 mM CaCl₂ in a total volume of 50 µL at 37 °C for 1 h unless otherwise stated. Temperature optimum was assayed using the FITC-casein assay. For the mutants, activity was assessed using EnzChek™ Protease Assay Kit (ThermoFischer).

10 $\mu\text{g/mL}$ BODIPY FL casein was prepared by resuspending the substrate in 50 mM Tris-HCl pH 8.5 (at RT) and 50 mM NaCl. 12.5 μL of BODIPY-FL casein was used per reaction, with 10 μL cleared extract in 50 mM Tris pH 8.5 (at RT), 50 mM NaCl and 1 mM CaCl_2 in a final volume of 100 μL . Samples were incubated at 37 $^\circ\text{C}$ for 1 h, and fluorescence was read.

pH optimum was determined using 1 μM Asn3-ISP, 350 μM N-succinyl-AAPF p-nitroanilide (Sigma-Aldrich) in 50 mM NaCl, 1 mM CaCl_2 , and 50 mM buffer (citrate buffer pH 3.0-6.0, acetate buffer pH 4.0-6.0, sodium phosphate buffer pH 6.0-8.0, Tris-HCl buffer pH 7.0-9.0 and glycine buffer pH 9.0-11.0). Reaction was run at 25 $^\circ\text{C}$ for 20 min, in presence of excess substrate.

2.9 | Determining the specific activity

Specific activity was determined using a protease colorimetric detection kit (Sigma-Aldrich). To avoid assay interference with amino groups from Tris, Asn3-ISP was dialyzed against 25 mM borate/NaOH pH 8.2, 50 mM NaCl before assaying. Casein was solubilized in water at pH 8.3. One unit is defined as the amount of enzyme that will hydrolyze casein to produce color (as determined by addition of Folin-Ciocalteu's Reagent) equivalent to 1.0 μmole tyrosine per minute at pH 8.3 at 37 $^\circ\text{C}$ in presence of 10 mM CaCl_2 .

2.10 | Differential scanning calorimetry

Prior to Differential Scanning Calorimetry (DSC) measurements, aliquots of Asn3-ISP at approximately 1 mg/mL were dialyzed into the following conditions overnight at 4 $^\circ\text{C}$: 50 mM Hepes pH 8.0, 50 mM NaCl (DSC buffer); DSC buffer with 2 mM CaCl_2 ; DSC buffer with 1 mM ethylenediaminetetraacetic acid (EDTA). Thermal unfolding transitions were measured using a Nano-Differential scanning Calorimeter-III (Calorimetry Sciences Corporation) from 5 to 75 $^\circ\text{C}$ with scan rates of 1 $^\circ\text{C/s}$. Buffer from the final dialysis step was used as a reference. Data were analyzed using the NanoAnalyze software (TA Instruments).

2.11 | Crystallization

Crystallization experiments were performed with a stock solution of purified Asn3-ISP at 30 mg/mL in 50 mM TrisHCl pH 7.5 (at RT), 50 mM NaCl. Initial crystallization conditions were screened using the vapor diffusion sitting drop method set up by a Phoenix crystallization robot (Art Robbins Instruments). The plates were set up with 60 μL reservoir solutions and sitting drops with equal amounts of reservoir solution mixed with protein stock solution in a total drop volume of 1 μL . The screens were incubated at 20 $^\circ\text{C}$. Diffraction-quality crystals were obtained from 6 conditions, as outlined in Supporting Information Table S3.

2.12 | X-ray data collection

Crystals grown in 0.25 M NH_4Ac , 22% PEG 1500, 0.1 M Na-Citrate pH 4.0, were transferred through a cryoprotectant solution (crystallization conditions with 20% (v/v) glycerol added, thereafter mounted in a nylon loop and flash-cooled in liquid N_2 . X-ray diffraction data were collected at the European Synchrotron Radiation Facility (ESRF; Grenoble, France) beamline ID23EH1. The data were integrated by

TABLE 1 Data collection and processing statistics^a

Diffraction source/Beamline	ESRF ID23EH1
Wavelength (\AA)	0.98
Detector	Q315R CCD (ADSC)
Crystal-to-detector distance (mm)	214.77
Rotation range pr. image ($^\circ$)	0.1
Total rotation range ($^\circ$)	130
Space group	$P2_12_12_1$
a, b, c (\AA)	70.17, 85.18, 104.58
α, β, γ ($^\circ$)	90, 90, 90
Mosaicity ($^\circ$)	0.2
Resolution range (\AA)	66.05-1.30 (1.32-1.30)
Total No. of reflections	675238 (36457)
No. of unique reflections	142753 (7606)
Completeness (%)	92.2 (99.1)
Multiplicity	4.7 (4.8)
$\langle I/\sigma(I) \rangle$	16.4 (2.2)
$R_{p,i.m.}$	0.026 (0.412)
Wilson B-factor (\AA^2)	17.39

^a Values in parentheses are for the outermost shell

XDS/XSCALE,³⁵ scaled and analyzed by programs in the CCP4 program suite³⁶ through autoPROC.³⁷ A summary of the data collection statistics is found in Table 1.

2.13 | Structure determination

The crystal structure was solved by molecular replacement using MolRep in the CCP4 program package³⁶ with 2XRM⁵ as search model (a representative structure of the homologous ISP from *B. clausii*). The initial refinement was executed in Refmac³⁸ followed by automated model improvement in Buccaneer.³⁹ The manual building was done in Coot⁴⁰ interspersed by cycles of refinement in Phenix⁴¹ and resulted in final $R_{\text{cryst}}/R_{\text{free}}$ values of 13.04/15.03. A summary of the refinement statistics is shown in Table 2. The atomic coordinates and structure factors have been deposited in the RCSB Protein Data Bank (www.rcsb.org) with the accession code 6F9M. Figures presented in the results section were generated using Chimera.⁴²

3 | RESULTS

3.1 | A new intracellular subtilisin protease with a conserved LIPY/F motif

A previously uncharacterized protease from *Planococcus* sp. AW02J18 was identified in an enzyme discovery initiative as a candidate for expression in *E. coli* (Supporting Information Table S1). According to sequence analysis, this protease contained a catalytic domain (Peptidase_S8/PF00082) as annotated by Pfam (residues 40-311, Figure 1A). Sequence analysis also revealed that it shared 53% sequence identity to the previously described intracellular subtilisin protease (ISP) from *B. clausii*⁶ (Supporting Information Figure S1). As expected from SignalP analysis, the ISP sequence does not contain a leader sequence to direct its export,⁴³ and is thus predicted to have

TABLE 2 Structure determination and refinement statistics

Resolution range (Å)	44.563-1.298
Completeness (%)	92.19
No. of reflection, working set	142740
No. of reflection, reference set	1992
Final R_{cryst}	13.04
Final R_{free}	15.03
MolProbity score	1.315
Clashscore	3.35
No. of non-H atoms	
Protein ^a	4548
Water	549
Other ^b	30
Total	5127
R.m.s. deviations	
Bonds (Å)	0.007
Angles (°)	0.962
Average B factors (Å ²)	
Overall	23.57
Protein	21.87
Water	37.20
Other*	32.01
Ramachandran plot (%)	
Preferred	96.87
Allowed	2.61
Outliers	0.52

^a Two molecules pr. asymmetric unit.

^b Two molecules of acetate, sodium and triethylene glycol (Peg 3) were identified in the electron density and modeled. These occupy similar positions around the 2 protein molecules in the asymmetric unit.

an intracellular localization. Instead, the *Planococcus* sp. AW02J18 ISP contains a short pro-peptide with a LIPY-sequence at the N-terminus, also identified in other homologues (Supporting Information Figure S1). Although the LIPY sequence has been reported as a conserved motif,⁶ evidence of its conservation has not previously been presented. To analyze the evolutionary conservation of the motif, sequences homologous to the *Planococcus* sp. AW02J18 ISP were collected. Using 152 UniRef90 sequences in a sequence alignment, we analyzed conservation of the motif in a context with 2 flanking residues on each side (8 residue window). A LIPY/F motif is derived from the alignment (Figure 1B). A hydrophobic leucine or valine, or in rare cases an isoleucine occurs at the first position. At the second position, the motif contains most often a hydrophobic isoleucine, but in certain sequences phenylalanine, leucine or valine. The third position is occupied by a highly conserved proline found in all but 2 sequences. This residue is structurally significant as part of the proline-centred curve in *B. clausii* ISP, which positions the scissile bond between proline and the previous residue out of reach for autocatalysis. At the fourth position, an aromatic tyrosine, phenylalanine or in rare cases histidine occurs. At flanking positions of these 4 residues some consensus occurs, such as a charged residues at proximate positions to the LIPY/F motif, and hydrophobic residues at positions 2 residues upstream and downstream (Figure 1B). A 4-residue motif can be expressed using the Prosite pattern syntax as [LVI]-[IFLV]-P-[YFH].

3.2 | The first 2 residues of the calcium-dependent ISP is processed

The full-length *isp* gene from *Planococcus* sp. AW02J18 was sub-cloned to a suite of expression vectors for heterologous expression. From SDS-PAGE analysis, we found that all recombinant constructs yielded soluble enzyme, but that solubility was further improved by use of fusion tags (Figure 1C). Since many serine proteases require calcium for proper folding and structural stability, activity was assessed on fluorescein isothiocyanate (FITC) conjugated casein in the absence or presence of calcium ions. Compared to extracts from strains carrying empty vectors, all recombinant enzymes were active, but required calcium for activity (Figure 1D). The p1-construct encoding an N-terminal deca-histidine (His) tag was chosen for in-depth characterization due to its potential to yield a recombinant enzyme that would mimic the native processed ISP, and ease downstream purification (Figure 1). In the absence of calcium, immobilized metal affinity chromatography (IMAC) was used for protein purification of His-ISP (approx. 38 kDa). In analogy to the ISP from *B. clausii*, the enzyme was incubated in presence of calcium to mature by autoproteolysis. From SDS-PAGE we obtained a "matured ISP", with an expected lower mass (~35 kDa) than full-length, of 95% purity (Figure 2A). With N-terminal sequencing we determined the starting residue on this protein entity to Asn3; we thus termed this protein Asn3-ISP. Using Asn3-ISP, we found that increasing concentrations of calcium had a positive effect on activity (Figure 2B), whereas EDTA inactivated the enzyme (Figure 2C). From SDS-PAGE analysis of the reaction products, we found that Asn3-ISP was further processed or degraded in presence of calcium (Figure 2C). In absence of calcium or in calcium-depleted reactions, the enzyme was however persistent against proteolysis (Figure 2C), and could be stored for 1 month without any effect on activity (Supporting Information Figure S2).

To further understand the processing, calcium chloride was added at various concentrations to the full-length recombinant enzyme (His-ISP) at a pH range 7.0-8.5. SDS-PAGE revealed that 2 processed ISP species < 37 kDa appeared in presence of 1 mM CaCl₂ (Figure 2D, protein bands numbered 2-3). N-terminal sequencing was performed on these 2 processed protein species, but data were only conclusive for the uppermost processed protein. In this protein, the artificial N-terminal residues (His-tag and 3C protease site) and the 2 first native residues of the ISP (residues Met1, Lys2) were processed. This confirms what is referred to as the Asn3-ISP. Increasing the concentration of calcium chloride up to 10 mM led to further processing as well as the appearance of degradation products (that is, fragments smaller than the 31 kDa peptidase domain). MS analyses were performed on various entities after calcium-induced activation, with identification of ISP peptides in all samples (Supporting Information Figure S3 and Table S5). No obvious sequential pattern between protein entities was identified. Tag-removal was confirmed by immunoblot analysis and compared to a catalytic mutant designed by replacing the catalytic Ser251 with Ala (Supporting Information Figure S4). The processing of the recombinant ISP from *Planococcus* sp. AW02J18 appears to occur in multiple steps.

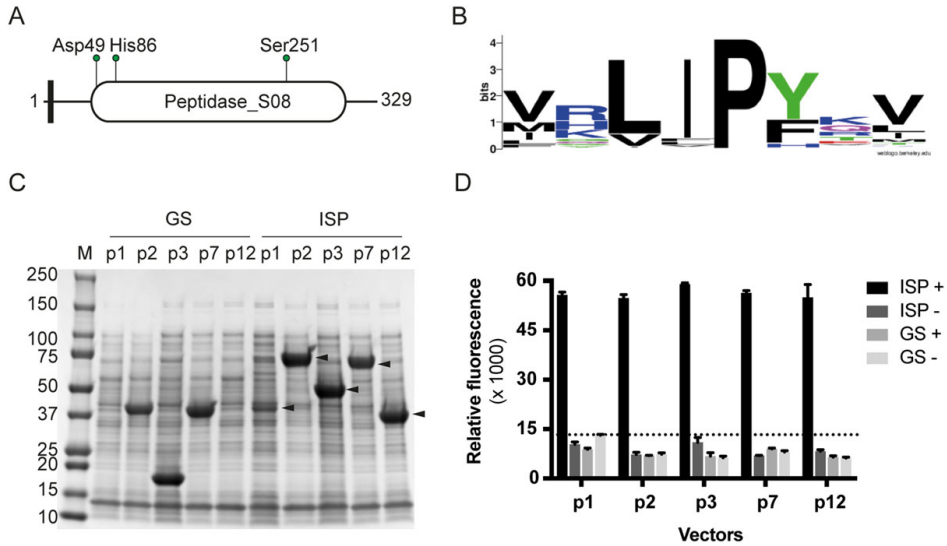


FIGURE 1 Overexpression and activity assessment of the recombinant ISP. A, A cartoon of the ISP architecture drawn to scale. Black box, LIPY/F motif; oval circle, Peptidase_S08 Pfam domain (PF00082); green pins point to residues involved in catalysis (catalytic triad). B, Sequence logo showing the evolutionary conservation of the LIPY/F motif based on an alignment with 152 ISP sequences. C, ISP constructs were produced from multiple vectors, and cleared lysates were inspected on SDS-PAGE for the presence of soluble overexpressed proteins. Arrows indicate soluble ISP proteins. Fusion partners from the various vectors are: p1, N-terminal His-tag; p2, N-terminal His-tag and MBP; p3, N-terminal His-tag and SUMO protein; p7, N-terminal MBP and C-terminal His-tag; p12, C-terminal His-tag. Empty vector controls (GS) will produce fusion partners only, wherein MBP and SUMO can be observed on SDS-PAGE. M, BioRad's Precision Plus Protein™ Dual Color Standard. D, Cleared lysates (see C for details), including empty vector controls (GS) were assayed overnight with FITC-casein, in the presence (+) or absence (-) of 1 mM CaCl_2 . Dotted line indicates the highest data point for background measurements (in the absence of calcium)

3.3 | *Planococcus* sp. AW02J18 ISP operates at moderate temperatures and alkaline pH

To identify its optimal conditions for further activity assessments, Asn3-ISP was characterized with respect to the specific activity, temperature and pH optimum in casein assays (Figure 3). It was found to operate optimally at pH 11.0, but was active across pH 7.0–11.0, whereas no activity was observed below pH 6.0 (Figure 3A). Precipitation was observed at pH 4.0 in both citrate and acetate buffers, likely explained by an estimated pI around 4. The temperature optimum was found to be around 45 °C (Figure 3B). No activity was identified above 60 °C, which indicates that the protein is destabilized at high temperatures. Using optimal temperature (45 °C) in alkaline conditions (pH 8.3) and 10 mM CaCl_2 the specific activity of the ISP was determined on casein to be 13 ± 1 U/mg.

To determine the thermal unfolding temperature of Asn3-ISP, DSC measurements were carried out (Figure 4). The enzyme unfolded as a single peak, which could be fitted to 2 two-state transitions with melting temperatures (T_m) separated by approximately 3.0 °C (Table 3). In the DSC data, the apparent T_m in absence of calcium and EDTA was around 60 °C, which is consistent with the data on temperature optimum and stability (Figure 4A). Addition of CaCl_2 increased the directly measured T_{max} by 1.7 °C, and the apparent T_m by up to 3.0 °C indicating that calcium has a stabilizing effect on the enzyme (Figure 4B). The presence of EDTA slightly increased the apparent T_m

(Figure 4C). Repeat scanning did not give rise to any subsequent unfolding transitions, indicating that ISP does not refold on the time-scale used for this experiment; therefore the thermodynamics of unfolding were not analyzed further. No exothermic signals indicative of aggregation were present in the raw data (not shown), and no visible precipitate was observed suggesting that these data can be used in a comparative manner to understand the effect of EDTA and calcium on the system.

3.4 | Structure of ISP with an intact catalytic triad and pro-peptide

ISPs are distinct from ESPs with regards to the N-terminal pro-peptide, their dimeric structure, and the sodium binding in the high affinity metal binding site,^{5,6} but details regarding their maturation are still unclear. To shed light on the latter, the crystal structure of the Asn3-ISP was determined by X-ray crystallography to a resolution of 1.3 Å (Figure 5). In addition to being the second unique structure of an ISP, it is the first structure of an ISP with a native catalytic triad, and it represents the highest resolution structure of this enzyme family to date. The structure of the ISP (residue 3–310) is dimeric, with each monomer including an almost intact pro-peptide bound across the active site. Using size exclusion chromatography (SEC) the molecular weight of Asn3-ISP in solution was estimated. The absence of calcium and presence of EDTA gave similar elution profiles in SEC and

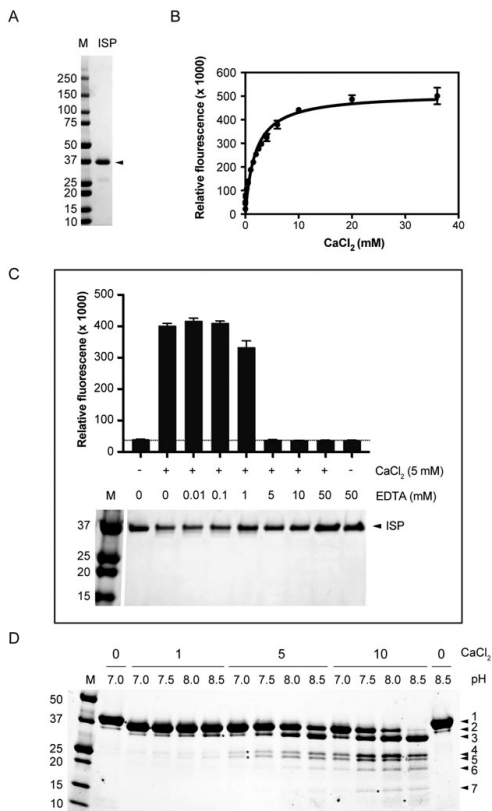


FIGURE 2 ISP activity, stability and processing. A, SDS-PAGE analysis of purified Asn3-ISP. Purity was estimated to 95% by inspection of lane intensity. B, Using the FITC-casein assay, 4 μ M Asn3-ISP was incubated with increasing concentration of CaCl₂ at 37 °C for 1 h. 50% activity is achieved with 2.5 mM CaCl₂. C, The activity of 1 μ g Asn3-ISP (as in A) was measured in the presence or absence of 5 mM CaCl₂ and concentrations of EDTA up to 50 mM (upper panel). Dotted line represents the average of buffer (37248 units) in presence of 5 mM CaCl₂ and 5 mM EDTA. Lower panel shows SDS-PAGE containing 0.5 μ g Asn3-ISP treated with CaCl₂ and EDTA as in the activity assay. M, BioRad's Precision Plus Protein™ Dual Color Standard. D, The His-ISP protein construct (p1-construct, numbered 1) was used to investigate calcium-induced maturation. 1, 5 or 10 mM CaCl₂ was added to 2 μ g enzyme at pH range 7.0-8.5 at room temperature for incubation overnight before analysis on SDS-PAGE. Theoretical mass of Asn3-ISP (numbered 2), 35 kDa. M, BioRad's Precision Plus Protein™ Dual Color Standard

mass estimates; here, 13.4 mL and 115 kDa, respectively (Supporting Information Figure S5). In presence of calcium, the Asn3-ISP eluted in 2 peaks at 13.2 and 14.1 mL corresponding to masses of approximately 128 and 88 kDa, respectively (referring to the regression line produced from known calibration proteins). Based on the theoretical mass of Asn3-ISP being 35 kDa, this suggests 3.3 monomers per oligomer in absence of calcium. In presence of calcium the oligomeric state shifts to a mixed population of both 3.7 and 2.5 monomers per oligomer. In the structure, there are 2 molecules of triethylene glycol

(Peg3) symmetrically bound at the dimer interface distant from the active site (Figure 5A), which may be adducts of Peg 1500 during crystallization or introduced during recombinant expression. In 3 structures of ISP from *B. clausii*, similar molecules are bound in this region: a strontium ion and a tetraethylene glycol molecule bound in an overlapping position (PDB ID: 2XRM); 3 water molecules bound in the same region (PDB ID: 2WWT); and Peg3 (PDB ID: 2X8J) almost perfectly overlapping the conformation observed in the *Planococcus* sp. AW02J18 ISP structure.

The structure contains a catalytic core (residues 20-310) overlapping the Pfam assigned Peptidase_S8 domain (residues 40-311). The first 2 residues, 2 loop regions (residues 184-191 and 217-223), and the C-terminal 20 residues are not defined in the electron density. Superpositioning of *Planococcus* sp. AW02J18 ISP (chain A) with the catalytic mutant *B. clausii* ISP (PDB ID: 2X8J, chain A) gave an RMSD of 0.67 Å across 282 atom pairs in an improved fit where far-apart residues are removed (across all 292 atom pairs of residues in the alignment: 1.21 Å), confirming that they have the same overall fold (Figure 5B). Superpositioning showed that catalytic triad residues are structurally conserved, although distances are slightly different in each monomer. Two distinct conformations were modelled in each monomer due to poor electron density: Monomer A, residues 248-252 (including the catalytic triad residue Ser251) and Monomer B, residues 16-20 (including parts of the pro-peptide). In monomer A the distances between Ser251O γ and His86Ne2 is 3.20 and 3.58 Å, respectively, whereas the corresponding distance for monomer B measures to 3.82 Å (Figure 5C). Superpositioning with the structure representing the active state of *B. clausii* ISP (PDB ID: 2XRM) has a shorter distance, although only estimated, as both *B. clausii* structures are Ser251Ala mutants. One surface loop (residues 97-104) is different, probably reflecting an insertion in the *Planococcus* sp. AW02J18 ISP (Supporting Information Figure S1). Although the side-chains of some residues in this loop (residues Asp100, Glu101, and Glu102) are visible only at low contour levels, a sodium ion in each monomer was putatively identified and modeled in electron density as for the *B. clausii* ISP structures (PDB IDs: 2XRM and 2X8J).

The 2 loop regions that are disordered (residues 184-191 and 217-223) in *Planococcus* sp. AW02J18 ISP are ordered in the structure that simulates the active state of the *B. clausii* ISP. Residues from both loops are contributors in the coordination of a calcium ion, in *B. clausii* ISP, these are: Asp186 (side-chain; SC), Arg188 (main-chain; MC), Thr191 (MC), Glu193 (SC), and Thr221 (SC). In *Planococcus* sp. AW02J18 ISP the residues contributing with specific side-chain contacts to the calcium ion are conserved, while 1 of the 2 unspecific main chain contacts are not conserved (Supporting Information Figure S1).

3.5 | Mutations in the LIPY/F motif of the pro-peptide relieve inhibition

Removal of the first 18 residues of *B. clausii* ISP by calcium treatment or by truncation released an ISP enzyme in an active conformation.⁵ The proteolytic site for cleavage is however not conserved among ISPs (Supporting Information Figure S1). As calcium seemed to improve activity (Figure 2B), but also further process the Asn3-ISP (Figure 2D), we aimed at identifying the second processing site for

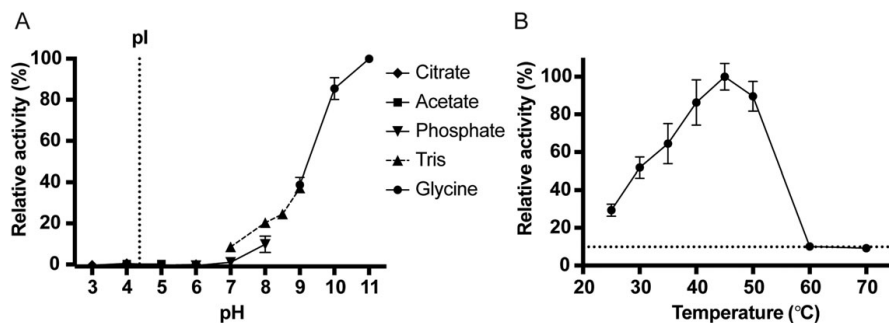


FIGURE 3 pH and temperature optimum of ISP. A, Using the N-succinyl-AAPF p-nitroanilide peptide, the activity of 1 μ M Asn3-ISP at pH 3.0–11.0 was measured in the initial rate of the reaction at 25 $^{\circ}$ C. Background from buffer was subtracted and data was made relative to measurement data at pH 11.0. Citrate buffer was used for pH 3.0–6.0 (diamonds), acetate buffer for pH 4.0–6.0 (square), sodium phosphate buffer from pH 6.0–8.0 (down-pointing triangles), Tris-HCl buffer for pH 7.0–9.0 (up-pointing triangles, dotted line between points) and glycine buffer (circles) for pH 9.0–11.0. Error bars represent deviation between 2 replicas in 1 representative experiment. The pI of the ISP is estimated to approximately 4.4 (vertical dotted line). B, Activity of 5 μ M Asn3-ISP was monitored in the FITC-casein assay across a temperature range of 25–70 $^{\circ}$ C. Background was subtracted and made relative to the measured data at 45 $^{\circ}$ C. CaCl₂ was added immediately before assaying. The assay took place for 1 h at the respective temperatures. Error bars represent deviation between data points from 3 independent experiments. The horizontal dotted line represents the highest background measurement

TABLE 3 Thermal denaturation measured by DSC^a

Treatment	ΔH_{cal}	T_{max}	ΔS	ΔH_{vH1}	T_{m1}	ΔH_{vH2}	T_{m2}	ΔD
None	120.1	60.7	0.356	105.0	57.2	209.1	61.0	0.49
2 mM CaCl ₂	145.1	62.4	0.421	142.0	60.1	266.0	62.8	0.70
EDTA	125.0	61.1	0.345	111.7	58.0	215.2	61.4	0.39

^a ΔH_{cal} (calorimetric enthalpy), ΔS (entropy of unfolding) and T_{max} are calculated directly from the unfolding transition. ΔH_{vH} and T_m are derived from fitting 2 two-stated scaled models to each transition after subtraction of buffer scans and a sigmoidal baseline

maturation. Despite repeated efforts, MS and N-terminal sequencing of various protein species isolated from SDS-PAGE gels did not reveal other processing than the removal of the 2 first residues (Supporting Information Figure S3 and Table S4). As an alternative approach, we designed various constructs where the N-terminal region of the *Planococcus* sp. AW02J18 ISP was truncated (Figure 6A). To design a close mimic of the N-terminus of native and processed enzyme, a p12-based construct was chosen (ISP-His, 38 kDa). This mimicked the full-length ISP sequence and respective truncation mutants with C-terminal His-tags albeit with 2 artificial residues at the N-terminus of recombinant enzyme (MS, Figure 6A). A Leu6 truncation construct was designed to remove the first 5 residues, not affecting the LIPY-sequence, to assay potential detrimental effects of removal of the β 1-strand of the antiparallel β -sheet required for structural stability (Figure 6B). An Arg10 truncation construct (that is, starting at Arg10) was designed to remove the LIPY-sequence from the native N-terminus, to release auto-inhibition induced by the motif. The Thr15-Arg20 truncations were designed to truncate the pro-peptide in search for an active enzyme that would mimic the processed *B. clausii* ISP. Truncations beyond Arg20 were considered to be destructive as these were anticipated to interfere with secondary structure elements in the core of the catalytic domain according to the *B. clausii* ISP structures.^{5,6} Positions of ISP truncations are summarized in Figure 6. None of the truncations were expected to impair the high affinity metal-binding site or dimerization, as previous reports have identified the

binding site and the dimer interface in other distant regions of the protein.²⁷ According to SDS-PAGE analysis recombinant enzymes were either not obtained or below our detection limits (Supporting Information Figure S6). Growth of *E. coli* was not affected by recombinant expression, suggesting that active enzymes, if present, were not lost due to cell death. In case the recombinant enzymes were present at undetectable levels, the truncated enzymes were assessed in an activity assay, but found not to present activity (Figure 6C).

The LIPY/F-motif (residues 6–9 in *Planococcus* sp. AW02J18 ISP) is conserved in pro-peptides of ISPs (Supporting Information Figure S1). In *B. clausii* ISP the LIPY-sequence is involved in binding the hydrophobic pocket at the active site, wherein Pro holds a critical position in displacing the scissile bond between Ile and Pro out of reach of the active site serine.⁶ According to structural data on *Planococcus* sp. AW02J18 ISP (Figure 5B) and *B. clausii* ISP⁶ the LIPY-sequence is involved in binding the active site, potentially having critical roles in inhibiting auto-proteolysis or cleavage of exogenous peptides. To investigate whether the LIPY/F-motif is required for inhibition, we designed point mutations in the motif by targeting the side chains of Leu6 and Ile7, which are protruding into the hydrophobic pocket. We designed Ala and Lys mutations at both sites and a double alanine mutant (substituting both positions with Ala). According to SDS-PAGE analysis, the Leu6Ala, and both Ile single mutants were successfully expressed, but gave lower yields than wild-type ISP (Figure 6D). Expression levels for the Leu6Lys single mutant and the double mutant were low, if any, and

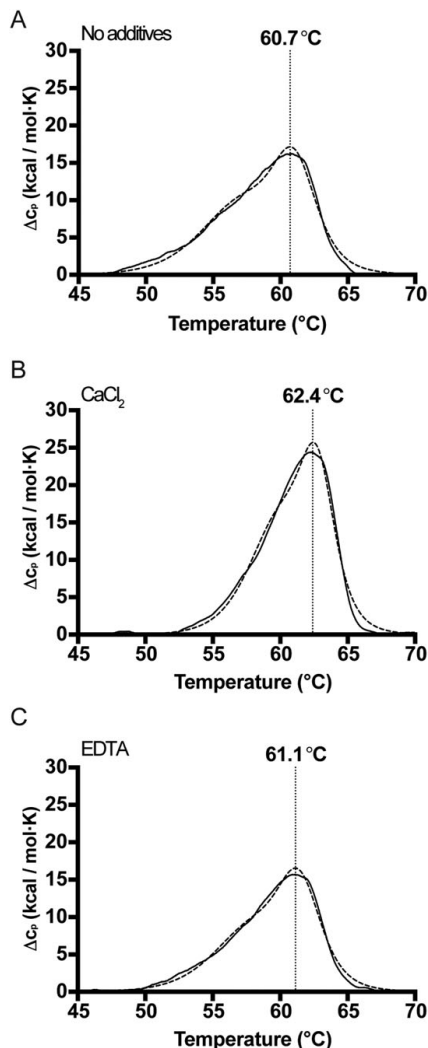


FIGURE 4 Thermal unfolding transitions of Asn3-ISP. Unfolding was measured in metal-depleted ISP in 3 conditions; A, without additives, B, in presence of 2 mM CaCl_2 , and C, in presence of 1 mM EDTA. Representative thermograms are shown after subtraction of buffer scans and fitting of a sigmoidal baseline (solid lines). The sum of the 2 two-state models fitted to each thermogram is shown with dashed lines. The T_m of the higher temperature transition is indicated with a vertical dotted drop-line for comparison

variation occurred in independent experiments. The ratio of soluble protein to expressed protein was generally higher for the mutants than for wild-type ISP (data not shown). Cleared lysates containing the wild-type ISP and mutants were assessed in an *in vitro* BODIPY-casein assay and compared to extracts from strains carrying the empty vector (Figure 6E). As expected, the wild-type ISP was found to be active upon calcium treatment as determined from an increase in fluorescent signal. Upon calcium addition, the Leu6Ala, and both Ile mutants showed a

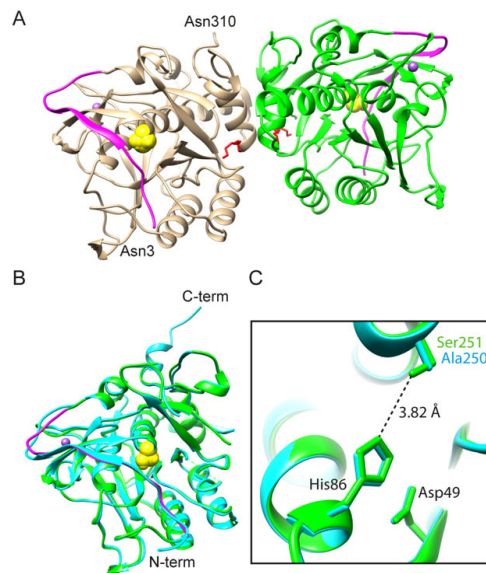


FIGURE 5 Structure of *Planococcus* sp. AW02J18 ISP. A, Dimer presented in ribbon. The pro-peptide (residues 3-20) is shown in magenta in both monomers (chain A in tan, chain B in green). The N-terminal and C-terminal residues are labelled in monomer A. The catalytic Ser251 is shown as a yellow sphere. The Peg molecules in the dimer interface are shown in red. B, Superposition of the ISP from *Planococcus* sp. AW02J18 (green, chain B) on the ISP template from *B. clausii* (blue, PDB ID: 2X8J, chain A). Ser251 in ISP from *Planococcus* sp. AW02J18 is shown as a yellow sphere, and its pro-peptide (residues 3-20) is shown in magenta. C, The catalytic triad of *Planococcus* sp. AW02J18 ISP (green, chain B) and the catalytic mutant of *B. clausii* ISP (cyan, chain A). Distance (Å) between Ser251 and His86 in *Planococcus* sp. AW02J18 ISP is given as a dashed line

similar response, but mutants showed a higher than baseline level of activity even in the absence of calcium. No activity was detected for the Leu6Lys mutant, probably because it was not expressed. The double mutant was however found to be active, despite the low expression levels. The activity of the double mutant was similar both in absence and presence of calcium, albeit low. In all cases, EDTA prevented activity, likely by chelating calcium at 1 or several binding sites.

4 | DISCUSSION

An ISP from *Planococcus* sp. AW02J18 is herein characterized in terms of its catalytic activity, stability and structure. For recombinant expression, we explored the utility of N-terminal His, His-SUMO, or His-MBP fusion tags to promote soluble expression of ISP, as previous data have shown that N-terminal tags can be used for both intracellular¹ and extracellular serine proteases.³² Expression trials showed that all fusion constructs were soluble (Figure 1). The ISP was active in the presence of calcium (Figure 2). The assumption that ISP requires pro-peptide processing for activation, for example, as in *B. clausii* ISP, allowed exploitation of its native protease activity for intrinsic tag

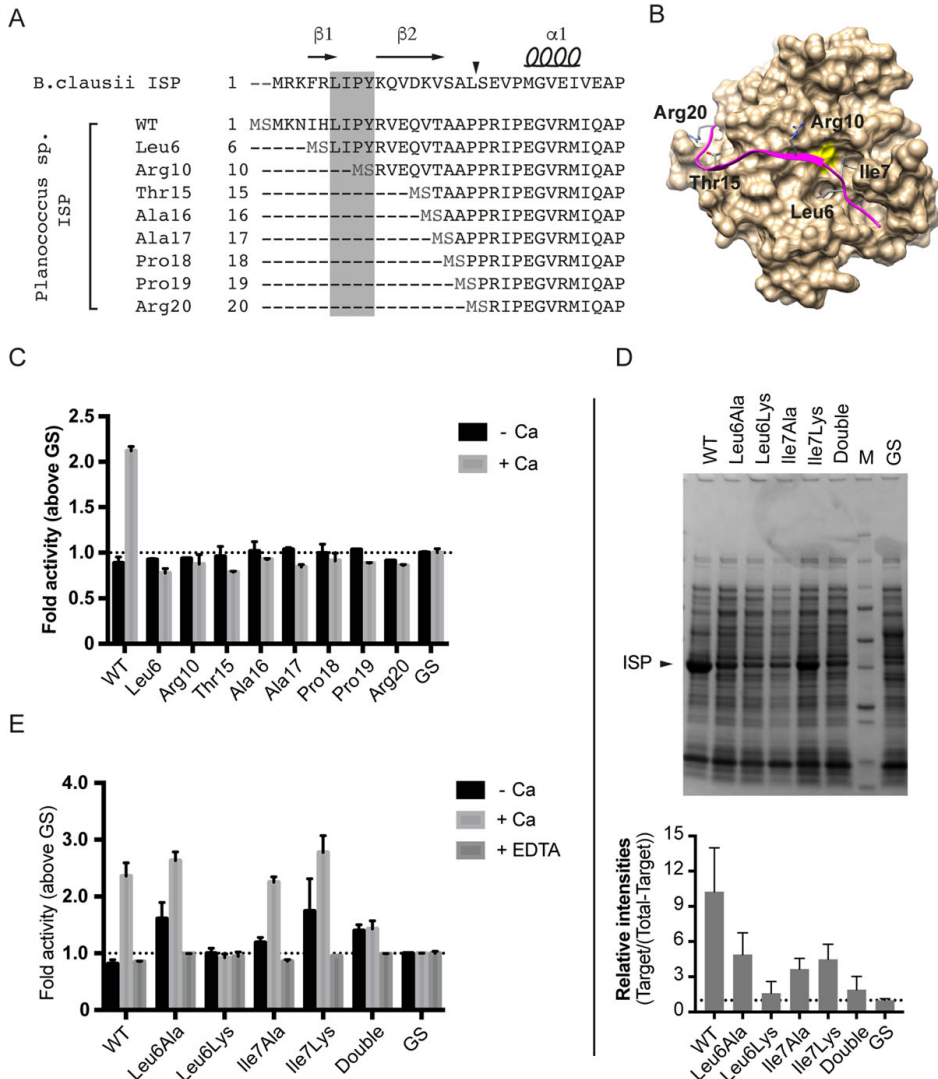


FIGURE 6 Engineering ISP at the N-termini. **A**, Alignment of the N-terminal region of ISP and the various truncated versions (indicated by starting residue given in 3-letter ambiguity codes and their sequential numbers). A grey box indicates the LIPY/F-motif. Beta-strands (β), alpha helix (α), and arrow that points to the site of maturation refers to information from *B. clausii* ISP (PDB ID: 2X8J). Residues in light grey (MS) are added to the recombinant enzymes. **B**, Solvent-accessible surface of *Planococcus sp.* AW02J18 ISP is shown (tan, chain A) with the catalytic serine residue in yellow. The pro-peptide (residues 3-20) is shown in magenta, with the Leu6 and Ile7 residues of the LIPY/F motif colored by atom. Additionally, defining residues used in the truncation experiment are indicated. **C**, Cleared lysates containing wild-type ISP-His or truncated versions (p12 mutated to Ala), were screened for activity against BODIPY-FL-casein in the absence and presence of 1 mM CaCl_2 (+Ca) for 1 h at 37 °C. Fluorescence was normalized to optical density of expression cultures, to account for any growth effects. Lysates with empty vectors (GS) was used as background, and samples were calculated as fold above control. Error bars represent standard deviation between parallels in 2 experiments. **D**, A representative SDS-PAGE analysis of cleared lysates containing wild-type ISP-His (WT) or mutant versions (double, both Leu6 and Ile7 mutated to Ala), M, BioRad's Precision Plus Protein™ Dual Color Standard; and GS, extracts with empty vector. Arrow points to the recombinant ISP variants. The lower panel shows intensities of target ISP proteins relative to the target corrected total lane intensity. Intensity data arise from 2 independent experiments. **E**, Cleared lysates from **D** analyzed as in **C**, in absence or presence of 1 mM CaCl_2 (Ca) or with 1 mM EDTA

removal. Indeed, the construct with an N-terminal His-tag facilitated creation of a processed ISP without artificial tags in the presence of calcium (Figure 2 and Supporting Information Figure S3).

The ISP operates at moderate temperatures, with optimal conditions at 45 °C (Figure 3), and unfolds at about 60 °C (Figure 4). The organism of which this ISP originates, *Planococcus sp.* AW02J18, was

isolated from a marine habitat, and is known to thrive at cold to moderate temperatures (data not shown). Although some ISPs are active at neutral pH,⁷ *Planococcus* sp. AW02J18 ISP, like the majority of ISPs,^{2,44–46} has optimal activity at alkaline pH (Figure 3). So far, 1 ISP has been structurally characterized, namely the ISP from *B. clausii*. This study provides structural information on a second unique ISP that originates from a phylogenetically and physiologically distinct genus.⁴⁷ The ISP crystallized mostly at acidic pH (Supporting Information Table S3), and calcium was not found in any of the crystals. The lack of activity and low processing below pH 7.0 (Figures 2 and 3) may partly explain why structures are in the inactive conformation. Whether lack of crystals at conditions above pH 7.0 is caused by degradation or because the active state does not promote crystal growth is impossible to say. Processing is not induced by pH shift alone (Figure 2D), but requires calcium. Both ISPs were found to crystallize in a dimeric state; thus, dimerization appears to be a generic feature of ISPs. According to size exclusion chromatography, the presence of calcium the Asn3-ISP lead to a mixed population of quaternary structures corresponding to approximately 2.5 and 3.7 monomers per oligomer (Supporting Information Figure S5). Whereas the dimeric form is confirmed in the crystal, the oligomeric state in solution was inconclusive. It appeared however that the presence of calcium induced 2 new states compared to the calcium-depleted enzyme solutions. In accordance with earlier observations, calcium depletion may lead to a more compact structure (here represented by the shift from 3.7 to 3.3 monomers per oligomer state). The presence of calcium may induce autoproteolysis, thereby reducing the apparent molecular weight (here represented by the shift from 3.3 to 2.5 monomers per oligomers). The higher molecular weight induced by the presence of calcium (here represented by the 3.7 monomer per oligomer state) may arise from a less compact structure or even from aggregation. The 2 monomers contained regions of poor electron density in proximity to each other. These are most likely partially flexible regions as a consequence of the structural reorganization caused by the insertion of the pro-peptide in the substrate-binding region. The C-terminal 20 residues were not defined in electron density, while in 2 different crystal forms representing structures of ISP from *B. clausii* (PDB ID: 2X8J and 2WWT), these residues are stabilized through interactions with symmetry mates. According to sequence alignments, the C-terminal region is not conserved (Supporting Information Figure S1), but the reason for this region being flexible in the structure of *Planococcus* sp. AW02J18 ISP is not clear. Ultimately, the requirement and role of the C-terminal residues in folding and dimerization of ISPs remains unclear.

From studies of *B. clausii* ISP, divalent metal ions, possibly calcium, bind close to the S1 pocket.^{5,6} In the crystals of *Planococcus* sp. AW02J18 ISP, calcium was not identified at any of the metal binding site. Two loop regions were not defined in the electron density of ISP, which is also the case for the *B. clausii* ISPs containing the intact pro-peptide (PDB IDs: 2X8J and 2WWT). These loop regions are however ordered in the *B. clausii* ISP structure that simulates the active conformation of the enzyme, albeit with a catalytic mutation (PDB ID: 2XRM). Residues from both loops contribute to the coordination of a calcium ion, and these residues are conserved in aligned sequences (Supporting Information Figure S1). This could indicate a specific role of calcium in the transition from inactive to active enzyme, not only

for the *B. clausii* ISP, but also for other ISPs. Asn3-ISP from *Planococcus* sp. AW02J18 was active in presence of calcium, but susceptible to self-degradation (Figure 2). The fact that ISPs were not active without exogenous addition of calcium suggests that available metal binding sites were not occupied after production. Due to conservation of calcium-coordinating residues (Supporting Information Figure S1), and the need for high EDTA concentrations to inhibit activity (Figure 2C), low affinity for calcium is likely not the case. DSC results suggest that additional calcium is only slightly stabilizing, and tightly bound calcium (removable with EDTA) is not essential for overall stability (Figure 4). DSC showed however that calcium does have a minor stabilizing effect; thus suggesting that the added calcium in our assays contribute to minor structural rearrangements.

It is likely that there are structural rearrangements, such as pro-peptide flip-out or removal, in order for the 2 loops to order and coordinate calcium. The IP residues of the LIPY/F motif in the pro-peptide are spatially close to residues in 1 of the loops that need to be re-oriented upon calcium binding. The 2 residues form hydrophobic interactions to the side-chain of Phe195 in our inactive structure and probably hinder this reorienting into the active conformation (this side-chain appears to be shifted almost 15 Å in the active state).

It is likely that the pro-peptide in the ISP from *Planococcus* sp. AW02J18 is removed, in analogy to several *Bacillus* ISPs.^{5,7} From the available structures of ISPs with intact pro-peptides (PDB ID: 6F9M, 2X8J, 2WWT, 2WVT, whereof 2 are shown in Figure 5B) and the sequence alignment (Supporting Information Figure S1), we found that 2 short beta-strands in the pro-peptide are likely structurally conserved. The secondary structure elements are stabilized by main chain interactions, which are sequence independent. A unique feature of the *Planococcus* sp. AW02J18 ISP that is not found in homologous ISPs is the presence of the 2 consecutive proline residues in the transition between the pro-peptide and the catalytic domain (Supporting Information Figure S1). The removal of the ISP pro-peptide in *Planococcus* sp. AW02J18 appears to be different, and possibly involves several steps (Figure 2D). In the first step the 2 first residues of the ISP (Met1, Lys2) are removed (protein band numbered 2, Figure 2D), as identified in the crystal and by N-terminal sequencing. Another product, which appears as the main product (around 30–35 kDa) at pH 8.5 in presence of 10 mM CaCl₂ (protein band numbered 3, Figure 2D), could possibly be functional. This product could in principle arise from processing of the C-terminal region of the protein, too, albeit not identified in the crystal or MS analyses (Supporting Information Figure S5 and Table S4). The N-terminal residues of this protein could not be identified. Unfortunately, MS analyses did not reveal obvious processing patterns at the N-terminal in the protein species from the SDS-PAGE analysis (Supporting Information Figure S5 and Table S4). This may partly be due to a lack of sequential degradation. Ultimately, we could not determine which ISP moiety that is responsible for or contribute to the activity identified in assays. A truncation experiment was conducted to trim the pro-peptide in the hunt for processing site(s). Two artificial residues (Met-Ser) are unavoidably added to the N-terminal end of these truncation constructs, which arise from fusion of the *isp* gene fragment to the start codon and the ligation seam added during sub-cloning (Figure 6A), and their negative interference on protein stability cannot be ruled out. Sequence analysis of *Planococcus*

sp. AW02J18 ISP, reveals that it contains 2 prolines in the transition from the pro-peptide to the catalytic domain (Supporting Information Figure S1). Whereas Pro at the P2 site is likely accepted, Pro at the P1 is highly unlikely due to the preference of hydrophobic residues at the S1 site.²⁷ Multiple prolines are normally not present in sites for autoproteolysis by serine proteases,⁴⁸ and the prolines may instead serve a structural role.⁴⁹ This does not however rule out that other proteases, for example proline-specific endopeptidases, could process and remove the pro-peptide in native conditions, or that processing site(s) are in other regions that were not included in this study.

Although it has been found that the pro-peptide of *B. clausii* ISP has a role in inhibition, the contribution of the conserved residues within the LIPY/F-motif has not been studied in detail. Due to the fact that Leu and Ile are conserved in the motif, and that the ISPs likely prefer hydrophobic amino acids at the S2 and S4 sites,²⁷ we studied point mutations of Leu6 and Ile7 in *Planococcus* sp. AW02J18 ISP. Data from 3 of the 4 single point mutations, which resulted in increased activity - even in the absence of excessive calcium, indicate that Leu6 and Ile7 have substantial roles in inhibition and support the involvement of calcium during activation. A closer inspection of the structural context suggests that substitution of Leu6 with Ala likely reduced the hydrophobic interaction to the active site, and thus relieves the inhibition (Figure 7). The substitution to lysine however seems to both reduce expression level (Figure 6D). It furthermore does not respond on

calcium addition in the activity assay (Figure 6E). Assuming that the mutant is properly folded, inhibition could be explained by the possibility that lysine can form hydrogen bond and/or salt bridge interactions with the catalytic Asp49 and the Asn84 residues, respectively (Figure 7). Structural explanations for the Ile7 mutants were not conclusive due to their proximity to the flexible region (183-193), but it is likely that both mutations cause reduced interactions with the pro-peptide. We thus conclude that the pro-peptide, with the LIPY/F motif in a central position, is involved in inhibition. Our data is in line with the proposed ISP model,²⁷ suggesting that calcium binding at the active site is prevented during pro-peptide inhibition.

AUTHOR CONTRIBUTIONS

G.E.K.B. designed the study, designed and supervised experiments and drafted the manuscript, Ø.L. carried out expression, purification, and biochemical assays, H.A. set up crystallization trials and performed the DSC experiment, I.L. performed data collection, processed data, determined the structure and refined it, A.G.M. performed phylogenetic analysis, A.W. supervised and designed the DSC experiment. P.P. supervised mutant design and performed bioinformatic analyses. All authors were involved in revision of the manuscript and approved the final version.

ACKNOWLEDGMENTS

We would like to thank Arne O. Smalås for sharing sequence data, Stefan Hauglid and Trine Carlsen for expert assistance on crystallization and support of staff during visit, and Hilde Eide Lien for excellent work on truncation and mutant studies. Provision of beam time at the European Synchrotron Radiation Facility (ESRF) is highly valued. The authors thank the Research Council of Norway for financial support (221568). Authors declare no conflict of interest.

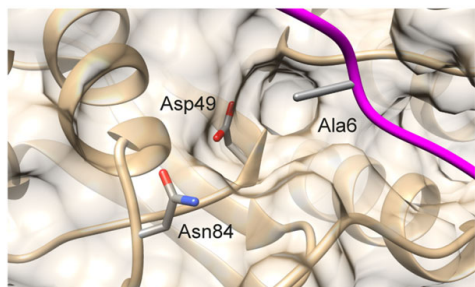
ORCID

Gro E. K. Bjerga  <http://orcid.org/0000-0001-5152-5139>

REFERENCES

- Lee AY, Goo Park S, Kho CW, et al. Identification of the degradome of Isp-1, a major intracellular serine protease of *Bacillus subtilis*, by two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionization-time of flight analysis. *Proteomics*. 2004;4(11):3437-3445.
- Sheehan SM, Switzer RL. Intracellular serine protease 1 of *Bacillus subtilis* is formed in vivo as an unprocessed, active protease in stationary cells. *J Bacteriol*. 1990;172(1):473-476.
- Nishino T, Shimizu Y, Fukuhara K, Murao S. Isolation and characterization of a proteinaceous protease inhibitor from *Bacillus subtilis*. *Agric Biol Chem*. 1986;50(12):3059-3064.
- Rigden DJ, Xu Q, Chang Y, Eberhardt RY, Finn RD, Rawlings ND. The first structure in a family of peptidase inhibitors reveals an unusual Ig-like fold. *F1000Research*. 2013;2:154.
- Gamble M, Künze G, Dodson EJ, Wilson KS, Jones DD. Regulation of an intracellular subtilisin protease activity by a short propeptide sequence through an original combined dual mechanism. *Proc Natl Acad Sci U S A*. 2011;108(9):3536-3541.

L6A



L6K

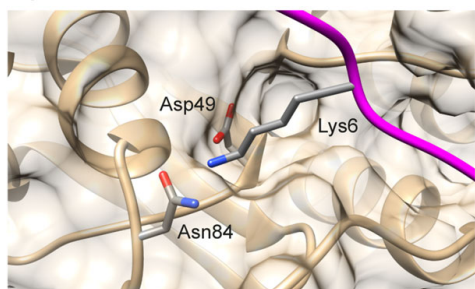


FIGURE 7 Mutation of Leu6 in the LIPY/F motif. Solvent-accessible surface of *Planococcus* sp. AW02J18 ISP (B-chain) is shown in opaque tan. The backbone and secondary elements is represented in ribbons, except for the pro-peptide which is shown in magenta. The side chains of mutants (Ala6, Lys6), Asp49 and Asn84 are colored by atomic elements

6. Vévodová J, Gamble M, Künze G, et al. Crystal structure of an intracellular subtilisin reveals novel structural features unique to this subtilisin family. *Structure*. 2010;18(6):744-755.
7. Jeong YJ, Baek SC, Kim H. Cloning and characterization of a novel intracellular serine protease (IspK) from *Bacillus megaterium* with a potential additive for detergents. *Int J Biol Macromol*. 2017;108:808-816.
8. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Res*. 2010;38(suppl_1):D211-D222.
9. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res*. 2014;42(D1):D503-D509.
10. Hedstrom L. Serine protease mechanism and specificity. *Chem Rev*. 2002;102(12):4501-4524.
11. Wong SL, Doi RH. Determination of the signal peptidase cleavage site in the preprosubtilisin of *Bacillus subtilis*. *J Biol Chem*. 1986;261(22):10176-10181.
12. Wells JA, Ferrari E, Henner DJ, Estell DA, Chen EY. Cloning, sequencing, and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillus subtilis*. *Nucleic Acids Res*. 1983;11(22):7911-7925.
13. Vasantha N, Thompson LD, Rhodes C, Banner C, Nagle J, Filpula D. Genes for alkaline protease and neutral protease from *Bacillus amyloliquefaciens* contain a large open reading frame between the regions coding for signal sequence and mature protein. *J Bacteriol*. 1984;159:811-819.
14. Bryan PN. Prodomains and protein folding catalysis. *Chem Rev*. 2002;102(12):4805-4816.
15. Power SD, Adams RM, Wells JA. Secretion and autoproteolytic maturation of subtilisin. *Proc Natl Acad Sci U S A*. 1986;83(10):3096-3100.
16. Ohta Y, Hojo H, Aimoto S, et al. Pro-peptide as an intramolecular chaperone: renaturation of denatured subtilisin E with a synthetic pro-peptide [corrected]. *Mol Microbiol*. 1991; 5(6):1507-1510.[PMC] [1686294][insertedFromOnline]
17. Ikemura H, Takagi H, Inouye M. Requirement of pro-sequence for the production of active subtilisin E in *Escherichia coli*. *J Biol Chem*. 1987;262(16):7859-7864.
18. Zhu XL, Ohta Y, Jordan F, Inouye M. Pro-sequence of subtilisin can guide the refolding of denatured subtilisin in an intermolecular process. *Nature*. 1989;339(6224):483-484.
19. Wright CS, Alden RA, Kraut J. Structure of subtilisin BPN' at 2.5 angstrom resolution. *Nature*. 1969;221(5177):235-242.
20. Neidhart DJ, Petsko GA. The refined crystal structure of subtilisin Carlsberg at 2.5 Å resolution. *Protein Eng*. 1988;2(4):271-276.
21. Betzel C, Klupsch S, Papendorf G, Hastrup S, Branner S, Wilson KS. Crystal structure of the alkaline proteinase Savinase from *Bacillus lentus* at 1.4 Å resolution. *J Mol Biol*. 1992;223(2):427-445.
22. Wells JA, Estell DA. Subtilisin: an enzyme designed to be engineered. *Trends Biochem Sci*. 1988;13(8):291-297.
23. Schechter I, Berger A. On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem Biophys Res Commun*. 1968;32(5):898-902.
24. Bode W, Papamokos E, Musil D. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech *Hirudo medicinalis*. Structural analysis, subtilisin structure and interface geometry. *Eur J Biochem*. 1987;166(3):673-692.
25. Bryan PN, Rollence ML, Pantoliano MW, et al. Proteases of enhanced stability: characterization of a thermostable variant of subtilisin. *Proteins*. 1986;1(4):326-334.
26. Pantoliano MW, Whitlow M, Wood JF, et al. The engineering of binding affinity at metal ion binding sites for the stabilization of proteins: subtilisin as a test case. *Biochemistry*. 1988;27(22):8311-8317.
27. Gamble M, Künze G, Brancala A, Wilson KS, Jones DD. The role of substrate specificity and metal binding in defining the activity and structure of an intracellular subtilisin. *FEBS Open Bio*. 2012;2:209-215.
28. De Santi C, Altermark B, de Pascale D, Willasson N-P. Bioprospecting around Arctic islands: marine bacteria as rich source of biocatalysts. *J Basic Microbiol*. 2016;56(3):238-253.
29. The UPProt Consortium UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45(D1):D158-D169.
30. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-780.
31. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188-1190.
32. Bjerga GEK, Arsin H, Larsen Ø, Puntervoll P, Kleivdal HT. A rapid solubility-optimized screening procedure for recombinant subtilisins in *E. coli*. *J Biotechnol*. 2016;222:38-46.
33. Geertsma ER, Dutzler R. A versatile and efficient high-throughput cloning tool for structural biology. *Biochemistry*. 2011;50(15):3272-3278.
34. Twining SS. Fluorescein isothiocyanate-labeled casein assay for proteolytic enzymes. *Anal Biochem*. 1984;143(1):30-34.
35. Kabsch W. XDS. *Acta Crystallogr D Biol Crystallogr*. 2010;66(Pt 2):125-132.
36. Winn MD, Ballard CC, Cowtan KD, et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*. 2011;67(Pt 4):235-242.
37. Vonrhein C, Flensburg C, Keller P, et al. Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr D Biol Crystallogr*. 2011;67(Pt 4):293-302.
38. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr*. 1997;53(Pt 3):240-255.
39. Cowtan K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr*. 2006;62(Pt 9):1002-1011.
40. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. 2004;60(Pt 12 Pt 1):2126-2132.
41. Adams PD, Afonine PV, Bunkóczi G, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*. 2010;66(Pt 2):213-221.
42. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera: a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612.
43. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*. 2011;8(10):785-786.
44. Yamagata Y, Ichishima E. A new alkaline serine protease from alkalophilic *Bacillus* sp.: cloning, sequencing, and characterization of an intracellular protease. *Curr Microbiol*. 1995;30(6):357-366.
45. Kirimura J, Shimizu A, Kimizuka A, Ninomiya T, Katsuya N. Contribution of peptides and amino acids to the taste of foods. *J Agric Food Chem*. 1969;17(4):689-695.
46. An S-Y, Ok M, Kim J-Y, et al. Cloning, high-level expression and enzymatic properties of an intracellular serine protease from *Bacillus* sp. WRD-2. *Indian J Biochem Biophys*. 2004;41(4):141-147.
47. Yoon J-H, Kang S-J, Lee S-Y, Oh K-H, Oh T-K. *Planococcus salinarum* sp. nov., isolated from a marine solar saltern, and emended description of the genus *Planococcus*. *Int J Syst Evol Microbiol*. 2010;60(Pt 4):754-758.
48. Kim JC, Cha SH, Jeong ST, Oh SK, Byun SM. Molecular cloning and nucleotide sequence of *Streptomyces griseus* trypsin gene. *Biochem Biophys Res Commun*. 1991;181(2):707-713.
49. Vanhoof G, Goossens F, De Meester I, Hendriks D, Scharpé S. Proline motifs in peptides and their biological processing. *FASEB J*. 1995;9(9):736-744.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Bjerga GEK, Larsen Ø, Arsin H, et al. Mutational analysis of the pro-peptide of a marine intracellular subtilisin protease supports its role in inhibition. *Proteins*. 2018;1-13. <https://doi.org/10.1002/prot.25528>



Graphic design: Communication Division, UIB / Print: Skjipes Kommunikasjon AS



uib.no

ISBN: 9788230845943 (print)
9788230864074 (PDF)