FACULTY OF
BIOSCIENCE ENGINEERING

A DUAL COMPUTATIONAL APPROACH TO MAP DOMAIN ARCHITECTURE IN PHAGE LYTIC PROTEINS

word count: 25.385

Steff Taelman
Student ID: 01406602

Promotor: Dr. ir. Michiel Stock
Co-promotors: Prof. dr. ir. Yves Briers and prof. dr. ir. Wim Van Criekinge

Tutors: Ir. Bjorn Criel and dr. ir. Michiel Stock

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of master in Bioscience Engineering: Bioinformatics.
Academic year: 2018 - 2019

GHENT
UNIVERSITY

Gent, June 14, 2019

The promotors,	The author,
Dr. ir. Michiel Stock, Prof. dr. ir. Yves Briers and prof. dr. ir. Wim Van Criekinge	Steff Taelman

PREFACE

It has been quite a journey. I initially chose Bioscience Engineering because of my chemistry and biology teachers in my last year of Secondary education. They both had a certain type of enthusiasm in their voice when talking about their fields. Five years later, I must say the vast majority of people I've met in this field have all had that same characteristic. I would like to dedicate this page to a few of them.

First of all, I would like to thank my promotors: dr. ir. Michiel Stock, prof. dr. ir. Yves Briers and prof. dr. ir. Wim Van Criekinge. You are quite diverse in your fields of expertise, but this offered plenty of perspectives and approaches and has ultimately made this a better dissertation. Wim, you are involved in a lot of interesting projects, so your signing off on this one really speaks to its potential and is part of the reason I chose this subject. Yves, you are really an authority in the field of phage lytic proteins and enzybiotics and your genuine interest in my results has been a big inspiration throughout the year. Michiel, you are a master of all trades in biology and bioinformatics, I have learned a lot from you on which strategies will work in different situations and which won't. It is really encouraging to see what triggers your mind and to see where ideas can go from there.

I would also like to thank my tutor: ir. Bjorn Criel. Not only did you create the database that has made almost all of my analyses possible, you spent a lot of time optimising and updating it through the year for the sake of these analyses. I would also like to thank you and Michiel for the input during our weekly sessions and the care taken in providing me with feedback.

Lastly I would like to thank my family and friends for their support and for reading this after I asked nicely.

CONTENTS

ABSTRACT

Antibiotic resistance is steadily turning into a global crisis. Recent efforts have put forward certain enzymes encoded by bacteriophages as a promising alternative to traditional antibiotics. There is a certain number of protein domains that has been identified in in these proteins, but not every possible combination of domains has been found in nature. This phenomenon has lead to the hypothesis of an underlying set of design rules on the basis of which functional phage lytic proteins are formed. PhaLP is a database constructed from various sources (UniProt, InterPro etc.) containing information on lytic proteins used during a bacteriophage's lytic life cycle. Quantitative analyses based on annotated protein domains in PhaLP show clear correlations between protein architecture and the bacterial host of the phage encoding them. This further substantiates the design rule hypothesis. A thorough understanding of these rules could facilitate the design of new effective enzyme-based antibiotics or enzybiotics.. A dual computational approach is employed to get an outline of these design rules. First of all, a cluster analysis is performed based on the pairwise similarity of the protein sequences. This points out broad host-ranges for clusters of similar proteins, which can be useful characteristic for an enzybiotic. Furthermore, it demonstrates evolutionary relations between sequences. A second approach uses several interpretable machine learning models to predict a host from sequence data and subsequently extracts the elements that are deemed important for the model's prediction. This approach is applied on every level of the bacterial taxonomy to map a narrowing path of design rules regarding host taxonomy.

Keywords: enzybiotics, phage lytic proteins, protein domains, sequence clustering, interpretable machine learning

SAMENVATTING

Antibioticaresistentie wordt langzaamaan een wereldwijde crisis. Recent zijn bepaalde enzymen, die gecodeerd worden door bacteriofagen, naar voor geschoven als veelbelovende alternatieven voor traditionele antibiotica. Een bepaald aantal eiwitdomeinen is reeds gevonden in deze enzymen, maar niet elke mogelijke combinatie daarvan wordt gevonden in de natuur. Door dit fenomeen is een hypothese onstaan van een onderliggende set van designregels op basis waarvan functionele lytische faageiwitten worden gevormd. PhaLP is een database met informatie over lytische eiwitten die worden gebruikt tijdens de lytische levenscyclus van bacteriofagen. De lytische faageiwitten in PhaLP werden onderworpen aan kwantitatieve analyses gebaseerd op geannoteerde eiwitdomeinen. Hieruit blijkt een duidelijke correlatie tussen de domeinarchitectuur van lytische eiwitten en de bacteriële gastheer van de faag die ze coderen. Deze correlatie staaft verder de hypothese van de onderliggende designregels. Een meer uitgebreide kennis van deze designregels is veelbelovend om het ontwerpen van nieuwe enzym-gebaseerde antibiotica of enzybiotica te vergemakkelijken. Een tweedelige computationele aanpak wordt gebruikt om deze regels in kaart te brengen. Ten eerste wordt een clusteranalyse uitgevoerd op basis van de paarsgewijze similariteit van de eiwitsequenties. Hieruit blijkt het brede gastheerspectrum van sommige clusters van gelijkaardige sequenties, wat een nuttig kenmerk kan zijn voor een enzybioticum. Hiernaast duidt het ook op evolutionaire relaties tussen sequenties. Een tweede aanpak maakt gebruik van interpreteerbare machine learning modellen om een gastheer te voorspellen o.b.v. sequentiegegevens en identificeert vervolgens de elementen die door het model belangrijk geacht worden in deze voorspelling. Deze aanpak wordt ook toegepast op elk niveau van de bacteriële taxonomie om een steeds specifieker wordende set van designregels met betrekking op gastheer uit te zetten.

Trefwoorden: enzybiotica, lytische faageiwitten, eiwitdomeinen, clusteranalyse, interpreteerbare machine learning

GLOSSARY

AA Amino Acid

AUC Area Under the ROC Curve

CBD Cell wall Binding Domain

CWA Cell Wall Amidase

CWG Cell Wall Glycosidase

CWP Cell Wall Peptidase

EAD Enzymatically Active Domain

FPR False Positive Rate

GlcNAc N-acetylglucosamine

GO Gene Ontology

HGT Horizontal Gene Transfer

MI Mutual Information

ML Machine Learning

MRSA Methicillin-Resistant *Staphylococcus aureus*

MSA Multiple Sequence Alignment

MurNAc N-acetylmuramic acid

PG Peptidoglycan

PGRP Peptidoglycan Recognition Particle

PhaLP Phage Lytic Protein

RF Random Forest

ROC Reciever Operating Characteristic

TPR True Positive Rate

VAPGH Virion-Associated Peptidoglycan Hydrolase

CHAPTER 1

INTRODUCTION

If no action is taken to counter or circumvent antibiotic resistance, 2.4 million people could die from infectious diseases in Europe, North America and Australia between 2015 and 2050 (OECD, 2018). Globally, 700,000 people die each year due to the recent appearance of superbugs, extensively or totally drug resistant bacteria that cause infections (nearly) untreatable by conventional antibiotics. If current trends continue, this figure is estimated to rise to around 10 million per year by 2050, a higher death rate than cancer currently has (Review on Antimicrobial Resistance, 2016). A promising alternative to traditional antibiotics can be found in bacteriophages. These are viruses that replicate within a bacterium before breaking out and infecting other bacteria. The proteins used in this mechanism are currently considered to be one of the most promising alternatives to conventional antibiotics (Czaplewski et al., 2016). This dissertation will provide an overview of the natural diversity of these enzymes and learn design rules regarding their domain architectures. This insight is essential for the development of new enzyme-based antibacterials, also called enzybiotics.

1.1 Antibiotic resistance

The discovery of antibiotics is considered one of the most important revelations of the 20th century and has drastically changed healthcare in the process. In the 75 years since their introduction, massive improvements in production have made them increasingly inexpensive, encouraging nonprescription and off-label uses (Davies and Davies, 2010). Combined with their easy use and effectiveness, it has caused the use of antibiotics to run rampant in recent years.

Antibiotic resistance is, simply put, the ability of a bacterium to successfully resist treatment with an antibiotic. It can be initiated by the introduction of one or more random mutations which provide the organism with a higher chance of survival against a certain antibiotic. As the only bacteria that will survive treatment are ones that have acquired the mutation(s), only these can reproduce. Consequently, the genetic

makeup of the species will drift towards one that is more resistant to the treatment (Normark and Normark, 2002). Each time a patient is put onto a regimen of antibiotics, any surviving organisms of the illness push towards a more resistant isolate. This is why doctors always ask to complete the full treatment, even if symptoms fade early.
The most commonly used antibiotics tend to be broad-spectrum, meaning they affect a wide range of bacteria. Although this makes for a quick cure of the illness without having to identify the pathogenic bacterium, it provokes resistance in several species at once (Carlton, 1999). In 2010, over 30% of the prescriptions for antibiotics in ambulatory care were found to be unnecessary (Fleming-Dutra et al., 2016). This kind of over- and misuse of antibiotics unnecessarily exposes bacteria to antibiotics, and thereby increases the overall chance of resistance development.

Resistance development through the acquisition of random mutations is a complex process with many variables that can drastically impact the rate at which it evolves (Martinez and Baquero, 2000), but it is hardly the only way antibiotic resistance can be obtained. Resistance developed by the acquisition of random mutations can spread to other bacteria through several mechanisms of horizontal acquisition of resistance genes. This Horizontal Gene Transfer (HGT) can even occur across species boundaries, allowing resistance genes against a broad-spectrum antibiotic to spread from a non-pathogenic bacterium to a pathogenic one. This results in a rapid spread of antibiotic resistance in a population of bacteria (Dzidic and Bedeković, 2003).

1.2 Bacteriophages

The word bacteriophage, derived from the Greek *phagein* and the word bacteria, literally translates to 'bacteria eater'. These viruses are among the smallest and most omnipresent biological entities on earth, estimated at 10^{31} entities on earth (which is 10 times more abundant than bacteria) (Hendrix, 2003). They are composed solely of a nucleotide string carrying their genetic information (i.e. DNA or RNA) surrounded by a protein shell or capsid. Bacteriophages are able to hijack a bacterium's reproduction machinery, making it their host. They latch onto the host and inject their genetic information into it. During their lytic life cycle, the viral DNA or RNA is amplified and new capsids are assembled around them. The newly made phages then escape from their host cell by lysis and spread out to target new hosts (see figure 1.1). Depending on the species and conditions, the number of phage progeny can be over 200 (Carlton, 1999).

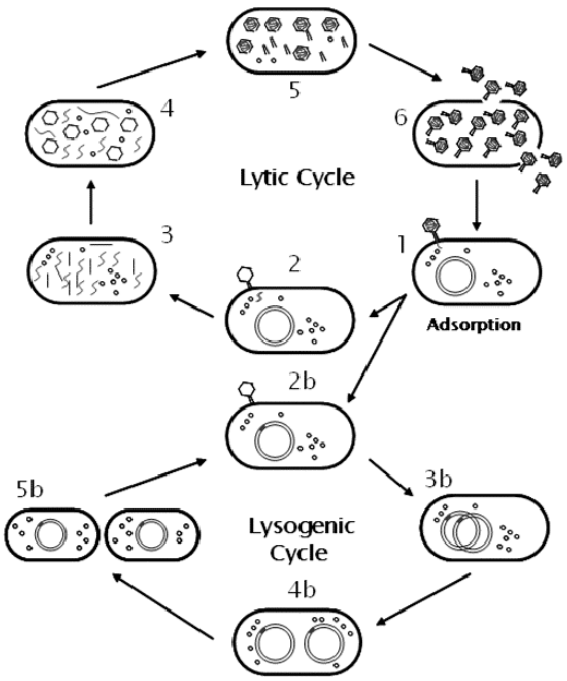


Figure 1.1: The lytic and lysogenic life cycle of a bacteriophage. Lytic cycle: the bacteriophage inserts its genetic information from inside the virus capsid into a bacterial cell (1 & 2), takes over the host's replication machinery and directs the synthesis of bacteriophage nucleic acids and proteins (3 & 4 and 5). The phage produces enzymes that break open the cell and the mature bacteriophage particles are released (6). Lysogenic cycle: infection occurs identically (1), but instead of replicating the viral DNA integrates into the bacterial genome (2b). When the bacterial cell then divides, the viral DNA is replicated along with the host cell's without expression of the lytic genes(3b and 4b). A copy of the viral DNA is transferred along with the host chromosome to the bacteria's offspring (5b). Following induction into the lytic phase, the integrated phage DNA is excised from the host cell genome, causing the lytic genes to be expressed. This brings about the start of a lytic cycle (Rees et al., 2012).

The therapeutic use of lytic bacteriophages to cure pathogenic bacterial infection is denoted as phage therapy. The fact that phages self-replicate and exponentially amplify in number may cause this method of combatting bacterial illnesses to work very fast and in small doses. They are highly specific towards hosts and are able to evolve with the bacterium (Abdelkader et al., 2019). These same characteristics, however, impose a requirement for a thorough understanding of the properties and behaviour of a certain phage for safe and controlled use (Hermoso et al., 2007). In general, a lack of a regulatory framework and standardised protocols has discouraged funding and advancements of clinical trials, causing attention to largely shift towards other phage-related antimicrobials.

1.2.1 Phage lytic proteins

More recently lytic proteins encoded by phages have been evaluated as putative antimicrobials (Hermoso et al., 2007). These enzymes are used by the phage to (i) infect

a bacterial cell and (ii) to lyse this cell after new bacteriophages have been assembled inside of it. The former are called Virion-Associated Peptidoglycan Hydrolases (VAPGHs) because they are part of the virus particle (virion), while the latter are endolysins (Rodríguez-Rubio et al., 2016). Both types of enzyme degrade peptidoglycan (PG), the main component of bacterial cell walls (see figures 1.1 and 1.3). VAPGHs only locally break down PG to form a pore through which the viral genome can be injected into the bacterial host. Endolysins, however, compromise the host's cell wall by digestion of its PG layer. At a certain point, this will cause the high internal pressure of the cytoplasm to take over, resulting in the bacterial cell exploding, i.e. lysis.

The use of phage lytic enzymes as antimicrobials, or enzybiotics, has multiple advantages over conventional antibiotics. As in phage therapy, enzybiotic treatment has high host specificity. In this case, however, the specificity is predominantly a result of the type of PG the enzyme can digest, rather than being determined by phage receptors and antiviral defence mechanisms (Abdelkader et al., 2019). This promotes a slightly broader specificity than phage therapy, as peptidoglycan type is often conserved at species level (Schleifer and Kandler, 1972), but still narrow enough to facilitate treatment of a disease without disturbing the normal flora (Fischetti, 2008). It has also been hypothesized that phage antimicrobials have lower risk concerning development of resistance since bacteriophages have co-evolved with their host bacteria to target conserved bonds in the PG layer (Rodríguez-Rubio et al., 2013). They are also capable of disrupting biofilms: matrices of multiple micro-organisms that are usually impenetrable by antibiotics (Meng et al., 2011; Sharma et al., 2018).

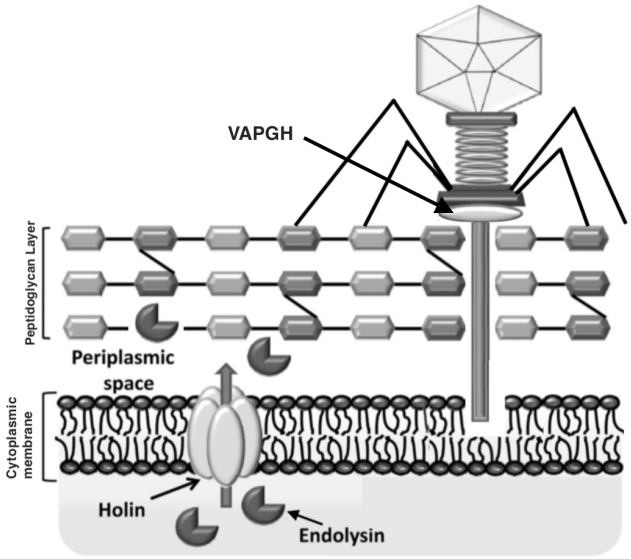


Figure 1.2: Schematic representation of the mode of action of Virion-Associated Peptidoglycan Hydrolases (VAPGHs) and endolysins on a Gram-positive bacterium. A VAPHG allows the phage to infect the bacterial cell from the outside in, while an endolysin is able to reach the peptidoglycan layer from the inside out by means of a holin (Modified from: Rodríguez-Rubio et al. (2016)).

While the use of a phage lytic enzyme as an enzybiotic rests on the same principle as their natural function, namely the degradation of PG, their mode of action is different (see figure 1.2). Naturally occurring phage lytic proteins are assisted in their journey to the site of action (the periplasmic space). VAPGHs are part of the virion that bind to the cell wall and bring the enzyme in the proximity of the PG (Rodríguez-Rubio et al., 2016). Endolysins gain access to the periplasmic space by means of holins, bacteriophage-encoded proteins that form pores through the cytoplasmic membrane. Depending on the type of phage this can cause (i) the endolysins to be able to cross the cytoplasmic membrane or (ii) the membrane itself to depolarise, provoking endolysins that have accumulated in the periplasmic space, but are still anchored in the cytoplasmic membrane, to be released and refold into an active conformation (Park et al., 2007). If phage lytic proteins are to be used as enzybiotics, they should work exogenously/from without. This implies that the enzybiotic must be able to reach the PG solely by diffusion. This can pose a problem if the PG layer is not directly accessible, as is the case in Gram-negative cells (cells that have an outer membrane surrounding the PG), as well as in Gram-positive cells which are decorated with (lipo)teichoic or teichuronic acids (Schleifer and Kandler, 1972) (see figure 1.3).

Nevertheless, some competent enzybiotics have been created for both Gram-positive and Gram-negative bacteria. Methicillin-Resistant *Staphylococcus aureus* (MRSA) is a Gram-positive pathogen that is resistant to many commonly-used antibiotics. Nonetheless, the ContraFect Corporation's lead enzybiotic, CF-301, has completed phase 2 clinical trials and is on its way to phase 3 (ContraFect, 2019). CF-301 is estimated to be brought onto the market as soon as 2022 (Czaplewski et al., 2016).

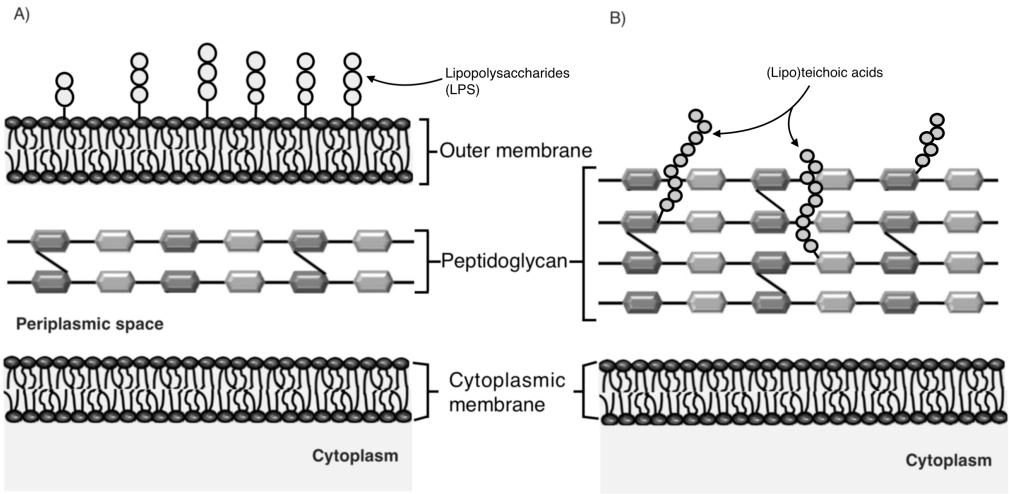


Figure 1.3: The difference in general cell wall structure between (A) Gram-negative and (B) Gram-positive bacteria. Gram-negative cells have an additional membrane surrounding the PG, while Gram-positive cells have a much thicker layer of PG. Gram-negative cell walls are often decorated with lipopolysaccharides and Gram-positives can carry teichoic or teichuronic acids.

Among Gram-negative bacteria, *Acetinobacter baumanii* has been observed to be more susceptible to phage lytic proteins (Gerstmans et al., 2018; Thummeepak et al., 2016). The reasons behind this are however still unknown. Various physical and chemical outer membrane permeabilization techniques have been put forward (Callewaert et al., 2011; Briers et al., 2008, 2011), but more promising is protein engineering. A possible strategy here is to modify the enzyme with a lipopolysaccharide (LPS)-destabilizing peptide that allows it to travel through the outer membrane of a Gram-negative cell. These types of bactericidal, outer membrane-penetrating engineered endolysins have been coined Artilysins® (Briers et al., 2014).

1.3 Protein Domains

Lytic enzymes, as all proteins, are strings of amino acids (AAs) folded in three-dimensional space. Within these sequences, regions that constitute separate functional entities and are able to autonomously fold are denoted protein domains. Proteins often encompass multiple domains that can interact with one another or contribute to a cooperative effect (Ponting and Russell, 2002). Thousands of these protein domains are known and have been catalogued in specialized libraries such as NCBI's CDD (DeWeese-Scott et al., 2010), Pfam (Bateman et al., 2007), SMART (Letunic et al., 2014) and TIGRFAMs (Harkins et al., 2012).

In naturally occurring phage lytic proteins, there are two types of domain architectures: globular and modular (see figure 1.4). Globular proteins contain a single domain responsible for the enzymatic digestion of the PG layer of the bacterial host. Such domain is aptly classified as an Enzymatically Active Domain (EAD). Modular enzymes contain multiple domains. Among these are (one or more) EADs, but also Cell wall Binding Domains (CBD)s which allow the protein to bind to a cell before cutting its PG layer (Oliveira et al., 2013). The trait of having multiple EADs further adds to the robustness of phage lytic proteins against bacterial resistance, as two lytic domains are predicted to be more resilient to resistance development than one (Rodríguez-Rubio et al., 2016; Schmelcher et al., 2012).
Lytic proteins with a globular structure are mostly found in phages infecting Gram-negative bacterial cells (Briers and Lavigne, 2015), while the modular kind is predominantly encountered in Gram-positive bacteria (Gerstmans et al., 2018). The hypothesis is that, as Gram-positive bacteria don't have an additional cell wall around the PG layer, the binding of the enzyme through a CBD prevents its diffusion after digestion, which would cause cells that have not yet been infected by the phage to rupture (Loessner et al., 2002). Nevertheless, exceptions to this rule have been identified (Briers et al., 2007; Walmagh et al., 2012, 2013).

The fact that protein domains are able to function and evolve independently (Pawelkowicz et al., 2016) brings about the opportunity of their synthetic recombination without loss of function. Given sufficient knowledge of the domains commonly found in phage lytic proteins, specific properties and functionalities could be cherry-picked to recombine into an enzybiotic with the desired characteristics (Briers et al., 2014; Gerstmans et al., 2018; Schmelcher et al., 2011). For instance, particular binding or enzymatic domains could be chosen to target either a highly specific or very broad spectrum of hosts. Additionally, domains could be optimised to the environmental conditions the enzybiotic would encounter.

1.4 Objectives of this research

Naturally occurring phage lytic proteins are the result of the grand experiment of evolution. Through millions of years of mutations and recombinations, competent domains have formed and have banded together into functional architectures. Presuming that HGT events are frequent (on an evolutionary time scale), the number of theoretical domain architectures that can be formed with the domains observed in natural phage lytic proteins is enormous. However, only a limited number of architectures have been observed in nature (Oliveira et al., 2013; Vidová et al., 2014). This leads to the hypothesis that there are fundamental design rules that determine which combinations will be functional, and will thus occur in nature, and which will not. Even though natural phage lytic proteins do not necessarily make good enzybiotics and vice versa, deeper knowledge of these design rules would give insight into which factors are important to engineer a phage lytic protein as a targeted (or intentionally un-targeted/broad-spectrum) antibiotic. This would allow current research on protein engineering of lytic enzymes to gradually shift strategy from directed evolution (Gerstmans et al., 2018; Heselpoth and Nelson, 2012; Linden et al., 2015) to a more direct method based on rational design.

The objective of this research is to examine the protein domains in naturally occurring phage lytic proteins and to identify their importance and function within the protein. Alongside various bioinformatic techniques, the predictive power of machine learning algorithms is used to extract crucial domains and rule-defining characteristics of phage lytic proteins. Chapter 2 will focus on exploratory analyses of the PhaLP database, the main resource used in this research. Chapter 3 will make use of common bioinformatic protocols to infer crucial domains and evolutionary relationships between lytic proteins. Finally, chapter 4 will explore how to take advantage of the predictive power of machine learning methods to extract domains that are decisive to a specific characteristic of phage lytic proteins.

It should be noted that the methods used in this inquiry will be mostly focussed on identifying domains and domain architectures that play a role in the targetting and binding of the lytic enzymes to a certain (spectrum of) bacterial host. This model characteristic was chosen because it is a main point of interest in the engineering of new enzybiotics and it has abundant data readily available. Given the right data, these methods should however be adaptable to other protein traits.

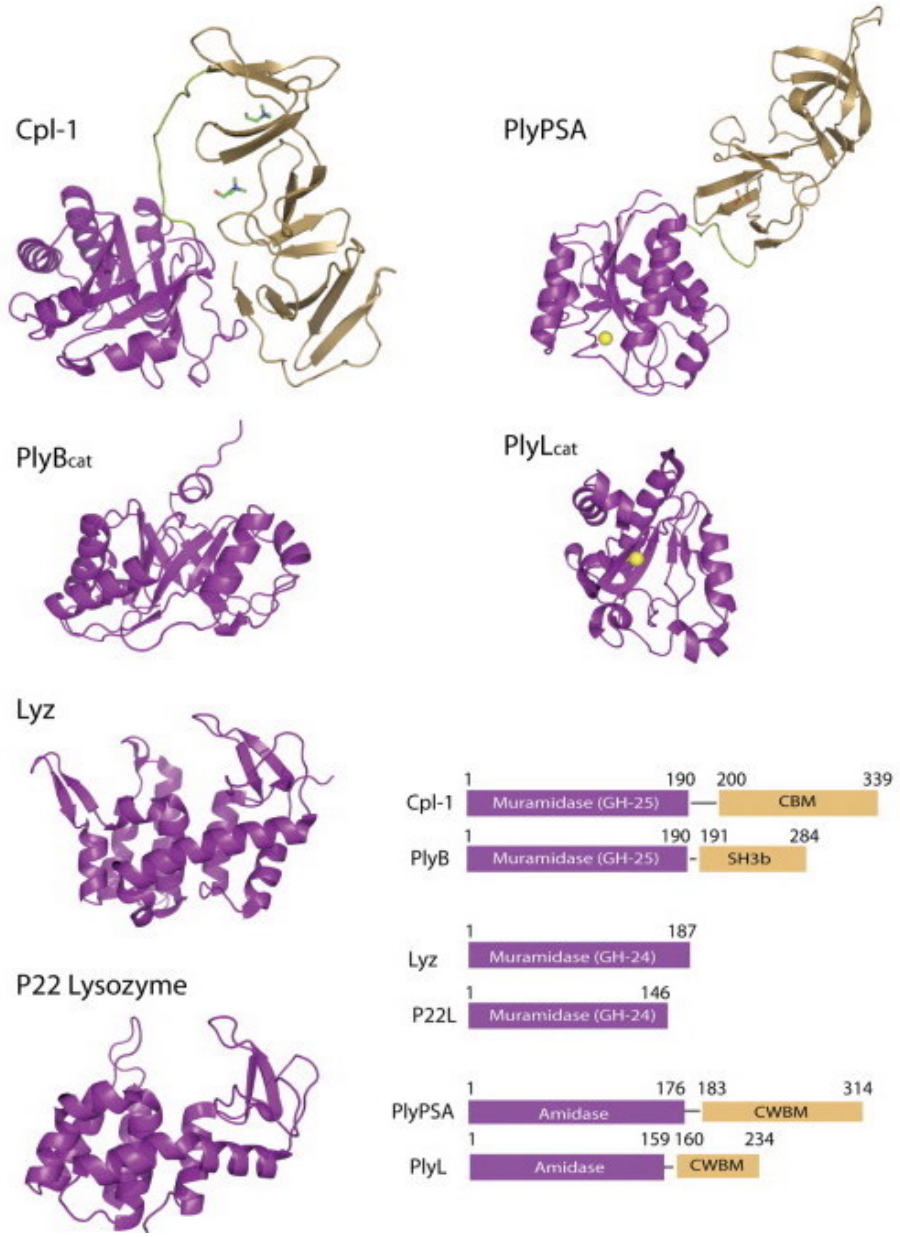


Figure 1.4: Graphical representation of the modularity of endolysins. Catalytic domains are coloured in purple and are sometimes linked to a binding domain (in orange) (Hermoso et al., 2007). Although there might be domain interactions or cooperative effects, these domains provide distinct functionalities to the protein.

CHAPTER 2

PHALP DATABASE

The data used in this research originates from PhaLP: a database of Phage Lytic Proteins (Criel, 2017). This database was assembled from a query of UniProt (version 2019_03), VirusHostDB, ExPASy, NCBI and InterPro(Scan) (version 74.1) based on specific viral taxonomy, keywords and gene ontologies relating to phage lytic proteins. The current version of the database (PhaLP version 2019_03[1]) hosts information on 3901 distinct proteins. This includes, but is not limited to, the protein's amino-acid sequence, corresponding coding DNA, phage and bacterial host taxonomy, experimental evidence, gene ontology, conserved domains etc. (see figure 2.1). The enzymes in this database are all naturally occurring phage proteins and thus do not include any experimental recombinations or fusions.

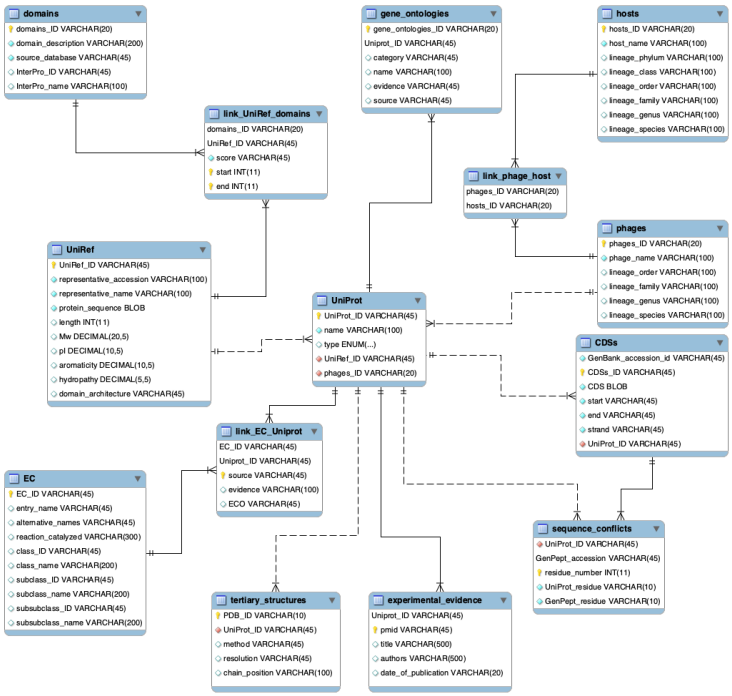


Figure 2.1: Enhanced-Entity Relationship model of PhaLP including table and field names. The relationships between tables in this database are all one-to-many, meaning that a unique entry in table A can link to multiple rows in table B. This is illustrated as an arrow (>|) at the table with the unique entry and an double line at the table with many linked rows (||).

[1]This version is provided as a dump file in the digital appendix.

2.1 Host Taxonomy

As specified in section 1.4, this research is mainly aimed at gathering domains that are important for targeting a certain bacterial host. PhaLP includes 117 unique host genera, whose full taxonomy is illustrated in figure 2.2. Among the bacterial phyla present in this dataset, the genera belonging to the Actinobacteria and Firmicutes are all Gram-positive, while those appertaining to the Bacteroidetes, Cyanobacteria, Fusobacteria and Proteobacteria are Gram-negative. The phylum Deinococcus-Thermus, for which only one genus is included in PhaLP, is left out of this classification as it shares characteristics with both groups (Gupta, 1998). Of the 3636 UniProt phage lysins that have a host genus linked to them, 2066 have a Gram-positive host, 1567 have a Gram-negative host and 4 have a host from the Deinococcus-Thermus phylum. The entry with the UniProt accession 'S6BFI4' is found in both Deinococcus-Thermus and Gram-positive hosts.

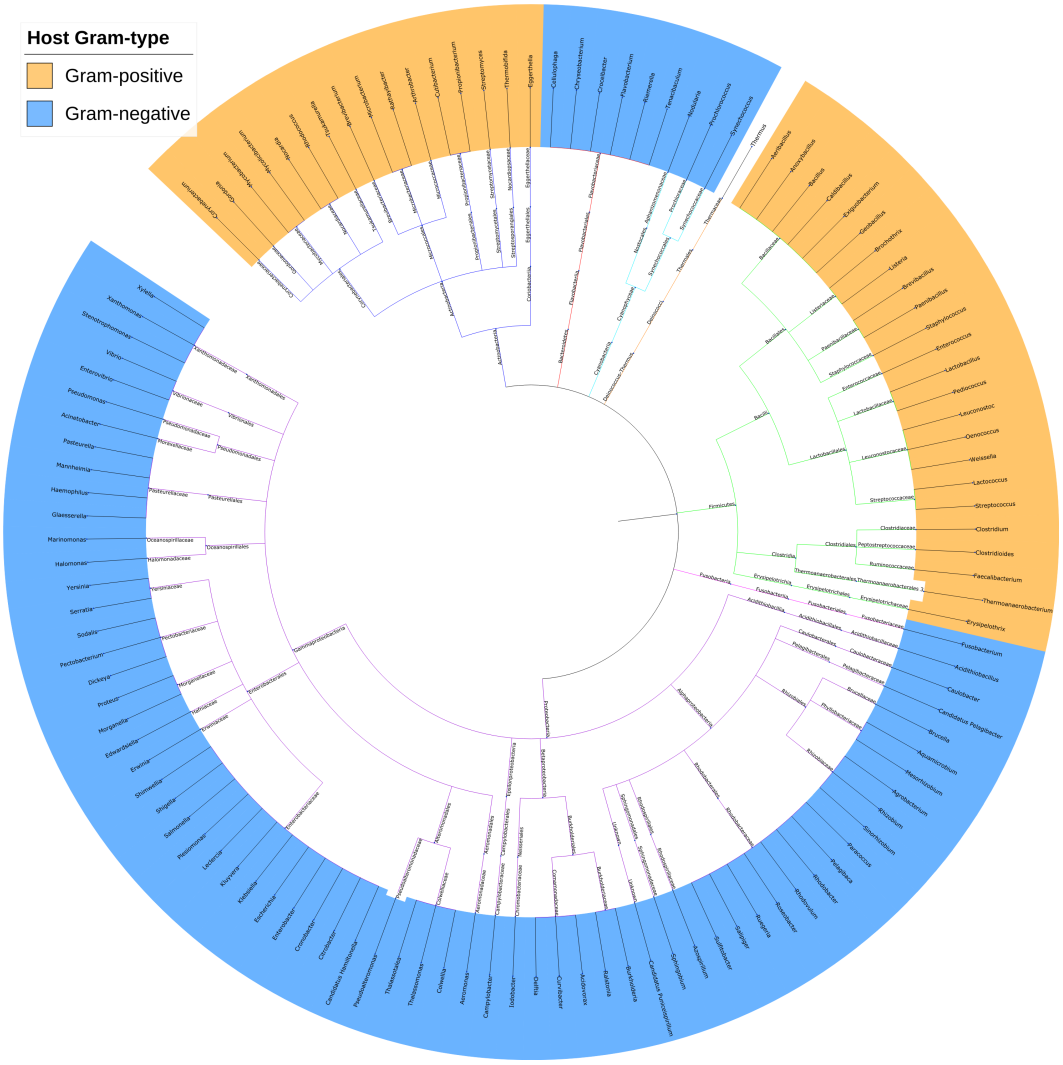


Figure 2.2: Phylogenetic tree of bacterial hosts for which lytic enzymes are included in PhaLP.

Amidst the 3636 host-linked entries in the current version of PhaLP, 231 are known to be capable of infecting multiple hosts. This host-spectrum is generally limited to two different bacterial genera. The main pair here is *Mycobacterium* and *Mycolicibacterium*, having 145 shared assailants. A likely explanation for this is that these two genera, both from the family of Mycobacteriaceae, have similar peptidoglycan structure (type A1γ, see section 2.2.1) and comparable compositions of cell wall and, up until recently, were classified as one single genus (Gupta et al., 2018). These similarities make them vulnerable to the same types of phage lytic proteins.

Propionibacterium and *Cutibacterium* have 58 common infecting phages. These genera also belong to the same taxonomic family, i.e. the Propionibacteriaceae, but unlike the Mycobacteriaceae, this family shows a wider spectrum of peptidoglycan types, as well as more sequence variation (Stackebrandt et al., 2014). Interestingly enough, these 58 phages produce 56 unique endolysin sequences with an identical domain composition. These sequences all relate to studies on bacteriophages targeting the deprecated *Propionibacterium acnes*[2] on the human skin (Farrar et al., 2007; Marinelli et al., 2012; Liu et al., 2015; Brown et al., 2016), hence a plausible explanation for this lack of diversity can be found in the lipid-rich anaerobic environment in which these bacterial hosts reside (Marinelli et al., 2012). As a consequence, this particular domain architecture should be interesting for the engineering of an enzybiotic to treat acne.

The entries in PhaLP with the broadest host spectrum emanate from the Enterobacteria phage PRD1. This phage has two known lytic proteins in PhaLP and can infect six different genera of bacterial host: *Escherichia*, *Salmonella*, *Proteus*, *Acinetobacter*, *Pseudomonas* and *Vibrio*. The P7 protein (UniProt accession 'P27380') is a VAPGH, while the P15 protein (UniProt accession 'P13559') plays a role in both injection of the phage-genome and lysis of the host, making it rather interesting as a putative broad-spectrum enzybiotic (Rydman and Bamford, 2002). Although these two entries do not have a uniform sequence or structure, both contain multiple domains with a transglycosylase functionality (e.g. LT GEWL: cd00254; SLT 1: IPR008258; Transglyc AS: IPR000189 etc.).

It must be noted that since biology and evolution are ongoing processes, taxonomic classification of bacteria is dynamic and rarely matches the biological truth perfectly. Inaccurately classified bacteria may thus cause some correlations and similarities to become unclear in the data exploration and analyses below. Accordingly, possible relations and causal inferences are open for interpretation.

[2] In late 2016, most *P. acnes* were reclassified as *Cutibacterium acnes* (Scholz and Kilian, 2016).

2.2 PhaLP domains

PhaLP contains a table of domains annotated to each phage lytic protein within it (see figure 2.1). As mentioned in section 1.3, there are two types of domain that can be found in these proteins that are involved in its lytic activity: EADs and CBDs. Despite their shared goal, i.e. digestion of PG, there is a large variety of EADs, CBDs and possible architectures in which they are arranged. This variety largely springs from the diversity of PG found in different bacterial cell walls (Schleifer and Kandler, 1972).

2.2.1 Peptidoglycan

Peptidoglycan (PG), also called murein, is a heteropolymer common to all bacterial cell walls. As the name suggests, it consists of glycan strands cross-linked through short peptides (see figure 2.3). Glycans are strands of monosaccharides linked with glycosidic bonds. In PG these are usually strands of alternating β-1,4-linked N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) residues. These strands can vary slightly in acetylation, phosphorylation and chain length depending on the bacteria in which they are present, but they are usually quite uniform in composition (Ghuysen, 1968). The stem peptide is bound from its N-terminus to the carboxyl group of MurNAc and consists of a few amino acids (AAs) in alternating L- and D-configurations. Furthermore, the peptide subunits of peptidoglycan are cross-linked to one another through interpeptide bridges, usually between the amino group of a diamino acid and the C-terminus of D-alanine (Schleifer and Kandler, 1972).

Not only the peptide subunits but the cross-links between them show great variety in chemical composition (Cummins and Harris, 1956). Peptidoglycan types are classified based on this variety. The cross-links can occur either between positions 3 and 4 of the peptide subunits or between positions 2 and 4, dividing peptidoglycan types into a group A and B, respectively. Subsequently, based on the components and biosynthesis of the interpeptide bridge, the groups are further divided into subgroups denoted by a digit, within which variations based on the third AA in the peptide subunit are categorised by a Greek letter (Schleifer and Kandler, 1972).
Within Gram-negative bacteria, little variation in peptidoglycan type is perceived and the most prominent type is A1γ. The Gram-positive bacteria, however, display great disparity in PG composition and structure (Schleifer and Kandler, 1972).

2.2.2 Enzymatically Active Domains (EADs)

The enzymatic domains that act upon the PG, i.e. the EADs, are ordinarily found at the N-terminus of the protein (Loessner, 2005). Among their enzymatic activities, there are generally regarded to be three main classes: Glycosidases, Amidases and Endopeptidases. These cell wall hydrolases will each act upon the PG layer of a host bacterium in a specific way (see figure 2.3).

Cell wall Glycosidases (CWGs) catalyse the hydrolysis of glycosidic β-1,4 linkages (Vermassen et al., 2019). The enzymes in this class generally belong to the EC classification 3.2.1 and according to the CAZyme database (Carbohydrate Active Enzymes database[3]), there are currently 162 known families. This group can be further differentiated into N-Acetylglucosaminidases and lysozymes. The former cleave the β-1,4 specifically between the GlcNAc and the MurNAc of the bacterial PG (Rodríguez-Rubio et al., 2016). Lysozymes and lytic transglycosylases both cleave the bond between MurNAc and GlcNAc, but the reaction mechanism and end-products they generate are different (Höltje et al., 1975). As the reaction meachanism in lytic transglycosylases does not involve water, they are not technically hydrolases and are classified under the EC numbers 4.2.2.n1 and 4.2.2.n2 (Herlihey and Clarke, 2017). For simplicity, these will however be grouped as CWGs in this study.

Cell wall Amidases (CWAs) hydrolyse the amide bond between MurNAc and L-alanine residues, effectively cleaving the glycan strand from the peptide moiety (Höltje et al., 1994). These enzymes can all be classified under the EC number 3.5.1 (Vermassen et al., 2019). A Cell Wall Peptidase (CWP) cleaves the bond between two amino acids within the PG layer (Höltje et al., 1994). These enzymes are restricted to EC numbers 3.4 and can be subdivided into endopeptidases and carboxypeptidases (Vermassen et al., 2019).

2.2.3 Cell wall Binding Domains (CBDs)

The often C-terminal CBDs are responsible for the binding of a phage lytic protein to a ligand in or on the bacterial cell wall or PG (Loessner et al., 2002). This highly specific binding, together with the specialized catalytic mechanisms of the EAD, brings about a well-defined spectrum of activity for lytic proteins (Eugster et al., 2011). Additionally, binding of the C-terminus to PG-associated ligands increases proximity of the N-terminal EAD to its substrate (Loessner, 2005).

[3] http://www.cazy.org

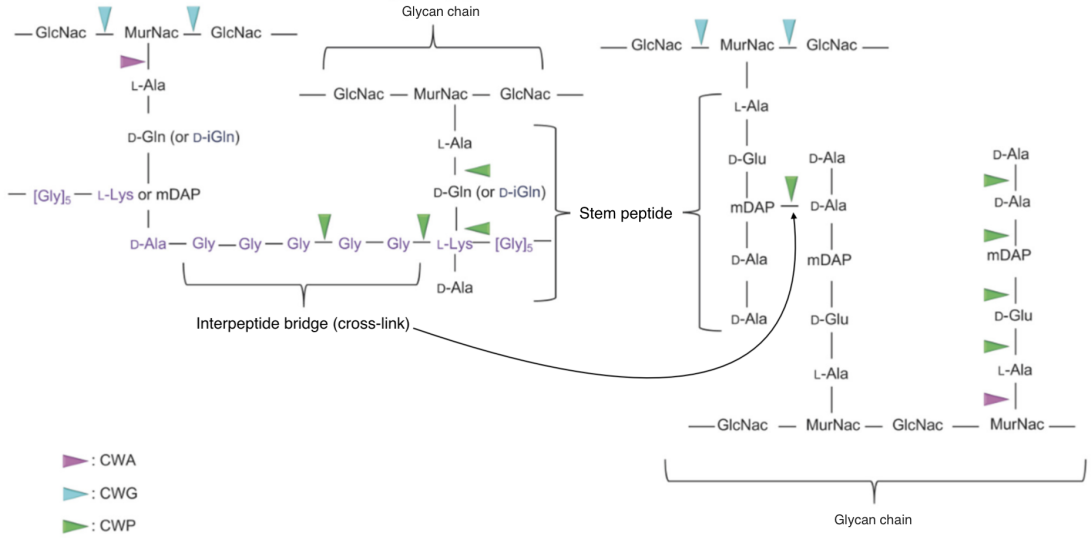


Figure 2.3: Schematic representation of the primary structures of two common peptidoglycan (PG) types. N-Acetylglucosamine (GlcNAc) and N-Acetylmuramic acid (MurNAc) make up the glycan strands, while the stem peptides are made up of several amino acids and meso-diaminopimelic acid (mDAP). Through cross-links between the several peptide subunits of peptidoglycan, the polymer gains its strength and rigidity. The structure on the right in peptidoglycan type A1γ, the most common one in Gram-negative bacteria. The structure on the left is a variation of type A3 (either α or γ depending on the presence of L-Lys or mDAP in the third position of the peptide subunit). A3 types are common to *Staphylococcus aureus*, a Gram-positive bacteria (Schleifer and Kandler, 1972).
The different classes of cell wall hydrolases and the location on the cell wall where they impact are also indicated on this figure. Cell Wall Amidases (CWAs) cleave bonds between the glycan strand and the peptide subunits. Cell Wall Glycosidases (CWGs) cleave within the glycan chain and can be further divided into N-Acetylglucosaminidases, transglycosylases and lysozymes dependant on which exact bond they target. Cell Wall Peptidases (CWPs) cleave bonds between AAs, this can be both in the peptide subunit as well as along the interpeptide bridge.

2.3 Quantitative domain analysis

The connection in PhaLP from UniProt accessions to bacterial host taxonomy allows for a preliminary examination of phage lytic protein design towards host-spectrum. To do this, domains were manually curated into categories through the use of Gene Ontology (GO) terms and EC numbers. To avoid redundancy between protein domains from different source databases, domain accessions from InterPro were used whenever possible.

2.3.1 Abundance

As initial analysis, each entry in PhaLP was queried for domain annotation to give insight into individual domain abundance. The EADs, subcategorised into hydrolytic

classes, are set out in table 2.1. The CBDs are set out in table 2.2. Both tables also include domain accessions from the various source databases incorporated in PhaLP that link to a certain domain.

A total of 82 EADs could be distinguished. Among these by far the most abundant are N-acetylmuramoyl-L-alanine amidase type 2 domain (Amidase 2; IPR002502) and the Peptidoglycan Hydrolase Recognition Particle superfamily (PGRP SF; IPR36505). These are both CWAs and are both found in around half of the phage lytic proteins in PhaLP.

Among the CWGs, lysozymes are the most abundant and also most diverse group. 19 different types could be identified, of which the lysozyme-like domain superfamily was observed in 1395 distinct phage lytics proteins.

The CWPs are the smallest and least abundant group in the database. Perhaps the most interesting domain from this class is CHAP (Cysteine Histidine-dependent Amidohydrolase/Peptidase; IPR007921), which is currently believed to cleave between stem-peptides and cross-links in certain PG-types (Sundarrajan et al., 2014). CHAP domains can be found in the protein architecture of 248 different entries in PhaLP.

Table 2.1: The conserved EADs present in PhaLP.

Enzymatic Class	Domain	Description	Linked accessions	Number of UniProt entries
Amidase	Amidase D sub1	Unnamed amidase AmiD subfamily	PTHR30417:SF1	39
	Amidase02 C	N-acetylmuramoyl-l-alanine amidase 02 C	IPR021976 and PF12123	44
	Amidase-r sub1	N-acetylmuramoyl-l-alanine amidase-related subfamily	PTHR30032:SF1	5
	Amidase D	N-acetylmuramoyl-l-alanine amidase AmiD	PTHR30417	556
	Amidase-r	N-acetylmuramoyl-l-alanine amidase-related	PTHR30032	5
	Amidase 30404 sub4	N-acetylmuramoyl-l-alanine amidase subfamily	PTHR30404:SF4	72
	SleB 1	Cell wall hydrolase SleB, domain 1	IPR042047 and G3DSA:1.10.10.2520	24
	Amidase C	N-acetylmuramoyl-l-alanine amidase AmiC	PTHR30404:SF0	148
	Endolysin T7 type	Endolysin T7 type	IPR034689 and MF_04111	134
	Amidase 30404	N-acetylmuramoyl-l-alanine amidase	PTHR30404	231
	Zn-exopept	Zn-dependent exopeptidases	G3DSA:3.40.630.40 and SSF53187	239
	PGRP SF	PGRP domain superfamily	IPR036505, G3DSA:3.40.80.1 and SSF55846	1935
	PGRP met/bac	Peptidoglycan recognition protein family domain, metazoa/bacteria	IPR006619 and SM00701	273
	PGRP	Peptidoglycan recognition protein	IPR015510 and PTHR11022	291
	Amidase 5	Bacteriophage lysin	IPR008044 and PF05382	12
	Amidase D sub11	N-acetylmuramoyl-l-alanine amidase AmiD	PTHR30417:SF11	125
	Anhydro amidase AMPD	1,6-anhydro-N-acetylmuramoyl-l-alanine amidase AMPD	PTHR30417:SF4	1
	Amidase 3	N-acetylmuramoyl-L-alanine amidase type 3	IPR002508, cd02696, PF01520 and SM00646	239
	Amidase 2	N-acetylmuramoyl-L-alanine amidase type 2	IPR002502, cd06583, PF01510 and SM00644	1841
Glycosidase	Glycosidases	Glycosidase superfamily	G3DSA:3.20.20.80	291
	Glycoside hydrolase SF	Glycoside hydrolase superfamily	IPR017853 and SSF51445	286
	Glyco hydro 19 cat	Glycoside hydrolase, family 19, catalytic	IPR000726 and PF00182	5
	Glyco hydro fam25 subgr	Glycoside hydrolase, family 25 subgroup	IPR018077 and SM00641	211
	Glyco hydro 25	Glycoside hydrolase, family 25	IPR002053 and PF01183	279
	Glyco hydro 25 AS	Glycoside hydrolase, family 25, active site	IPR008270 and PS00953	16
	Glyco hydro 25 AtlA-like	GH25_AtlA-like	cd06522	1
	Glyco hydro 25 Cpl1-like	GH25_Cpl1-like	cd06415	44
	Glyco hydro 25 Lyc-like	GH25_Lyc-like	cd06525	3
	Glyco hydro 25 LysA-like	GH25_LysA-like	cd06417	1
	Glyco hydro 25 LytC-like	GH25_LytC-like	cd06414	2
	Glyco hydro 25 muramidase	GH25_muramidase	cd00599	76
	Glyco hydro 25 PlyB-like	GH25_PlyB-like	cd06523	25
	Glyco hydro 66	Glycosyl hydrolase family 66	IPR025092 and PF13199	1
	GLUCO	Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase-like domain	IPR002901, PF01832 and SM00047	11

Continued on next page

Table 2.1: The conserved EADs present in PhaLP.

Enzymatic Class	Domain	Description	Linked accessions	Number of UniProt entries
Glycosidase (continued)	LT-GEWL	Lytic Transglycosylase and Goose Egg White Lysozyme domain	cd00254	168
	Gp16	Internal virion protein Gp16	IPR038994 and MF_04121	102
	Mur transglyc D	Membrane-bound lytic murein transglycoylase D	PTHR33734:SF14	98
	SLT 1	Transglycosylase SLT domain 1	IPR008258 and PF01464	170
	Transglyc AS	Prokaryotic transglycosylase, active site	IPR000189 and PS00922	144
	Transglyc F	Membrane-bound lytic murein transglycoylase F	PTHR35936:SF19	6
	Endolysin/autolysin	Endolysin/autolysin	IPR033907 and cd00737	375
	Endolysin lambda type	Endolysin lambda type	IPR034691, cd00736 and MF_04109	87
	Endolysin T4 type	Endolysin T4 type	IPR034690 and MF_04110	147
	T4-like	bacteriophage_T4-like_lysozyme	cd00735	235
	T4-type lysozyme	T4-type lysozyme	IPR001165 and PR00684	386
	Lysozyme 23208	Lysozyme protein	PTHR23208	41
	Lysozyme RrrD-r	Lysozyme RrrD-related	PTHR38107:SF2	171
	Lysozyme 40	Unnamed lysozyme	G3DSA:1.10.530.40	1108
	Lysozyme 23208 sub38	Lysozyme protein subfamily	PTHR23208:SF38	41
	Phage lysozyme2	Phage tail lysozyme	IPR041219 and PF18013	6
	Lysozyme-like SF	Lysozyme-like domain superfamily	IPR023346 and SSF53955	1395
	Glyco hydro 24	Glycoside hydrolase, family 24	IPR002196 and PF00959	1202
	Phage PRD1 P15 lysozyme	Bacteriophage PRD1, P15, lysozyme	IPR016284 and PIRSF001069	1
	IraD/Gp25-like	IraD/Gp25-like	IPR007048 and PF04965	1
	Lysozyme 280	Unnamed lysozyme	G3DSA:2.40.10.280	1
	Muramidase	N-acetylmuramidase	IPR024408 and PF11860	1
	Gp5 OB N	Protein Gp5, N-terminal OB-fold domain	IPR009590 and PF06714	197
	Chitinase	Unnamed chitinase	G3DSA:1.10.530.70	7
	DUF847	Protein of unknown function DUF847	IPR008565 and PF05838	2
Peptidase	Peptidase U40	Peptidase U40	IPR019505 and PF10464	1
	Peptidase M15A C	Peptidase M15A, C-terminal	IPR013230 and PF08291	2
	Peptidase-r	Peptidase-related	PTHR21666	249
	Mur hydro NLPD	Murein hydrolase activator NLPD	PTHR21666:SF263	1
	Peptidase-r sub266	Unnamed subfamily with metalloendopeptidase activity	PTHR21666:SF266	1
	Mur DD endopept MEPM	Murein DD-endopeptidase MEPM	PTHR21666:SF270	3
	Peptidase-r sub271	Unnamed subfamily with metalloendopeptidase activity	PTHR21666:SF271	2
	Cys proteinase SF	Cysteine proteinases	G3DSA:3.90.70.10	19
	NLP P60	Endopeptidase, NLPC/P60 domain	IPR000064 and PF00877	2
	Peptidase C39	Peptidase C39-like	IPR039564 and PF13529	7
	Peptidase M23	Peptidase M23	IPR016047 and PF01551	252
	Papain-like cys pep SF	Papain-like cysteine peptidase superfamily	IPR038765 and SSF54001	249
	endopeptidase-like	Endopeptidase domain like	G3DSA:3.90.1720.10	263
	CHAP	CHAP domain	IPR007921, PF05257 and PS50911	248
	Gp5 SF	Peptidoglycan hydrolase Gp5 superfamily	IPR038288 and G3DSA:1.10.530.50	1
Unknown	Cell wall hydrolase SleB	Cell wall hydrolase, SleB	IPR011105 and PF07486	23
	Hydro 34135 sub2	Unnamed subfamily with hydrolase activity	PTHR34135:SF2	98
	Hydro 34135 sub1	Unnamed subfamily with hydrolase activity	PTHR34135:SF1	41
	Dup hybrid motif	Duplicated hybrid motif	IPR011055, G3DSA:2.70.70.10 and SSF51261	263
	PG exotransglyc lys	Unnamed superfamily	G3DSA:1.10.530.10	281
	Hydro 38107 sub3	Unnamed subfamily with hydrolase activity	PTHR38107:SF3	45
	DUF3597	Domain of unknown function DUF3597	IPR022016 and PF12200	7
	Glygly endopept	Cell wall targeting domain of glycylglycine endopeptidase	G3DSA:2.30.30.410	283

Despite the advantages CBDs serve to the efficacy of lytic proteins, they are not essential to the lytic functionality. Consequently, only about one third of the entries in PhaLP possesses a known CBD (See table 2.2). Furthermore, as will be discussed in section 2.3, the vast majority of phage lytic proteins containing a CBD targets a bacterial host of positive Gram-type. Noteworthy CBDs are SRC Homology 3 domains (IPR003646, SSF82057 and G3DSA:2.30.30.40), Peptidoglycan Binding domains (IPR002477, IPR036366 and IPR036365) and Lysin Motif domains (IPR036779,

Table 2.2: The conserved CBDs present in PhaLP.

Domain	Description	Linked accessions	Number of UniProt entries
Peptidoglycan BD-like	Peptidoglycan binding-like	IPR002477 and PF01471	256
PGBD SF	PGBD superfamily	IPR036366 and G3DSA:1.10.101.10	290
Invasin/intimin cell adhesion	Invasin/intimin cell-adhesion fragments	IPR008964 and SSF49373	1
SPOR-like	Sporulation-like domain	IPR007730, PF05036 and PS51724	20
LysM SF	Lysin motif domain superfamily	IPR036779 and G3DSA:3.10.350.10	169
SH3-like bac-type	SH3-like domain, bacterial-type	IPR003646, PF08239, PF08460, PS51781 and SM00287	322
PSA CBD	PSA endolysin, cell wall binding domain	IPR041341 and PF18341	8
Cell wall/Cho-BD repeat	Cell wall/choline-binding repeat	IPR018337, PF01473, PS51170 and G3DSA:2.10.270.10	97
Cpl-7 lyso C	Cpl-7 lysozyme, C-terminal	IPR013168, PF08230 and SM01095	25
LGFP	LGFP repeat	IPR013207 and PF08310	66
SH3-like SF	SH3-like domain superfamily	IPR036028 and SSF50044	2
SPOR-like SF	Sporulation-like domain superfamily	IPR036680, G3DSA:3.30.70.1070 and SSF110997	16
SH3-r pro SF	Prokaryotic SH3-related domain superfamily	SSF82057	8
SH3 SF	SH3 domains superfamily	G3DSA:2.30.30.40	68
LysM dom SF	Lysin motif domain superfamily	SSF54106	165
LysM	Lysin motif domain	IPR018392, cd00118, PF01476, PS51782 and SM00257	169
Vir attach	Virus attachment protein , globular domain	G3DSA:2.60.90.20	1
Attachment protein shaft SF	Attachment protein shaft domain superfamily	IPR009013 and SSF51225	1
Big 2	Bacterial Ig-like, group 2	IPR003343, PF02368 and SM00635	1
LysM GPI 2	Lysin motif domain-containing GPI-anchored protein 2	PTHR33734:SF11	65
Peptidoglycan-BD 3	Peptidoglycan binding domain	IPR018537 and PF09374	1
CWB repeat SF	Cell wall binding repeat superfamily	SSF69360	97
Vir attach sigma1 reovir	Viral attachment sigma 1, reoviral	IPR002592 and PF01664	1
PGBD-like SF	PGBD-like superfamily	IPR036365 and SSF47090	296
Any CBD			976

IPR018392 and SSF54106). These domains can be found in the majority of the entries in PhaLP with a known CBD.

2.3.2 Occurrence and distribution

Narrowing down on host-specificity, the occurrence of all domains specified in tables 2.1 and 2.2 was mapped against the host taxonomy on genus level (although these are sorted on higher levels as well, see figure 2.4). The inverse relation, i.e. the ratio of occurrence of all host genera for a particular domain, was charted in figure 2.5.

Quantitative analyses of this type have been conducted before on datasets of phage lytic proteins. Oliveira et al. (2013), for instance, analysed the appearance of 35 domains in 727 endolysins. Since then, the amount of known and annotated phage lytic proteins has risen tremendously, allowing for a more accurate estimation of the domain distribution in nature. The use of PhaLP version 2019_03 now accommodates for a study of the occurrence of 106 domains across 3636 unique phage lytic proteins.

In figure 2.4, a clear distinction can be made between domains that appear in hosts of different Gram-types. The Gram-negative genera seem to be a lot less likely to contain a CBD, with the phyla Bacteroidetes, Cyanobacteria and Fusobacteria containing none at all. This further supports the hypothesis of Loessner et al. (2002) that

the endolysins of Gram-negative bacteria are mostly globular because their cell wall beyond the PG layer already prevents diffusion of the endolysin after digestion, nullifying the need for a binding domain. Surprisingly, the only common Gram-negative CBD identified by Oliveira et al. (2013), namely Peptidoglycan-BD 3 (IPR018537), is only found in a single entry in PhaLP[4]. This discrepancy can however be traced back to the fact that not all phage lytic proteins studied by Oliveira et al. (2013) are also incorporated in PhaLP and will likely be resolved with future versions of the database. It however points to an important caveat that a quantitative analysis is not an exact representation of the occurrence and distribution in nature and is highly dependent on the database.

Among the Proteobacteria, some CBDs do appear, but rather infrequently. The SRC Homology 3 (SH3) domain family (G3DSA:2.30.30.40) is one of the most frequent CBDs among Proteobacteria. It is detected in all phage lytic proteins for hosts of the genera *Pelagibaca*, *Salipiger* and *Sphingobium* and more scattered in *Agrobacterium*, *Ruegeria* and *Sinorhizobium*. These six genera all belong to the class of the Alphaproteobacteria, with half of them (*Pelagibaca*, *Salipiger* and *Ruegeria*) sharing the same taxonomic family, the Rhodobacteraceae. In fact, almost all occurrences of CBDs in Gram-negative hosts are located within a small cluster of genera belonging to the class of Alphaproteobacteria in the figure (between *Agrobacterium* and *Sphingobium*). Furthermore, peptidoglycan binding domains (IPR036366, IPR036365 & IPR002477) are also found in phage lytics proteins targeting Pseudomonadales and *Escherichia*, albeit rather infrequently. Since there is no overlap in lytic proteins targeting Gram-positive bacteria and these Proteobacteria, this could indicate a unique need for CBDs when targeting certain Proteobacteria.

About half of the phage lytic proteins for Gram-positive hosts in PhaLP contain CBDs. The most abundantly detected ones in PhaLP are Peptidoglycan Binding Domains (PGBD SF; IPR002477, IPR036365 & IPR036366) and Lysin Motif domains (LysM; IPR018392). These two domain families were also deemed prominent in the study by Oliveira et al. (2013). Most CBDs are shared between the two Gram-positive phyla, but there are exceptions: the sporulation-like (SPOR-like; IPR007730 & IPR036680) and SRC Homology 3-like (SH3-like; IPR003646 & SH3; G3DSA:2.30.30.40) domains are unique to the Firmicutes, although both are also prominent among the Alphaproteobacteria.

EADs execute the main function of phage lytic proteins and are evidently found in all entries regardless of Gram-type or modularity. Some domains are almost universally incorporated, such as the Peptidoglycan Recognition Particle domain superfamily

[4]The single entry in PhaLP containing Peptidoglycan-BD 3 has Pseudomonas as a host. Seeing as it only appears for one entry out of the 132 that have this host (resulting in an occurrence fraction of 0.0075), it is indistinguishable in figure 2.4.

(PGRP SF; IPR036505) and the N-acetylmuramoyl-L-alanine amidase type 2 (Amidase 2; IPR002502) domains. While other domains seem rather strictly confined to a Gram-type, phylum or even lower taxonomic level. The endolysin T7 type (IPR034689), for instance, is solely detected in bacterial hosts from the class of Gammaproteobacteria, in which it is only rarely spotted outside of the order of Enterobacterales. Other CWAs, such as Amidase 30404 sub4 (PTHR30404:SF4) are restricted to genera within the Firmicutes, although its overarching domain family Amidase 30404 (PTHR30404) is also detected in multiple Bacteroidetes genera. Analogously, the Amidase D sub11 (PTHR30417:SF11) is exclusive to the Firmicutes as well, while its larger domain family (Amidase D; PTHR30417) is frequently present in all Gram-positive hosts even some Alphaproteobacteria. This unique distribution could possibly point to a divergent evolution of lytic protein domains towards ones specialized for specific host-ranges. Also notable within the CWAs is the cell wall hydrolase SleB 1 domain (IPR042047), whose occurrence is entirely limited to the order of Synechococcales. Vice versa, in the 57 entries targeting a Synechococcales host, 24 contained this specific domain. Nevertheless, further research with larger sample sizes sequenced from nature should be conducted to understand the full scope of this correlation.

As can be seen in the bottom right corner of figure 2.4, CWGs are the most abundant EAD for Gram-negative hosts, having a much more sparse distribution within Gram-positive ones, where CWAs have the upper hand. Among the CWGs, lysozyme and transglycosylase domains are most prominent in Gram-negative hosts, while glucosaminidases are confined to Gram-positive ones (see rightmost columns in figure 2.5). Transglycosylase domains themselves seem significantly more present in hosts from the Enterobacterales class, this can be clearly seen as a cluster in both figure 2.4 and figure 2.5.

As also observed by Oliveira et al. (2013), the enzymatic domains belonging to the lysozyme-like superfamily (e.g. autolysin; IPR033907, T4-type lysozyme; IPR001165 & glycoside hydrolase family 24; IPR002196) almost exclusively occur in Gram-negative genera. From these, the T4-type lysozyme (IPR001165) is even limited to the Proteobacteria. The glycoside hydrolase superfamily domains (G3DSA:3.20.20.80, IPR018077 & IPR002053) are solely found in phage lytic proteins targeting Gram-positive or Gram-ambiguous[5] genera. Unlike the study by Oliveira et al. (2013), however, domains from the glycoside hydrolase family 25 (e.g. IPR002053, IPR018077 & cd00599) are found in both Gram-positive phyla rather than being mostly restricted to the Firmicutes (see figure 2.6). Nevertheless, some domains within this family (e.g. IPR008270, cd06414, cd06415, cd06522, cd06523

[5]Bacteria in the phylum Deinococcus-Thermus possess a secondary cell envelope (primary characteristic of Gram-negative bacteria), but still show a Gram-positive stain (Gupta, 1998). These are furthermore classified as Gram-ambiguous in this research.

& cd06525) do seem specialized to hosts from this phylum, although these domains are quite rare and appear in too little entries within PhaLP to make substantial conclusions (see table 2.1).

CWPs seem less common overall and almost completely absent from the Gram-negative hosts. A possible hypothesis for this relation could be that, as PG-types are quite uniform among Gram-negative bacteria (see section 2.2.1), more general lytic protein architectures will function on most of these hosts, creating no increased fitness for bacteriophages capable of specialised hydrolysis in the peptide moiety. The only relatively abundant peptidase domain for Gram-negative hosts is the cysteine proteinase superfamily (G3DSA:3.90.70.10), which is found in 14 out of the 57 lytic proteins for hosts of the genus *Synechococcus*. This might even be a specialised domain for this host as this domain is overall only observed in 19 entries in PhaLP (see table 2.1). The Actinobacteria show only one peptidase domain of note, i.e. Peptidase M23 (IPR016047), a family of endopeptidases whose occurrence within PhaLP is almost entirely restricted to the order of the Corynebacteriales.

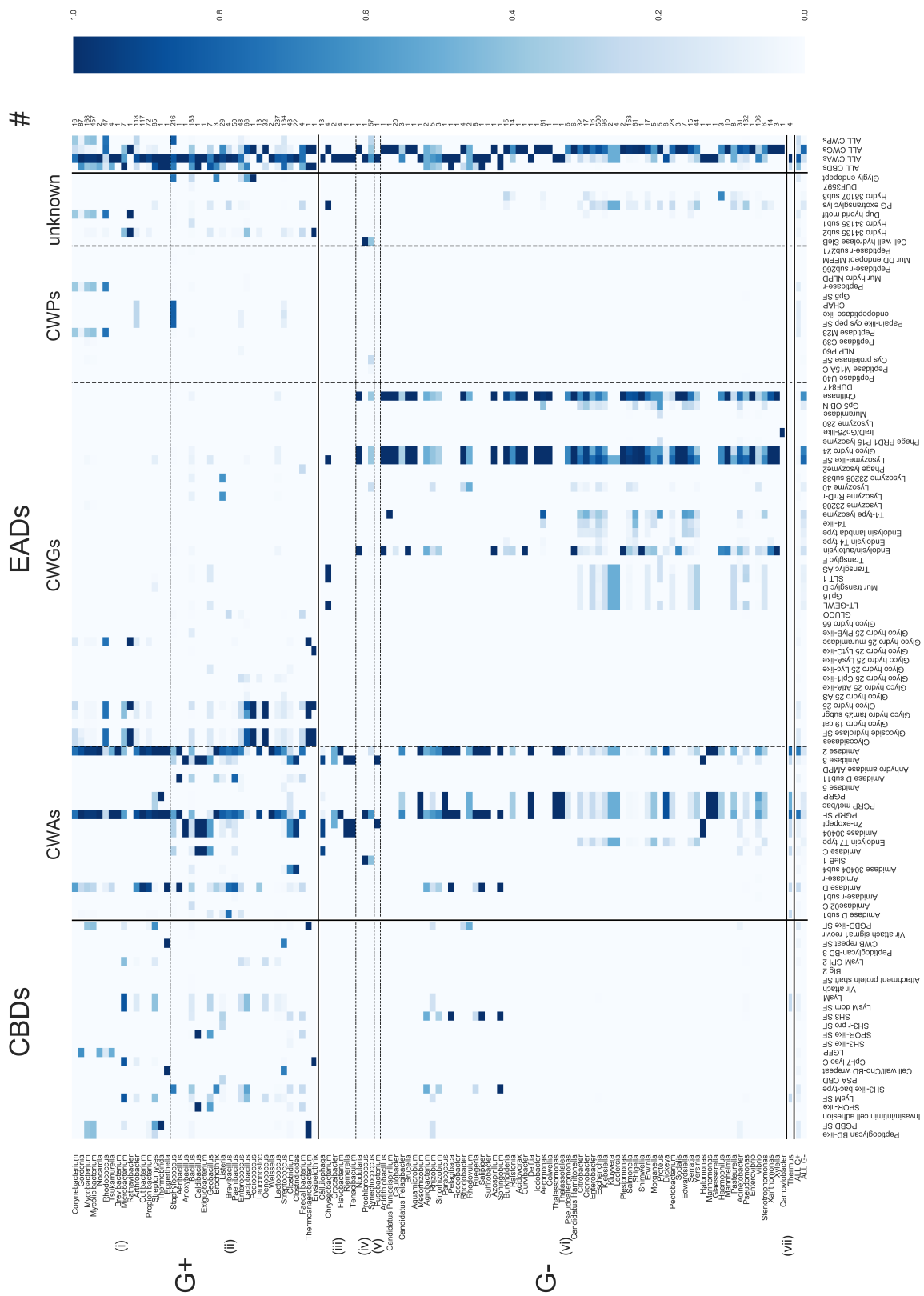


Figure 2.4: Occurrence and distribution of a domain across bacterial hosts. The colour bar on the right denotes the fraction of phage lytic proteins in PhaLP with a specific host that contain a certain domain. The horizontal lines divide different phyla, with the full horizontal lines indicating divides between Gram-types. The examined phyla from top to bottom are: (i) Actinobacteria, (ii) Firmicutes, (iii) Bacteroidetes, (iv) Cyanobacteria, (v) fusobacteria, (vi) Proteobacteria and (vii) Deinococcus-Thermus.

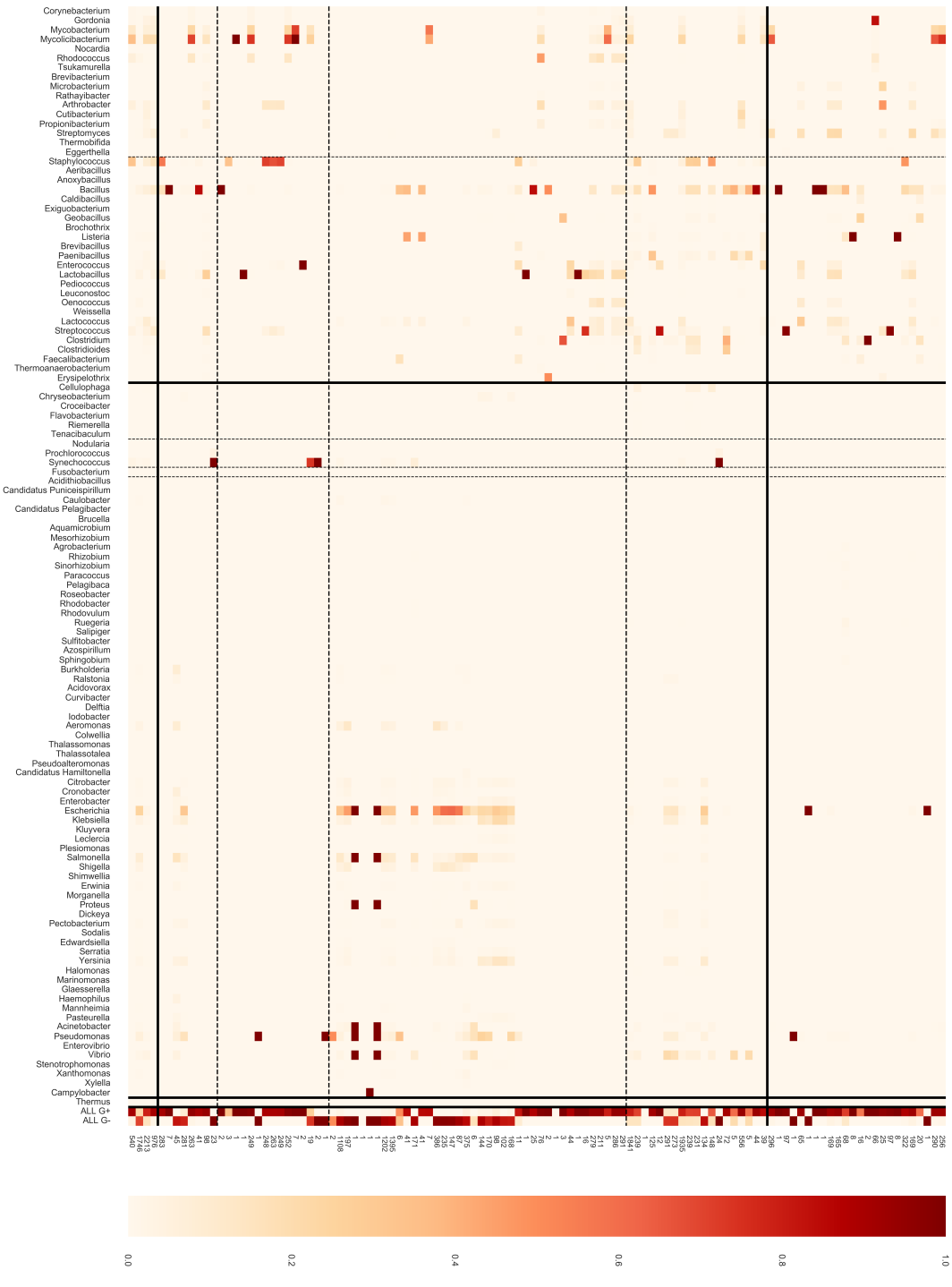


Figure 2.5: Occurrence and distribution of target hosts for phage lytic proteins with specific domains. The colour bar on the right denotes the fraction of entries in PhaLP that contain a certain domain that are also able to target a certain bacterial host. Vertical lines divide different phyla and horizontal lines divide types of domains. The EADs are subdivided into (i) Cell Wall Amidases, (ii) Cell Wall Glycosidases, (iii) Cell Wall Peptidases and (iv) hydrolases of unknown class.

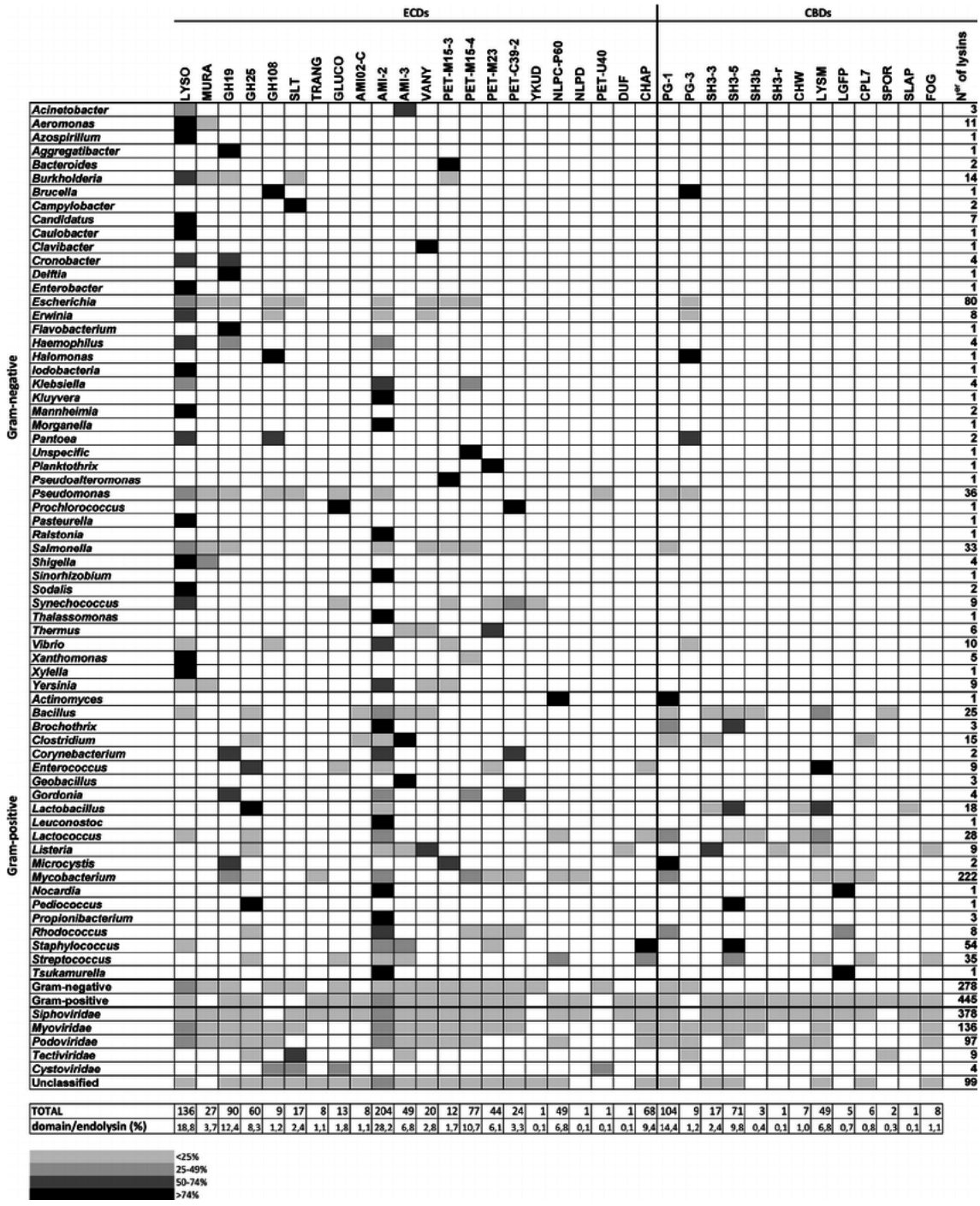


Figure 2.6: The occurrence and distribution of 35 domains from 727 endolysins across phage families and bacterial hosts as analysed by Oliveira et al. (2013).

CHAPTER 3

CONSERVATION APPROACH

As specified in section 1.4, the overarching goal of this dissertation is to find underlying design rules in the architectures of phage lytic proteins. An important part of this process is learning from nature. The natural phage lytic proteins that are discovered, are the result of the massive experiment of natural evolution. They have evolved through duplications, mutations and recombinations and only those that resulted in an increase of fitness have been assimilated into the gene pool. This means that by looking at a set of proteins of a certain function, we gain a unique window into what makes the execution of that function feasible. With that in mind, this chapter will focus on finding commonalities in the amino acid (AA) sequences of phage lytic proteins and will try to root these commonalities in evolution.

3.1 Conserved Domains

Conservation of a region in an AA-sequence means this sequence has (for the most part) withstood the various mechanisms driving evolution. If such region is conserved across many proteins of a similar function, this signifies a certain degree of importance to that function, as conservation is achieved through a fitness advantage. According to the NCBI handbook, a conserved domain is a recurring unit in polypeptide chains that can be discovered by comparative analysis. Molecular evolution uses these domains as building blocks of modular proteins. They are recombined in different architectures to make proteins of various functions (Sayers and Bryant, 2002).

While a conserved region does not necessarily constitute a protein domain, it has been proven that sequence identity of over 40% generally means a correspondence of function (Wilson et al., 2000; Todd et al., 2001). The hypothesis thus is that through alignment of different phage lytic proteins, conserved domains and even domain architectures can be found. Not only would this allow to quickly identify host-ranges for which phage lytic domain composition is similar, it can provide insights on horizontal and vertical gene transfer based on the distance measure between sequences that an alignment score generates. It could also be possible to annotate novel conserved

domains, although recent efforts by Pfam, CDD etc. make this less likely (Bateman et al., 2007; DeWeese-Scott et al., 2010).

3.1.1 Local alignment

The most straightforward way of sequence comparison is through alignment algorithms. These algorithms, often based on dynamic programming, cycle through every character and return a score on how much the compared sequences are alike (Needleman and Wunsch, 1970). There are two main variants of this algorithm, global and local alignments. While global alignments optimise the overall alignment of two sequences, local alignments search for conserved subsequences within the given input sequences (Smith and Waterman, 1981). Since global alignments may include stretches of little similarity, local alignments are often preferred for homology-based domain annotation (Altschul et al., 1990; DeWeese-Scott et al., 2010).

Alignments performed in the analyses in this chapter were all local alignments on the full AA sequences of phage lytic proteins. AAs have several advantages over nucleotides in sequence comparisons. First of all, because there are 20 different AAs and only 4 different nucleotides, an AA match has a lot more significance than a nucleotide match would[1]. Additionally, many triplets of nucleotides (codons) code for the same amino acid, causing certain substitutions to only introduce noise and accordingly lower the sensitivity and reach of the algorithm. Lastly, the likelihood of AA substitutions occurring during evolution varies substantially according to the particular AA, unlike for nucleotides where substitution probabilities are more uniform. This added degree of variability vastly improves the performance of an alignment as it allows to discern between possible mutations and non-aligning sections (Koonin and Galperin, 2003).

Alignment algorithms score sequence similarity based on a substitution matrix and gap penalties. A substitution matrix includes scores to be added or subtracted from the total for each relation of characters. This matrix returns positive scores if characters from the input sequences line up and negative or zero scores for mismatches. The BLOSUM62 matrix was chosen in this case because it has been optimised for more distantly related sequences (average similarities of 20 to 40%), which is generally the case for the phage lytic proteins in PhaLP (Henikoff and Henikoff, 1992) (see appendix A.1). Gap penalties are scores subtracted from the total in case of insertions or deletions. These were set as 10 for the opening of a gap and 1 for the extension of a gap, which are standard values for BLOSUM62 matrices (Reese and Pearson, 2002).

[1] An amino acid match carries over four bits of information, while a nucleotide match carries only two. This means statistical significance of a matching segment can be reached for much shorter sequences.

	-	T	A	Y	R	D	G	A	G	I	W	T
-	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	7	0	0	0	0	0	0	2	0
R	0	0	0	0	12	2	0	0	0	0	0	1
G	0	0	0	0	2	11	8	0	6	0	0	0
N	0	0	0	0	0	3	11	6	0	3	0	0
G	0	0	0	0	0	0	9	11	12	2	1	0
I	0	0	0	0	0	0	0	8	7	16	6	0
W	0	0	0	2	0	0	0	0	6	6	27	17
T	0	5	0	0	1	0	0	0	0	5	17	32
W	0	0	2	2	0	0	0	0	0	0	16	22
V	0	0	0	1	0	0	0	0	0	3	6	16
I	0	0	0	0	0	0	0	0	0	4	0	6

Sequence 1: T A Y R D G A G I W T

Sequence 2: Y R - G N G I W T W V I

Figure 3.1: An example of an alignment matrix for a local alignment on two AA-sequences using the BLOSUM62 substitution matrix and a gap penalties of 10 and 1 for opening and extending, respectively. The red arrows indicate the path of back-tracing taken by the algorithm to find the resulting alignment, which is pictured on the right.

The score for each possible alignment between two AA-sequences $\mathbf{v} = v_1, v_2, \ldots, v_n$ and $\mathbf{w} = w_1, w_2, \ldots, w_m$ is stored in an alignment matrix $\mathbf{A}$ of size $n \times m$. The best possible alignment at each position is then gathered from the maximal score obtained by either insertion, deletion or (mis)match:

$$a_{i,j} = \max \begin{cases} a_{i-1,j-1} + BLOSUM62(v_i, w_j) \\ a_{i-1,j} - \sigma \qquad \text{(gap in w)} \\ a_{i,j-1} - \sigma \qquad \text{(gap in v)} \end{cases}, \tag{3.1}$$

with σ denoting the gap penalty (which is either 1 or 10 depending on opening or extension). The optimal alignment can then be found by the highest score in $\mathbf{A}$ (Smith and Waterman, 1981). An example of how this score is calculated for two sequences can be found in figure 3.1.

Alignments were all performed pairwise, meaning an optimal alignment was computed for every possible pair of sequences and stored in a score matrix. Since comparative analysis on identical sequences is redundant, duplicate proteins were excluded from the alignments. Annotations from omitted alignments (e.g. phage, protein type and bacterial host) were added onto the accession of their matching sequence. Similarity was assessed for 2591 unique phage lytic proteins, yielding a symmetric 2591×2591 score matrix $\mathbf{S}$.

Since alignment scores add up for every match, the optimal score for each pairwise alignment is highly dependent on the length of the conserved segment. Therefore, score matrix **S** was scaled:

$$\hat{s}_{i,j} = \frac{s_{i,j}}{\sqrt{s_{i,i}}\sqrt{s_{j,j}}}, \tag{3.2}$$

for each element $s_{i,j}$ in **S** where $0 \leq i, j \leq 2591$.
This brings about a 2591×2591 matrix $\hat{\mathbf{S}}$ of similarity scores between 0 and 1 (with 0 for completely different and 1 for identical sequences).

3.2 Clustering

To isolate highly conserved regions across a multitude of sequences, a Multiple Sequence Alignment (MSA) could be used. However, if employed directly on the entire set of sequences, it would yield a huge, vaguely related region with a lot of gaps. A more valuable approach would be to group similar sequences together and to carry out a MSA on the individual groups.

Gathering sequences into highly similar groups can be achieved through agglomerative clustering. This unsupervised machine learning technique starts from a partition of singleton data points and iteratively merges the mutually closest pairs into clusters until all data points have been joined into one (Müllner, 2011). This iterative merging also allows the algorithm to construct a dendrogram as a visual representation of the (dis)similarity, or distance, between data points (see figure 3.2). The scaled similarity $\hat{\mathbf{S}}$ was used in this case as a distance measure between sequences to build linkages upon. The result of this analysis set out as a clustered heatmap can be seen in figure 3.3. Some aspects of this figure will be discussed below, but readers are encouraged to scope out any other details from the full figure available in the digital appendix.

The most similar cluster of phage lytic proteins can be seen in the upper-left corner of this figure (cluster A). The 90 sequences to be found in this cluster all target a bacterial host in the Propionibacteriaceae family, the majority targets both *Propionibacteria* and *Cutibacteria*. The specific bacteriophages that encode the sequences in this cluster show less consistency apart from all being Siphoviridae.
The sequences in this cluster contain a C-terminal Amidase 2 domain (IPR002502) and an N-terminal domain of unidentified function and are mostly identical apart from some scattered point-mutations and a variable linker region between the two domains

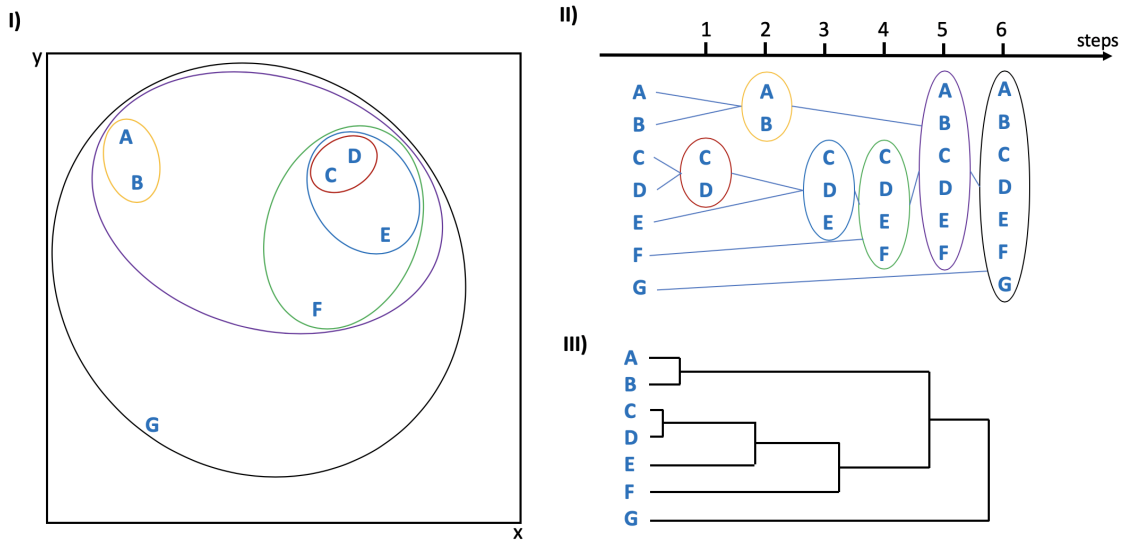


Figure 3.2: Schematic representation of how an agglomerative, hierarchical clustering algorithm joins datapoints and constructs a dendrogram. (I) Representation of data points A through G in a 2D-space at certain distances from each other. (II) At each step the pair closest to each other is merged into one cluster. (III) Based on the order of merging, data points and their individual distances from each other are visualised in a dendrogram.

(which was identified as such by DFLPred; (Meng and Kurgan, 2016)). As already mentioned in section 2.1, the high sequence similarity for these phage lytic proteins likely relates to the unique evolutionary constraints imposed by the habitat of the bacterial hosts they infect.

Perhaps one of the most interesting clusters in this figure is cluster B, as can be seen isolated in figure 3.4. While most sequences in this cluster have a scaled similarity of around 0.7, certain subgroups within it are almost exact copies. Furthermore, as can be seen in figure 3.3, almost no other sequences in PhaLP are similar to them. This cluster, containing only VAPGHs, targets bacteria from 11 different genera from the Enterobacterales clade. A single sequence targeting *Stenotrophomonas* (A gammaproteobacteria belonging to the order of Xanthomonadales) is also found in this cluster.

Although MSA of these sequences shows various segments that are unique to one or several UniProt entries, a query of the conserved domains within these sequences through NCBI's CD-Search tool (Marchler-Bauer and Bryant, 2004) shows evidence for only one domain, i.e. internal virion protein D (PHA00368[2]). Further analysis through DFLPred determined that the variable regions between conserved segments were likely linker segments, causing a break in the alignments, hence the separated subclusters. Since this cluster has variable linker regions as the sole source of sequence variation and is able to target such a large span of bacteria (including many known

[2]PHA00368 is not included in PhaLP/InterPro.

pathogens, e.g. *Escherichia*, *Klebsiella*, *Yersinia*, *Salmonella* etc.), it could be highly interesting as a broad-spectrum enzybiotic.

3.2.1 Evolutionary relations

The isolated dendrogram of the the cluster analysis can be seen in figure 3.6. This figure was further annotated with information gathered from PhaLP regarding protein

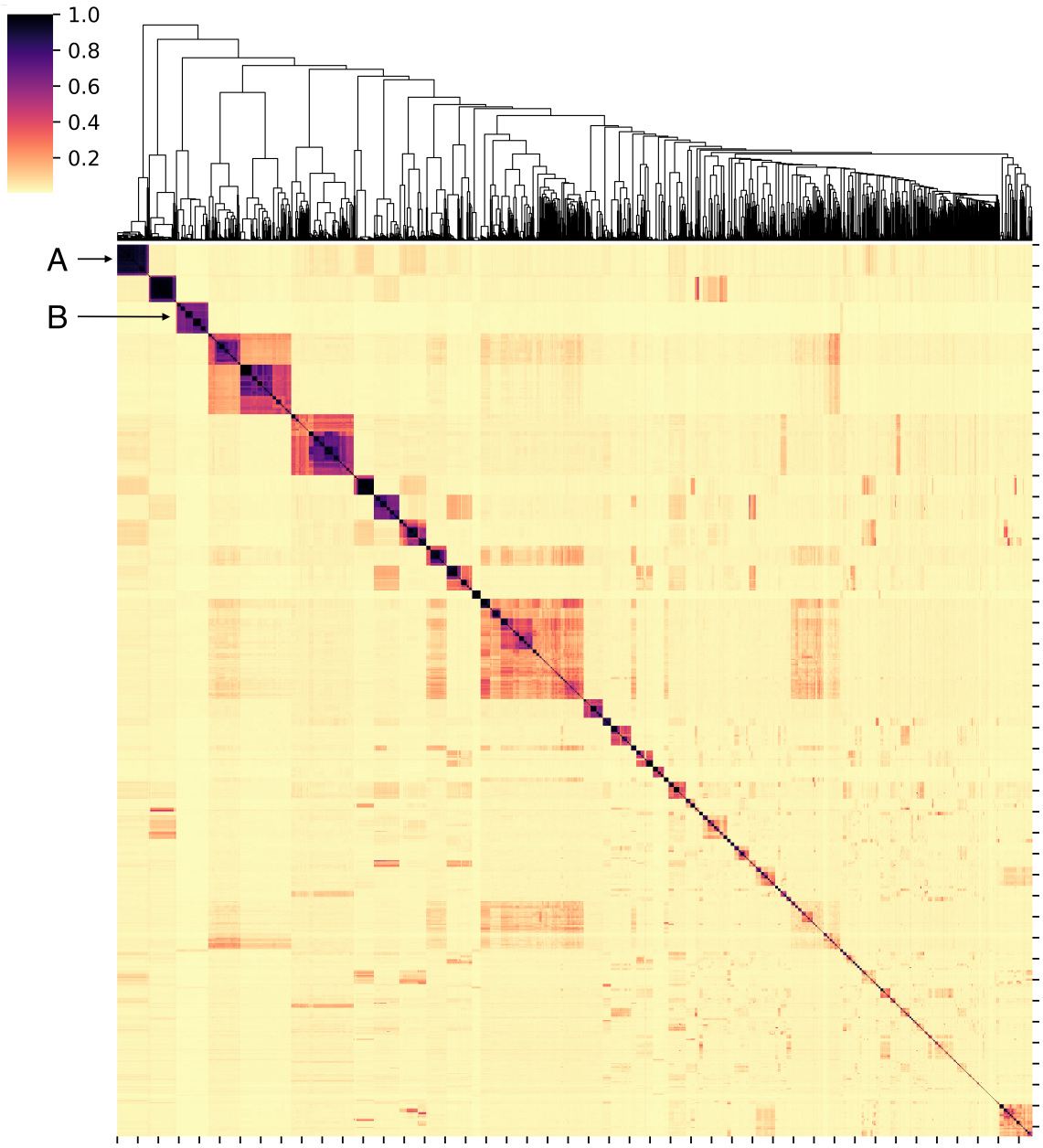


Figure 3.3: Heatmap of the score matrix $\hat{\mathbf{S}}$ hierarchically clustered through the simple nearest points algorithm (Müllner, 2011). The darker an area, the more similar the sequences of those phage lytic proteins. Sequences are identically ordered on both axes, resulting in perfect similarity on the diagonal. Due to overlapping labels at this size, the sequence accessions on the axis were left out of this figure. The fully annotated figure can be found in the digital appendix.

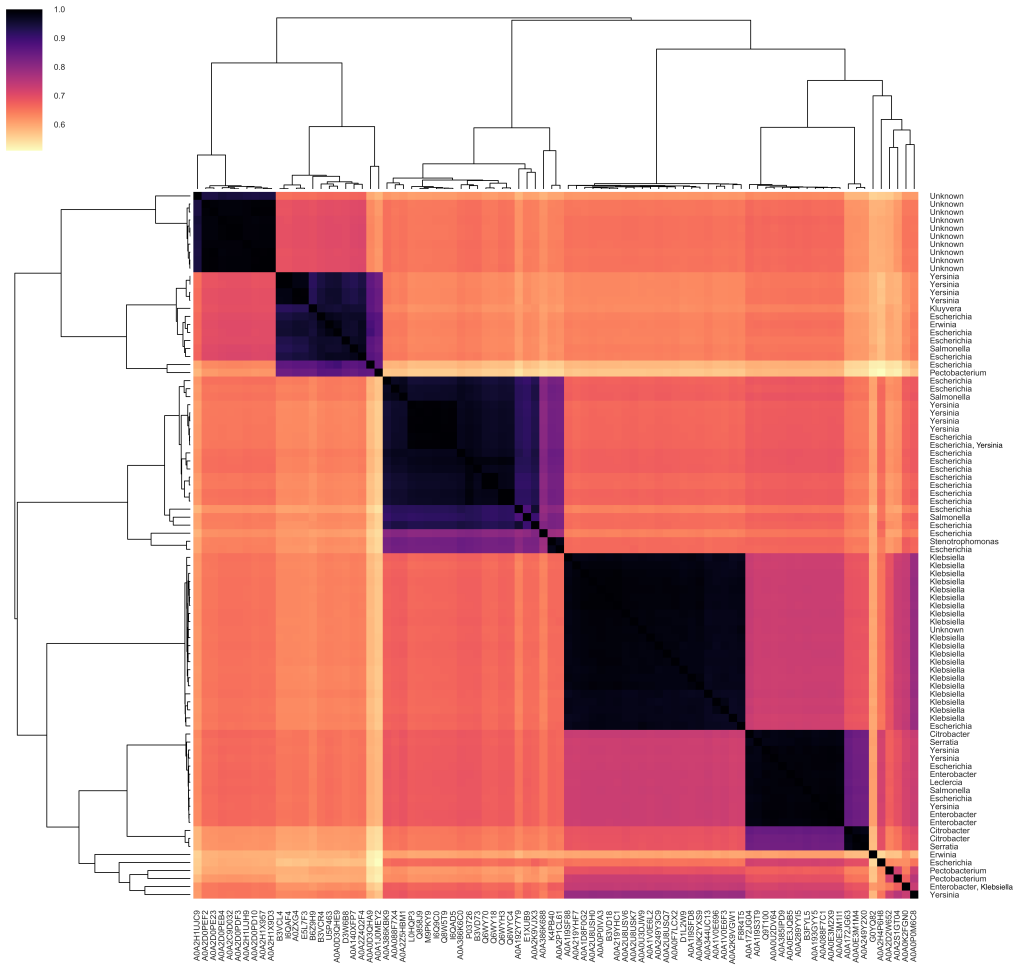


Figure 3.4: Close-up version of cluster B in figure 3.3 with bacterial host annotations on the y-axis where available.

type and bacterial host. One of the first things that can be noticed in this figure is that sequences targeting Gram-positive hosts almost never cluster together with sequences targeting Gram-negatives and vice versa. This separation between phage lytic proteins from phages infecting Gram-negative and Gram-positive host is also reflected in the domain composition and distribution analysis (see section 2.3.2) and is likely rooted in the same concept. While substrate specificity can vary within a family of conserved domains, the reaction chemistry is usually preserved (Todd et al., 2001). Therefore it is reasonable to assume that the differing enzymatic mechanisms of phage lytic proteins, for instance glycosidase activity and lysozyme activity (which occur more in Gram-positive and Gram-negative hosts, respectively (see figure 2.4)) would manifest themselves as separate clusters.

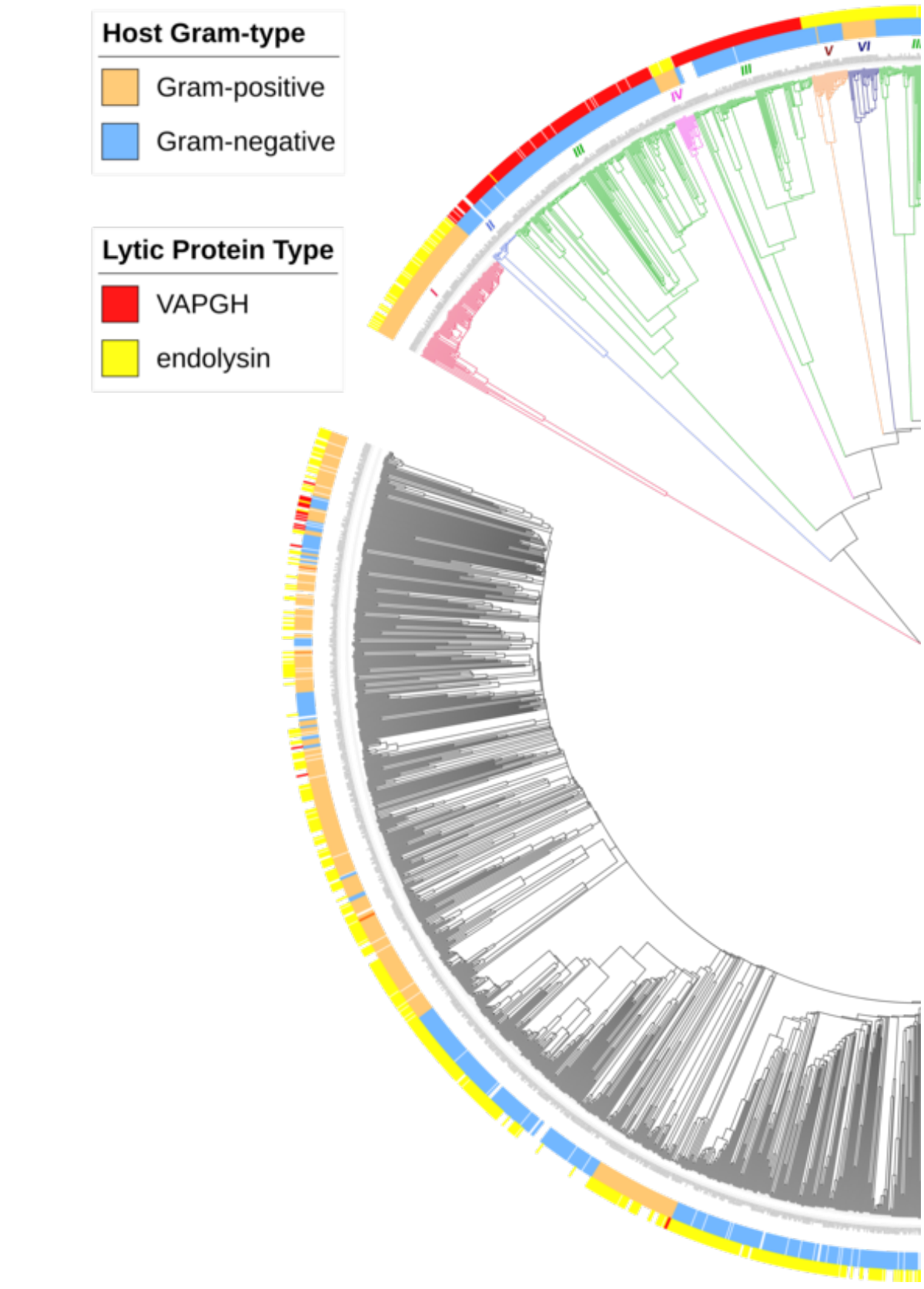
To a degree, the clusters of phage lytic proteins also correlate to certain taxonomic levels of the host. While many small clusters of high-similarity sequences that target a particular species can be explained away as variants that only differ in a couple of point-mutations. (e.g. clusters IV, VII and VIII), larger clusters typically relate to

specific clades as well (e.g. clusters I, III, VI, IX, X and XII). It is therefore probable that these derive from a common viral ancestor.

The clusters that target seemingly unrelated hosts can convey significant information as well. Cluster II, for instance, contains 24 phage lytic proteins of which some target *Pseudomonas aeruginosa*, which belongs to the phylum Gammaproteobacteria, while others target *Chryseobacterium*, a bacterium belonging to the Bacteroidetes phylum. Apart from its Gram-type and common habitats, the latter is not related to *Pseudomonas*. A MSA was performed for the sequences in this cluster and was illustrated in figure 3.5. It shows that the sequences from phages that infect *Chryseobacterium* contain an alternative startcodon. Combined with the fact that both these sequences contain the LT GEWL domain (cd00254), a domain that is very uncommon among Bacteroidetes (see figure 2.4), the occurrence of a horizontal gene transfer event could be hypothesised. Malki et al. (2015) examined some of these sequences in depth and determined that this was however not the case. It was discovered that two frame shift mutations had occurred in the phage genome coding for the sequence targeting *Chryseobacterium*. These frame shifts disturbed the coding regions of a putative minor head protein and a putative structural protein. According to the study, this phage, called ϕFenriz, can also infect *E. Coli ATCC 8739*, *Arthrobacter sp.* and *Microbacterium sp.*. The width of this host-range is unprecedented and could thus be interesting in the scope of a broad-range enzybiotic. The current hypothesis for this range proposed by Malki et al. (2015) is that this generalism is of benefit to phages inhabiting an oligotrophic environment.

Figure 3.6 also contains information on the types of lytic protein. Although PhaLP contains a lot more endolysins than VAPGHs, the VAPGHs all cluster towards relatively high similarities and only rarely group together with endolysins. Notable here is the separation for the clusters targeting Gammaproteobacteria (cluster III), the VAPGHs are arranged entirely separately from the endolysins.

Figure 3.5: Segments from the MSAs of both the AA-sequences and the encoding DNA of a few of the sequences from the Chryseobacterium/Pseudomonas cluster (cluster II in figure 3.6). The phage lytic protein targeting *Chryseobacterium* (A0A0P0ILQ2) is a lot shorter and has an alternative start codon, i.e. TTG. Malki et al. (2015) discovered that this protein was truncated due to two frame shift mutations upstream of the gene.


Tree scale: 1
Host Gram-type
Gram-positive
Gram-negative
Lytic Protein Type
VAPGH
endolysin
I
II
III
IV
V
VI

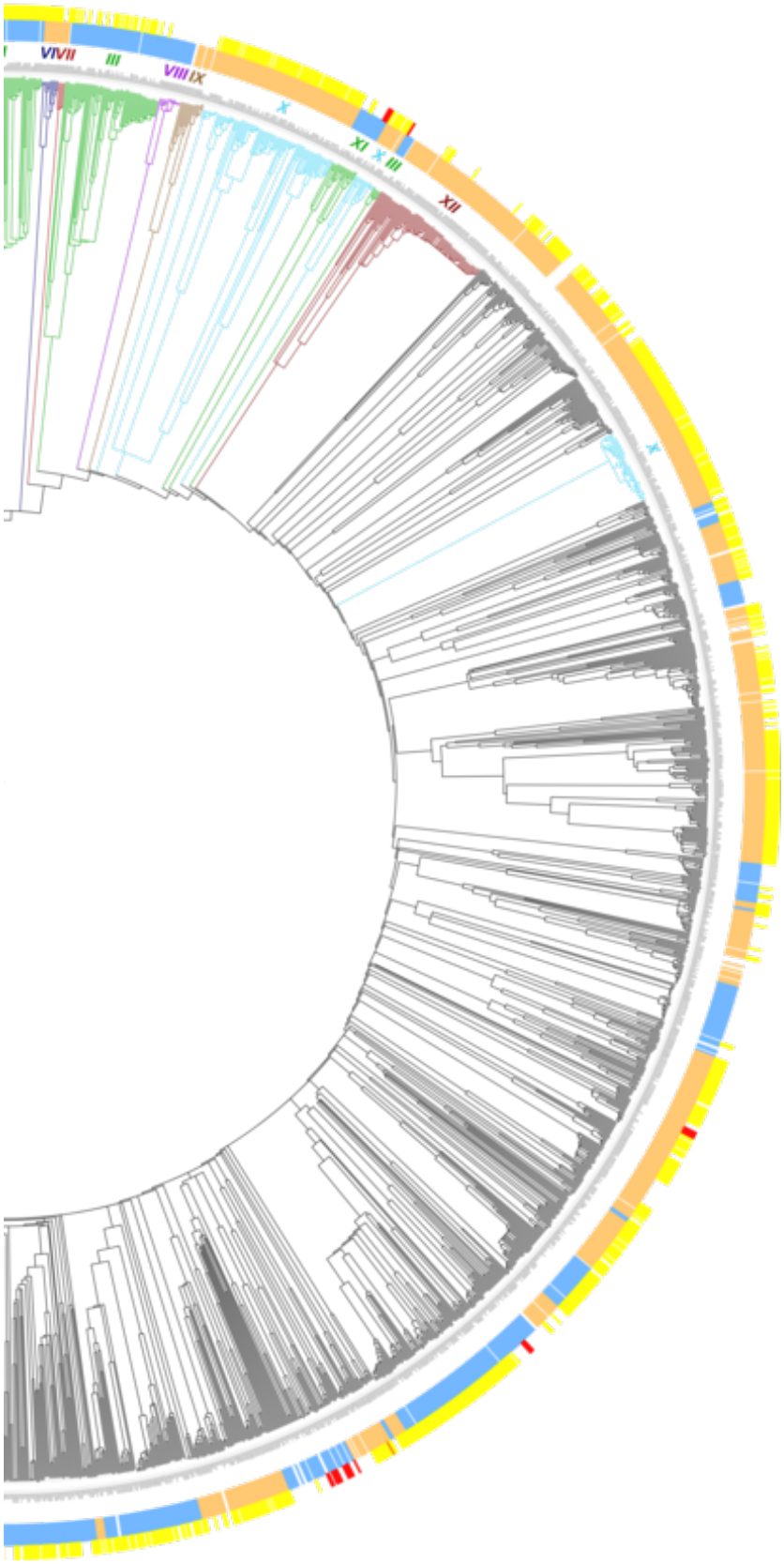

VI VII
III
VIII IX
X
XI
XIII
XII
X

Figure 3.6: Isolated dendrogram from the similarity clustering analysis annotated with protein-types and bacterial host information where available. The roman numerals refer to specific groupings of bacterial host that can be targeted: (I) Propionibacteriaceae, (II) *Chryseobacterium* & *Pseudomonas*, (III) Various Gammaproteobacteria, (IV) *Lactococcus lactis*, (V) Enterobacteriaceae & *Staphylococcus*, (VI) *Arthrobacter sp. ATCC 21022*, (VII) *Staphylococcus aureus*, (VIII) *Synechococcus sp. WH 7803*, (IX) *Bacillus*, (X) Mycobacteriaceae, (XI) *Escherichia* and (XII) *Streptococcus pneumoniae (Ik ga nog zoeken naar een mooiere manier om deze figuur te splitsen over 2 pagina's in Latex)*. This figure was made through the interactive Tree of Life tool (Letunic and Bork, 2019).

CHAPTER 4

INTERPRETABLE MACHINE LEARNING APPROACH

Machine Learning (ML) is the scientific study of computational and statistical models that seek patterns within large amounts of data (Hastie et al., 2017). This is often used in predictive modelling to infer an outcome or label for a given set of features after being trained on a dataset with known outcomes. This is called supervised learning and could, for instance, be used to predict the bacterial host that can be targeted given the set of domains present in a phage lytic protein (see section 4.1.2). These algorithms are often black box models, making it very difficult to grasp why certain predictions are made. However, through the use of interpretable ML models (Molnar, 2019), the design rules necessary for engineering a phage lytic protein with a desired characteristic, e.g. the ability to target a certain bacterial host, can be obtained. This could then be expanded to the engineering of an targeted enzybiotic. This chapter will focus on the feature engineering, model selection and optimal design of ML algorithms capable of inferring these design rules.

4.1 Supervised Machine Learning

Within ML, there are three approaches: supervised learning, unsupervised learning and reinforcement learning. They all look for patterns in the data, but the task at hand is different. Unsupervised learning methods, such as the agglomerative cluster analysis in chapter 3, find associations and group data based on the absence or presence of commonalities, independent of any label. Reinforcement learning is a branch of ML where appropriate actions within an environment are predicted based on the optimization of a reward function. This method will however not be further discussed in this research.

Supervised methods rely on a set of given input-output pairs to learn a function to label unannotated examples. Here, the data is generally first split into two sets, the training and the test set. The training set is a subset of which both input data

(features) and corresponding output (labels) are supplied to the algorithm. Based on this subset, a prediction model is built that conditions itself on the given input-output relations to learn to predict the outcome for new unseen inputs (Hastie et al., 2017). Predictions can then be made on the test set, which is kept separate from this learning process, to validate the accuracy of the predictions.

In this chapter, supervised ML methods will be used to predict host information given a phage lytic protein. The end goal here will not be the prediction task itself, but the extraction of the rationale behind each prediction. First, an appropriate model should be selected and features should be engineered to represent the phage lytic proteins to the model. These two steps are crucial, since they also impact the quality of interpretations to be extracted.

4.1.1 Feature engineering

Feature engineering is a fundamental step in the ML process where variables are constructed to describe a point in the dataset. For p features, this should be a p-dimensional vector for every phage lytic protein in the dataset, in this case a phage lytic protein. In certain cases, competent feature vectors can be generated directly from the input data through preprocessing functions, e.g. word or character counters for text-based inputs, but as model interpretation is often extracted from the values of features, this negates the possibility for meaningful interpretation (Molnar, 2019). Furthermore, using domain knowledge to construct appropriate features can usually improve upon the learning method (Hastie et al., 2017). To obtain the most accurate prediction, features should not only be relevant, but redundancy between them should be avoided for the sake of maximizing information over computational burden, i.e. the time and memory it takes to execute and store the model. Maximising the non-redundant information also reduces the risk of overfitting, which is when a model describes instances from the training set too strictly, causing it to not generalize well and fail to accurately perform predictions on test data (Hawkins, 2004).

As protein domains have demonstrated high correlation to a lytic protein's host spectrum (see chapters 2 and 3), a binary vector referencing the presence or absence of domains seems a natural choice for a feature vector. Additionally, this presence or absence is also a characteristic that can be altered in the lab by synthetic recombination (Gerstmans et al., 2018). The interpretation of a model that predicts host characteristics on this basis could thus be valuable in the recombination of a lytic protein into one with a desired characteristic. To optimise the amount of information, and in turn optimise the accuracy of the ML model, a few different feature sets were compared in redundancy and predictive power. The redundancy within a feature set

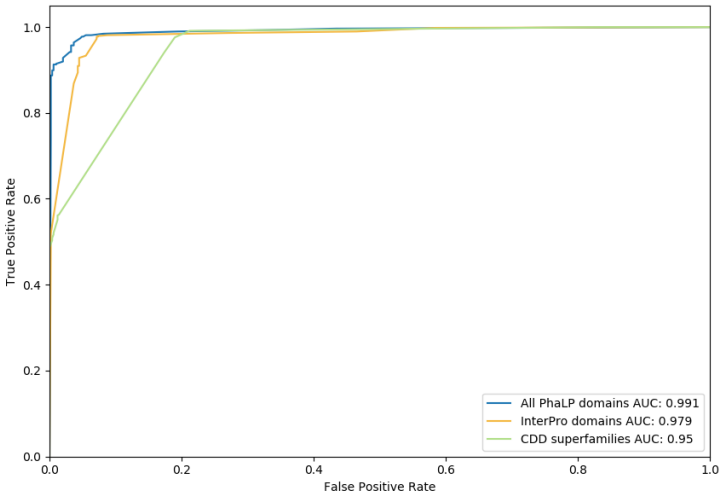


Figure 4.1: The ROC curves of three RF models using the three feature set described in section 4.1.1. The number in the legend refers to the AUC score of the individual models, a statistic often used to summarize the efficacy of a ML algorithm (see section 4.1.3).

was measured by the Mutual Information (MI) between features in a set. This is a measure of similarity between two features of the same data point. High MI thus signifies that two features explain the data in a very similar manner, meaning only a relatively low amount of information is gained from the inclusion of both features in the model (Pedregosa et al., 2011). The predictive power was evaluated by the capability of a standard Random Forest (RF) model to predict the Gram-type of phage lytic proteins based on each feature set.

The binary vector representing the absence or presence of the domains was based on a BLAST of the domains against a sequence. The significance of hits is quantified as the E-value, which is the expectation value of occurrence of the obtained pairwise alignment score (Pearson, 2013). These E-values were adopted from RPS-BLAST and significant hits under a chosen threshold were counted as present (Criel, 2017).

Three feature sets were considered. The first makes use of every domain found in the PhaLP database. A total of 180 relevant domains and 2591 unique sequences are taken into account. PhaLP collects domains from a myriad of different protein databases (Pfam, SMART, CATHDB etc.), hence some conserved domain accessions can describe the same protein domain. Accordingly, the MI between some features here is as high as 0.61.

The second set of features maps PhaLP domains onto CDD conserved domain superfamilies, which are sets of evolutionarily related single-domain models (DeWeese-Scott et al., 2010). While not every domain can be mapped onto a superfamily, mean-

ing loss of information, this guarantees a minimum of redundancy between features (maximum MI between features is only 0.29). This yields a 2591×41[1] feature matrix.

The final set of features makes use the broad integration of domains and sequence motifs into InterPro accessions (Mitchell et al., 2018). Although not every PhaLP domain has a corresponding InterPro accession, the majority of information is retained, while also decreasing redundancy. The maximal MI between features here is 0.58, which is lower than that in the PhaLP feature set, while almost doubling the size of the feature matrix to 2591×75[2].

Although arguments can be made for any of the three feature sets, ultimately the InterPro features were chosen. Even though the RF model predicting Gram-type based on these features does not perform quite as well as one based on all PhaLP domains (see figure 4.1), this feature set offers a middle ground between extremes of redundancy and predictive power that should prove more robust against extension to more than two classes of prediction. Additionally, InterPro offers clear descriptions and functional annotations to most of the domains included in their database, thereby facilitating the interpretation of the results. Domain annotations for some of the sources of PhaLP, e.g. PantherDB, are oftentimes sparse.

4.1.2 Model selection

Although there are some methods of extracting post-hoc interpretations from complex machine learning models (Ribeiro et al., 2016), the most straight-forward way of achieving interpretability is by using so-called interpretable models. Because of the low complexity of these algorithms, it is still possible to inspect model components directly, a characteristic called translucency (Molnar, 2019). The most common interpretable models rely on decision trees, linear or logistic regression or probabilistic classification. In this chapter, the bacterial host of a phage lytic protein will be predicted at different levels. As this is a categorical value, the models that will be discussed are all classifiers.

Naive Bayes

A highly interpretable, yet still powerful model can be found in the Naive Bayes classifier. This algorithm makes use of Bayes' theorem of conditional probabilities to calculate the probability of a class given a value of a certain feature. The Naive part

[1] Out of the 180 domains in PhaLP, 73 domains can be mapped onto 41 CDD superfamilies.
[2] Not all PhaLP domains have a linked InterPro accession and some have the same InterPro accession.

refers to the model's assumption of conditional independence of the features[3], as probabilities are calculated for each feature independently. Naive Bayes models the probability of a class C_k as:

$$P(C_k|\mathbf{x}) = \frac{1}{Z} \times PC_k \prod_{i=1}^{p} P(x_i|C_k)\,, \tag{4.1}$$

with $\mathbf{x} = [x_1, x_2, \ldots, x_p]$ a vector of values for p features and Z a scaling factor ensuring that the sum of the probabilities for all classes is normalised.

Despite the very straight-forward calculation of probabilities and consequent translucency of this model, it is not always interpretable on a global level. Inherently, any feature set larger than three dimensions is incomprehensible for the average person and sets of more than a few dozen features can no longer be held in our working memory simultaneously. Models like Naive Bayes can, however, be understood on a modular level, meaning the effect of a variation in value of a specific feature can be interpreted should all other values stay constant (Molnar, 2019).

Decision Rules

Linear models tend to fail in instances where features interact with each other, as can be the case for protein domains (Ponting and Russell, 2002). An interpretable alternative in that case is a decision rule classifier. Rule-based classifiers split the data in subsets based on feature cut-off values learned by the algorithm. The various cut-offs culminate in a specifically bounded region in the feature space which relates to a particular class of outcome. The prediction of a certain class then corresponds to an evaluation of whether or not an input is located in that region:

$$r_k(\mathbf{x}) = \prod_{i=1}^{p} I(x_i \in s_{ik})\,, \tag{4.2}$$

where I is the indicator function, yielding 1 if its argument is true and a 0 otherwise, x_i is the value of feature i and s_{ik} is a specified subset of the set of possible values x_i can take on $\mathbf{S_i}$. This base learner will return a 0 or 1 corresponding to whether class k is predicted or not (Friedman and Popescu, 2008).

Through the use of decision cut-offs, these types of classifiers are able to discretize the high-dimensional feature space into a set of interpretable "IF conditions THEN response" statements. Their predictive power borders on that of RF models, while

[3]Dependence of features does however not render this model ineffective as long as the dependences distribute evenly among the classes or cancel each other out (Zhang, 2005).

being more interpretable because they select the most relevant features and return them in a format that semantically resembles natural language (Molnar, 2019).

4.1.3 Performance metrics

Even though the interpretation of the predictions is the main goal here instead of the predictions themselves, it stands to reason that the better a model's predictive performance is, the more valuable its interpretations will be. To get an understanding of how well a certain model performs, various metrics are commonly used. These metrics are calculated based on the predictions made on the test set, which contains data that is new to the model and thus creates a reliable validation of predictive power. Metrics can however vary in the extent to which they accurately describe model performance. Some prominent performance metrics are discussed below.

Accuracy

The classification accuracy is simply defined as the fraction of correct predictions out of all the predictions made:

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \tag{4.3}$$

Where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives (see table 4.1). Although this intuitively seems like a good measure of performance in a single number, this metric is often not used due to the accuracy paradox. This phenomenon refers to the fact that a highly imbalanced dataset for which all predictions are made in the dominant class, will still have a high accuracy (Valverde-Albacete and Peláez-Moreno, 2014).

Precision

The precision is the fraction of correctly predicted instances from a certain class (true positives) out of all the instances that were predicted as this class (true positives and false positives) (see table 4.1):

$$\text{precision} = \frac{TP}{TP + FP}. \tag{4.4}$$

This metric, however, says nothing about how many instances belonging to a class were predicted as such.

Recall

The recall refers to the fraction of correctly predicted instances from a certain class (true positives) out of all the instances that actually belong to this class (true positives and false negatives):

$$\text{recall} = \frac{TP}{TP + FN}\,. \tag{4.5}$$

This metric, however, says nothing about the fraction of correct predictions for that class like the precision does. In a binary classification, the recall is also called the True Positive Rate (TPR) as it is equal to the ratio of correctly predicted positives over the total number of positive labels (see table 4.1).

Area Under the ROC Curve (AUC)

The AUC is a frequently-used metric measuring the discrimination, i.e. the ability of a learning algorithm to correctly classify instances in a binary problem. As the name suggests, it represents the area under a Reciever Operating Characteristic (ROC) curve, a curve which sets out the TPR, or recall, against the False Positive Rate (FPR) at various discrimination thresholds (e.g. figure 4.1). Analogously to the TPR, the FPR is the ratio of incorrectly predicted positives over the total number of negative labels (see table 4.1):

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{4.6}$$

$$FPR = \frac{FP}{N} = \frac{TP}{FP + TN}\,. \tag{4.7}$$

The higher the overall TPR is relative to the FPR, the more accurate the classification. This is then reflected in a large AUC. The difficult expansion for use in multi-class problems and an inability to assign weighted misclassification costs are however notable weaknesses to this metric (Halligan et al., 2015). Accordingly, AUC is generally only used in the early stages of model assessment.

Table 4.1: Confusion matrix of a binary classification problem. Instances that are correctly predicted are commonly referred to as true, while incorrectly predicted instanced are referred to as false. This nomenclature is often used to describe different performance metrics.

	Positive label	Negative label
Positive prediction	True Positives (TP)	False Positives (FP)
Negative prediction	False Negatives (FN)	True Negatives (TN)

f_1-score

The f_1-score, also called F-score or F-measure, is the harmonic average between precision and recall:

$$f_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4.8}$$

As this metric can be calculated for each class and subsequently weighted according to the amount of instances in that class to produce and average weighted f_1-score, it provides an overview of performance that is robust enough for the purposes of this research.

4.2 Gram-type prediction

Chapters 2 and 3 have shown that there are clear distinctions in the composition of phage lytic proteins targeting Gram-positive and Gram-negative bacteria. Therefore, it is expected to be possible to predict the Gram-types of the hosts of these proteins based on their domain composition.

Proteins with Gram-ambiguous hosts were removed from the data since only four instances of this type were present and this would therefore likely not generate any usable predictions. Furthermore, 180 proteins for which no host annotation was available, were also omitted. A feature set using 75 relevant InterPro domains for 2407 phage lytic proteins remained. The data was split into a train and a test set, with 80% going into the training set and 20% going into the test set. As the number of proteins of each Gram-type is not fully balanced within the dataset (1104 G- and 1303 G+), this split was stratified on Gram-type, meaning the fraction of each type remained equal in both train and test set.

A Naive Bayes model and a rule-based classifier were trained and tested on this dataset. As the feature values are binary (0 for absence of a domain and 1 for presence), a Naive Bayes model was chosen that calculates its conditional probabilities based on multivariate Bernouilli distributions. The smoothing hyperparameter α was optimized through a grid search on a 10-fold cross-validation. This means the training data itself was split into 10 subsections and was trained and tested on all subsections individually and results of the test predictions were averaged for each split to give a more accurate estimate of the model's general performance. This process was repeated for a range of values of α, the optimal one was ultimately determined to be 0.001.

	precision	recall	f1-score	support
Gram-negative	0.96	0.76	0.85	221
Gram-positive	0.82	0.97	0.89	261
micro avg	0.87	0.87	0.87	482
macro avg	0.89	0.86	0.87	482
weighted avg	0.89	0.87	0.87	482

(a) Performance statistics for the Bernouilli-distributed Naive Bayes model.

	precision	recall	f1-score	support
Gram-negative	0.96	0.90	0.93	221
Gram-positive	0.92	0.97	0.94	261
micro avg	0.94	0.94	0.94	482
macro avg	0.94	0.93	0.93	482
weighted avg	0.94	0.94	0.94	482

(b) Performance statistics for the Skope Rules decision model.

Table 4.2: Classification reports on the performances of the Naive Bayes and Skope Rules models for Gram-type prediction. These set out precision, recall and f_1-score for the prediction of each class and also provide the support, the number of instances of this class in the test set.

For the rule-based classifier, a python implementation called Skope Rules was used (Gardin et al., 2018). The minimum recall and precision necessary for a rule were both set as 0.75 and rules were generated up to 10 features in length.

The predictive potency of both models is somewhat similar (see table 4.2), although the Skope Rules algorithm performs a little better overall. The most probable explanation for the higher predictive power of the rule-based classifier, lays in the interactions that can occur between protein domains (Ponting and Russell, 2002). As a Naive Bayes model calculates conditional probabilities independently for all features, these cannot be taken into account. Decision-based classifiers can however take these into consideration, since the cut-off for each interacting feature only defines one boundary in the feature space. What this means in a purely biologic sense is that the interactions between domains in a phage lytic protein have only a minor effect on the Gram-type of the host it is able to target. This can be concluded from the fact that the Naive Bayes model, which only takes presence or absence of a domain into consideration, can still make accurate Gram-type predictions.

In order to grasp the rationale behind the predictions of the Naive Bayes model, the individual feature importances were extracted in the format of conditional probabilities $P(x_i|C_k)$ with x_i the value of the i-th feature and C_k the k-th class. The result of this extraction visualised as a heatmap can be seen in figure 4.2. This figure indicates which domains are the most significant in the prediction of each Gram-type, i.e. the domains whose presence is most indicative of a certain Gram-type.

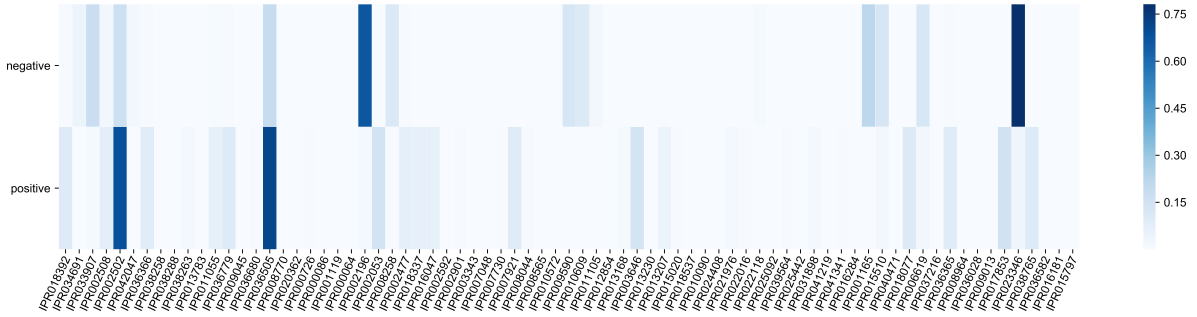


Figure 4.2: A heatmap of the conditional probabilities of each InterPro domain given a Gram-type.

For the Gram-negative bacteria, the most decisive domains are estimated to be lysozymes (Lysozyme-like SF: IPR023346 and Glyco hydro 24: IPR002196). This is quite plausible, since figure 2.4 indicated that these domains are very abundant in lytic proteins targeting Gram-negative bacteria, while only appearing rarely in those targeting Gram-positive bacteria.

For Gram-positive targeting proteins, the most significant domains identified by the Naive Bayes model are amidases (e.g. PGRP SF: IPR036505 and Amidase 2: IPR002502), although these appear to also be somewhat influential for Gram-negative lytic proteins. This is compatible with the results of the quantitative analysis, as both of these domains were abundant in proteins targeting both Gram-types, albeit more so for Gram-positives. Considering the overlap in occurrence and the fact that the model allocates its largest conditional probabilities for Gram-positive predictions to these domains, it makes sense that the precision for these predictions is marginally lower (see table 4.2a). The biggest drawback of interpretations on conditional probabilities is that no context is given, i.e. each feature importance is calculated independently.

For the rule-based model, the rules that define the base learners described in equation 4.2 can be extracted directly. These rules can combine multiple domains, but only incorporate AND-relationships, meaning every condition in the rule must be met for an instance to be classified in the class it defines. Different rules are completely separate, meaning these correspond to OR-relationships. The resulting precision and recall are also supplied for each rule individually. The top three performing rules (based on f_1-score) that the Skope Rules algorithm generates for the prediction of a Gram-positive host are as follows:

1. IF NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

precision: 0.954106
recall: 0.941597

2. IF NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR042047 THEN G+

   precision: 0.94856
   recall: 0.943302

3. IF NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR041219 and NOT IPR042047 THEN G+

   precision: 0.947929
   recall: 0.941246

Because a prediction on Gram-type is a binary classification and the set of domains is of reasonable size, most rules created at this level are based solely on the exclusion of particular domains. The top three rules above are thus solely based on the absence of the specified domains. What is surprising is the performance this achieves. The best performing rule culminates into an f_1-score of 0.95 based on the absence of the PGRP (IPR015510), Tail accessory factor GP4 (IPR020362), Lysozyme-like SF (IPR023346), Gp5 SF (IPR038288) and SleB 1 (IPR042047) domains. Conditions on domains that only appear very rarely in the dataset, e.g. SleB 1 and Gp5 SF, are most likely only included for the sake of maximising the performance metrics. Those that appear very abundantly, however, like the PGRP and the Lysozyme-like SF domains, convey a lot more meaning and can thus be of high interest in the synthesis of targeted enzybiotics at Gram-type level. The latter in this case were also determined to be valuable in the Naive Bayes analysis. The full list of rules generated by this algorithm can be found in appendix B.1.1. Remarkably, not a single one of these rules mentions a CBD that appears in more than one entry in PhaLP. Note that the inverse of the rules above should mostly hold true for a Gram-negative prediction, although performance will not necessarily be equal. A full list of rules generated specifically for Gram-negative host classification is included in appendix B.1.2.

4.3 K-mer approach

Although protein domains have been assumed as single functional entities in the analyses up until this point, single AAs can have an impact on a protein's function, as well. The extent of this impact, however, varies greatly. Through the use of interpretable ML models, it can be deduced which AAs have the most effect towards a certain pro-

tein characteristic. In theory, this information could then be used in to tailor a protein to a specific need through a rational design experiment (Lutz, 2010).

Table 4.3: Performance statistics for the k-mer-based Naive Bayes model.

	precision	recall	f1-score	support
M. smegmatis	1.00	0.95	0.97	93
Other	0.92	1.00	1.00	635
micro avg	0.99	0.99	0.99	728
macro avg	1.00	0.97	0.98	728
weighted avg	0.99	0.99	0.99	728

Accordingly, a Naive Bayes model was constructed to predict a particular host species for phage lytic proteins. To construct a feature set capable of conveying single AA information without losing interpretability, a count of the specific k-mers that occur within each protein sequence was used. K-mers refer to all possible substrings of length k that can be found in a string. The frequencies at which they appear in a certain sequence or genome has been extensively used as a unique microbial signature (Jiang et al., 2012; Siranosian et al., 2015; Wang et al., 2018), making for an easily constructed, relevant and highly interpretable feature set. A k-mer size of three was chosen. As sequences containing the unidentified AA 'X' were excluded from this analysis, there are 20 different characters that can appear in an AA-sequence, generating a feature vector of size 3638×8000.

In this example, the host label to be predicted was chosen as *Mycolicibacterium smegmatis* as (i) phage lytic proteins targeting this species are very abundant in PhaLP and (ii) because these lytic proteins showed high similarity in figure 3.5 and it would thus be interesting to find out if this level of conservation is caused by an increased fitness towards host targeting. The algorithm[4] can be altered to predict a species of choice, although performance may vary.

As the set of hosts is binarized into either *M. smegmatis* and any other host, the resulting dataset is rather imbalanced, i.e. 464 positives versus 3174 negatives. To somewhat mitigate this imbalance, the 80/20 train test split was stratified on the host label.

Due to the discrete nature of a count-based feature set, a Naive Bayes classifier based on a multinomial distribution was used. Once again the hyperparameter α was optimised through a grid search on a 10-fold cross-validation. Despite the imbalance in the dataset, the classifier is able to perform exceptionally well (see table 4.3), with only five lytic proteins for *M. smegmatis* being wrongly classified.

[4]This aglorithm is available at https://github.ugent.be/bw10master/2018_Taelman_Steff

The conditional probabilities of the features given a class computed by the Naive Bayes model are again used for interpretation. In this case however, not exactly the impact of the full features, i.e. the k-mers, are the most relevant for follow-up studies in rational design, but the individual AAs. To make this conversion, sequences were iterated over each AA they contain and the feature importances for all k-mers that AA is found in, are stacked (see figure 4.3).

Complete phage lytic protein sequences could then be visualised through a sequence logo, where the size of the character for each AA represents its importance in the prediction of *M. smegmatis* as the protein's host. Part of such a sequence logo generated for the UniProt accession 'Q856D3', a phage lytic protein targeting *M. smegmatis*, can be seen in figure 4.4. Interestingly enough, the important AAs seem to somewhat line up with the active site of Amidase 2 (IPR002502), the main EAD in this protein.

4.4 Hierarchical classification

The previous sections both describe flat classifications, i.e. a direct prediction on the Gram-type or species level, but antimicrobials are often desired for specific host-ranges. While this could be achieved by a multi-class model with a set of taxonomic clades as labels, it could be interesting to take into consideration the inherent hierarchy of bacterial taxonomy (see section 2.1). Not only could this improve the overall quality of the predictions, it enables interpretation on each level, allowing researchers to pick and choose a desired host-range for the synthesis of an enzybiotic.

As definitions vary, hierarchical classification in this research will refer to a classification approach that can cope with a pre-defined class hierarchy. This means there is only one root of the hierarchy and every other element within it can be defined through a "IS A" relationship. This type of relationship is asymmetric (e.g. every Gammaproteobacteria is a Proteobacteria, but not every Proteobacteria is a Gammaproteobacteria), anti-reflexive (i.e. the hierarchic tree flows in one direction

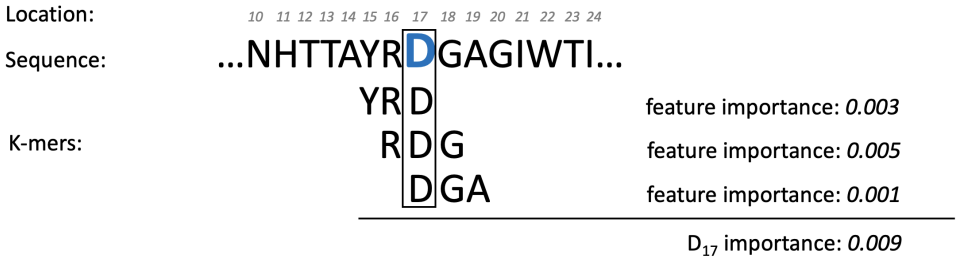


Figure 4.3: An example of how k-mer feature importances were stacked to achieve importance for each single AA in a protein sequence.

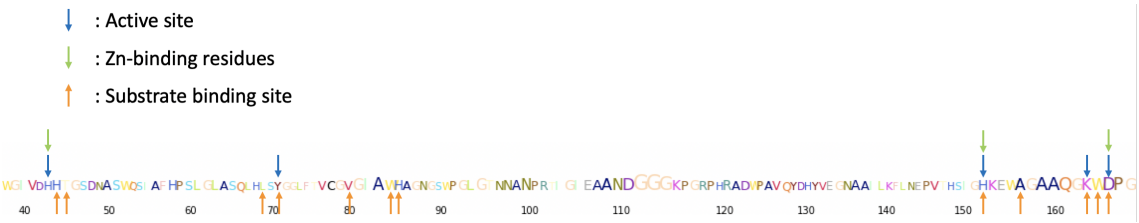


Figure 4.4: A sequence logo indicating AA-importance for host classification for the phage lytic protein with UniProt accession 'Q856D3'. Due to its size, this figure was cropped to only the Amidase 2 domain (IPR002502).

and does not loop) and transitive (e.g. if every Pseudomonadales bacteria is a Gammaproteobacteria and every Gammaproteobacteria is a Proteobacteria, then every Pseudomonadales bacteria is a Proteobacteria) (Silla and Freitas, 2011).

There are generally three types of models that can manage data with a pre-defined class hierarchy: flat classifiers, local classifiers and global classifiers. Flat classifiers predict the leaves of the hierarchic tree directly. This makes use of the asymmetrical "IS A" relationship to infer higher level classes from the predicted leaf class. Accordingly, this strategy is also called the bottom-up approach (Barbedo and Lopes, 2006). Local classifiers are modular combinations of several models that propagate the prediction throughout the tree from the top down. Several subtypes exist in this category. Global Classifiers, also called big-bang methods, consider the entire class hierarchy at once to make predictions (Silla and Freitas, 2011). Two distinct local classification approaches were chosen for their compatibility with the high-performant, interpretable models discussed in section 4.1.2.

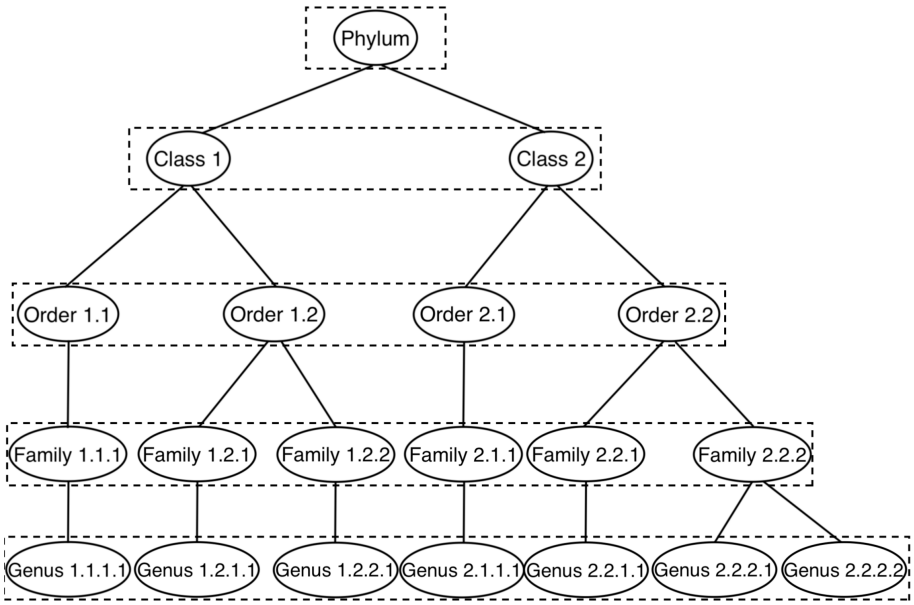


Figure 4.5: A schematic representation of a level-based local classifier for the prediction of bacterial taxonomy. The circles represent classes and the dashed rectangles represent multi-class classifiers. These models are trained separately and subsequently combined to make predictions.

4.4.1 Level-based local classifier

A level-based local classifier trains separate models on each level of the class hierarchy. In the case of classification on bacterial host, this means a separate model on phylum, class, order, family and genus level (see figure 4.5). To then combine the models, a prediction can be made top-down, iteratively picking the most probable prediction and restricting deeper levels to the child branches of the predicted class node (Silla and Freitas, 2011). As this method only presents a greedy optimization of prediction probability and risks propagating high-level errors downwards, a brute-force approach was chosen wherein probabilities are calculated for each leaf node based on all prediction probabilities preceding it. The prediction is then made based on the leaf with the highest probability (see section 4.4.1).

The level-based local classifier was built on five multivariate, Bernouilli-distributed Naive Bayes models. As there are more than two classes to be predicted at each level, the models in this case make multi-class predictions instead of binary ones. Bacterial clades for which less than 10 phage lytic proteins are known, were omitted from the dataset. Although this means the exclusion of the majority of hosts (see figure 4.6), the rationale behind this was that (i) classes with this little support are highly unlikely to be correctly classified and (ii) misclassification of those instances would bring down certainty of the predictions for other classes. The resulting dataset contains 2201 phage lytic proteins targeting 38 different bacterial genera. The training and testing was performed on an 80/20 split stratified on host genus as in section 4.2. The hyperparameter α was tuned through a grid search on a 10-fold cross-validation.

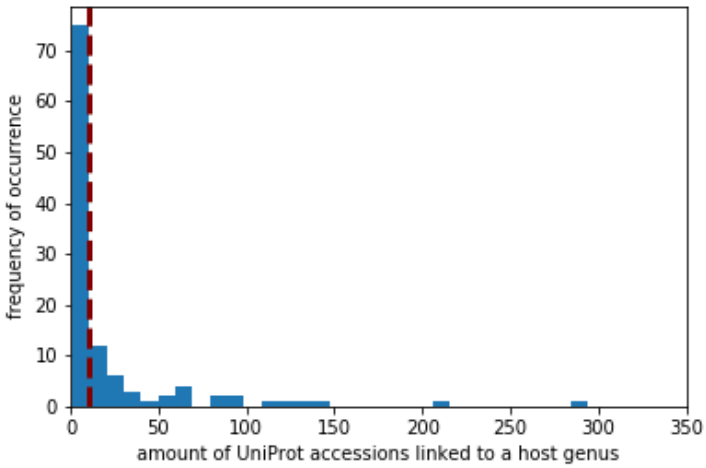


Figure 4.6: A histogram of the frequency of occurrence of different amounts of phage lytic proteins being targeting a bacterial genus. In the pursuit of maximization of the model's quality, all phage lytic proteins for which less than 10 are known to target a certain genus, were excluded from this classification.

Probability propagation

Each of the five models generates a matrix of prediction probabilities of size $n \times k$ where n is the number of instances in the test set and k is the number of classes to predict on that level. For each instance, this set of probabilities adds up to one. For example, in the order level model in figure 4.5, the probabilities could be: $P(\text{Order 1.1}) = 0.2$, $P(\text{Order 1.2}) = 0.35$, $P(\text{Order 2.1}) = 0.15$ and $P(\text{Order 2.2}) = 0.3$. To streamline these probabilities for a combination towards each leaf, the probabilities of these orders given their respective taxonomic class were calculated using Bayes' theorem (see appendix B.2). For the above example, this would mean:

$$\begin{aligned} P(\text{Order 1.1}|\text{Class 1}) &= P(\text{Order 1.1}) \times \frac{P(\text{Class 1}|\text{Order 1.1})}{P(\text{Class 1})} \\ &= 0.2 \frac{1}{0.2 + 0.35} \\ &= 0.36\,, \end{aligned} \tag{4.9}$$

where $P(\text{Class 1}|\text{Order 1.1}) = 1$ due to the asymmetrical nature of a taxonomic tree. Analogously, $P(\text{Order 1.2}|\text{Class 1}) = 0.64$, $P(\text{Order 2.1}|\text{Class 2}) = 0.33$ and $P(\text{Order 2.2}|\text{Class 2}) = 0.67$. Once probabilities given the parent nodes at each split are calculated, the genus probability is calculated by propagating downward for each branch, i.e. multiplying at each node in the path to a genus. For instance:

$$\begin{aligned} P(\text{Genus 1.1.1.1}) &= P(\text{Genus 1.1.1.1}|\text{Family 1.1.1}) \times P(\text{Family 1.1.1}) \\ &= P(\text{Genus 1.1.1.1}|\text{Family 1.1.1}) \times P(\text{Family 1.1.1}|\text{Order 1.1}) \\ &\quad \times P(\text{Order 1.1}) \\ &= P(\text{Genus 1.1.1.1}|\text{Family 1.1.1}) \times P(\text{Family 1.1.1}|\text{Order 1.1}) \\ &\quad \times P(\text{Order 1.1}|\text{Class 1}) \times P(\text{Class 1}) \\ &= P(\text{Genus 1.1.1.1}|\text{Family 1.1.1}) \times P(\text{Family 1.1.1}|\text{Order 1.1}) \\ &\quad \times P(\text{Order 1.1}|\text{Class 1}) \times P(\text{Class 1}|\text{Phylum}) \times P(\text{Phylum})\,. \end{aligned} \tag{4.10}$$

An example of how these probabilities propagate through the taxonomic tree to make a prediction is illustrated in figure 4.7.

For every phage lytic protein in the test set, the genus with the highest prediction probability is chosen. Some phage lytic proteins can however target multiple hosts (see section 2.1). The algorithm will therefore count all genera that can be targeted by a lytic protein as correct predictions. The classification report on this prediction can be seen in table 4.4.

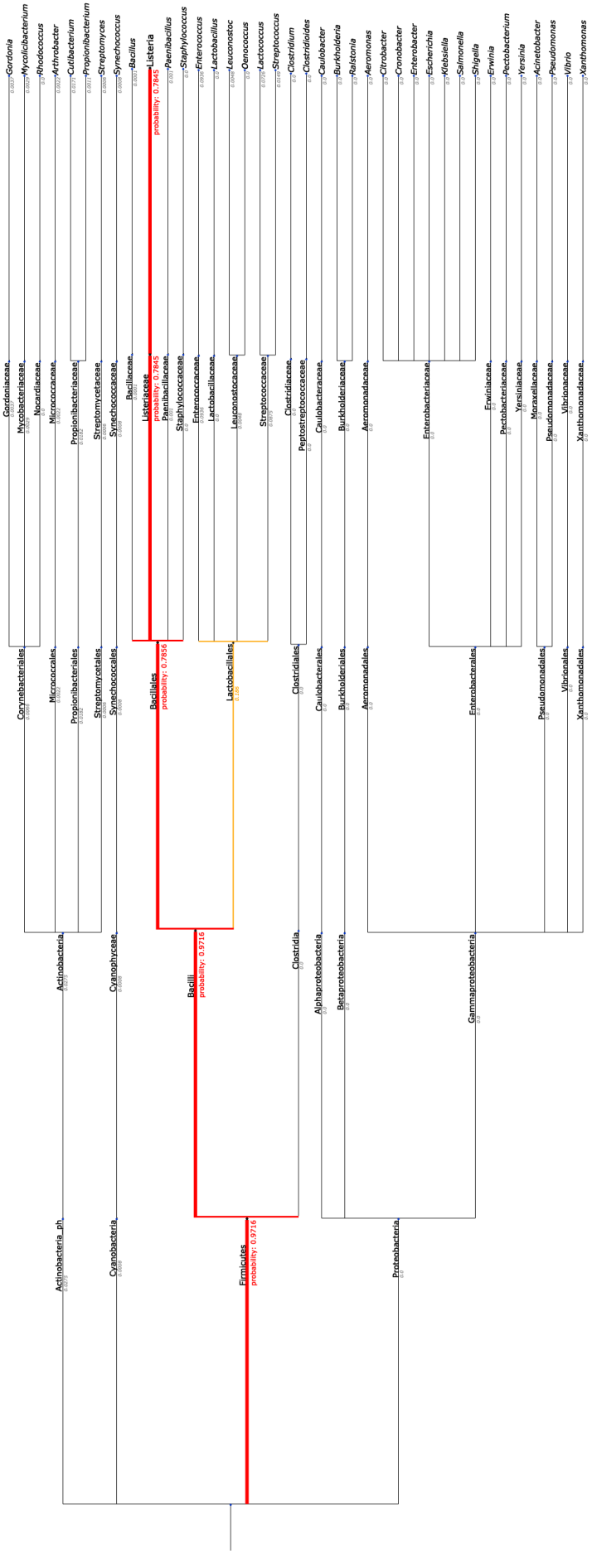


Figure 4.7: The prediction of the bacterial host of the phage lytic protein with the UniProt accession 'A0A0B5CTW5' through level-based local classification with probability propagation. Prediction probabilities are annotated in italics for each clade. The red line indicates the predicted taxonomy, orange lines signify other high probability (P > 0.1) clades in case of ambiguity.

Table 4.4: Classification report of the level-based local classification of phage lytic proteins into 38 different bacterial host genera. These set out precision, recall and f_1-score for the prediction of each class and also provide the support, the number of instances of this class in the test set.

	precision	recall	f_1-score	support
Acinetobacter	0.00	0.00	0.00	5
Aeromonas	0.16	0.78	0.27	9
Arthrobacter	0.50	0.08	0.14	12
Bacillus	0.79	0.54	0.64	28
Burkholderia	0.00	0.00	0.00	3
Caulobacter	0.00	0.00	0.00	3
Citrobacter	0.00	0.00	0.00	5
Clostridioides	0.33	1.00	0.50	3
Clostridium	0.33	0.40	0.36	5
Cronobacter	0.00	0.00	0.00	3
Cutibacterium	0.17	1.00	0.29	17
Enterobacter	0.00	0.00	0.00	2
Enterococcus	0.25	0.12	0.16	8
Erwinia	0.00	0.00	0.00	2
Escherichia	0.26	0.19	0.22	59
Gordonia	0.83	0.77	0.80	13
Klebsiella	0.00	0.00	0.00	17
Lactobacillus	0.41	0.85	0.55	13
Lactococcus	0.00	0.00	0.00	25
Leuconostoc	0.00	0.00	0.00	2
Listeria	1.00	0.25	0.40	4
Mycolicibacterium	0.67	0.24	0.35	42
Oenococcus	0.00	0.00	0.00	3
Paenibacillus	0.00	0.00	0.00	4
Pectobacterium	0.00	0.00	0.00	4
Propionibacterium	0.00	0.00	0.00	2
Pseudomonas	0.21	0.26	0.23	19
Ralstonia	0.00	0.00	0.00	3
Rhodococcus	1.00	0.40	0.57	5
Salmonella	0.17	0.10	0.13	20
Shigella	0.00	0.00	0.00	8
Staphylococcus	0.84	0.81	0.82	26
Streptococcus	0.90	0.78	0.84	23
Streptomyces	0.19	0.36	0.25	14
Synechococcus	0.88	0.64	0.74	11
Vibrio	0.00	0.00	0.00	10
Xanthomonas	0.03	0.50	0.06	2
Yersinia	0.00	0.00	0.00	7
micro avg	0.34	0.34	0.34	441
macro avg	0.26	0.26	0.42	441
weighted avg	0.38	0.34	0.33	441

Table 4.5: Classification report of flat classification of phage lytic proteins into 38 different bacterial host genera. These set out precision, recall and f_1-score for the prediction of each class and also provide the support, the number of instances of this class in the test set.

	precision	recall	f_1-score	support
Acinetobacter	0.00	0.00	0.00	5
Aeromonas	0.00	0.00	0.00	9
Arthrobacter	0.00	0.00	0.00	12
Bacillus	0.55	0.61	0.58	28
Burkholderia	0.00	0.00	0.00	3
Caulobacter	0.00	0.00	0.00	3
Citrobacter	0.00	0.00	0.00	5
Clostridioides	0.00	0.00	0.00	3
Clostridium	0.00	0.00	0.00	5
Cronobacter	0.00	0.00	0.00	3
Cutibacterium	0.18	0.94	0.30	17
Enterobacter	0.00	0.00	0.00	2
Enterococcus	0.67	0.25	0.36	8
Erwinia	0.00	0.00	0.00	2
Escherichia	0.29	0.73	0.42	59
Gordonia	0.91	0.77	0.83	13
Klebsiella	0.00	0.00	0.00	17
Lactobacillus	0.34	0.77	0.47	13
Lactococcus	0.00	0.00	0.00	25
Leuconostoc	0.00	0.00	0.00	2
Listeria	0.00	0.00	0.00	4
Mycolicibacterium	0.45	0.24	0.31	42
Oenococcus	0.00	0.00	0.00	3
Paenibacillus	0.00	0.00	0.00	4
Pectobacterium	0.00	0.00	0.00	4
Propionibacterium	0.00	0.00	0.00	2
Pseudomonas	0.00	0.00	0.00	19
Ralstonia	0.00	0.00	0.00	3
Rhodococcus	1.00	0.80	0.89	5
Salmonella	0.00	0.00	0.00	20
Shigella	0.00	0.00	0.00	8
Staphylococcus	0.77	0.88	0.82	26
Streptococcus	1.00	0.70	0.82	23
Streptomyces	1.00	0.07	0.13	14
Synechococcus	1.00	0.45	0.62	11
Vibrio	0.17	0.40	0.24	10
Xanthomonas	0.00	0.00	0.00	2
Yersinia	0.00	0.00	0.00	7
micro avg	0.37	0.37	0.37	441
macro avg	0.22	0.20	0.52	441
weighted avg	0.34	0.37	0.31	441

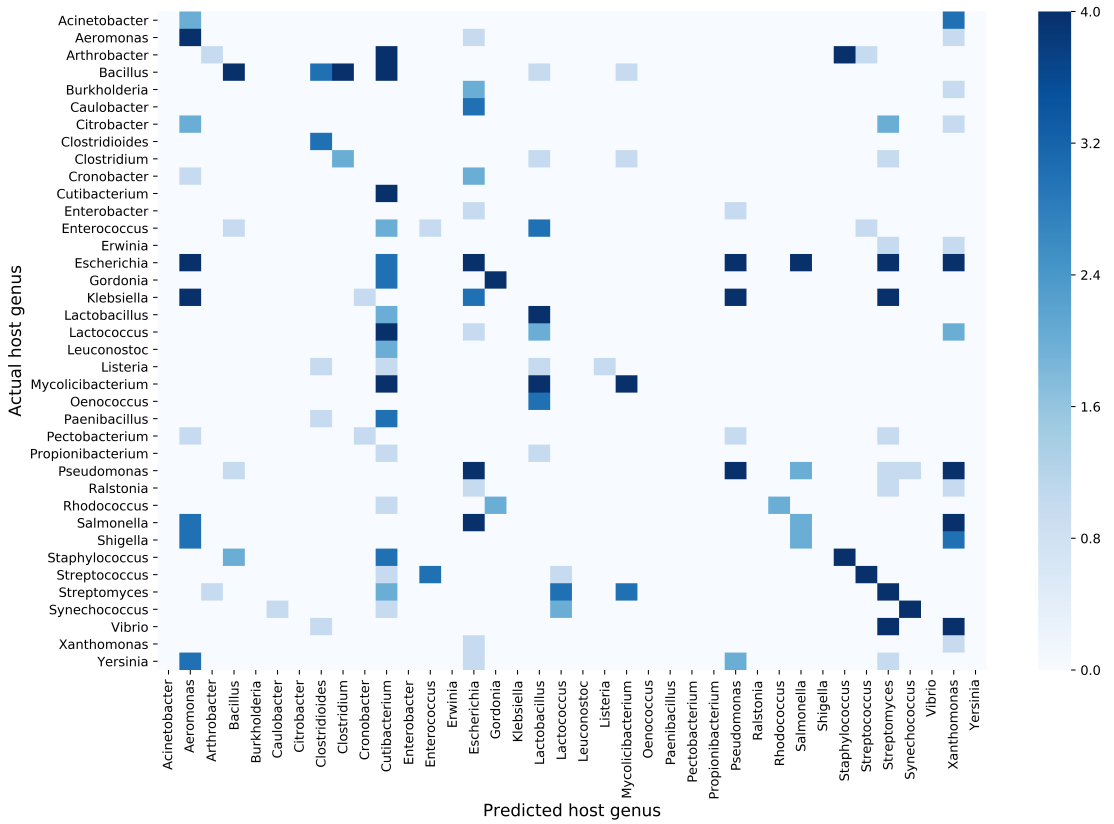


Figure 4.8: A confusion matrix for the level-based local classification classification of phage lytic proteins into 38 different bacterial host genera. This kind of plot sets out the actual hosts of the instances in the test set versus the hosts that were predicted for those instances.

Performance-wise, this model seems on average only slightly better than a flat model, i.e. a model directly classifying the lowest hierarchical level (see tables 4.4 and 4.5). In the confusion matrices in figures 4.8 and 4.9, it can be seen that both models make frequent incorrect predictions in the *Cutibacterium* and *Escherichia* genera, likely when average prediction probabilities are quite low. This results in a weighted average f_1-score of 0.33. At first glance, this appears low, but the model makes a prediction in 38 classes. A random classification of the same number of classes would only generate an average f_1-score of 0.03. The disadvantage of this type of layered hierarchical classification is that the lower level models have increasing numbers of classes to predict and will thus be only marginally effective. This problem is further exacerbated if some of the higher level models are already error-prone, as the low prediction probabilities of misclassified instances will trickle down.

Interpretation

In a similar way to section 4.2, interpretations can be extracted from the Naive Bayes models through the conditional probabilities $P(x_i|C_k)$. In figure 4.10, the top three

deciding domains and their respective probabilities are plotted for each split in the hierarchy. The PGRP SF domain(IPR036505) appears to be the most important domains in most branches of this figure, except in the Proteobacteria, where it no longer makes the top three. In Proteobacteria, the most crucial domain by far is the Lysozyme-like SF (IPR023346). These results are congruent with those from the quantitative analysis in section 2.3.

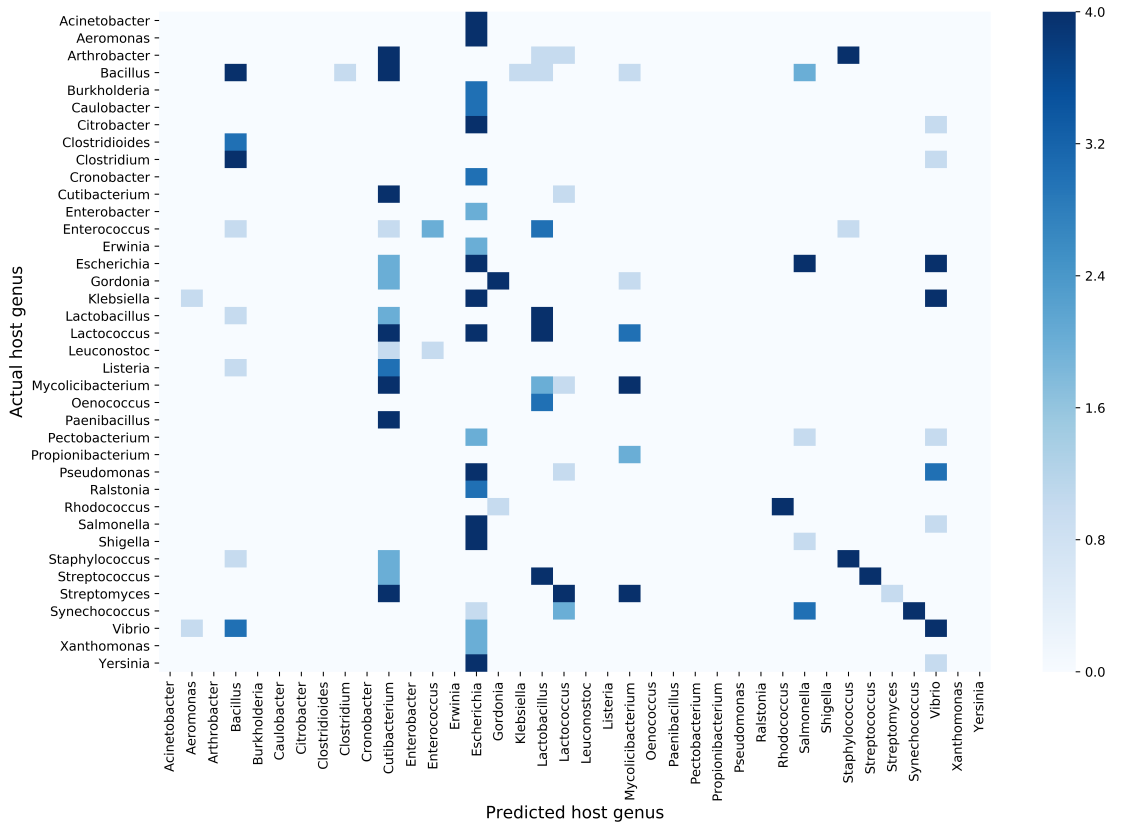


Figure 4.9: A confusion matrix for the flat classification of phage lytic proteins into 38 different bacterial host genera.

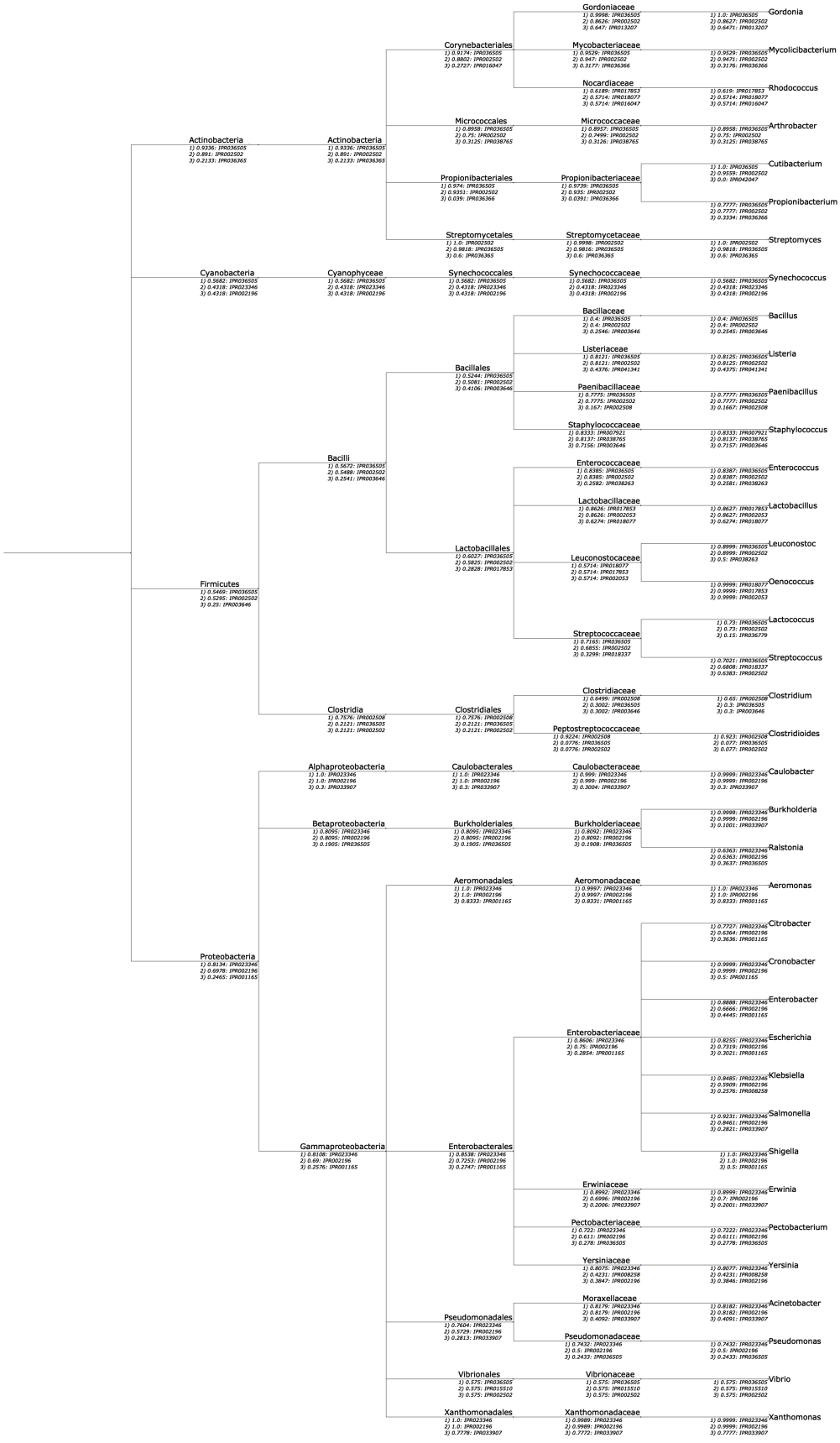


Figure 4.10: The taxonomic tree of bacterial hosts of phage lytic proteins with for each split, the top three decisive protein domains and their probabilities as calculated by a Naive Bayes model.

4.4.2 Parent node local classifier

An even more localized approach to hierarchic classification is by parent node local classifiers. This strategy employs a separate model at each point in the class hierarchy where a node splits into two or more classes (see figure 4.11). Accordingly, the further down the tree, the less data the model will be trained and validated on. This method enjoys the advantages of a limited computational strain while still utilising very specialised classifiers.

The Skope Rules algorithms described in section 4.1.2 were used as parent node classifiers. Skope Rules is, however, strictly a binary classifier. Since some nodes in the class hierarchy have more than two child branches, a One-Versus-All mechanism was implemented at these nodes. This means that for every child branch, a classifier is built to distinguish that child branch from all other branches (see figure 4.12). Analogously to the level-based local classifier, low-abundance hosts, i.e. genera for which less than 10 unique lytic proteins are known, and instances without known hosts were omitted from the data. The minimum precision and recall necessary for a rule to be created was set at 0.5 and rules of up to 5 conditions were generated.

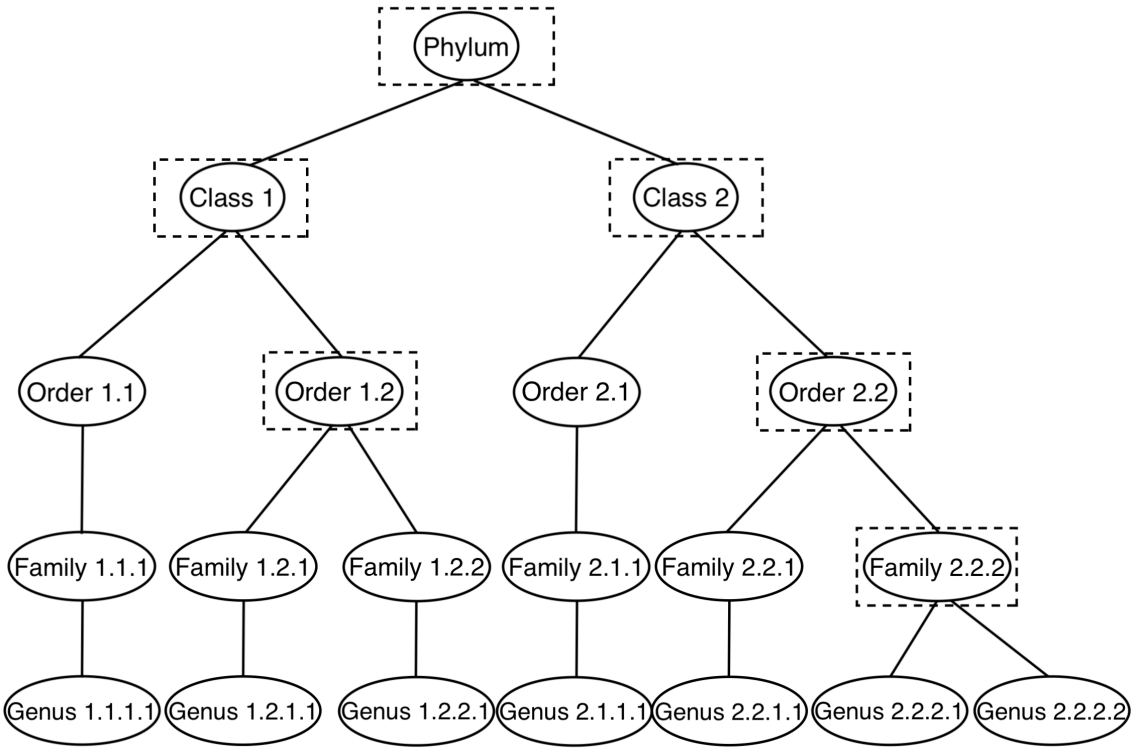


Figure 4.11: A schematic representation of a parent node local classifier for the prediction of bacterial taxonomy. The circles represent classes and the dashed rectangles represent multi-class classifiers.

The decision rules can be extracted from the algorithm for each branch separately, generating a comprehensive path of rules for each leaf (see figure 4.13). To facilitate easy interpretation, domains that should not occur in a lytic protein for it to be classified in that branch were crossed out. Branches without printed rules either didn't have a classifier built due to fact that the parent node only has one child and no split occurs or it was impossible to generate a rule with a precision and recall above the threshold. This is mostly the case at higher taxonomic levels, since these can contain very diverse bacteria with many different PG-types and cell wall compositions, requiring a more diverse set of phage lytic proteins to target all hosts in this group. To supply a certain degree of quality assurance, rules per branch are ranked on performance (f_1-score).

Due to the thresholds set for inclusion in this figure, some branches do not have associated domains. This low performance can be caused by the fact that these genera include many different species. The sequences of the phage lytic proteins targeting some of these species are highly similar (see for instance the *Synechococcus sp. WH 7803* cluster in figure 3.5), which could point to highly specialised architectures on species and even subspecies level. This would bring down their f_1-score in rule-based methods and would lead to them being overshadowed by more universally present domains in probabilistic classifiers. The results will be discussed in more depth in chapter 5.

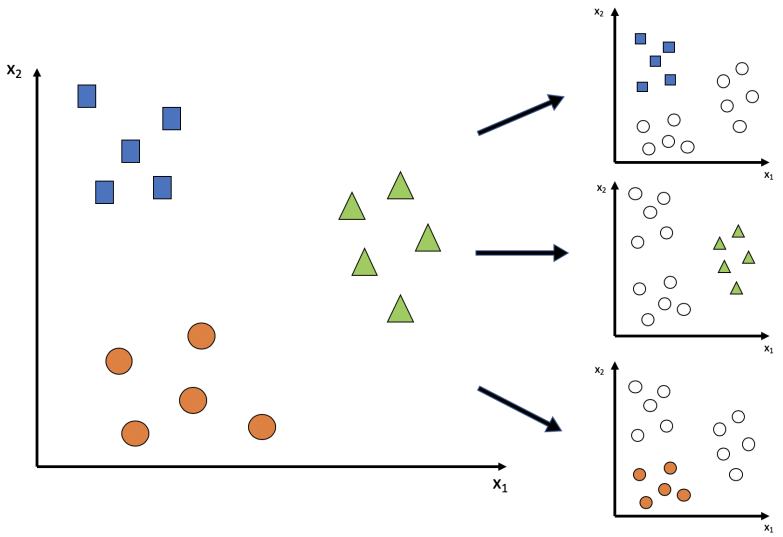


Figure 4.12: A schematic representation of how One-Versus-All classifiers can be used to break down multi-class problems into several binary problems. The axes represent two features x_1 and x_2. The algorithm will always model the distinction between one class and all other data points, e.g. the blue squares versus all other instances. Afterwards the models can be combined into one set of predictions.

Figure 4.13: The taxonomic tree of bacterial hosts of phage lytic proteins with for each split, the a set of rule-defining protein domains as calculated by a Skope Rules model. They are ranked on the f_1-scores they generate for instances belonging to that branch. InterPro domain accessions that are struck through are domains that should not occur in a lytic protein for it to be classified in that branch. Regularly displayed domain accessions should occur.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

We have shown that nature has numerous cell wall binding and degrading domains available. Nature has combined those in various compositions, as described in the PhaLP database. While an almost infinite number of combinations of these domains could be created through the natural evolutionary process of horizontal transfer, natural selection has withheld only a limited number of combinations, dependent on the phage and its host. In this work, the existing variation in PhaLP is described and subsequently used in interpretable ML models to map the design rules that have been established throughout natural evolution and selection. Despite the relatively small amount of data and often incomplete annotation on this subject, some promising observations and prospectives can be put forward.

5.1 Domain composition

The results from the quantitative analyses in chapter 2 indicate a clear relation between the types of domains in a phage lytic protein and the host this is able to target. CBDs were shown to be the most common in Gram-positive-targeting lytic proteins, although they were also found in some proteins targeting Alphaproteobacteria, Pseudomonadales and even *Escherichia*. These occurrences were, however, infrequent or involved in hosts with only very few known lytic proteins and should thus be further investigated in depth. All other Gram-negative hosts were found to be targeted by strictly EAD-containing proteins.

Out of the EADs, lytic protein architectures for Gram-positive hosts seemed to be mostly based on the action of PGRP (IPR036505, IPR006619 & IPR015510) and Amidase 2 domains (IPR002502). The CWGs that occur here are mostly restricted to N-acetylglucosaminidases. Many of the architectures for Gram-negative hosts also contain the PGRP and Amidase 2 domains, but more frequent are lysozyme and trans-

glycosylase domains. Apart from the occurrence of CHAP domains (IPR007921) in *Staphylococcus*, CWPs appeared to be generally uncommon.

5.2 Host-ranges

Regarding host-ranges, some interesting overlap is found in the exploratory analyses in section 2.1 and the cluster analysis in section 3.2. Globular endolysins targeting *Propionibacterium* and *Cutibacterium acnes* indicate high similarity, likely due to the lipid-rich anaerobic environment in which their bacterial hosts reside (Marinelli et al., 2012). These sequences could prove useful in an antimicrobial compound targeting these pathogens on the human skin and this also demonstrates the potential of similar analyses based on environment rather than host. Additionally, figure 4.13 also indicates some candidates for further distinguishing between *Propionibacterium* and *Cutibacterium*.

VAPGHs have shown a particular potential as broad-spectrum antibacterials. The VAPGHs encoded by Enterobacteria phage PRD1 demonstrate a wide range of hosts from the Gammaproteobacteria and as illustrated in figure 3.6, VAPGHs targeting Gammaproteobacteria in general have relatively similar sequences. Other interesting VAPGHs are those found in cluster II of this figure. Some phages from which these lytic proteins emanate display unprecedented width in host-range, with the ϕFenriz phage even crossing Gram-type. The standing hypothesis here was again due to the conditions in which this phage operates (Malki et al., 2015).

The lytic proteins targeting *Synechococcus* demonstrate an especially narrow host-range. High similarity was found among sequences targeting the *WH 7803* strain. These sequences contain the peculiar peptidase C70 domain (IPR022118), a domain not included in the initial quantitative analyses on the basis of its lack of confirmed GO-terms for cell wall hydrolysis. The SleB domains (IPR011105 & IPR042047) were also found exclusive to *Synechococcus* hosts, which is reflected in the conditional probability of IPR042047 in the *Synechococcus*-branch of figure 4.10.

5.3 Design rules

The agglomerate of the interpretable ML analyses can be found in figure 5.1. This figure displays a guideline of domain composition in natural phage lytic proteins based on the data assembled in PhaLP. The non-crossed out symbols represent domains that were determined by the level-based or parent node local classifiers to be highly

correlated to their particular branch. This was visualised for domains with conditional probabilities of the Naive Bayes classifier above 0.8 or an f_1-score equal to or higher than 0.95 in the Skope Rules model. Furthermore, domains required by the parent node model to be absent for classification were cross-referenced with the results from the quantitative analysis (see section 2.3) and those absent in every downstream branch were signified by a crossed-out symbol.

Many of the unannotated branches in this figure belong to Gram-negative bacteria, e.g. *Synechococcus*, *Pseudomonas* and *Vibrio*. Due to the highly conserved PG composition in these bacteria (see section 2.2.1), it could also be possible that the domain composition of phage lytic proteins targeting these hosts is more determined by other factors. For instance, peptidoglycan composition has been observed to vary within species and even strains due to environmental factors and growth conditions (Schleifer and Kandler, 1972; Schleifer et al., 1976). Although this variation is generally low, this could mean that domain architecture is less directly associated with the bacterial host, causing performance of the ML models to drop.

Because the mechanism behind the annotation in figure 5.1 is known, some interpretations can still be made for the branches without explicit domains assigned. In figure 4.13, the Skope Rules model points to distinctions between *Cutibacterium* and *Propionibacterium* largely made on groups of glycoside hydrolases (IPR002053, IPR018077 & IPR017853). After cross-reference with the quantitative analysis, it was found that these are absent in all *Cutibacterium*-targetting proteins in PhaLP. While some phage lytic proteins target both of these genera, and these glycoside hydrolase domains are thus not present in every *Propionibacterium*-targeting protein, they are present in one-third of the lytic proteins that solely target *Propionibacterium*. Incorporation of such domain may thus help narrow down the host-spectrum.

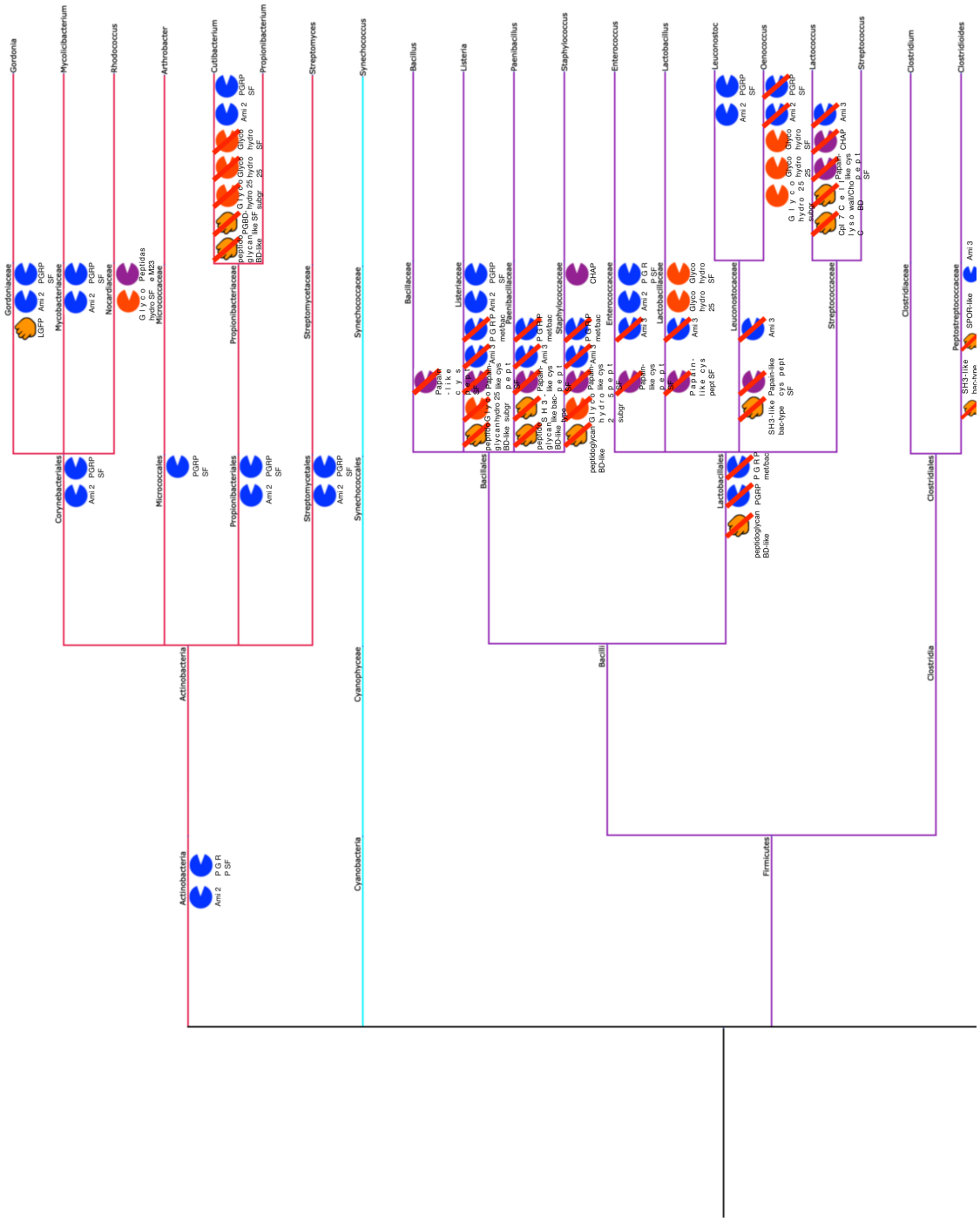
5.4 Future prospects

While nature has evolved towards the requirements of the natural function of phage lytic proteins, protein engineers nowadays try to mimic evolution on a laboratory scale to re-engineer proteins towards the application the protein engineer has in mind. This strategy is called directed evolution. In the case of phage lytic proteins, the application at hand is the engineering of enzybiotics. This thesis may give the engineer indications which domains should be (minimally) included in an enzybiotic targeting the desired genus. However, to construct a functional enzybiotic, additional parameters have to be taken in mind. The conditions in which an enzybiotic is applied,

differ from those that prevail in the infected bacterial cell. Additionally, an enzybiotic requires stability and an applicable shelf life.

To work towards the full design of an enzybiotic, the methods described in this thesis should be expanded to include information on the order and amount of the domains found in a phage lytic protein. Domains themselves can vary in sequence as well, hence an adaptation on the method used in section 4.3 should be implemented to find the optimal variation. Furthermore, pH- and thermostability should be considered, as well as other environmental factors. As new phage lytic proteins are discovered and known ones are better characterised and annotated, these methods should also be expanded accordingly, as this can further substantiate or disprove any of the hypotheses made, as well as improve the quality of predictive models and the design rules deduced from them.

While antibiotic resistance is shaping up to be a global crisis and solutions so far are sparse, interpretability might just cast a light on the solution to a very predictable trend.


Gordonia
Mycolicibacterium
Rhodococcus
Arthrobacter
Cutibacterium
Propionibacterium
Streptomyces
Synechococcus
Bacillus
Listeria
Paenibacillus
Staphylococcus
Enterococcus
Lactobacillus
Leuconostoc
Oenococcus
Lactococcus
Streptococcus
Clostridium
Clostridioides
Actinobacteria
Cyanophyceae
Bacilli
Clostridia
Actinobacteria
Cyanobacteria
Firmicutes

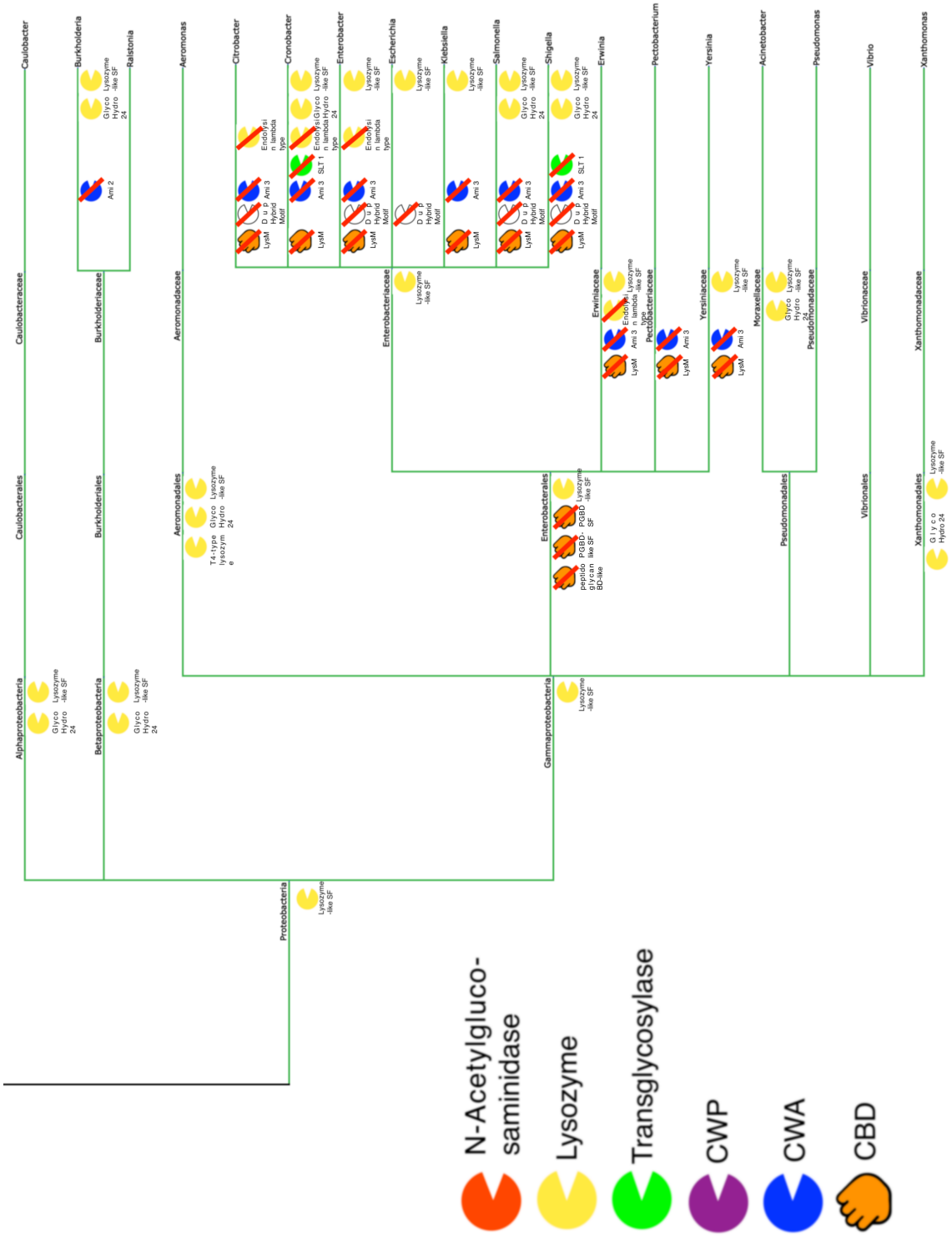

Proteobacteria
Alphaproteobacteria
Caulobacterales
Caulobacteraceae
Caulobacter
Betaproteobacteria
Burkholderiales
Burkholderiaceae
Burkholderia
Ralstonia
Gammaproteobacteria
Aeromonadales
Aeromonadaceae
Aeromonas
Enterobacteriales
Enterobacteriaceae
Citrobacter
Cronobacter
Enterobacter
Escherichia
Klebsiella
Salmonella
Shigella
Erwiniaceae
Erwinia
Pectobacteriaceae
Pectobacterium
Yersiniaceae
Yersinia
Pseudomonadales
Moraxellaceae
Acinetobacter
Pseudomonadaceae
Pseudomonas
Vibrionales
Vibrionaceae
Vibrio
Xanthomonadales
Xanthomonadaceae
Xanthomonas
Lysozyme -like SF
Glyco Hydro 24
T4-type lysozyme
Ami 2
Ami 3
LysM
Dup Hybrid Motif
Endolysin lambda type
SLT 1
PGBD-like SF
peptidoglycan BD-like
N-Acetylgluco-saminidase
Lysozyme
Transglycosylase
CWP
CWA
CBD

Figure 5.1: A schematic aggregate of figures 4.10 and 4.13 conveying a rough guide to domain architecture in natural phage lytic proteins. Striked out symbols represent domains that are not found in nature in the specified branch, while regular symbols represent domains that strongly correlated with a specific branch. The tree itself is coloured on phylum.

BIBLIOGRAPHY

Abdelkader, K., Gerstmans, H., Saafan, A., Dishisha, T., and Briers, Y. (2019). The preclinical and clinical progress of bacteriophages and their lytic enzymes: The parts are easier than the whole. *Viruses*, 11(2):1–16.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Barbedo, J. G. s. and Lopes, A. (2006). Automatic genre classification of musical signals. *EURASIP Journal on Advances in Signal Processing*, 2007(1):1–12.

Bateman, A., Sonnhammer, E. L. L., Ceric, G., Hotz, H.-R., Mistry, J., Tate, J., Forslund, K., Coggill, P. C., Eddy, S. R., Sammut, S. J., and Finn, R. D. (2007). The Pfam protein families database. *Nucleic Acids Research*, 36:281–288.

Briers, Y., Cornelissen, A., Aertsen, A., Hertveldt, K., Michiels, C. W., Volckaert, G., and Lavigne, R. (2008). Analysis of outer membrane permeability of *Pseudomonas aeruginosa* and bactericidal activity of endolysins KZ144 and EL188 under high hydrostatic pressure. *FEMS Microbiology Letters*, 280(1):113–119.

Briers, Y. and Lavigne, R. (2015). Breaking barriers: Expansion of the use of endolysins as novel antibacterials against gram-negative bacteria. *Future Microbiology*, 10:377–390.

Briers, Y., Volckaert, G., Cornelissen, A., Lagaert, S., Michiels, C. W., Hertveldt, K., and Lavigne, R. (2007). Muralytic activity and modular structure of the endolysins of pseudomonas aeruginosa bacteriophages ϕkz and el. *Molecular Microbiology*, 65(5):1334–1344.

Briers, Y., Walmagh, M., and Lavigne, R. (2011). Use of bacteriophage endolysin el188 and outer membrane permeabilizers against pseudomonas aeruginosa. *Journal of Applied Microbiology*, 110(3):778–785.

Briers, Y., Walmagh, M., Van Puyenbroeck, V., Cornelissen, A., Cenens, W., Aertsen, A., Oliveira, H., Azeredo, J., Verween, G., Pirnay, J.-P., Miller, S., Volckaert, G., and Lavigne, R. (2014). Engineered endolysin-based “artilysins” to combat multidrug-resistant gram-negative pathogens. *mBio*, 5(4):1–10.

Brown, T. L., Petrovski, S., Dyson, Z. A., Seviour, R., and Tucci, J. (2016). The formulation of bacteriophage in a semi solid preparation for control of propionibacterium acnes growth. *PLOS ONE*, 11(3):1–16.

Callewaert, L., Walmagh, M., Michiels, C. W., and Lavigne, R. (2011). Food applications of bacterial cell wall hydrolases. *Current Opinion in Biotechnology*, 22(2):164–171.

Carlton, R. (1999). Phage therapy: past history and future prospects. *Archivum Immunologiae et Therapiae Experimentalis*, 47:267–274.

ContraFect (2019). *ContraFect Presents Additional Positive Data from the Phase 2 Trial of Exebacase at the 29th European Congress of Clinical Microbiology and Infectious Diseases (ECCMID)*.

Criel, B. (2017). A bioinformatic approach to infer horizontal gene transfer events with the aim to improve endolysin engineering. Master's thesis, UGent.

Cummins, C. S. and Harris, H. (1956). The chemical composition of the cell wall in some gram-positive bacteria and its possible value as a taxonomic character. *Microbiology*, 14(3):583–600.

Czaplewski, L., Bax, R., Clokie, M., Dawson, M., Fairhead, H., Fischetti, V. A., Foster, S., Gilmore, B. F., Hancock, R. E. W., Harper, D., Henderson, I. R., Hilpert, K., Jones, B. V., Kadioglu, A., Knowles, D., Ólafsdóttir, S., Payne, D., Projan, S., Shaunak, S., Silverman, J., Thomas, C. M., Trust, T. J., Warn, P., and Rex, J. H. (2016). Alternatives to antibiotics—a pipeline portfolio review. *The Lancet Infectious Diseases*, 16(2):239–251.

Davies, J. and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews*, 74(3):417–433.

DeWeese-Scott, C., Zheng, C., Lanczycki, C. J., Robertson, C. L., Zhang, D., Hurwitz, D. I., Chitsaz, F., Lu, F., Marchler, G. H., Song, J. S., Fong, J. H., Anderson, J. B., Jackson, J. D., Geer, L. Y., Gwadz, M., Omelchenko, M. V., Mullokandov, M., Derbyshire, M. K., Zhang, N., Thanki, N., Gonzales, N. R., Geer, R. C., Yamashita, R. A., Lu, S., Bryant, S. H., Ke, Z., and Marchler-Bauer, A. (2010). Cdd: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Research*, 39:225–229.

Dzidic, S. and Bedeković, V. (2003). Horizontal gene transfer-emerging multidrug resistance in hospital bacteria. *Acta Pharmacologica Sinica*, 24(6):519–526.

Eugster, M. R., Haug, M. C., Huwiler, S. G., and Loessner, M. J. (2011). The cell wall binding domain of Listeria bacteriophage endolysin PlyP35 recognizes terminal GlcNAc residues in cell wall teichoic acid. *Molecular Microbiology*, 81(6):1419–1432.

Farrar, M. D., Howson, K. M., Bojar, R. A., West, D., Towler, J. C., Parry, J., Pelton, K., and Holland, K. T. (2007). Genome sequence and analysis of a propionibacterium acnes bacteriophage. *Journal of Bacteriology*, 189(11):4161–4167.

Fischetti, V. A. (2008). Bacteriophage lysins as effective antibacterials. *Current Opinion in Microbiology*, 11(5):393–400.

Fleming-Dutra, K. E., Hersh, A. L., Shapiro, D. J., Bartoces, M., Enns, E. A., File, Thomas M., J., Finkelstein, J. A., Gerber, J. S., Hyun, D. Y., Linder, J. A., Lynfield, R., Margolis, D. J., May, L. S., Merenstein, D., Metlay, J. P., Newland, J. G., Piccirillo, J. F., Roberts, R. M., Sanchez, G. V., Suda, K. J., Thomas, A., Woo, T. M., Zetts, R. M., and Hicks, L. A. (2016). Prevalence of inappropriate antibiotic prescriptions among us ambulatory care visits, 2010-2011. *JAMA*, 315(17):1864–1873.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954.

Gardin, F., Gautier, R., Goix, N., Ndiaye, B., and Schertzer, J.-M. (2018). *Skope Rules*. MIT.

Gerstmans, H., Criel, B., and Briers, Y. (2018). Synthetic biology of modular endolysins. *Biotechnology Advances*, 36(3):624–640.

Ghuysen, J. M. (1968). Use of bacteriolytic enzymes in determination of wall structure and their role in cell metabolism. *Bacteriological Reviews*, 32(4):425–464.

Gupta, R. S. (1998). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes. *Microbiology and Molecular Biology Reviews : MMBR*, 62(4):1435–1491.

Gupta, R. S., Lo, B., and Son, J. (2018). Phylogenomics and comparative genomic studies robustly support division of the genus mycobacterium into an emended genus mycobacterium and four novel genera. *Frontiers in Microbiology*, 9:1–41.

Halligan, S., Altman, D. G., and Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European Radiology*, 25(4):932–939.

Harkins, D., Beck, E., Selengut, J. D., Basu, M. K., Richter, R. A., and Haft, D. H. (2012). Tigrfams and genome properties in 2013. *Nucleic Acids Research*, 41:387–395.

Hastie, T. J., Friedman, J. H., and Tibshirani, R. J. (2017). *The elements of statistical learning*. Springer.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12.

Hendrix, R. W. (2003). Bacteriophage genomics. *Current Opinion in Microbiology*, 6(5):506–511.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919.

Herlihey, F. A. and Clarke, A. J. (2017). *Controlling Autolysis During Flagella Insertion in Gram-Negative Bacteria*, pages 41–56. Springer, Singapore.

Hermoso, J. A., García, J. L., and García, P. (2007). Taking aim on bacterial pathogens: from phage therapy to enzybiotics. *Current Opinion in Microbiology*, 10(5):461–472.

Heselpoth, R. D. and Nelson, D. C. (2012). A new screening method for the directed evolution of thermostable bacteriolytic enzymes. *JoVE*, 69:1–8.

Höltje, J.-V., Kopp, U., Ursinus, A., and Wiedemann, B. (1994). The negative regulator of *β*-lactamase induction AmpD is a N-acetyl-anhydromuramyl-l-alanine amidase. *FEMS Microbiology Letters*, 122(2):159–164.

Höltje, J. V., Mirelman, D., Sharon, N., and Schwarz, U. (1975). Novel type of murein transglycosylase in escherichia coli. *Journal of Bacteriology*, 124(3):1067–1076.

Jiang, B., Song, K., Ren, J., Deng, M., Sun, F., and Zhang, X. (2012). Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, 13(1):1–17.

Koonin, E. and Galperin, M. (2003). *Sequence — Evolution — Function*. Kluwer Academic.

Letunic, I. and Bork, P. (2019). Interactive tree of life (itol) v4: recent updates and new developments. *Nucleic Acids Research*, pages 1–4.

Letunic, I., Doerks, T., and Bork, P. (2014). Smart: recent updates, new developments and status in 2015. *Nucleic Acids Research*, 43:257–260.

Linden, S. B., Zhang, H., Heselpoth, R. D., Shen, Y., Schmelcher, M., Eichenseher, F., and Nelson, D. C. (2015). Biochemical and biophysical characterization of plygrcs, a bacteriophage endolysin active against methicillin-resistant staphylococcus aureus. *Applied Microbiology and Biotechnology*, 99(2):741–752.

Liu, J., Yan, R., Zhong, Q., Ngo, S., Bangayan, N. J., Nguyen, L., Lui, T., Liu, M., Erfe, M. C., Craft, N., Tomida, S., and Li, H. (2015). The diversity and host interactions of propionibacterium acnes bacteriophages on human skin. *The ISME Journal*, 9:2078–2093.

Loessner, M. (2005). Bacteriophage endolysins—current state of research and applications. *Current Opinion in Microbiology*, 8:480–487.

Loessner, M. J., Kramer, K., Ebel, F., and Scherer, S. (2002). C-terminal domains of listeria monocytogenes bacteriophage murein hydrolases determine specific recognition and high-affinity binding to bacterial cell wall carbohydrates. *Molecular Microbiology*, 44(2):335–349.

Lutz, S. (2010). Beyond directed evolution—semi-rational protein engineering and design. *Current Opinion in Biotechnology*, 21(6):734–743.

Malki, K., Kula, A., Bruder, K., Sible, E., Hatzopoulos, T., Steidel, S., Watkins, S. C., and Putonti, C. (2015). Bacteriophages isolated from lake michigan demonstrate broad host-range across several bacterial phyla. *Virology Journal*, 12(1):1–5.

Marchler-Bauer, A. and Bryant, S. H. (2004). Cd-search: protein domain annotations on the fly. *Nucleic Acids Research*, 32(2):327–331.

Marinelli, L. J., Fitz-Gibbon, S., Hayes, C., Bowman, C., Inkeles, M., Loncaric, A., Russell, D. A., Jacobs-Sera, D., Cokus, S., Pellegrini, M., Kim, J., Miller, J. F., Hatfull, G. F., and Modlin, R. L. (2012). Propionibacterium acnes bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *mBio*, 3(5):1–13.

Martinez, J. L. and Baquero, F. (2000). Mutation frequencies and antibiotic resistance. *Antimicrobial Agents and Chemotherapy*, 44(7):1771–1777.

Meng, F. and Kurgan, L. (2016). Dflpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*, 32(12):341–350.

Meng, X., Shi, Y., Ji, W., Meng, X., Zhang, J., Wang, H., Lu, C., Sun, J., and Yan, Y. (2011). Application of a bacteriophage lysin to disrupt biofilms formed by the animal pathogen Streptococcus suis. *Applied and Environmental Microbiology*, 77(23):8272–8279.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C., Yong, S.-Y., and Finn, R. D. (2018). Interpro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):351–360.

Molnar, C. (2019). *Interpretable Machine Learning*.

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *Computing Research Repository*, abs/1109.2378:1–29.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Normark, B. H. and Normark, S. (2002). Evolution and spread of antibiotic resistance. *Journal of Internal Medicine*, 252(2):91–106.

OECD (2018). *Stemming the Superbug Tide*. OECD Publishing.

Oliveira, H., Melo, L. D. R., Santos, S. B., Nóbrega, F. L., Ferreira, E. C., Cerca, N., Azeredo, J., and Kluskens, L. D. (2013). Molecular aspects and comparative genomics of bacteriophage endolysins. *Journal of Virology*, 87(8):4558–4570.

Park, T., Struck, D. K., Dankenbring, C. A., and Young, R. (2007). The pinholin of lambdoid phage 21: Control of lysis by membrane depolarization. *Journal of Bacteriology*, 189(24):9135–9139.

Pawelkowicz, M., Osipowski, P., Wojcieszek, M., Kowalczuk, C., Plader, W., and Przybecki, Z. (2016). *Bioinformatic investigation of the role of ubiquitins in cucumber flower morphogenesis*. SPIE 10031.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, 42(1):1–8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Ponting, C. and Russell, R. (2002). The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure*, 31:45–71.

Rees, C., Botsaris, G., and Cardona, P.-J. (2012). *The Use of Phage for Detection, Antibiotic Sensitivity Testing and Enumeration*, chapter 14, pages 293–306. IntechOpen, Rijeka.

Reese, J. and Pearson, W. (2002). Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics*, 18(11):1500–1507.

Review on Antimicrobial Resistance (2016). *Tackling Drug-resistant Infections Globally: Final Report and Recommendations*. Review on antimicrobial resistance.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *CoRR*, abs/1606.05386:91–95.

Rodríguez-Rubio, L., Gutiérrez, D., Donovan, D. M., Martínez, B., Rodríguez, A., and García, P. (2016). Phage lytic proteins: biotechnological applications beyond clinical antimicrobials. *Critical Reviews in Biotechnology*, 36(3):542–552.

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-phiipla88 display multiple active catalytic domains and do not trigger staphylococcal resistance. *PLOS ONE*, 8(5):1–7.

Rydman, P. S. and Bamford, D. H. (2002). The lytic enzyme of bacteriophage prd1 is associated with the viral membrane. *Journal of Bacteriology*, 184(1):104–110.

Sayers, E. and Bryant, S. (2002). Macromolecular structure databases. In McEntyre, J. and Ostell, J., editors, *The NCBI Handbook [Internet]*, chapter 3. National Center for Biotechnology Information, Bethesda (MD).

Schleifer, K., Hammes, W., and Kandler, O. (1976). Effect of endogenous and exogenous factors on the primary structures of bacterial peptidoglycan. In Rose, A. and Tempest, D., editors, *Advances in Microbial Physiology*, volume 13, pages 245–292. Academic Press.

Schleifer, K. H. and Kandler, O. (1972). Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriological Reviews*, 36(4):407–477.

Schmelcher, M., Donovan, D. M., and Loessner, M. J. (2012). Bacteriophage endolysins as novel antimicrobials. *Future Microbiology*, 7(10):1147–1171.

Schmelcher, M., Tchang, V. S., and Loessner, M. J. (2011). Domain shuffling and module engineering of listeria phage endolysins for enhanced lytic activity and binding affinity. *Microbial Biotechnology*, 4(5):651–662.

Scholz, C. F. P. and Kilian, M. (2016). The natural history of cutaneous propionibacteria, and reclassification of selected species within the genus propionibacterium to the proposed novel genera acidipropionibacterium gen. nov., cutibacterium gen. nov. and pseudopropionibacterium gen. nov. *International Journal of Systematic and Evolutionary Microbiology*, 66(11):4422–4432.

Sharma, U., Vipra, A., and Channabasappa, S. (2018). Phage-derived lysins as potential agents for eradicating biofilms and persisters. *Drug Discovery Today*, 23(4):848–856.

Silla, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72.

Siranosian, B., Perera, S., Williams, E., Ye, C., de Graffenried, C., and Shank, P. (2015). Tetranucleotide usage highlights genomic heterogeneity among mycobacteriophages. *F1000Research*, 4:1–17.

Smith, T. F. and Waterman, M. S. (1981). Comparison of biosequences. *Advances in Applied Mathematics*, 2(4):482–489.

Stackebrandt, E., Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E., and Thompson, F. (2014). *The Family Propionibacteriaceae: Genera other than Propionibacterium*, pages 725–741. Springer, Berlin, Heidelberg.

Sundarrajan, S., Raghupatil, J., Vipra, A., Narasimhaswamy, N., Saravanan, S., Appaiah, C., Poonacha, N., Desai, S., Nair, S., Bhatt, R. N., Roy, P., Chikkamadaiah, R., Durgaiah, M., Sriram, B., Padmanabhan, S., and Sharma, U. (2014). Bacteriophage-derived chap domain protein, p128, kills staphylococcus cells by cleaving interpeptide cross-bridge of peptidoglycan. *Microbiology*, 160(10):2157–2169.

Thummeepak, R., Kitti, T., Kunthalert, D., and Sitthisak, S. (2016). Enhanced antibacterial activity of acinetobacter baumannii bacteriophage Øabp-01 endolysin (lysabp-01) in combination with colistin. *Frontiers in Microbiology*, 7:1–8.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307(4):1113–1143.

Valverde-Albacete, F. J. and Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLOS ONE*, 9(1):1–10.

Vermassen, A., Leroy, S., Talon, R., Provot, C., Popowska, M., and Desvaux, M. (2019). Cell wall hydrolases in bacteria: Insight on the diversity of cell wall amidases, glycosidases and peptidases toward peptidoglycan. *Frontiers in Microbiology*, 10:1–27.

Vidová, B., Sramkova, Z., Tišáková, L., Oravkinová, M., and Godany, A. (2014). Bioinformatics analysis of bacteriophage and prophage endolysin domains. *Biologia*, 69(5):541–556.

Walmagh, M., Boczkowska, B., Grymonprez, B., Briers, Y., Drulis-Kawa, Z., and Lavigne, R. (2013). Characterization of five novel endolysins from gram-negative infecting bacteriophages. *Applied Microbiology and Biotechnology*, 97(10):4369–4375.

Walmagh, M., Briers, Y., Santos, S. B. d., Azeredo, J., and Lavigne, R. (2012). Characterization of modular bacteriophage endolysins from Myoviridae phages OBP, 201ϕ2-1 and PVP-SE1. *PLOS ONE*, 7(5):1–10.

Wang, Y., Fu, L., Ren, J., Yu, Z., Chen, T., and Sun, F. (2018). Identifying group-specific sequences for microbial communities using long k-mer sequence signatures. *Frontiers in microbiology*, 9:872–872.

Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology*, 297(1):233–249.

Zhang, H. (2005). Exploring conditions for the optimality of naive bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(02):183–198.

APPENDIX A

A.1 BLOSUM 62 substitution matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	0	-1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	-2	5	C
		2	0	-2	0	-1	0	0	0	1	0	0	0	1	0	1	-1	1	1	-1	S
C	9		2	-1	-1	-1	0	0	0	0	0	0	-1	0	-1	1	0	1	1	3	T
S	-1	4		2	-2	-1	-1	0	0	-1	-1	-1	1	1	0	-1	0	0	2	1	P
T	-1	1	5		2	-1	-2	-2	-1	0	0	1	1	0	0	1	0	1	1	2	A
P	-3	-1	-1	7		2	0	-1	-2	0	1	1	0	0	-1	0	-1	1	2	4	G
A	0	1	0	-1	4		3	-1	-1	0	0	1	-1	0	-1	0	-1	0	0	0	N
G	-3	0	-2	-2	0	6		2	-1	-1	-1	0	-1	0	0	0	0	2	1	3	D
N	-3	1	0	-2	-2	0	6		1	0	0	2	2	1	-1	0	0	2	2	4	E
D	-3	0	-1	-1	-2	-1	1	6		0	-2	0	1	1	-1	0	0	1	3	3	Q
E	-4	0	-1	-1	-1	-2	0	2	5		2	-1	0	1	0	-1	0	1	2	2	H
Q	-3	0	-1	-1	-1	-2	0	0	2	5		-1	-1	0	-1	1	0	1	3	-4	R
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8		1	-2	-1	1	1	2	3	1	K
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5		-2	-1	-1	0	1	2	4	M
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5		-1	1	0	0	1	3	I
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5		-1	0	-1	1	2	L
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4		0	1	2	4	V
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4		-1	-2	1	F
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4		-1	2	Y
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		-1	W
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Figure A.1: The scores for matches and mismatches as given by the BLOSUM62 substitution matrix as assembled by Henikoff and Henikoff (1992)

APPENDIX B

B.1 Gram-type prediction rules

B.1.1 Gram-positive

All rules and corresponding performance statistics generated for a Gram-positive host prediction by a Skope Rules model.

1. IF NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

   precision: 0.954106
   recall: 0.941597

2. IF NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR042047 THEN G+

   precision: 0.94856
   recall: 0.943302

3. IF NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR041219 and NOT IPR042047 THEN G+

   precision: 0.947929
   recall: 0.941246

4. IF NOT IPR015510 and NOT IPR023346 and NOT IPR042047 THEN G+

   precision: 0.944106
   recall: 0.943607

5. IF NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR042047 THEN G+

   precision: 0.943216
   recall: 0.944337

6. IF NOT IPR011105 and NOT IPR015510 and NOT IPR023346 THEN G+

   precision: 0.941858
   recall: 0.945556

7. IF NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

   precision: 0.942584
   recall: 0.942584

8. IF NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR040471 and NOT IPR042047 THEN G+

   precision: 0.944578
   recall: 0.938922

9. IF NOT IPR002508 and NOT IPR007048 and NOT IPR008964 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR041219 and NOT IPR042047 THEN G+

   precision: 0.969399
   recall: 0.869761

10. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.963612
    recall: 0.870889

11. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.964798
    recall: 0.869565

12. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038288 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.97047
    recall: 0.864833

13. IF NOT IPR002508 and NOT IPR007048 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.960422
    recall: 0.872902

14. IF NOT IPR002508 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.962865
    recall: 0.870504

15. IF NOT IPR002508 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.962985
    recall: 0.869837

16. IF NOT IPR002508 and NOT IPR007048 and NOT IPR015510 and NOT IPR023346 and NOT IPR03825 and NOT IPR042047 THEN G+

    precision: 0.958165
    recall: 0.873309

17. IF NOT IPR002508 and NOT IPR003343 and NOT IPR007048 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.968835
    recall: 0.864571

18. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.970706
    recall: 0.862722

19. IF NOT IPR002508 and NOT IPR007048 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.966287
    recall: 0.86604

20. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR042047 THEN G+

    precision: 0.966353
    recall: 0.86506

21. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.962633
    recall: 0.868061

22. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038288 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.965879
    recall: 0.864865

23. IF NOT IPR002508 and NOT IPR007048 and NOT IPR008964 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.956175
    recall: 0.872727

24. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.964973
    recall: 0.865163

25. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR040471 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.966443
    recall: 0.863309

26. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR042047 THEN G+

    precision: 0.959264
    recall: 0.869048

27. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.963952
    recall: 0.864671

28. IF NOT IPR002508 and NOT IPR011105 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 THEN G+

    precision: 0.960973
    recall: 0.866939

29. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.961538
    recall: 0.866189

30. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR02334 and NOT IPR038258 and NOT IPR042047 THEN G+

    precision: 0.959677
    recall: 0.867558

31. IF NOT IPR002508 and NOT IPR003343 and NOT IPR007048 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.963838
    recall: 0.86374

32. IF NOT IPR002508 and NOT IPR011105 and NOT IPR015510 and NOT IPR023346 THEN G+

    precision: 0.955979
    recall: 0.870068

33. IF NOT IPR002508 and NOT IPR007048 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR042047 THEN G+

    precision: 0.961108
    recall: 0.865853

34. IF NOT IPR002508 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR042047 THEN G+

    precision: 0.961275
    recall: 0.865635

35. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR042047 THEN G+

    precision: 0.960054
    recall: 0.866197

36. IF NOT IPR002508 and NOT IPR011105 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 THEN G+

    precision: 0.962766
    recall: 0.863962

37. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.958778
    recall: 0.867108

38. IF NOT IPR002508 and NOT IPR008964 and NOT IPR011105 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR041219 THEN G+

    precision: 0.967828
    recall: 0.859524

39. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.959239
    recall: 0.866258

40. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR042047 THEN G+

    precision: 0.964683
    recall: 0.861385

41. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.96438
    recall: 0.861013

42. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR040471 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.963061
    recall: 0.861865

43. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR042047 THEN G+

    precision: 0.958467
    recall: 0.865053

44. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR040471 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.966395
    recall: 0.858616

45. IF NOT IPR002508 and NOT IPR007048 and NOT IPR015510 and NOT IPR023346 and NOT IPR038288 and NOT IPR042047 THEN G+

precision: 0.967785
recall: 0.857313

46. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.959891
    recall: 0.8635

47. IF NOT IPR002508 and NOT IPR008964 and NOT IPR011105 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR040471THEN G+

    precision: 0.967828
    recall: 0.856465

48. IF NOT IPR002508 and NOT IPR002901 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR040471 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.966033
    recall: 0.85766

49. IF NOT IPR002508 and NOT IPR002901 and NOT IPR003343 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and IPR038288 and NOT IPR042047 THEN G+

    precision: 0.965147
    recall: 0.858164

50. IF NOT IPR002508 and NOT IPR002901 and NOT IPR007048 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and IPR042047 THEN G+

    precision: 0.965333
    recall: 0.85782

51. IF NOT IPR002508 and NOT IPR007048 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR042047 THEN G+

    precision: 0.958868
    recall: 0.862719

52. IF NOT IPR002508 and NOT IPR003343 and NOT IPR007048 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and IPR042047 THEN G+

    precision: 0.957276
    recall: 0.863855

53. IF NOT IPR002508 and NOT IPR007048 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and IPR042047 THEN G+

    precision: 0.9629
    recall: 0.858643

54. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.962213
    recall: 0.859036

55. IF NOT IPR002508 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR042047 THEN G+

    precision: 0.961126
    recall: 0.859712

56. IF NOT IPR002508 and NOT IPR003343 and NOT IPR011105 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 THEN G+

    precision: 0.959839
    recall: 0.860744

57. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.960055
    recall: 0.860494

58. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and IPR041219 and IPR042047 THEN G+

    precision: 0.962213
    recall: 0.858002

59. IF NOT IPR002508 and NOT IPR011105 nd NOT IPR015510 and NOT IPR020362 and NOT IPR023346 THEN G+

    precision: 0.954606
    recall: 0.863527

60. IF NOT IPR002508 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.958865
    recall: 0.859603

61. IF NOT IPR002508 and NOT IPR003343 and NOT IPR007048 and NOT IPR015510 and NOT IPR020362 and NOT IPR023346 and NOT IPR038288 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.965673
    recall: 0.85375

62. IF NOT IPR002508 and NOT IPR007048 and NOT IPR008964 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.964217
    recall: 0.854216

63. IF NOT IPR002508 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR040471 and NOT IPR042047 THEN G+

    precision: 0.956815
    recall: 0.859394

64. IF NOT IPR002508 and NOT IPR003343 and NOT IPR007048 and NOT IPR011105 and NOT IPR015510 and NOT IPR020362 and IPR023346 and NOT IPR038288 and NOT IPR041219 THEN G+

    precision: 0.960437
    recall: 0.856448

65. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR041219 and NOT IPR042047 THEN G+

    precision: 0.962213
    recall: 0.853892

66. IF NOT IPR002508 and NOT IPR003343 and NOT IPR015510 and NOT IPR023346 and NOT IPR038258 and NOT IPR038288 and NOT IPR042047 THEN G+

    precision: 0.959459
    recall: 0.851319

B.1.2 Gram-negative

All rules and corresponding performance statistics generated for a Gram-negative host prediction by a Skope Rules model.

1. IF NOT IPR002502 and NOT IPR016047 and IPR023346 and NOT IPR036779 THEN G-

precision: 0.980668
recall: 0.785915

2. IF NOT IPR002502 and NOT IPR003646 and IPR023346 and NOT IPR036779 THEN G-

   precision: 0.978648
   recall: 0.785714

3. IF NOT IPR002502 and NOT IPR003646 and NOT IPR016047 and NOT IPR018392 and IPR023346 THEN G-

   precision: 0.980523
   recall: 0.782885

4. IF NOT IPR003646 and NOT IPR010090 and IPR023346 and NOT IPR036505 and NOT IPR036779 THEN G-

   precision: 0.979594
   recall: 0.783096

5. IF NOT IPR002502 and NOT IPR003646 and NOT IPR010090 and IPR023346 and NOT IPR03677 THEN G-

   precision: 0.981131
   recall: 0.781398

6. IF NOT IPR002502 and IPR023346 and NOT IPR036779 THEN G-

   precision: 0.976147
   recall: 0.784069

7. IF NOT IPR002502 and NOT IPR003646 and NOT IPR010090 and NOT IPR018392and IPR023346 THEN G-

   precision: 0.98144
   recall: 0.780476

8. IF NOT IPR002502 and NOT IPR010090 and NOT IPR018392 and IPR023346 THEN G-

   precision: 0.979635
   recall: 0.781191

9. IF NOT IPR002502 and NOT IPR016047 and NOT IPR018392 and IPR023346 THEN G-

   precision: 0.974453
   recall: 0.784141

10. IF NOT IPR002502 and NOT IPR010090 and IPR023346 and NOT IPR036779 THEN G-

    precision: 0.98168
    recall: 0.779296

11. IF NOT IPR003646 and NOT IPR010090 and NOT IPR018392 and IPR023346 and NOT IPR036505 THEN G-

    precision: 0.981731
    recall: 0.779204

12. IF NOT IPR002502 and NOT IPR018392 and IPR023346 THEN G-

    precision: 0.974399
    recall: 0.783337

13. IF NOT IPR010090 and IPR023346 and NOT IPR036505 and NOT IPR036779 THEN G-

    precision: 0.976619
    recall: 0.780172

14. IF NOT IPR010090 and NOT IPR018392 and IPR023346 and NOT IPR036505 THEN G-

    precision: 0.979951
    recall: 0.777869

15. IF NOT IPR018392 and IPR023346 and NOT IPR036505 THEN G-

    precision: 0.974979
    recall: 0.780204

16. IF NOT IPR002502 and NOT IPR003646 and NOT IPR018392 and IPR023346 THEN G-

    precision: 0.977535
    recall: 0.77836

17. IF IPR023346 and NOT IPR036505 and NOT IPR036779 THEN G-

    precision: 0.973674
    recall: 0.779803

18. IF NOT IPR003646 and IPR023346 and NOT IPR036505 and NOT IPR036779 THEN G-

    precision: 0.978533
    recall: 0.770423

B.2 Bayes' theorem

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{B.1}$$