FACULTY OF BIOSCIENCE ENGINEERING

PREDICTING INTERACTIONS BETWEEN BACTERIA AND THEIR PROPHAGES BASED ON GENOMIC SEQUENCE DATA

word count: 19,659

Tristan Vanneste
Student ID: 01508143

Supervisor(s): Prof. dr. ir. Yves Briers and dr. ir. Michiel Stock
Tutor: Ir. Dimitri Boeckaerts

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of master in Bioscience Engineering: Cell and Gene Biotechnology.
Academic year: 2019 - 2020

GHENT UNIVERSITY

DANKWOORD

Toen ik vijf jaar terug voor het eerst op onze faculteit toekwam, keek ik met volle bewondering naar de oudere studenten die toen hun thesis maakten. Ik kon zeer veel respect hebben voor de studenten die een heel jaar lang werkten om dat ene document af te werken. Nu vijf jaar later ben ik zelf mijn thesis aan het schrijven en eerlijk gezegd is het allemaal voorbij gevlogen. In tegenstelling tot sommige andere studenten, heb ik altijd met veel plezier en enthousiasme gewerkt aan mijn thesis. Hoe verder ik me verdiepte in de faagwereld, hoe meer ik mij er toe aangetrokken voelde en hoe enthousiaster ik werd. Ik zou graag mijn begeleiders en vrienden bedanken voor de steun die ze mij gaven tijdens het jaar wanneer ik het wat moeilijker had met het uitschrijven van mijn thesis.

Eerst en vooral wil ik graag mijn promotoren, Yves en Michiel, bedanken om deze thesis in goeie banen te leiden. Hun kennis en creativiteit hielpen mij geïnteresseerd te blijven tot op het laatste moment. Bedankt voor het vele verbeterwerk en de snelle antwoorden op mijn e-mails vol met vragen. Dankzij jullie kon ik rond dit onderwerp werken en ben ik in een volledig nieuwe faagwereld terecht gekomen.

Uiteraard wil ik ook mijn begeleider Dimi niet vergeten te bedanken. Hij was mijn rots in de branding. Hij beantwoordde mijn veel te talrijke en soms ook overbodige vragen altijd met het nodige enthousiasme op gelijk welk moment van de dag. Zijn enthousiasme zorgde ervoor dat ik voor dit uitzonderlijk interessant onderwerp koos en hij liet mij kennismaken met de wonderbaarlijke wereld van bacteriofagen. Ik zal onze wekelijkse, soms zelf nog frequenter, brainstorm-momenten zeker missen. Ook een bedanking voor mijn ouders, familie en buurjongen Bart, voor de vele wandelmomenten die voor de nodige inspiratie zorgden voor het schrijven van mijn thesis en het broodnodige geduld als ik ergens mee worstelde.

CONTENTS

SAMENVATTING

Antibiotica-resistente bacteriën vormen een steeds groter wordend probleem voor zowel ontwikkelings- als ontwikkelde landen. ESKAPE-bacteriën spelen hierin een significante rol omdat deze bacteriën vaak resistent zijn tegen de gebruikelijke antibiotica. Het vinden van middelen die tegen deze bacteriën werken is de laatste jaren aanzienlijk afgenomen. Een mogelijk alternatief voor het gebruik van antibiotica is faagtherapie. Faagtherapie is het gebruik van bacteriofagen voor de bestrijding van bacteriële infecties. Bacteriofagen, kortom fagen, zijn virussen die bacteriën infecteren en op het einde van hun levenscyclus kunnen doden. Fagen zijn eerder kieskeurig qua gastheer, ze infecteren niet zomaar alle bacteriën van hetzelfde species. De meeste fagen infecteren bacteriën met stam-specificiteit. Momenteel wordt deze bacterie-faag interactie bepaald in het lab. Dit is een kostelijke en vooral tijdrovende activiteit. In deze thesis worden machine learning technieken gebruikt om deze interacties *in silico* te voorspellen op basis van hun genoom.

Om deze interacties te kunnen voorspellen, werden ESKAPE genomen verzameld uit publieke databases. Daarna werden profaagsequenties, met behulp van PHASTER, gedetecteerd in deze ESKAPE genomen. Voor elke sequentie, zowel de bacteriën als de fagen, werden frequenties berekend van alle 3-meren. Deze frequenties dienden als features voor verschillende machine learning modellen. Uiteindelijk werden de machine learning modellen vergeleken met elkaar via Precision-Recall curves en F1-score curves. Ons beste model, een Support Vector Machine, gaf voorspellingen op het stam-niveau met een nauwkeurigheid van 85.84% en een F1-score van 84.84%.

Trefwoorden: bacterie-faag interactie, faagtherapie, machine learning, kernel methoden, Random Forest, K-nearest neighbors, Support Vector Machine, Linear Discriminant Analysis, Two-step Kernel Ridge Regression.

SUMMARY

Antibiotic-resistant bacteria have become increasingly problematic in both developing and developed countries. ESKAPE bacteria play a significant role in this because these bacteria are often resistant to commonly used antibiotics. Finding drugs that work against these bacteria has decreased considerably in recent years. A possible alternative to the use of antibiotics is phage therapy. Phage therapy is the use of phages to fight bacterial infections. Bacteriophages, in short phages, are viruses that infect bacteria and kill them at the end of their replication cycle. Phages are picky in terms of their host, they do not just infect all bacteria of the same species. Most phages infect bacteria with strain specificity. Currently, bacterium-phage interactions are being determined in the lab. This is a costly and time-consuming activity. In this thesis, machine learning techniques are used to predict these interactions *in silico* based on genomic sequence data.

To predict these interactions, ESKAPE genomes were collected from public databases. Prophage sequences were detected, in the bacterial genomes, using PHASTER. For each of the phage sequence and bacterial sequences, frequencies were computed for all 3 mers. These frequencies served as features for various machine learning models. Ultimately, the machine learning models were compared with each other via Precision-Recall and F1 score curves. Our best model, a Support Vector Machine, gave predictions at the strain level with an accuracy of 85.84% and an F1 score of 84.84%.

Keywords bacterium-phage interaction, phage therapy, machine learning, kernel methods, Random Forest, K-nearest neighbors, Support Vector Machine, Linear Discriminant Analysis, Two-step Kernel Ridge Regression.

CHAPTER 1

INTRODUCTION AND OUTLINE

According to the World Health Organisation (WHO), one of the biggest present-day threats to global health, food security and development is antibiotic resistance. It leads to longer hospital stays, higher medical costs and increased mortality. But what exactly is the phenomenon, antibiotic resistance, and how can we try to combat it? For that, we first have to briefly explain what antibiotics are. In 1928, Alexander Fleming discovered the first antibiotic, penicillin, by serendipity. During the Second World War, penicillin saved millions of lives by controlling bacterial infections that soldiers got on the battlefield. Since then, many other classes of antibiotics were discovered. They were soon after employed to combat bacterial infections in several ways (Ventola, 2015). The definition of an antibiotic is any organic molecule that inhibits growth or kills microbes by specific interactions with bacterial targets (Davies and Davies, 2010). An antibiotic kills or affects bacterial cells and not human cells. It can do so by affecting the cell wall, the cell membrane, the DNA-copying machinery that is unique for bacteria or many others targets (Ventola, 2015). Antibiotic resistance is a phenomenon where bacteria develop resistances against certain types of antibiotics. The spread of antibiotic resistance increasingly leads to bacterial infections that are difficult or even impossible to treat. This is causing a huge problem in the medical world. According to the WHO, one of the causes is that antibiotics are frequently overprescribed by health workers and veterinarians, leading to their overuse by the public. Human overuse but also misuse e.g., in the case of viral infections or the use of antibiotics for human medicine as a growth promoter in livestock has stimulated the rise of multidrug-resistant superbugs. For example, extremely drug-resistant *Mycobacterium tuberculosis* is almost impossible to treat with existing medicines (O'Neill, 2016).

In 2014, a total of 700,000 people died because of antibiotic resistance. This number is probably a big underestimation due to poor surveillance and reporting. The antibiotic-resistant *Mycobacterium tuberculosis* alone kills over 200,000 people every year and without further action, this number is only going to increase in the future. By 2050, superbugs could kill up to 10 million people every year in a worst-case scenario (O'Neill, 2016). What would happen if routine medical procedures are threatened by antibiotic resistance? If procedures like blood transfusion or childbirth become dangerous due to a high risk of post-procedure infections? For these reasons, scientists

are urgently searching for effective alternatives to antibiotics.

The interest in bacteriophages (or phages), as a promising alternative for antibiotics, is beginning to re-emerge into the scientific world. Phages are viruses that invade bacterial cells and, at the end of their lytic cycle, phages can disrupt the host cell wall, resulting in the bacterial cell to undergo lysis (Sulakvelidze et al., 2001b). A decade before the discovery of penicillin, doctors applied phages to treat bacterial infections (Schmidt, 2019). However, the limited knowledge and understanding of phage biology made this treatment less popular than antibiotics when the latter became available. This delayed the widespread adaptation of phage therapy (Mansour, 2017). In the Soviet-Union and some East-European countries, researchers continued to study and use phages while in the West-European countries phages were put on the sideline for a long time. However, with the ever-increasing threat of antibiotic resistance, phage therapy is regaining popularity within Western countries (Pirnay, 2014).

A disadvantage of using phages as a treatment against bacterial infections is that phages, in general, infect bacteria with strain level specificity (Hesse and Adhya, 2019). Indeed, phages are picky: a phage can, for example, infect a *Pseudomonas aeruginosa* PA96 but not necessarily a *Pseudomonas aeruginosa* UCBPP-PA14 so it does not infect all bacteria from the same species. This makes it very hard to find the matching phage against a particular infection (Leite et al., 2018). Additionally, infections are sometimes caused by multiple bacteria so doctors need multiple phages to treat a single patient (Kutter and Sulakvelidze, 2005). Nowadays, researchers and medical professionals are looking for bacterium-phage interactions in laboratories, which is a time-consuming activity. The aim of this thesis is to develop machine learning models that can accurately predict interactions between a bacterium and phages to speed up the process of finding and characterizing bacterium-phage interactions.

In the second chapter, bacteriophages will be described along with their characteristics. Also, the interactions between bacteria and phages will briefly be discussed. The concept of the evolutionary arm-race that is going on between bacteriophages and bacteria for millions of years is going to be introduced. Finally, the advantages and disadvantages of phage therapy are given together with the most important differences between phage therapy and antibiotics. In the third chapter, the database construction is going to be explained from scratch. Complete bacterial genomes were collected from NCBI from within Python. After collecting these sequences, they were used as input for PHASTER, to collect active prophage sequences. We computed the 3-mer frequencies using Python. Afterwards, the 3-mer frequencies were used as features for the bacterial and prophage sequences. PCA and t-SNE were used to explore the prophage sequences. Finally, a multiple sequence alignment was performed on a

subsample of 50 prophages. In Chapter four, the different constructed modelling approaches will be described, this will be preceded with a short introduction to pairwise learning and the four possible prediction settings. In Chapter five, a conclusion will be presented and the work in this thesis will be summarized. Future perspectives will be elaborated on together with some alternative methods.

CHAPTER 2

PHAGES AND PHAGE THERAPY

2.1 What are bacteriophages?

Phages, also called bacteriophages, are viruses that infect bacteria. Phage means to eat or to devour in Greek. Phages are the most ubiquitous viruses on the planet. They are found in the ocean, deep-sea vents, the soil, the food that we eat and the water we drink (Prescott, 1993). An estimated total number of between 10^{30} and 10^{32} bacteriophages exist on Earth, more than any other organism (Prescott, 1993). Frederick Twort (Twort, 1915) and Felix d'Herelle (Brock, 1998) independently discovered the existence of bacteriophages in 1915 and 1917. Since then, phages have been used in a variety of practical applications such as detection and identification methods of bacteria, as well as in the food industry to kill *Listeria monocytogenes* on cheese (Atamer et al., 2013). Nowadays, the interest of using phages as antimicrobial agents is strongly increasing (Hesse and Adhya, 2019).

Essentially, a phage contains a DNA or RNA genome. The genome can be double-stranded or single-stranded and is covered with a coat of proteins, a capsid, and these capsids are made up of virus-encoded proteins (Kutter and Sulakvelidze, 2005). Together, the combination of the genome and capsid is called the nucleocapsid. In addition, a virus can be covered with an envelope, also called the lipoprotein membrane. The entire structure, the genome, the capsid, and sometimes the envelope, is called a virion or a virus particle. The virion moves from bacterium to bacterium often killing the bacteria they encounter. The exterior of the virus particle is covered with viral proteins, called spikes. These spikes can interact with bacterial proteins or bacterial receptors that are present on the bacterial cell surface.

One way to classify viruses is according to symmetry. There are two big classes: viruses with cubic symmetry and viruses with helical symmetry (Kutter and Sulakvelidze, 2005). Some viruses are not classified into these two categories, they belong to the third category, namely the complex viruses. Most known bacteriophages (around

95 %) belong to the order of the *Caudovirales* and are complex viruses (Roux et al., 2015). Their structure is a mix of the previous classes. They have an icosahedral head and a helical tail. Members of the *Caudovirales* are all tailed DNA viruses with their genetic material inside their head structure. A tail can contain tail fibers, tail spikes or tail tips (Figure 2.1) and these are all important for the specific interaction between phage and host (Nobrega et al., 2018).

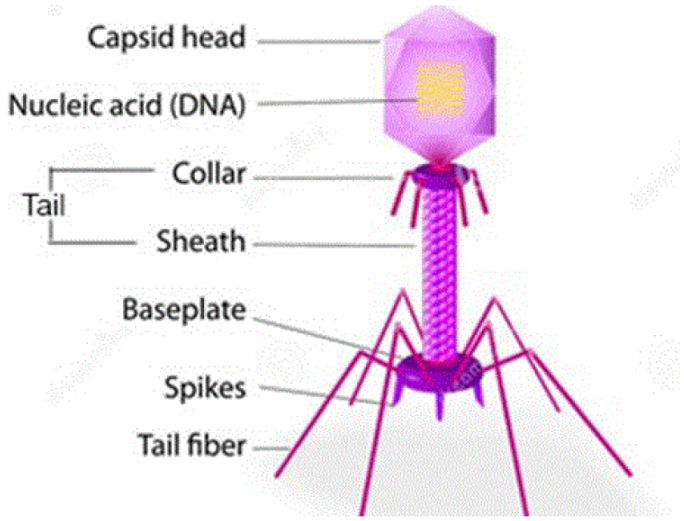

Figure 2.1: A typical structure for the *Caudovirales* (Mansour, 2017).

The *Caudovirales* are divided into three families based on the morphologies of their tail. The first family is the *Siphoviridae*, they account for 60 % of the total number of *Caudovirales* and have long, flexible tails. Around 25% of the *Caudovirales* are phages of the *Myoviridae* family with double-layered, contractile tails. The third family is the *Podoviridae* (15%), characterized by short, stubby tails. They can extend their tail due to some key infection proteins enclosed inside the head, upon contact with their host (Kutter and Sulakvelidze, 2005). The *Caudovirales* group is the most important group within phages and to know how and why they interact with certain bacteria is of utmost importance for applications in biotechnology and medicine. The tail structures are key determinants of the host specificity and infection process of the respective phages (Nobrega et al., 2018). These tailed phages are using receptor-binding proteins (RBPs), like tail fibers, tail tips and tail spikes, to interact with their host surface receptors such as lipopolysaccharide (LPS).

2.2 Replication cycles of bacteriophages

Phages are also classified based on their replication strategy. The two most frequently followed replication cycles are the lytic and lysogenic cycle (Kutter and Sulakvelidze, 2005). Phages that strictly follow the lytic cycle are called virulent phages. In this

lytic cycle, phages infect and rapidly kill their host cells (Figure 2.2). The cycle starts with a phage tail accidentally making contact with a matching bacterial surface receptor. Subsequently, phages inject their genetic material in the bacterial genome through the phage tail (Kutter and Sulakvelidze, 2005). In general, the phage tail has an enzymatic mechanism that can penetrate the peptidoglycan matrix and the inner membrane to release the genetic material into the bacterial cell.

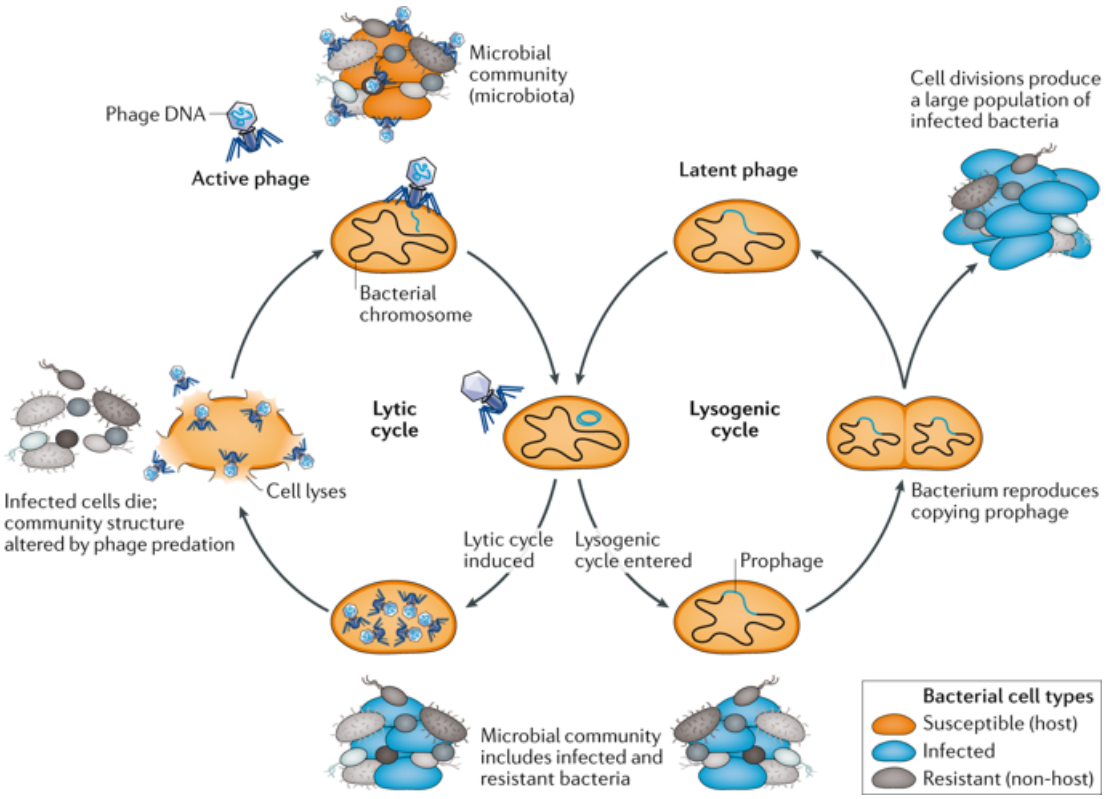

Figure 2.2: The lytic and lysogenic replication cycles of phages (Garretto et al., 2019).

Immediately after penetration, the genetic material of the phage synthesizes early proteins. The translated proteins protect the phage genome and restructure the host metabolism towards the needs of the phage. They inhibit protease activity, block restriction enzymes, destroy several host proteins and immediately terminate various host macromolecular biosyntheses (Kutter and Sulakvelidze, 2005). This causes the phage to take over the host metabolism and molecular machinery to produce a vast number of new phages. Phages start to synthesize new sigma factors or DNA-binding proteins, resulting in hijacking the host RNA polymerase complex to ensure phage transcription or they encode their own RNA polymerase. These transitions lead to specific modifications of host chaperones (macromolecular structure folding proteins) and eventually to the production of late proteins. These late proteins are subcomponents of the phage tail and head. These specific modifications ensure that the produced phage proteins are properly folded. Subsequently, the phage proteins are

folded and assembled into a complete virus particle. The complete virus particle can be spread out by bursting out of the cell membrane due to the enormous intracellular pressure or due to viral enzymes. This process is called lysis (Kutter and Sulakvelidze, 2005). *Caudovirales*, the tailed phages, use two components for lysis: an endolysin, an enzyme that cleaves in the peptidoglycan matrix of the host bacteria and a holin, a protein that creates pores in the inner membrane so that lysins can reach the peptidoglycan layer and precipitate lysis.

The second frequently observed replication cycle is the lysogenic cycle (Figure 2.2). This is more of a dormant phase where the phage integrates into the host genome. When this happens the phage is called a prophage. The inserted phage can also form a plasmid that is not integrated into the bacterial genome but is replicated alongside the genome. In this quiescent state, the phage gets replicated as its host cell replicates. The lysogenic cycle can last for over a thousand generations and changes the phenotype of the bacterium due to expression of genes that normally are not expressed. This is called lysogenic conversion (Monteiro et al., 2019). An example is *Vibrio cholerae* which encodes the toxins that cause cholera symptoms due to infection from CTXφ bacteriophage (Clokie et al., 2011).
Phages in the lysogenic lifecycle can switch towards a lytic lifecycle. This phenomenon is called prophage induction and this process does not happen randomly (Owen et al., 2020). This event is precisely controlled by several factors, including damage to the host DNA, external conditions and others (Wang et al., 2003). If this event happens too early, too few bacteria carrying the prophage will have reproduced. If the lysis is delayed for too long, opportunities for infecting new cells are lost (Abedon, 1990). Phages that can switch from the lysogenic cycle into the lytic cycle are called temperate phages. These cells are called lysogenic or lysogenized because of the ability of the prophage to turn into the lytic cycle and lyse. Temperate phages protect their host from subsequent infections by encoding a repressor protein. This repressor also blocks transcription of other phage genes (Kutter and Sulakvelidze, 2005). This mechanism is an example of superinfection immunity.

A less frequently observed third type of replication cycle is the chronic cycle, followed by mostly archaeal phages and some filamentous bacteriophages. In essence, this cycle is also redirecting the bacterial metabolism towards the assembly of new virions but instead of breaking the cell wall and precipitate lysis, the progeny of the phages is continuously released from the host by budding or extrusion (Weinbauer, 2004). According to Cenens et al. (2013), there are two more replication cycles, namely pseudolysogeny and carrier-state lifecycles. Phages following these replication cycles

are carried inside a host without being in the lytic or lysogenic lifecycle, which gives benefits to the phage such as protection against new infections from other phages and preventing a lytic cycle when the host resources are limited. They are generally used as synonyms but the carrier state more often refers to bacteria with a plasmid-like prophage (Weinbauer, 2004).

2.3 Interactions between phages and bacteria

Phages are obligate parasites, they cannot complete their lifecycle and produce offspring without infection of a host, in this case, a bacterium. This means that if the host does not have enough resources to survive, the phages cannot replicate within the host (Weinbauer, 2004). As mentioned before, the temperate phages are not the only ones that benefit from the interaction, the bacterial host can as well. Firstly, the temperate phage protects their host from new infections by other phages, which can be done by changing the bacterial surface receptors (Cenens et al., 2013). Secondly, they also provide horizontal gene transfer and encode additional genes that have a beneficial effect on the host, resulting in increased microbial diversity (Cenens et al., 2013). Thirdly, temperate phages presence results in pathogenicity for numerous bacterial strains, e.g., *Vibrio cholerae* (Clokie et al., 2011) and *E. coli* 0157 (Ross et al., 2016).

The property of facilitating horizontal gene transfer is very important for bacterial evolution. Antibiotic resistance is a direct consequence of this phenomenon. Horizontal gene transfer can be very beneficial to the recipient host cell by spreading virulence or resistance properties throughout the community (Kutter and Sulakvelidze, 2005). In bacteria, horizontal gene transfer can occur through transformation, conjugation and transduction. Transformation is the direct uptake, incorporation and expression of foreign genetic material out of the environment through the cell membrane. Conjugation is the process that involves the transfer of the genetic material via plasmids from a donor cell to a recombinant recipient cell. The last form of horizontal gene transfer, namely transduction, involves phages. Transduction is the ability of phages to mobilize bacterial genes and carry them to another bacterial cell. There are two kinds of transduction: the generalized and the specialized transduction. In generalized transduction, any part of the bacterial genome can be transferred to a new bacterial host by the phage (Trevors, 1999). More specifically, when a phage lyses the host the bacterial chromosome is broken into small pieces. Sometimes, phage packaging proteins can erroneously incorporate a piece of bacterial DNA instead of phage DNA into the phage head. The assembled phage particle, now carrying bacte-

rial DNA, can transfer the DNA into a new host.
Conversely, specialized transduction only carries restricted parts of the bacterial chromosome (Trevors, 1999). This is because of the specific mechanism of specialized transduction. A prophage always integrates at a specific point in the bacterial chromosome with the help of an enzyme system. Normally, only the prophage sequence is excised from the bacterial chromosome. However, occasionally the excision is abnormal leading to some bacterial genes in the phage DNA (Griffiths et al., 2000). Again, these bacterial genes can subsequently be transferred into a new bacterial host. Generalized transduction can be carried out by virulent and temperate phages while specialized transduction can only be carried out by temperate phages and not by virulent phages (Griffiths et al., 2000).

2.4 Coevolution of bacteria and phages

Phages outnumber bacteria by up to a factor of 10 (Stern and Sorek, 2011). Still, bacteria have evolved numerous mechanisms to avoid getting infected by phages. Conversely, phages have evolved to infect the bacteria that have previously become resistant to phage infection. This has led to an evolutionary arms-race that has been going on for millions of years. This arms-race led to an increase in diversity in bacteria and phages and will lead to continuous variations and selection towards the adaptation of the host and the counter-adaptation in the phage. As a result of a constantly changing balance between prey and predator, species have to evolve to stay at the same fitness level (Stern and Sorek, 2011). Temperate phages increase diversity by transferring genetic material from one host to another. Coevolution is not only happening between bacteria and phages but also in other host-parasite interactions. Because of the fast replication and turnover of bacteria and phages, this process is happening faster than in other host-parasite interactions.

The three most important and well known bacterial defences are inhibition of phage attachment to cell surface receptors, restriction-modification systems and clustered regularly interspaced short palindromic repeats and CRISPR-associated genes (CRISPR/-Cas) (Stern and Sorek, 2011). These bacterial defence systems have high genetic variability, which is the consequence of the coevolutionary arms-race with phages. This battle resulted in a huge variety of restriction-modification systems and many subtypes of the CRISPR/Cas system. Coevolving lytic phages can lead to increased diversity within a bacterial community by selecting for multiple modes of resistance (Stern and Sorek, 2011). Another characteristic is the propensity of undergoing lateral gene transfer, sometimes even with distantly-related prokaryotes. This mobility

allows the host to quickly counteract the invading phage and thus contribute to enormous diversity in phages and bacteria (Stern and Sorek, 2011).

2.4.1 Cell surface receptors

For the host to be infected by the phage, the phage must first attach to the surface of the host through a process that is called adsorption. As mentioned in Section 2.1, the phage uses its RBPs to recognize bacterial surface receptors of the host. These RBPs include tail fibers, tail spikes and tail tips (Nobrega et al., 2018). Bacterial receptors are presented by polysaccharides, LPS and surface proteins. The interaction between RBPs and bacterial receptors constitutes the primary determinant of host specificity.

Therefore, a trivial way of avoiding the infection by phages is to modify these surface receptors or simply make them inaccessible. Phages, on the other hand, can modify their RBPs to acquire novel interaction capabilities. For example, normally phage *lambda* (*Siphoviridae*) targets the receptor LamB in its *Escherichia coli* host. When the expression of the original receptor decreases through mutation the phage *lambda* mutates as well to acquire the ability to target a new surface receptor, namely, OmpF in addition to LamB (Samson et al., 2013). This coevolutionary process is observed in multiple hosts and with various phages.
A second way of masking the surface receptors is by producing a surface component such as an exopolysaccharide (EPS) (Samson et al., 2013). This is literally masking the surface receptors for the phages. Phages can counteract this mechanism by producing hydrolases that cleave the EPS to get to the surface receptor.

The third mechanism of the bacterial surface defence is to only express their surface receptors under specific environmental conditions or in response to specific conditions, such as quorum sensing (the ability to detect and respond to cell population density by regulating genes (Miller and Bassler, 2001)). If the expression of the surface receptor is variable, e.g., increased or decreased in the growth phase, it is beneficial for the phages to have multiple RBPs. Phages can counter this mechanism by encoding RBPs with variable specificities, which is achieved by RBP gene mutation that results in a variety of RBPs and thus expansion of the host-range (Figure 2.3).

2.4.2 Restriction-modification systems

The restriction-modification (RM) system is probably the best-studied phage defence mechanism and occurs in over 90% of the sequenced bacterial and archaeal genomes

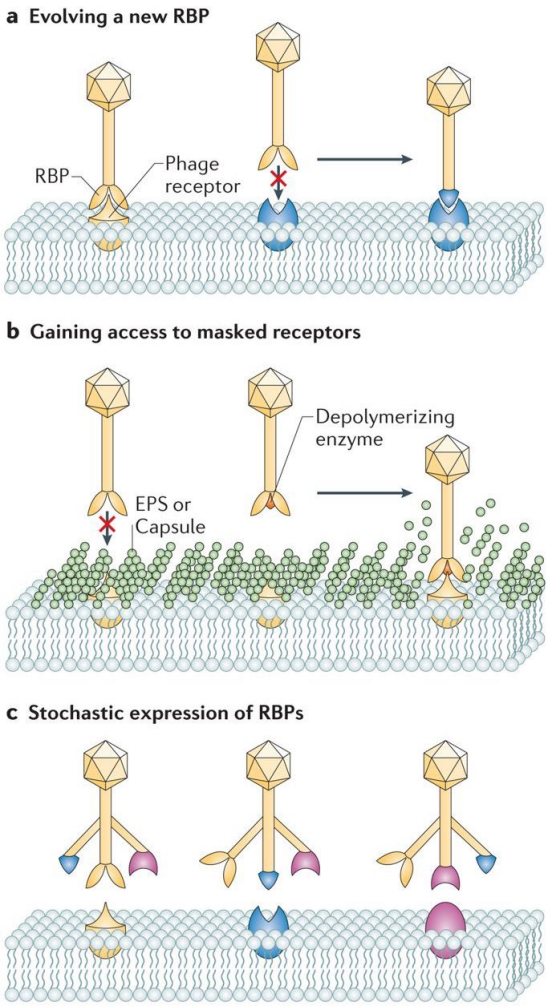

Figure 2.3: An overview of the three bacterial mechanisms to avoid infection by phages and the counteraction of phages (Samson et al., 2013). In (a) the bacterial surface receptor changed, leading to a change in the RBP. In (b) the bacterial surface receptor was masked with EPS or a capsule, the phage reacted by expressing a depolymerizing enzyme. In (c) a single phage has multiple RBPs for interacting with different bacterial receptors (Samson et al., 2013).

(Samson et al., 2013). The system consists of two components: the first one restricts new incoming foreign genetic material and the second one protects the host genetic material from getting restricted. Both activities are regulated by the recognition of a specific DNA sequence that is four to eight base pairs long. Protection is mostly achieved by modification (methylation) of this specific DNA sequence such that foreign genetic material can be recognized. The RM system typically encodes a methyltransferase gene that regulates the defence activity and a restriction endonuclease gene that regulates the foreign restriction activity (Figure 2.4).

Phages have developed mechanisms to evade the RM system in numerous ways (Samson et al., 2013). Firstly, phages have acquired a methyltransferase gene themselves, or they stimulate the host methyltransferases to offer protection to the phage

genome. Secondly, phages can avoid this RM system through inhibition of the endonuclease activity by producing specific proteins, e.g., Ocr protein from phage T7, which blocks the active site of the restriction endonuclease. Another example is a *Bacillus subtilis* phage that incorporates unusual bases in their genome such as 5-hydroxymethyluracil instead of thymine and thus avoiding action by the restriction endonuclease (Stern and Sorek, 2011).

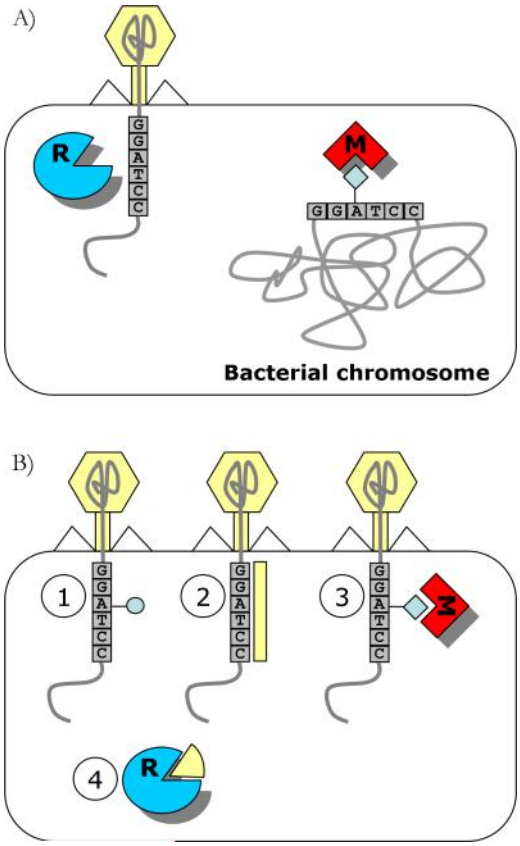

Figure 2.4: The restriction-modification defence system: A) a general overview of the methylasetransferase (M) and the restriction endonuclease (R) activity. B) Examples of evading the bacterial system 1: Incorporation of unusual bases 2: Masking the restriction site with phage proteins 3: Using the methyltransferase for masking the phage genetic material 4: inactivation of the restriction endonuclease (Stern and Sorek, 2011).

2.4.3 CRISPR/Cas defence system

CRISPR/Cas is an adaptive, widespread mechanism that archaea and bacteria use for protection against viral infections by breaking down the foreign DNA. It is used in 40% of the bacteria and 90% of the archaea (Stern and Sorek, 2011). The mechanism is similar to RNA interference (RNAi) in eukaryotes. CRISPR/Cas uses small RNA molecules for specific sequence detection and neutralization of foreign genetic material. Although this is a powerful mechanism to target phage sequences, phages have

found different ways to keep the arms-race going.

Firstly, a mutation or recombination in the target sequence of the phage can prevent the CRISPR/Cas from recognizing the phage because the original DNA sequence stored in the CRISPR loci does not match the mutated sequence any longer (Stern and Sorek, 2011). Secondly, anti-CRISPR genes are preventing the activity of the CRISPR/Cas system through several mechanisms. Bondy-Denomy et al. (2013) found anti-CRISPR genes that prevent the activity of the CRISPR/Cas system. However, they did not unravel the actual mechanisms of these anti-CRISPR genes. Two years later, Bondy-Denomy et al. (2015) found three anti-CRISPR/Cas genes and unveiled their mechanisms. Two of the translated proteins blocked the DNA-binding activity of the CRISPR-Cas complex by interacting with different protein subunits leading to steric or non-steric modes of inhibition. The third protein binds with the Cas helicase-nuclease and prevents its recruitment to the DNA-bound CRISPR-Cas complex. Even anti-anti-CRISPR genes have recently been discovered and they inhibit the anti-CRISPR system (Tang, 2019).

2.4.4 Are defence systems a burden or a benefit?

In this section, the pros and cons of the microbial immune systems are shortly reviewed. Evidently, the phage infecting the host and killing it is a strong disadvantage to the bacterial host. However, phages also benefit host fitness by supplying new possible beneficial genes by enabling horizontal gene transfer (Stern and Sorek, 2011). Although, two disadvantages come with encoding phage defence mechanisms. The first disadvantage is the energy that is needed to carry additional genetic cargo. A second disadvantage is autoimmunity, this is the immune system that recognises its own cells and tissues as foreign, resulting in an immune response against itself. For example, in RM systems the protecting methyltransferase is less stable than the restriction enzyme, which can cause the restriction enzyme to work on the host genetic material (Samson et al., 2013). The CRISPR/Cas system is also not perfect and the periodic acquisition of bacterial genetic material leads to spacers that target the bacterial genome (Stern and Sorek, 2011). Paradoxically, phages themselves also carry anti-phage defence mechanisms. This is to defend their already infected host against competitors.

2.5 Phage therapy

2.5.1 Historical aspects

As an ancient proverb states: 'The enemy of my enemy is my friend' and that is the essence of phage therapy. In this therapy, doctors use phages to combat bacterial pathogens. Phages can certainly be called enemies of our enemies as they are natural predators of bacteria (Hesse and Adhya, 2019). Phage therapy is the application of phages in clinical or veterinary context to combat bacterial infections. Additionally, phages can also be used as biological control agents that reduce the number of bacteria in food (Kutter and Sulakvelidze, 2005).
It was in 1915 that Frederick W. Twort made a rather odd observation that is now known as the discovery of bacteriophages. When he was growing *Vaccinia virus* on agar media in the absence of living cells, Twort noted that many colonies of *Micrococcus species* grew. Only these colonies appeared watery or glassy. He noted under the microscope that these colonies had degenerated into small granules that were coloured red with Giemsa stain. Twort himself said that he could not draw definitive conclusions from this (Kutter and Sulakvelidze, 2005). He thought it was a living protoplasm or an enzyme with the power of growth. It was Félix d'Herelle, in 1917 that concluded it had to be a microbe that was antagonistic to bacteria and that caused them to lyse. d'Herelle worked at the Pasteur Institute in Paris and from 1919 on, the institute started applying phages in therapy but abandoned its use once cheap and broad-host-range antibiotics became available (Kutter and Sulakvelidze, 2005). Only in the east of Europe, researchers kept applying phage therapy to treat wound infections, gastroenteritis, sepsis and other ailments (Pirnay, 2014). Because of the lack of scientific knowledge and the advent of antibiotics phage therapy was left behind until now. The recent and rapid emerge of antibiotic-resistant superbugs has forced scientists and companies to search for alternatives, leading to, among others, the founding of AmpliPhi Biosciences and Adaptive Phage Therapeutics (Schmidt, 2019). The former developing phage therapeutics for drug-resistant bacterial infections and the latter one is working on delivering phage therapy to hospitals and others.

2.5.2 Three important aspects of phage therapy

There are three important aspects of phage therapy. Firstly, practitioners can employ either natural or genetically modified phages in phage therapy (Figure 2.5). Natural phages are phages that are found in nature and can just be collected from vari-

ous sources. Genetically modified phages are phages that are modified to enhance some therapeutic characteristic(s). For example, companies hunt for broad-host-range phages and then engineer them for the desired attributes such as improved penetration of a biofilm or greater efficiency of killing bacteria (Schmidt, 2019). Engineered phages have some advantages over natural ones, particularly in commercial development. Additionally, engineered modifications are patentable while natural phages are not (Schmidt, 2019).

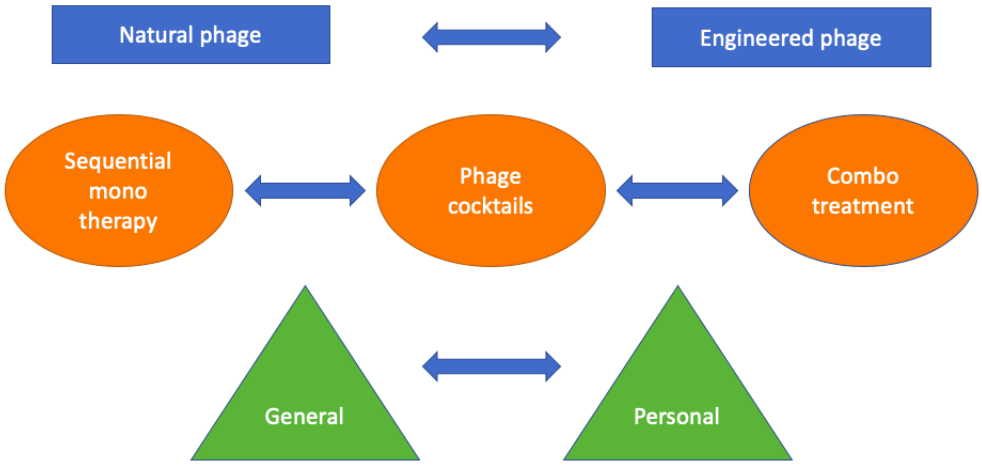

Figure 2.5: The three general aspects of phage therapy. Firstly, natural or engineered phages can be deployed. Secondly, phage cocktails can consist of only one single type of phage or a combination of different phages. Thirdly, general phage preparations or personalized preparations can be used (Schmidt, 2019).

Secondly, one can add only a single type of phage or a combination of different phages, called a phage cocktail. The resistance onset can partially be delayed by adding several phages in one cocktail, each targeting a different bacterial receptor (Schmidt, 2019). Depending on the genetic diversity of the bacterial species that has to be combated there will be more or fewer phages needed in the phage cocktail (Hesse and Adhya, 2019). If the bacteria have low genetic diversity and thus a limited number of different phage receptors, like *Staphylococcus aureus*, it is sufficient to treat the infection with a few different phages. Conversely, *Acinetobacter baumannii* is a pathogen exhibiting high genetic diversity and therefore is better treated with a large number of different phages (Schmidt, 2019).
Phages in the phage cocktail mostly have a complementary feature like their host-range. This is to improve the efficacy of the cocktail and to minimize the resistance onset.

A third important aspect of phage therapy is the use of general phage preparations versus personalized phage preparations. In some East-European regions, pharmacies offer fixed cocktails against the most common bacterial infections. Treatment can also be personalized. This means that the pathogens causing the infection need to be identified beforehand, after which phages that are most effective against a patient's specific bacterial strains are applied. It goes without saying that the personalized cocktail is more expensive than the fixed cocktail. Additionally, personalized cocktails are only used for chronic infections to ensure that there is enough time to select for the specific phage. Also, according to Örmälä and Jalasvuori (2013), fixed cocktails could induce resistance towards those phages in the fixed cocktail, resulting in a need to constantly change those phages in the cocktails.

2.5.3 Comparing antibiotics with phage therapy

It is of utmost importance that people understand the difference between antibiotics and phage therapy so that previous mistakes are avoided. This includes overuse and misuse of the therapeutic agent. Phages and bacteria are constantly in an ongoing evolutionary arms-race, this is why phages could be one of the potential solutions towards the post-antibiotic era (Gordillo Altamirano and Barr, 2019). Phages can be administered to humans in many ways, for example, orally, rectally, locally, with aerosols and intravenously (Sulakvelidze et al., 2001a). The way of administrating phages influences the effectiveness of the treatment, but this will not be further discussed in detail.

Firstly, the most distinguishing characteristic is that most antibiotics have a broad-host-range while phages are very specific (Hesse and Adhya, 2019). In general, most phages are known to be strain-specific. As a result, they can precisely target pathogenic bacteria, while leaving beneficial microbiota unaffected. Because of this characteristic, it is important to know which bacteria cause the infection before physicians can treat a patient (Pirnay et al., 2018). Thus, phage specificity can be considered a big advantage and disadvantage at the same time. Phage therapy requires to identify the bacterial pathogen at strain level before applying the correct phage cocktail to the patient. This can cost a lot of time in wet-lab circumstances and is a major reason why predicting interactions *in silico* between phages and bacteria can be a big step forward in phage therapy (Monteiro et al., 2019).

Secondly, phages replicate exponentially after adding them to the site of infection. Therefore, phage preparations can be administered at lower concentrations compared

to antibiotics (Sulakvelidze et al., 2001a). However, the exponential behaviour of phage replication also poses regulatory challenges for practical application. Because of the size of phages, relatively big, phages are quickly removed from circulating in the body. If this happens too fast and thus no bacterial cells were infected, the replication cycle ends.

Thirdly, bacteria can develop resistance against antibiotics as well as phages. However, in contrast to antibiotics, phages can coevolve to regain the ability to infect their host (e.g., by targeting a new surface receptor). But when resistance occurs it should be possible to select another phage that still has its effectiveness against the phage-resistant bacteria or use a phage that is specially trained against the resistant bacterium (Pirnay et al., 2010). Finally, developing a new antibiotic may take around thirteen years while selecting new appropriate phages is a relatively fast process that is mostly completed in weeks or months (Sulakvelidze et al., 2001a). Although, an antibiotic that is already on the market and is still effective can be prescribed immediately.

The comparison above is made between lytic phages and antibiotics, lysogenic phages are not useful for killing bacteria immediately and are thus not included in this comparison. However, it is possible to convert lysogenic phages into a lytic variant (Schmidt, 2019). It has to be mentioned that bacterial resistance to phages is not inevitable but according to Carlton (1999), it happens at a tenfold lower rate than with antibiotics. Since phage resistance is not correlated with antibiotic resistance, it could also be added in combination, a strategy that is also being employed at the Queen-Astrid military hospital here in Belgium. Scientific research shows that bacteria that are getting resistant to phage therapy are sometimes resensitized to antibiotics, which leads to a synergetic effect (Hesse and Adhya, 2019). This is because antibiotics and phage therapy use alternative mechanisms to target bacteria.

2.5.4 Phage engineering

Several companies are trying to enhance the ability of phages to interact with certain pathogens by engineering the natural phage. They hunt for natural phages with a broad-host-range and then engineer them for the desired attributes, for example, more immunogenicity, greater access to biofilms and improved pharmacology. Furthermore, researchers can engineer phages to overcome host-range limitations. Another example is transforming lysogenic phages into lytic phages by deleting the

phage repressor (Dedrick et al., 2019). By doing so scientists eliminated the chance that the phage goes into the lysogenic cycle and that genes are transferred by transduction (Monteiro et al., 2019). A variety of ways exist to convert phages from a lysogenic lifestyle to a lytic lifestyle (creating so-called vir mutants), as reviewed by Monteiro et al. (2019).

In addition, synthetic biology can be used to improve the efficiency and safety of the temperate phages in such a way that the temperate phages can have a similar effect as using strictly virulent phages (Monteiro et al., 2019). Moreover, due to the integration of the temperate phage, the bacterium gets insensitive for further phage infections. Superinfection immunity causes problems in phage therapy. Not only are temperate phages very abundant in nature, but they are also easy to find and isolate because of the advances in high-throughput sequencing. Since bacterial DNA usually also contains the DNA of one or a few lysogenic phages, marked by an integrase, they are very easy to recognize in genome databases where the isolation of the genetic material of a strictly lytic phage is rather difficult. Park et al. (2017) have succeeded in engineering a temperate phage that delivers a synthetic gene network which interferes with the bacterial CRISPR/Cas system and eventually kills the bacterium. Yosef et al. (2015) engineered a temperate phage to resensitize the bacterium to antibiotics. Strictly lytic phages will probably remain the preferred choice for the following years in phage therapy, although vir mutants can be used to target different bacterial surface receptors than the strictly lytic phages.

2.5.5 Disadvantages of phage therapy

Phage therapy does have some disadvantages, of course, and it would be naive to believe otherwise. Firstly, Sarker et al. (2012) pointed out that there is an enormous phage loss during gastric passage conditions. Applying a larger dose could be a solution to this problem.
Secondly, most bacterial infections involve several different pathogens. Therefore, targeting only one of these pathogens will not resolve the infection. This problem is solved by using complex phage cocktails containing phages against many strains. However, the genetic variability in the bacterial species is also becoming a problem, for example, phage cocktails against *E. coli* contain ten different phages but only target 50% of the bacterial strains (Brüssow, 2019). When more phage strains are added, interferences problems start to appear.
Thirdly, sometimes pathogens are present in such a low concentration that the replication threshold is not exceeded. Because phages are relatively big entities they are

quickly removed out of the body. Therefore, phages are sometimes not able to infect enough bacterial host cells and then no productive phage infection chain can occur (Brüssow, 2019).
Fourthly, collecting and plating the specific phage in time is sometimes difficult when the infection is acute since short disease durations need an early phage intervention. Even when the cocktail is given on time phages still need to have access to the pathogen.
Finally, because phages and bacteria have been co-evolving for millions of years this implies that a phage will never fully eliminate a bacterium due to phage-resistant species. Indeed, phages need their bacterial host to replicate, so they have no advantage of completely eliminating their means of reproduction. This is why the combination of phage therapy with classical antibiotics can prove useful in cases where both have a synergetic effect (Monteiro et al., 2019). The combined use of the two can lead to complete elimination of the bacterial species.

A major drawback of phage therapy is that phage preparations were classified as medical products (European Union) or drugs (US), based on the literal implementation of definitions (Pirnay et al., 2018). As a result, a lot of costly and time-consuming procedures are needed for the development of phage-based antibacterials. However, the characteristic of phages that makes them so useful, to overcome the bacterial resistance so fast, is completely worthless in this case (Pirnay, 2014). For example, fifty German patients were suffering from an infection of *E.coli* O104:H4. No effective alternatives were available and because of the strict medicine regulation, phages could not be used (Pirnay, 2014). It would take years to make all the arrangements for the medicine regulation for the new O104:H4 phages. The Declaration of Helsinki in June 1964 is a way to avoid this strict medicine regulation (Association, 2013). It says that humans can be used for experiments if no other option is available and if the patient agrees. With the Declaration of Helsinki, physicians can apply phage therapy and override any national or local law.

Another drawback is the limitation in intellectual property for natural entities such as genes or phages. All these drawbacks make it almost impossible to manufacture customized phages. In Belgium, there is since recently a new law that allows physicians and pharmacists to personalize the patient treatment and produce medicines that are not commercially available, better known as magistral preparations (Pirnay et al., 2018). In the Belgian and European law, the notion of a magistral preparation is defined as "any medicinal product prepared in a pharmacy in accordance with a medical prescription for an individual patient" (Pirnay et al., 2018). Magistral preparations are

mixed from their constituent ingredients by a pharmacist for a specific patient according to the prescription of a physician. Magistral preparations are increasing because of the demand for personalized therapies and rare diseases. In the face of this looming antibiotic crisis, it could be very helpful to consider phage therapy applications in the medical world (Pirnay, 2014).

CHAPTER 3

AN ESKAPE-BASED GENOME DATABASE

3.1 Introduction: Why do we focus on ESKAPE organisms?

Several highly problematic bacterial pathogens have been grouped in a category that has received the name ESKAPE. This acronym denotes the following bacterial species: *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumanii*, *Pseudomonas aeruginosa* and *Enterobacter* species. Simultaneously, this acronym symbols their ability to escape the effects of antibiotics through evolutionarily developed mechanisms (Rice, 2008). Moreover, these are six of the most important pathogens when it comes to antimicrobial resistance. The ESKAPE group consists of both Gram-positive and Gram-negative bacteria. They mostly occur in the human gut, on the skin and in the environment (Santajit and Indrawattana, 2016). In particular, ESKAPE pathogens differentiate themselves from other pathogens due to their increased resistance to commonly used antibiotics such as vancomycin, carbapenems and penicillin (Santajit and Indrawattana, 2016). The increased resistance and clinical significance result in a necessity to understand their mechanisms of resistance.

The ESKAPE organisms are commonly found in hospitals both in developing and developed countries (Rice, 2008). Because of the reasons mentioned in the previous chapter, the emergence and escalation of resistance among ESKAPE pathogens is increasing (Natarajan and Usha, 2018). Especially for these pathogens, the increase in resistance is alarming, because they cause two-third of all the healthcare-associated infections, also called nosocomial infections. Nosocomial infections are infections that patients contract during hospital care or other healthcare facilities that were not present or incubating at the time of admission (Khan et al., 2017). These infections lead to an overall increase in mortality and morbidity in hospitals and intensive care

units. Most strains are able to facilitate horizontal gene transfer, allowing them to pass resistance genes from one pathogen to another (Pendleton et al., 2013). This is especially problematic with nosocomial infections, where selection pressure is even higher due to constant exposure to antibiotics.
Multidrug resistance is amongst the top three threats to global public health and is caused by an inappropriate use or unneeded prescription of antibiotics (Santajit and Indrawattana, 2016). In 2011, Magill et al. (2014) conducted a survey in the United States about nosocomial infections which counted 75,000 deaths in that year alone.

3.2 Data collection and preprocessing

In this thesis, the goal is to accurately predict interactions between bacteria and their prophages using machine learning models. This prediction problem will be represented as a binary classification (a binary outcome that will be represented by a one or a zero). For this purpose, bacterial sequences were collected from the National Center for Biotechnology Information (NCBI), a database which is publicly available (NCBI Resource, 2018). There are databases publicly available, with known interactions between a bacterium and its (pro)phage(s). However, these databases mostly annotate the (pro)phage's host at the species level, not at the strain level (Leite et al., 2018). In practice, bacterium-phage interactions are specific at the strain level. Consequently, these databases are not suited to construct machine learning models to aid practical phage therapy. To circumvent the problems of annotation, we chose to collect complete bacterial genomes and detect prophages in these genomes. By doing so, the detected prophages together with each corresponding bacterial genome by definition constitute an interaction at the strain level. We collected bacterial genomes of the ESKAPE organisms because they are the priority in research towards alternative antimicrobial solutions.
One could argue whether this is relevant since prophages have a lysogenic replication cycle and in most applications phages with a lytic replication cycle are used but it is possible to convert temperate prophages into virulent phages (Monteiro et al., 2019). Additionally, the gained knowledge can be used for treatment with virulent phages since they use the same infection machinery (tail fibers, tail spikes and tail tips).

3.2.1 ESKAPE genome collection and filtering

Firstly, sequenced genomes of the ESKAPE organisms were collected from NCBI (NCBI Resource, 2018). Data collection was restricted to completely sequenced genomes in order to increase the relevance of subsequent processing steps and analyses. Sec-

Table 3.1: The number of unique bacterial strains and the total collected genomes per bacterial species.

Organism	Unique strains	Total genomes
Enterococcus faecium	81	85
Staphylococcus aureus	474	478
Klebsiella pneumoniae	369	376
Acinetobacter baumannii	161	164
Pseudomonas aeruginosa	145	149
Enterobacter cloacae	26	26
Enterobacter aerogenes	15	15

ondly, raw collected data was filtered. Duplicate entries and genomic sequences with unknown nucleotides were removed. The raw ESKAPE genome database consists of 1,515 complete genome sequences. After filtering for the sequences that contain unknown nucleotides, 1,457 genomes were left. We did not apply a filter for plasmids because we reasoned that plasmids can also contain prophages. 1,293 of the 1,457 bacterial genomes contained at least one prophage (see further below). 1,271 unique bacterial strains were found amongst the 1,293 bacterial genomes (Table 3.1), i.e., a strain that only appears once in the database. The largest bacterial species is *Staphylococcus aureus* with 474 unique strains and the smallest group is the *Enterobacter aerogenes* with 15 unique strains. A Python script was used to automatically access the NCBI database and collect ESKAPE genomes. Furthermore, we have collected information on the bacterial strain annotation, sequencing method and description for the bacterial genomes.

3.2.2 Prophage detection

After the collection of bacterial genomes, the next step is to detect the prophages. This was done with PHASTER using the complete bacterial genomes as input. PHASTER is a tool to identify and annotate phages within bacterial genomes or plasmids. Hence the name PHAge Search Tool- Enhanced Release (Arndt et al., 2016). PHASTER is the upgraded version of the popular web server PHAST (Zhou et al., 2011). Of the many tools to detect prophages, we chose to work with PHASTER because PHASTER can work with an API and thus automate our prophage detections. PHASTER works with a relatively straightforward pipeline. Bacterial genome sequences can be supplied in a FASTA or GenBank format and then a BLAST search is performed against a custom (pro)phage database that combines protein sequences from the NCBI phage database and the prophage database made by Srividhya et al. (2006). Phage-like genes are then clustered into prophage regions using DBSCAN (Ester et al., 1996). Again, a BLAST search is performed but for the non-phage genes against a non-redundant bacterial protein database. Every detected prophage region is assigned a complete-

Table 3.2: The number of genomes for each organism from NCBI that contained at least one active prophage and the total genomes found.

Organism	Genomes with active prophage(s)	Total genomes (Percentage)
Enterococcus faecium	85	92 (92.4%)
Staphylococcus aureus	478	546 (87.6%)
Klebsiella pneumoniae	376	412 (91.2%)
Acinetobacter baumannii	149	178 (83.7%)
Pseudomonas aeruginosa	164	182 (90.1%)
Enterobacter cloacae	26	28 (92.9%)
Enterobacter aerogenes	15	19 (79.0%)

Table 3.3: The number of active prophages for each bacterial species and the average number of active prophages per organism.

Organism	Active prophages	Number of active prophage(s) / total genomes
Enterococcus faecium	192	2.1
Staphylococcus aureus	1,083	2
Klebsiella pneumoniae	1,299	3.2
Acinetobacter baumannii	357	2
Pseudomonas aeruginosa	468	2.6
Enterobacter cloacae	94	3.4
Enterobacter aerogenes	39	2.1

ness score based on the proportion of phage genes in the identified region: If this score is higher than 90, the prophage region gets the label of 'active'. A score between 70-90 receives the label of 'questionable' phage region and below 70 is the 'incomplete' zone. We filtered for active prophages with a genome size larger than 10 Kb, as nucleotide sequences smaller than 10 Kb can be difficult to distinguish from other integrative elements. These steps in the database construction are based on methods used before by (Costa et al., 2018) and (Shen et al., 2020). The former used PHAST to discover prophages in the *Acinetobacter baumannii* genomes and the latter explored the *Klebsiella pneumoniae* genomes with the PHAST web server as well (Zhou et al., 2011). The PHASTER API was accessed by a Python script to automatically upload the bacterial genome sequences (as FASTA files) to the PHASTER server for processing and detecting active prophages.

The database consists of 1,293 unique bacterial genomes and 3,532 prophage genomes. After having detected the prophage sequences, we have analyzed their GC content and length (Figures 3.1 and 3.2). The average GC% lays between 33.15% (*Staphylococcus aureus*) and 62.62% (*Pseudomonas aeruginosa*). In Figure 3.2 one can see the boxplot that represents the length of the prophages per bacterial species. The smallest prophage genomes has a length of 10,006 basepairs (*Pseudomonas aeruginosa*) and the largest sequence has 134,280 basepairs (*Staphylococcus aureus*) while the length of the bacterial host varies between 2.73 Mbp (*Enterococcus faecium*) and 6.63 Mbp (*Pseudomonas aeruginosa*). It was found that 1,293 of the 1,457 bacterial

genomes contained active prophages (Table 3.2). A total of 3,532 active prophages were found in these 1,293 bacterial genomes (Table 3.3). On average, every bacterial genome contained between 2 and 3.4 active prophages. *Enterobacter cloacae* had with an average of 3.4 active prophages per genome the highest number of prophages and *Staphylococcus aureus* and *Acinetobacter baumannii* with only 2 active prophages per genome the lowest number.

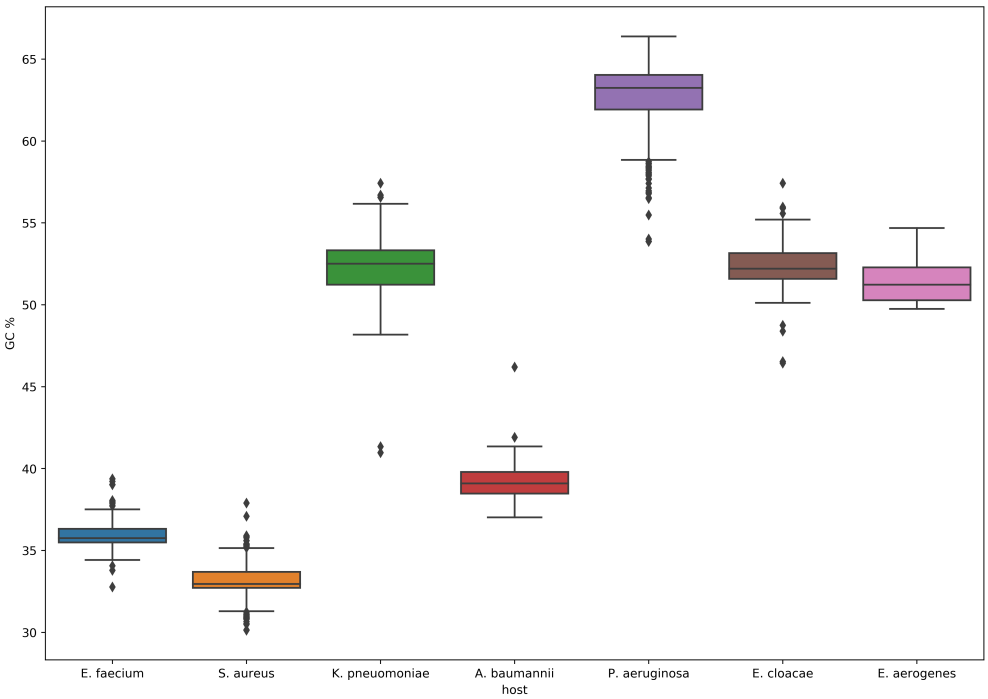

Figure 3.1: A boxplot of the prophages GC% per bacterial species.

3.2.3 Prophage processing and final database construction

Since the goal is to use supervised learning methods to accurately predict bacterium-phage interactions, an interaction matrix containing known interactions (also called labels) is needed. In this case, this is a positive interaction (one) or a negative interaction (zero) between a bacterium and phage. Since the database consists of 1,293 bacteria and 3,532 prophages, the interaction matrix has the same dimensions, namely 1,293 by 3,532. An interaction is indicated with a one and no interaction with a zero. Each row represents a bacterium and each column a prophage. Note that there is only one interaction per prophage (column) but that one bacterium (row) can contain mul-

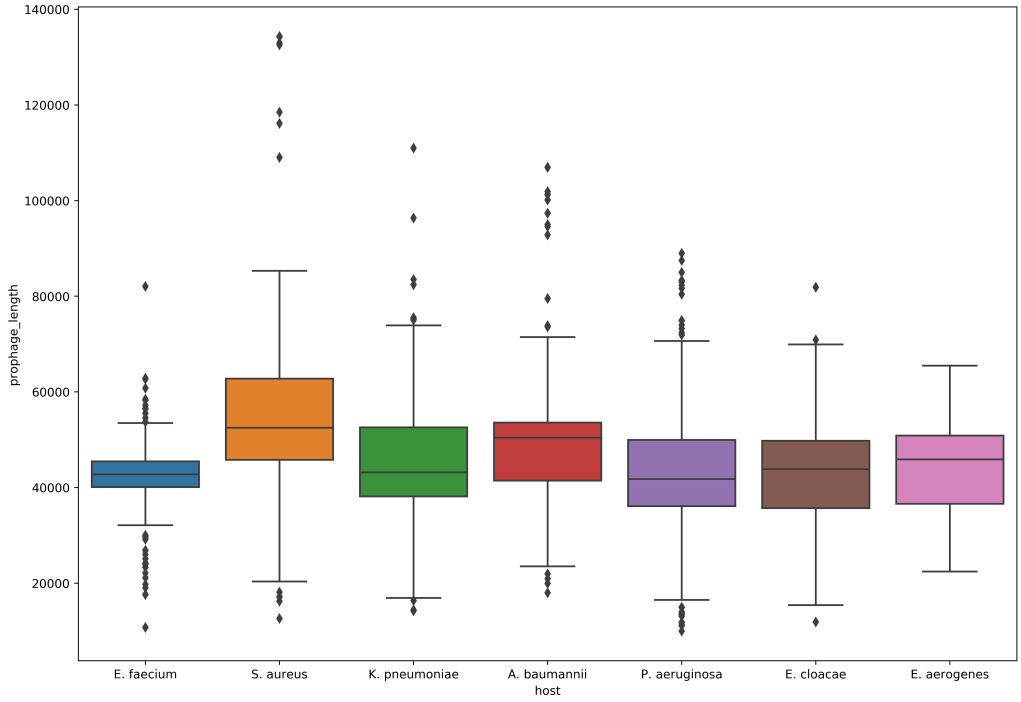

Figure 3.2: A boxplot of the prophages length per bacterial species.

tiple interactions. Most common machine learning methods require both instances (data points) from a positive class, as well as a negative class. This, by definition, results in a binary classification in which both classes are tried to be separated as well as possible. In this case, both known positive interactions between bacteria and phages as well as known non-interacting bacterium-phage pairs (negative interactions) are needed. A negative interaction implies that there is no interaction between a given phage and a bacterium. Which is not the same as not knowing whether there is interaction. However, as our data collection approach focused on detecting prophages in bacterial genomes, these constitute only positive interactions. Besides, negative interactions are scarce in most publicly available databases. For simplicity, we will start with assuming that all non-observed interactions are negative interactions but keep in mind that this is not identical. One could say that this is a naive approach but considering the strain specificity of bacterium-phage interactions, this is not such a terrible idea.

To summarize, the final database consists of an interaction matrix with the observed and constructed labels and both the bacterial genomes and prophage sequences grouped in two separate files (for storage capacity reasons and ease of use). This is the starting point for further feature engineering methods.

3.3 Constructing the primary feature matrices

All 3-mer frequencies were computed and normalized for every bacterial and prophage sequence and stored in a feature matrix. The features were constructed separately for the bacteria and the prophages. This gives us for every bacterial sequence and prophage a number of 64 features, namely, the 3-mer frequencies. We chose to start with the 3-mers frequencies because of simplicity. Since we are the first to predict these interactions based on the *k*-mer features with machine learning, there is no literature that describes what number of *k* is ideal. To summarize, the bacteria feature matrix has 1,293 rows with 64 columns and the prophage feature matrix has 3,532 rows with 64 columns. There is thus a separate feature matrix for both the bacteria and the prophages.

3.4 Data exploration and visualization

To visualize and understand our database, the Principal Component Analysis (PCA) method (Paul et al., 2013) together with a t-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008) was performed on the prophage sequences. These methods are both used to reduce high dimensionality in datasets.

PCA was performed on the features of the prophages to explore and visualize the 3-mer frequencies for the prophages with only two components. We looked at the species level instead of at the strain level because there were too many unique strains to visualize properly. The initial PCA could group the prophages with only two components relatively well (Figure 3.3). Only the prophages that infect *Enterobacter aerogenes* and *Enterobacter cloacae* had overlap with the prophages that infect *Klebsiella pneumoniae*. This is probably because *Enterobacter* is our smallest bacterial species and PCA tries to maximize the variability for the entire dataset, thus potentially focussing less on the smallest bacterial species. According to the first principal component, one could almost say that the PCA distinguished between prophages that infect gram-positive and gram-negative bacteria (Figure 3.3). Since *Staphylococcus aureus* and *Enterococcus feacium* are the only two gram-positive species, it appears trivial that they are more different than the other prophages. Surprisingly, the prophages that infect *Acinetobacter baumannii* showed more similarity to the prophages that infect gram-positive species than with the other prophages from gram-negative bacteria, represented as two principal components at least. One could also say that according to the first principal component the prophages that infect *Staphylococcus*

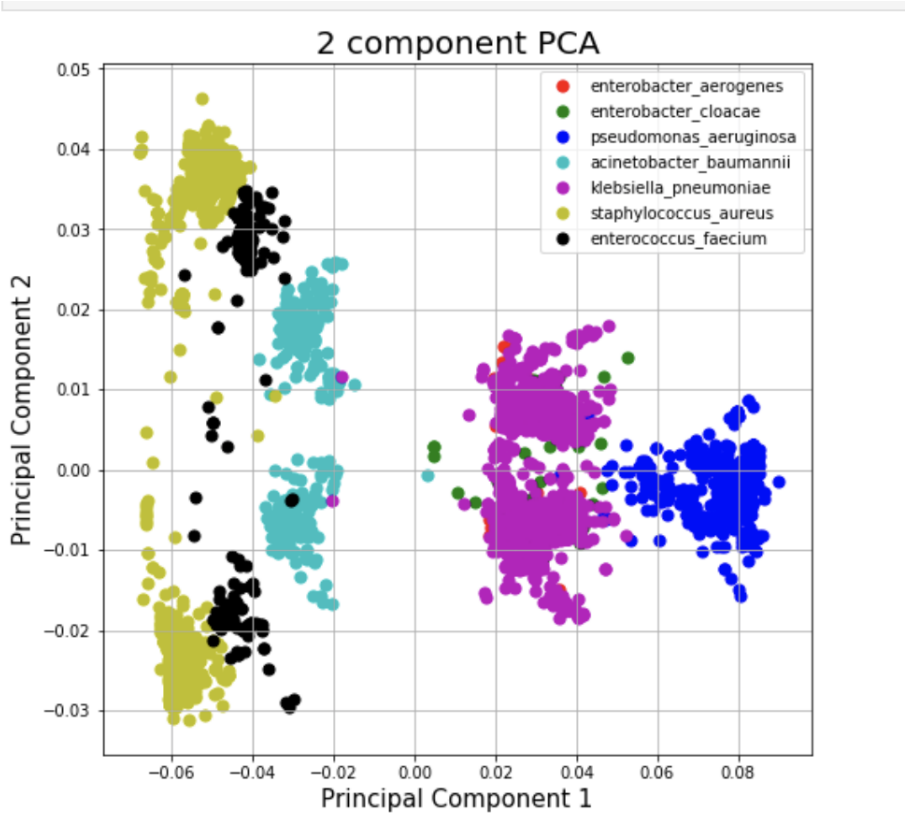

Figure 3.3: A Principal Component Analysis of the prophage features

aures and *Pseudomonas aeruginosa* are the least similar to each other. The first principal component explained 77.49% of the variance, while the second principal component explained 11.24%. Analysis of the loadings showed that the most important features for the first principal component are the frequencies of the GCG and TTT 3-mers with weights of respectively, 0.23 and -0.32. In addition, the fact that the PCA could distinguish pretty good between different prophages with only two components, potentially signals high conservation among prophages related to the same species. We hypothesize that representing an entire genomic sequence by only the 3-mer frequencies leads to less diversity and is probably too simplistic as representation.

Looking at the second principal component, we hypothesized that some of the prophage genomes are present as reverse complements in our database. This finding was strengthed by the fact that the largest loadings, in absolute value, of the second principal component are the frequencies of the AAA 3-mer and the frequencies of the TTT 3-mer. To investigate this, we took a subsample of 50 prophage sequences that infect *Staphylococcus aureus*. 25 of the 50 prophages had a second principal component that was higher than 0.01 and 25 of the prophages had a second principal component that was lower than 0.00. The original 50 prophages sequences were analyzed with PCA seen in Figure 3.4. We computed the reverse complements for the

25 prophages that had originally a second principal component of below 0.00 and computed a second PCA for the subsample of 50 prophages that now included 25 reverse complement sequences (Figure 3.5). In this plot, the PCA could not distinguish the prophage sequences very well. The first principal component showed, in absolute values, the highest loadings for ACG (0.23) and AAA (-0.43). This further supports our observation that the second principal component reflects the variation between forward and reverse complemented phage sequences. In addition, the presence of reverse complement sequences was also confirmed by sending a toy example to PHASTER in which an already detected prophage was replaced by its reversed complement sequence in its host bacterial genome. Here again, PHASTER predicted the presence of the prophage as before.

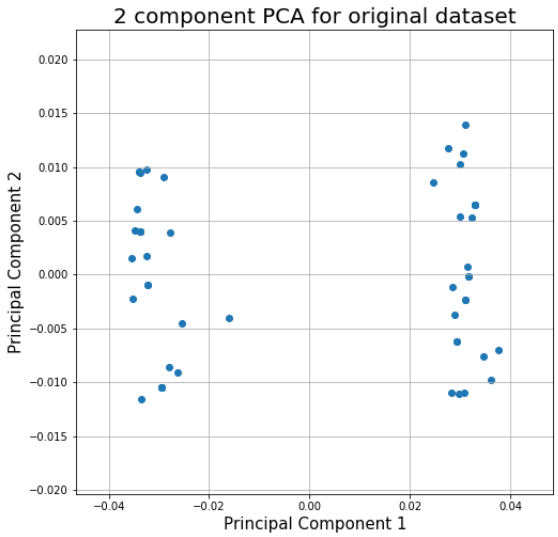

Figure 3.4: A Principal Component Analysis on a subset of 50 prophages that infect *Staphylococcus aureus*.

Since a prophage does not have one correct orientation (genes can occur in both directions because the genomic material of these phages is dsDNA) (Zeldis et al., 1973), we added the reverse complement of each prophage sequence to the prophage sequence itself. Every prophage sequence is now represented by the originally detected sequences, appended by the reverse complement of each sequence. Again, we computed the 3-mer frequencies for these new sequences in Python and did a third PCA (Figure 3.6). We added these reverse complements to each prophage to eliminate the variability that is explained by the difference between the prophages in their forward orientation and reversed orientation. The first principal component explained 92.19% of the variance and the second principal component 1.62%. The fact that PCA can explain 92.19% of the variance of our dataset in only one Principal Component hints to redundancy of the data. The largest loadings, in absolute value, for the first prin-

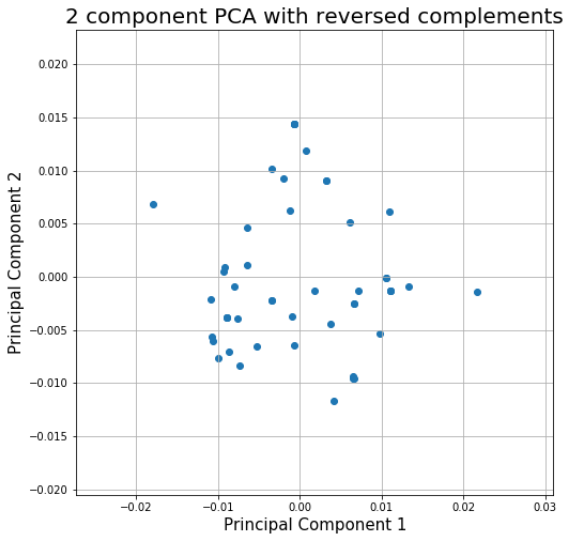

Figure 3.5: A Principal Component Analysis on a subset of 50 prophages that infect *Staphylococcus aureus* with computed reverse complements.

cipal component are the 3-mers GCG (0.22) and AAA (-0.29). The largest loadings, in absolute value, for the second principal component are the 3-mers CTG (0.28) and CGA (-0.27). There is no symmetrical element anymore that splits the prophages according to the y-axis. Adding the reverse complement of every prophage to the original prophage sequence handled the problem regarding the prophage sequences that could be present in both orientations. A pair plot for the largest loadings for the first principal component and the second principal component can be seen in Figure 3.7. On the diagonal one can see the densities for the largest loadings. For example, *Pseudomonas aeruginosa* has a noticeable higher frequency of the 3-mer CGA and a lower frequency of the 3-mer AAA than other prophages related to other bacterial hosts. High and narrow peaks are again hinting to redundancy in the data.

Another interesting note is that both AAA and TTT frequencies are negatively correlated for the first principal component, respectively -0.294 and -0.293. This means that prophages from species like *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* and *Enterobacter* contain a lower frequency of AAA and TTT 3-mers in their sequences than the prophages related to other bacterial hosts. Interesting is that these three bacterial species also have the highest GC% (Figure 3.1), resulting in a negative correlation between the GC% and the frequencies of both AAA and TTT 3-mers. Because of this similarity in GC% and both 3-mers the *Enterobacter* had overlap with the *Klebsiella pneumoniae* in the PCA (Figure 3.3).

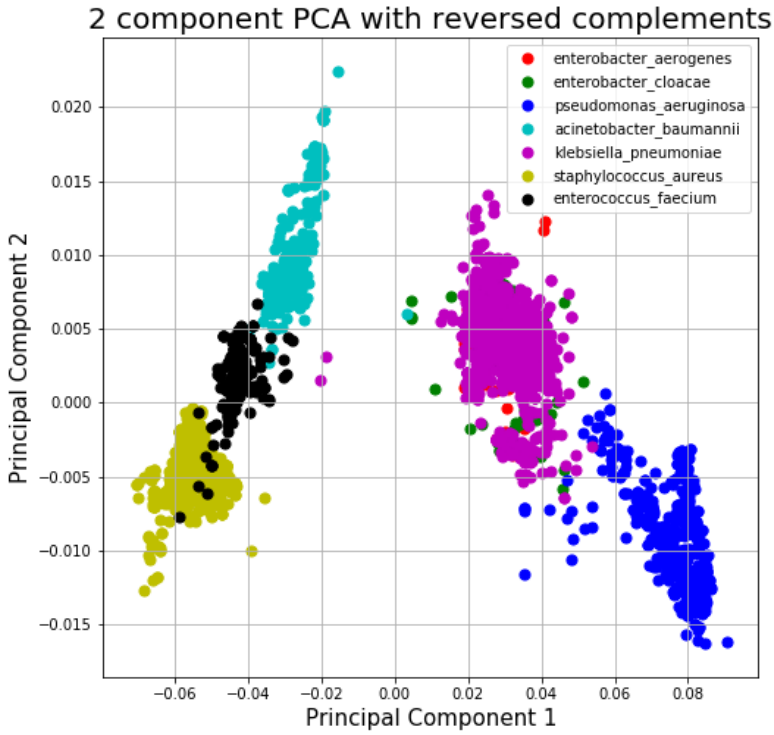

Figure 3.6: A Principal Component Analysis (PCA) on the prophage sequences with the reverse complements.

A t-SNE was also performed on the new prophage sequences (Figure 3.8). The idea of a t-SNE is to embed high-dimensional points in low dimensions in a way that respects the similarity between points so that clusters in the high-dimensional space can be preserved. Nearby points in the high-dimensional space correspond to the nearby points in the low dimensional space, and distant points in high-dimensional space correspond to distant embedded low-dimensional points (van der Maaten and Hinton, 2008). The t-SNE plot confirms that the prophages from *Enterobacter* are similar to prophage from *Klebsiella pneumoniae* since they are again plotted together. The exceptionally good separation between the different prophages suggests that the features are potentially too simplistic to represent complex phage sequences. If it is possible to differentiate that good with only two components for the t-SNE, there can not be that much diversity between the prophages. Representing entire genomes by their overlapping 3-mer frequencies potentially leads to an oversimplification of the prophage sequences, especially to distinguish intraspecies variability among them. However, through evolution phages have adjusted their 3-mers profiles to their host profile. This similarity could lead to accurate predictions at the species level and is probably the reason why t-SNE and PCA can distinguish that good for the prophages only using two components. Also, the t-SNE showed some outliers such as the black

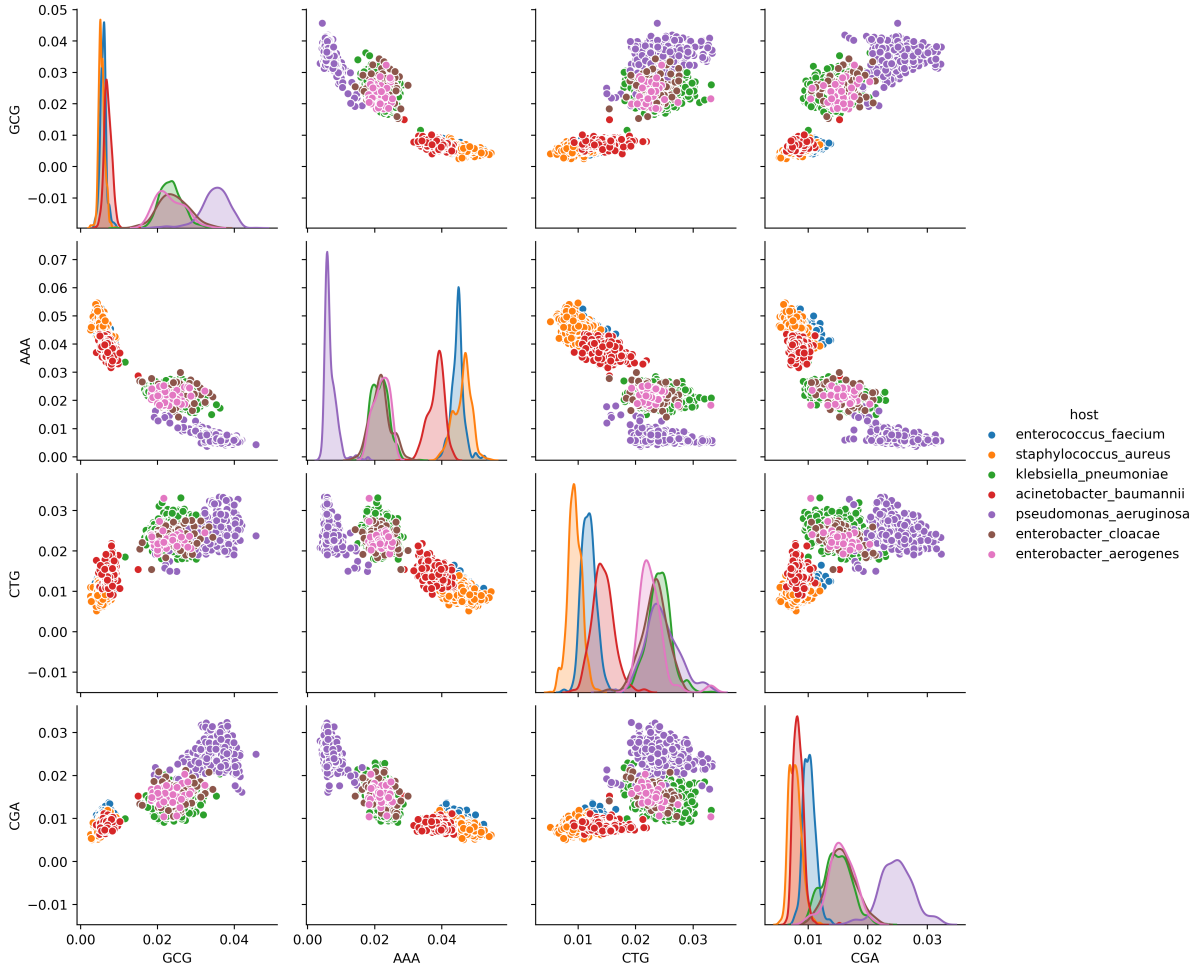

Figure 3.7: A pairplot of the largest loadings for the first principal component (GCG and AAA) and the second principal component (CTG and CGA). The diagonal represents the densitities for a given *k*-mer.

dots (*Enterococcus faecium*) in the yellow group (*Staphylococcus aureus*) but these were not further investigated.

3.5 Multiple sequence alignment

To further zoom in on the diversity (or similarity) between the prophage sequences, a whole-genome alignment was computed. However, because of limited compute resources, we were unable to align all of the prophage sequences per species. Typically, online tools for multiple sequence alignment limit the input size to between 2 and 4 MB. A subsample of 50 prophage sequences from *Staphylococcus aureus* was already 2.6 MB. This prophage subsample was the same subsample that was discussed in the previous subsection (3.4). 50 prophage sequences were collected from the original dataset. The reverse complements were computed for the prophages that had a second principal component lower than 0.00 since we hypothesize that the second

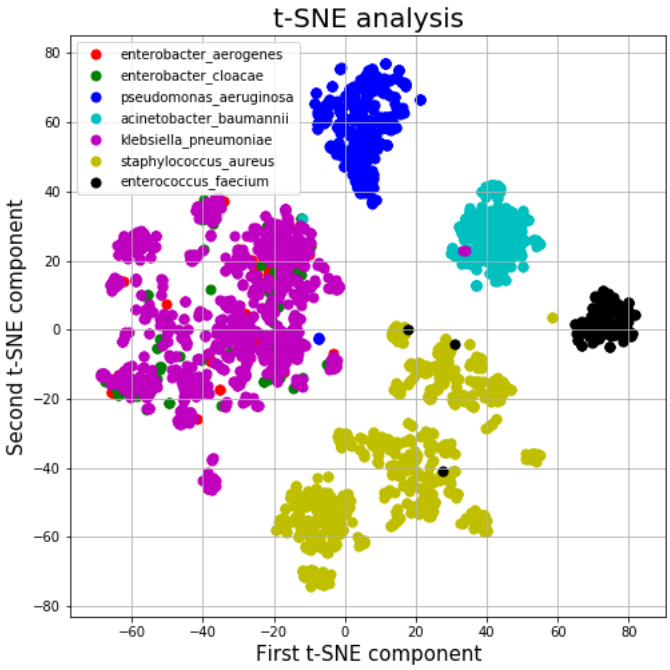

Figure 3.8: A t-distributed Stochastic Neighbor Embedding (t-SNE) on the prophages sequences with the reverse complements.

principal component is distinguishing between normal prophages and their reverse complement. This resulted in 25 original prophage sequences and 25 reverse complements of prophage sequences, all of which were related to *Staphylococcus aureus*. An MSA was done using Mauve (Darling et al., 2004). Mauve is a program for constructing MSAs in the presence of large-scale evolutionary events such as inversion and rearrangement. It employs algorithmic techniques that scale well in the lengths of sequences being aligned. All 50 sequences can be seen in Figure 3.10. Since it is a rather complex figure, we have zoomed in on the first 10 sequences in Figure 3.9. In this figure, we can see that each row is a different prophage genome. Each of the colored blocks surrounds a region of the genome sequence that aligned to a part of another genome and is presumably homologous and internally free from genomic rearrangement. When the block lies above the center black line, the aligned region is in the forward orientation. Consequently, when the block lies under the center black line, the aligned region is in the reverse orientation. Colored blocks in the first genome are connected by lines to similarly colored blocks in the other genomes. These lines indicate which regions in each genome are homologous. Inside each block, Mauve draws a similarity profile of the genome sequence. The height of the similarity profile corresponds to the average level of conservation in that region of the genome sequence. Regions outside the blocks lack detectable homology among the input genomes. In this figure, one can see that the prophages are kind of similar to each

other and contain a lot of horizontal gene transfer. This is as expected because all the 50 prophages are infecting the same bacterial species, e.g. *Staphylococcus aureus*. One can see some conserved regions (the red block and the purple block) and some more variable regions (the yellow block). Extrapolating these results to all prophages has to be done with caution since this is a small subset of the data. Ideally, an MSA is performed for more prophages but there is not yet an easy available system that can align many complete prophage sequences. Since performing an MSA on all the entire dataset would be too complex, an alternative is to perform a clustering analysis on the data and then align the clusters. But again, this is a computationally intensive process and was not further explored.

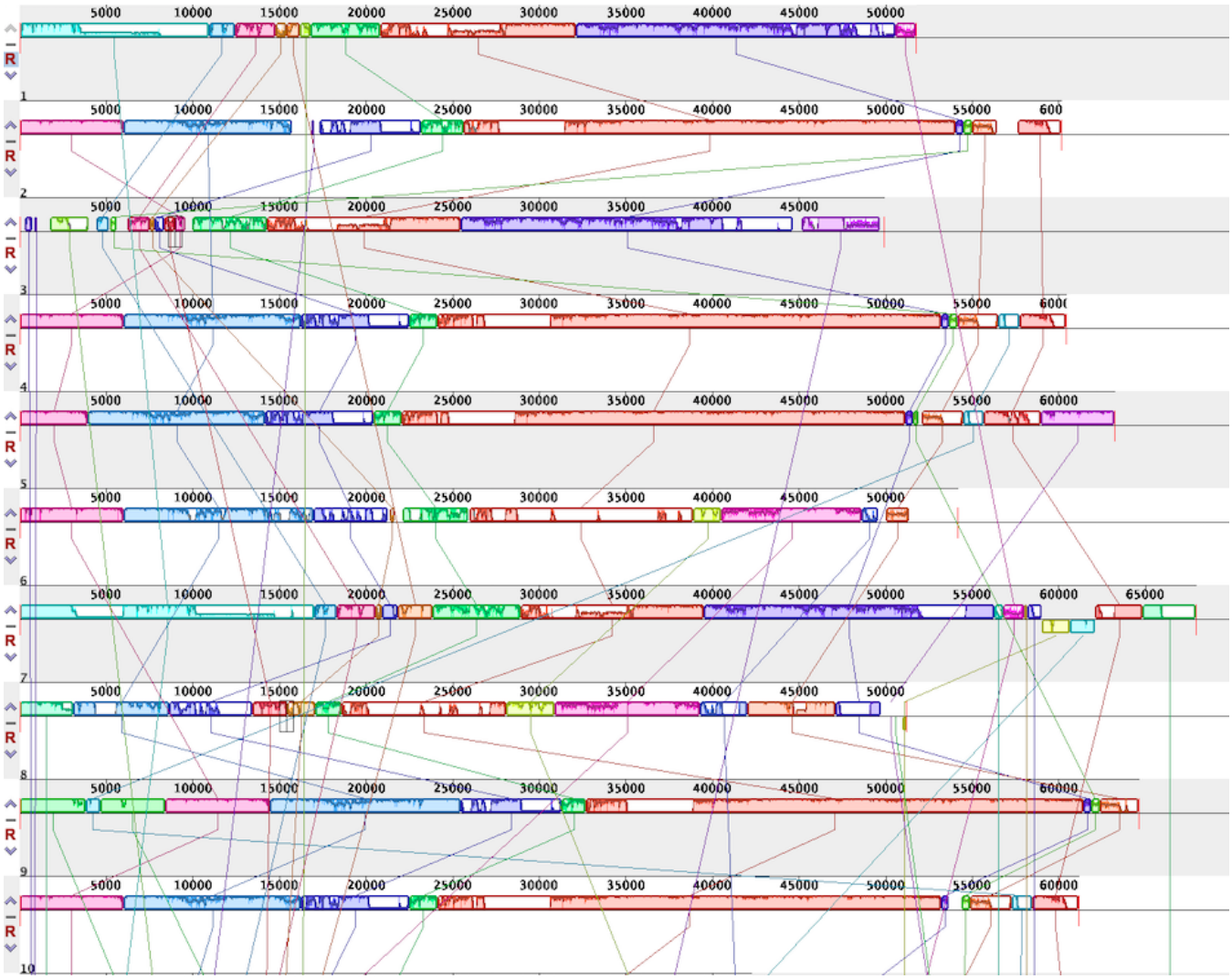

Figure 3.9: The first 10 prophages that infect *Staphylococcus aureus* for a multiple sequence alignment for 50 prophages sequences with 25 of them in normal orientation and 25 of them in reversed orientation.

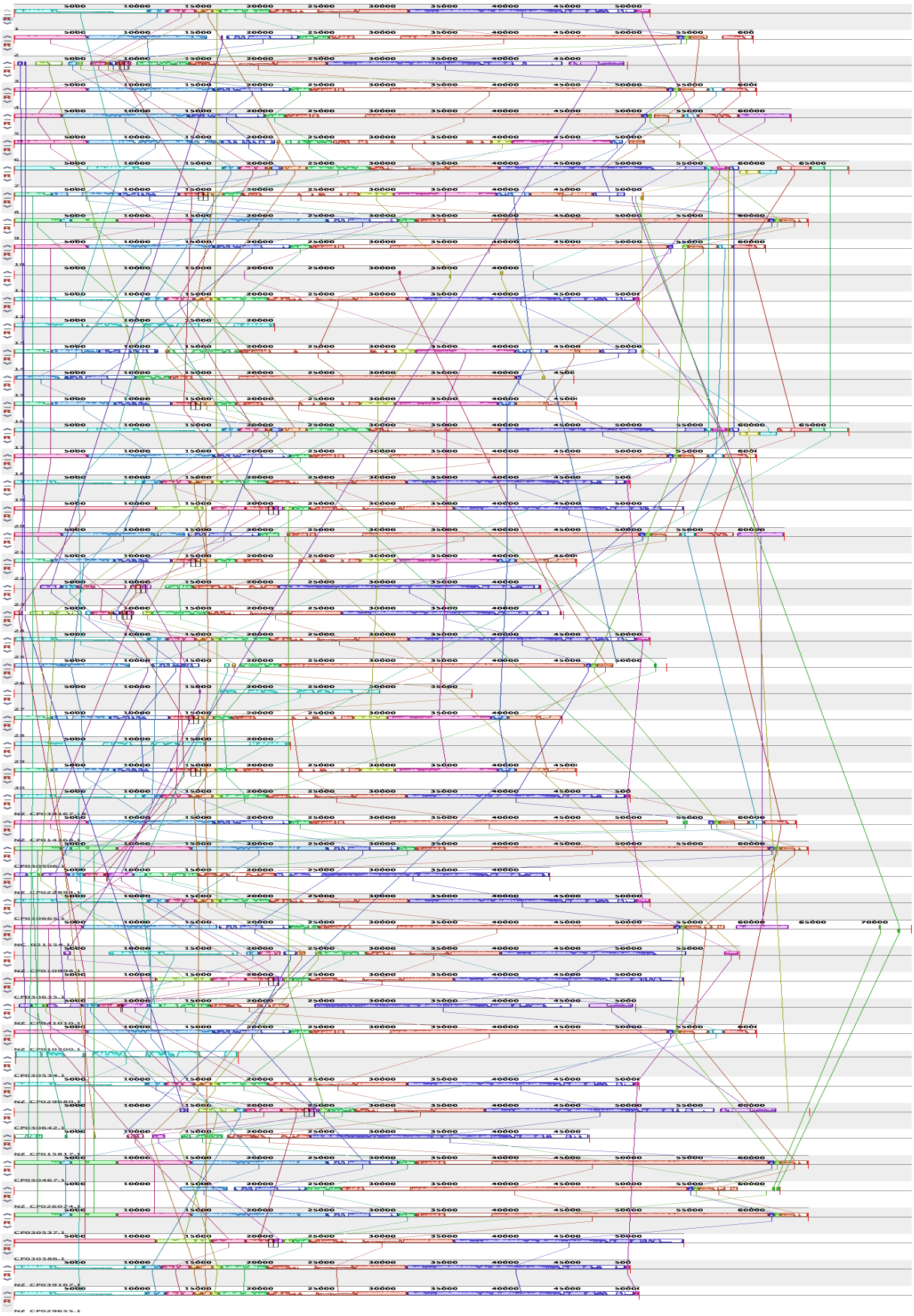

Figure 3.10: Multiple sequence alignment for 50 prophages sequences that infect *Staphylococcus aureus* with 25 of them in normal orientation and 25 of them in reversed orientation.

CHAPTER 4

PREDICTIVE MODELS TO INFER BACTERIUM-PHAGE INTERACTIONS

4.1 Introduction and pairwise learning definition

To recapitulate, this work aims to construct machine learning models that can accurately predict whether or not a specific bacterium-phage pair will interact. Carvalho Leite et al. (2019) stated that the main challenges to train classification models able to predict bacterium-phage interactions at the species and at the strain level is the need for both types of samples, namely bacterium-phage pairs that both interact as well as do not interact. They discuss two approaches to tackle this problem. The first one is the use of one-class classification methods. These are techniques that use only one labelled class to be trained and are used for the detection of outliers in a dataset. The second one is the generation of putative non-interacting data and uses single and ensemble-learning approaches, to predict bacterium-phage interactions at the strain level. Carvalho Leite et al. (2019) constructed feature vectors from the molecular weight of the proteins, the chemical composition and the amino acid frequency to predict the interactions. They collected their positive interactions from public annotated database like NCBI (NCBI Resource, 2018) and PhagesDB (Russell and Hatfull, 2016). The generated putative non-interactions were based on the following hypothesis: most known phages are specific at the strain level (except some rare exceptions that infect and kill a wide range of bacteria, e.g., bacteriophage Mu). Generating the negative interactions was based on two rules: 1) if a phage interacts with a species, no negative interaction between this specific phage and bacteria from this particular species will be generated 2) a phage only attacks one bacterial species. Using this approach they created 20,586 negative pairs for the 2,297 positive interactions at the strain level. Also, to maintain an equilibrium in the data, they generated the same amount of negative interactions per species as there were positive interac-

tions for that species. Their best multiclass models gave an accuracy of 95.7%, while their best one-class classification model resulted in an accuracy of 76.5% at the strain level.

Besides predicting bacterium-phage interactions based on machine learning models, Edwards et al. (2015) reviewed several computational tools and methods for predicting the host of a given phage based on their genomic sequences. Their database contained 820 phages with 153 different bacterial hosts. They examined different computational signals such as abundance profiles, genetic homology, CRISPRs, exact matches and oligonucleotide profiles to identify phage-host relationships. They reviewed a method that was similar to the method in this thesis. Edwards et al. (2015) computed *k*-mer profiles of lengths three to eight and took the host that had the smallest Euclidean distance between the phage nucleotide usage profile and the bacterial profile. They also used 3-mers and computed the Euclidian distance of phage's profile to the host's profiles to identify the appropriate host. They predicted between 8% and 17% of the hosts correctly at the species level with *k*-mer profiles differing between three and eight base pairs. Edwards et al. (2015) found out that the Euclidian distance of the 4-mers provided the strongest signal to identify the correct host. The 3-mer usage predicted approximately 10% of the host correctly at the species level.

Practically, this thesis focuses on the ESKAPE organisms, this in contrast with Carvalho Leite et al. (2019) that did not focus on any particular bacterial species. We considered two different approaches: a Two-step Kernel Ridge Regression, implemented in R and specifically tailored for pairwise learning (Section 4.2.2); and a selection of other widely used machine learning methods (Section 4.2.3). Therefore, a short introduction to pairwise learning is necessary. Afterwards, the methods to predict interactions between bacterial genomes and their prophages will be discussed. Finally, in the last part of this chapter, the results of both approaches will be examined.

In pairwise learning, one wants to predict the properties of pairs of objects. For example, given a ligand and a protein, one wants to predict if they interact with each other or not. The same idea applies to bacteria and prophages: given a bacterium and a prophage, one wants to predict if they interact or not. For each pair of objects there is one label, in this case, interaction (labelled as a binary one) or no interaction (labelled as a binary zero). Therefore, the goal in pairwise learning is to find a function such that given a pair of objects the output of the function can approach the true label as close as possible (Stock, 2017). A pair of objects is called a dyad. We will call the first object of the dyad an instance and the second object a task (Stock et al., 2016).

Pairwise settings occur in multiple domains of science, such as chemistry (predicting binding affinity between two types of molecules), medicine (design of personalized drugs) and ecology (predicting host-parasite interactions) (Stock, 2017).

In pairwise learning, one can distinguish four possible settings for making predictions. The first setting is when information about both objects are known during training but as parts of different dyads. This is the easiest setting in which to predict the label and is shown in Figure 4.1 as setting A. In our case, this is the same as not knowing the interaction between a specific bacterium (x) and prophage (y). But we do know the interactions between x and other prophages (except y). Also, the interactions between y and other bacteria (except x) are known. In Figure 4.1, this entails predicting the grey interactions. The second setting is when only one of both objects is known during training. Here, we can distinguish setting B and C in Figure 4.1 and this is equivalent to predicting the blue interactions or respectively the yellow interactions. In this thesis, this corresponds to predicting interactions between a bacteria and a prophage, of which one was never observed during training. Fourthly, setting D is when none of the objects is seen during training. This is the most difficult case and is equivalent to predicting the red interaction in Figure 4.1. In this case, this is equal to predicting the interaction between a new prophage and a new bacterium. This thesis is not particularly focused on any one setting. It must be said however that in practice setting B is the major focus. More specifically, setting B would represent a hospitalized patient infected with a novel bacterial strain that is not yet present in the database but still belongs to the ESKAPE group.

4.2 Methods

First, the problem of negative interactions is shortly addressed again with some possible solutions. Secondly, the Two-step Kernel Ridge Regression (TSKRR) method is introduced and fitted to the data. Thirdly, different models like Random Forest (RF), *K*-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) were fitted to the data. Hyperparameters were tuned and model evaluation was performed with Precision-Recall curves and F1-score plots. The F1-score is computed as the product of the precision and the recall, divided by the sum of the precision and recall multiplied by two. Formally, $F_1 = 2 * Precision * Recall/(Precision + Recall)$.

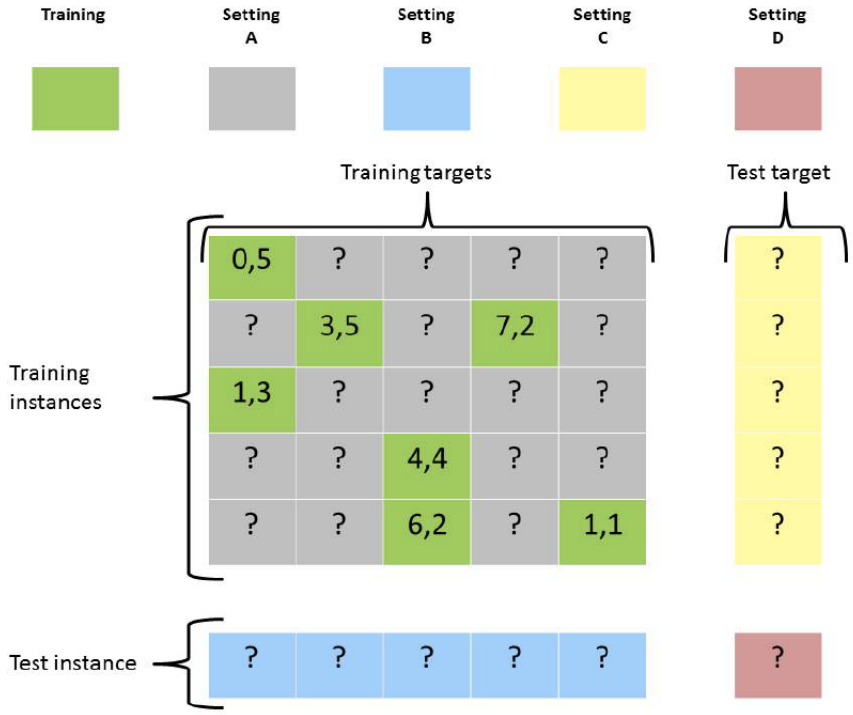

Figure 4.1: The four settings for prediction in pairwise learning. Setting A: both objects are observed during training but as parts of different dyads. Setting B: only one object is observed during training. Setting C: only one object is observed during training. Setting D: none of the objects are observed during training (Waegeman, 2019).

4.2.1 Handling negative interactions

As Carvalho Leite et al. (2019) already explained, the main challenge to train commonly used classification models able to predict bacterium-phage interactions at the species or at the strain level is to deal with both types of samples, namely the interactions and non-interactions between bacterium-phage pairs. The database contains 3,532 positive interactions but no confirmed negative interactions. Thus, in order to train a binary classifier, a strategy to construct negative interactions needed to be developed. Our strategy consisted of four separate approaches to handle negative interactions. The first approach was to impute the interactions within a bacterial species and for prophages unknown to interact. Since interspecies interactions seldom happen, these interactions were put to zero. The intraspecies interactions were put to Not Available (NA) and were imputed using the TSKRR method. Secondly, all the non-observed interactions were put to zero, this means the intraspecies and the interspecies interactions. Thirdly, we argued that setting the intraspecies interaction to zero was too strict. The intraspecies interactions were set to one instead of zero. This causes a shift in the interpretation of the dataset. Instead of predicting at the strain level, the model is now predicting at the species level. Since all the interactions within a species are now one, the model is now trained to recognize which prophage

can infect a specific bacterial species.

Fourthly, for the other widely used machine learning methods, the dataset slightly changed. Methods like Random Forest (RF), *K*-Nearest Neighbors (KNN) and Support Vector Machine (SVM) do not need two kernels (Section 4.2.2) and an interaction matrix. However, these methods also need both positively and negatively labelled interactions. Therefore, we have equally appended our 3,532 known positive interactions with 3,532 negative interactions, making a total of 7,064 interactions. The negative samples were collected but not randomly. Negative sampling was weighted based on the number of occurrences of each bacterial species. For example, the dataset consists of 1,083 positive prophages-bacterium interactions that infect *Staphylococcus aureus* so 1,083 negative prophages-bacterium interactions were selected that did not infect *Staphylococcus aureus*. In addition, to avoid a biased performance due to a particularly fortunate or unfortunate sampling, the sampling of negative interactions was repeated 200 times. In each repetition, RF, KNN and SVM models were fitted and the accuracy was computed.
The accuracy was also computed with shuffled labels. This was done to identity random patterns in our data. Accuracies with the shuffled labels, for binary classification problems, should typically be around 50%, equivalent to random guessing. If there would be a random pattern in the data the accuracy would be significantly higher than 50%.

4.2.2 Two-step Kernel Ridge Regression

Kernel methods can be used to create complex, nonlinear representations of objects by computing a dot product between implicit feature representations (Stock, 2017). A kernel function is a mathematical tool to represent and manipulate objects in high-dimensional feature spaces. The main idea in kernel methods is that in this high-dimensional space, a simple linear model could describe the patterns in the data. The problem is that mapping these features to a high-dimensional space is often computationally intensive. With kernels, one can perform algebraic operations in this high dimensional space without performing the mapping and thus avoiding the computational problem. Kernel methods are very popular in bioinformatics and they can easily be employed for pairwise learning settings. This can be done by defining so-called pairwise kernels, which measure the similarity between two dyads (pair of objects), in this thesis, a bacterium and prophage (Stock et al., 2016). These pairwise kernels were computed as the dot product of the feature matrix and the transpose of the feature matrix. This resulted in a bacteria kernel matrix of 1,293 by 1,293 and a

prophage kernel matrix of 3,532 by 3,532.

Predicting the interaction between a bacterium and prophage was done with TSKRR. This method is implemented in the xnet package in R (Stock et al., 2020). TSKRR uses two kernels (bacteria and prophages) as feature matrices. TSKRR is the combination of two ordinary Kernel Ridge Regressions, one for generalizing to new instances (first object of a dyad) and one that generalizes to new tasks (second object of a dyad), to indirectly predict new dyads (Stock et al., 2016). The hyperparameters of the TSKRR are the different regularization parameters for the kernel matrices. After optimizing the hyperparameters with a 2D grid-search, the optimal TSKRR was fitted to the data and with Leave-one-out cross-validation (LOOCV) were interactions predicted. Additionally, a precision-recall curve was plotted. Since TSKRR is a regression method, the predicted output values are not probabilities or classes but output values laying between 0 and 1. To convert these predicted output values into predicted classes a threshold needs to be set on the predicted output values. In this way, the threshold results in a binarization of the output values, which is the desired output format.

In addition to fitting a multi-species TSKRR model, a second TSKRR model was trained only with data corresponding to the largest species, *Staphylococcus aureus*, to compare performances and explain these in relation to the computed kernels. This species consists of 1,083 active prophages and 478 bacteria. Again, the hyperparameters were optimized and the kernel matrices were computed and used to train a TSKRR model. With LOOCV were the interactions predicted and the precision-recall curve plotted. Finally, a third TSKRR model was fitted to the new interaction matrix in which the intraspecies interactions were put to one instead of zero. Again, the hyperparameters were tuned, LOOCV was used to predict the interactions and a precision-recall curve was plotted together with an F1-score plot.

4.2.3 Other widely used machine learning methods

Besides training TSKRR models, we also trained several widely used machine learning models like RF (Breiman, 2001), KNN (Cunningham and Delany, 2007), SVM (Wang, 2005) and LDA (Rayens, 2012). These other widely used machine learning methods do not need two kernels matrices, but rather one data frame in which the features for the prophages and bacteria are combined. As explained in Section 4.2.1, 7,064 interactions were collected, both positive and negative, and a data frame was made with the features (3-mers) for these specific bacteria and prophages. As a result, the data frame has 7,064 rows, where each row represents a specific interaction between

one bacterium strain and one prophage. The features for a specific bacterium were collected together with the features for a specific prophage. This resulted in a data frame that has 7,064 rows and 131 columns. The first two columns describe the host and the host accession number, followed by the 64 features for the prophages and 64 features for the bacteria and finally a binary label that represents the interaction (one or zero).

4.3 Results and discussion

In this section, the results will be discussed for both the TSKRR and the other widely used machine learning models. First, the kernels were visualized by means of a heatmap. Secondly, the results of the TSKRR models will be described. Finally, the other widely used machine learning models will be evaluated and discussed.

4.3.1 Kernel representations of bacterial and prophage sequences

After computing the 3-mer frequencies for both prophages and bacteria, we had two feature matrices: one for the bacteria and one for the prophages. The dot product was computed of each of the feature matrices and its transpose. This resulted in a kernel matrix for the bacterial genomes and a kernel matrix for the prophage sequences. These kernels were plotted by means of a heatmap, Figures 4.2 and 4.3 visualize both kernels. In Figure 4.2 one can see the similarity within and the differences between bacterial species very clearly. The first 41 bacteria are from the genus *Enterobacter*, from row 41 to 205 are *Pseudomonas*, from row 205 to 354 are *Acinetobacter*, from row 354 to 730 are *Klebsiella*, then the large group from row 730 to 1,208 are *Staphylococcus* and finally from row 1,208 to 1,293 are *Enterococcus*. The groups on the diagonal represent the similarity within the species and have, in most cases, the highest values. The species that have the largest differences between each other have the lowest scores. The difference between *Pseudomonas* and *Staphylococcus* bacterial genomes is the largest as they appear almost dark in the plot. On the contrary, intraspecies differences between the genomes appear to be unobservable in the heatmap. This observation already hints at potential difficulties to predict interactions at the strain level and limited diversity in sequences or an oversimplistic representation of them. One can imagine that seeing no differences within species leads to difficult prediction tasks at the strain level. The heatmap can not distinguish between different prophages that infect the same bacterial species, this is probably an effect of our too simplistic feature engineering. Representing a genome only by

the 3-mer frequencies masks the diversity between the bacterial and prophage sequences. This was already hinted in the t-SNE (Figure 3.8) and PCA (Figure 3.6) were one can see that with only two components the t-SNE and PCA can group the different prophages quite well. In addition, from the MSA plot (Figure 3.9) one can see that a subset of prophages is quite similar to one another, making it even harder to distinguish within prophages that infect the same bacterial species. An alternative is to compute other *k*-mers instead of 3-mers. *K*-mers such as 4-mers, 5-mers and 6-mers will probably have more differences within one bacterial species or prophages that infect the same bacterial species, leading to more intraspecies differences than the 3-mers.

The kernel matrix constructed from the prophage sequences also shows the intraspecies similarity and the interspecies differences although some intraspecies differences can be observed as similar patterns within each block (representing phages related to one particular bacterial species). Again, one can see that the differences between the prophages of *Pseudomonas* and *Staphylococcus* are the largest. This could already be seen in the previous boxplots and PCA plot and is now confirmed again. For prophages related to some of the bacterial species (e.g. *Staphylococcus aureus*), the heatmap suggests that there is some variation between the prophages infecting the same bacterial species. This could be further examined with clustering methods or a phylogenetic tree. Due to limited time and computational resources, these analyses could not be done in this thesis.

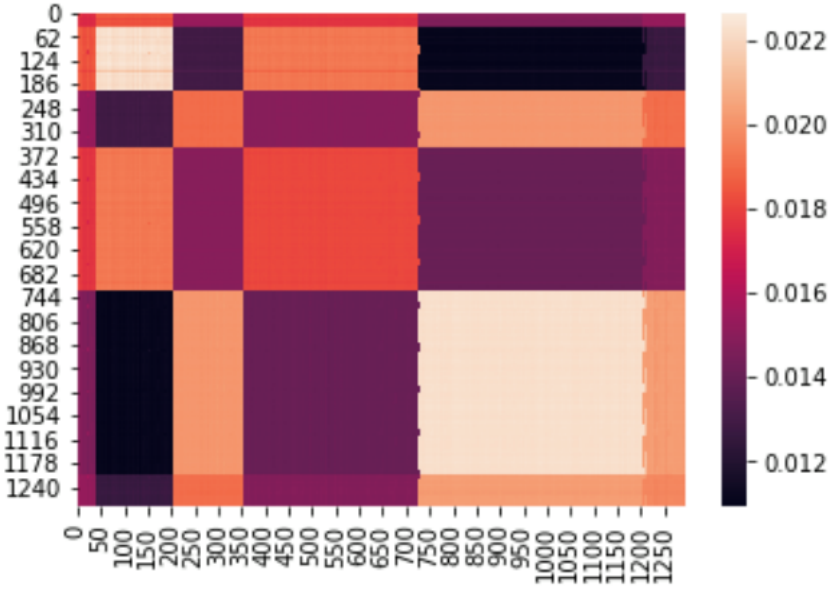

Figure 4.2: A heatmap visualization of the bacteria kernel.

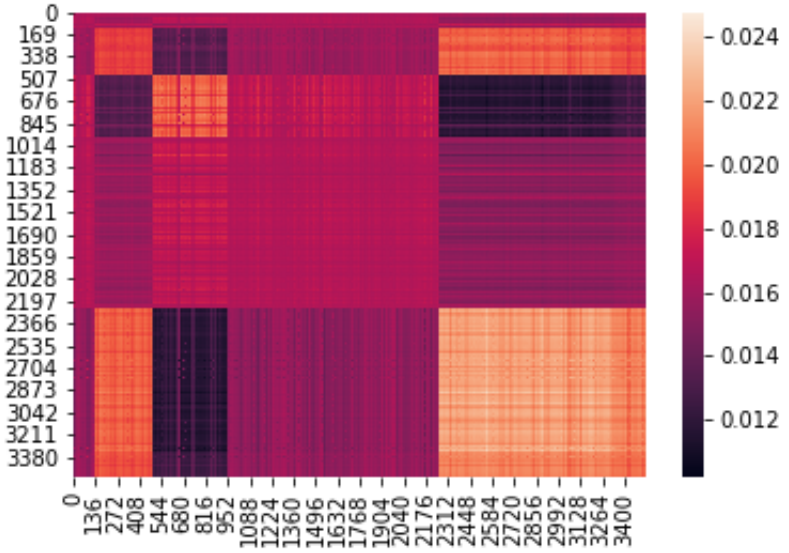

Figure 4.3: A heatmap visualization of the prophage kernel.

4.3.2 Imputation of the negative interactions for the TSKRR

First, the imputation of the intraspecies interactions was done. Again, these were computed because there are no confirmed negative interactions in our dataset. We assume that all unknown interactions are negative interactions but this is a hypothesis. As a result, the interaction matrix only consists of 3,532 positive interactions, in contrast to the 4,563,344 (total number of interactions in the matrix, minus 3,532) negative interactions. Since our interaction matrix is very sparse, all the imputed values were close to zero. This is probably because we only have around 0.077% positive interactions in the total interaction matrix. Nevertheless, a model was fitted (k = 0.01 and g = 0.01) to the imputed interaction matrix and a precision-recall curve was plotted (Figure 4.4). As one can see the area under the curve is approximately zero and thus represents bad predictions at the strain level.

4.3.3 Two-step Kernel Ridge Regression results

As the TSKRR model with imputed missing values did not give good predictions, the intraspecies interactions were put back to zero. A TSKRR model was trained based on the two constructed kernel matrices representing respectively, the bacterial genomes and the prophage sequences. After finetuning the hyperparameters with 2D grid-search, namely the weights of the kernels (k = 0.001 and g = 0.000359), a precision-recall curve was plotted (Figure 4.5) with the predicted interactions from LOOCV. From

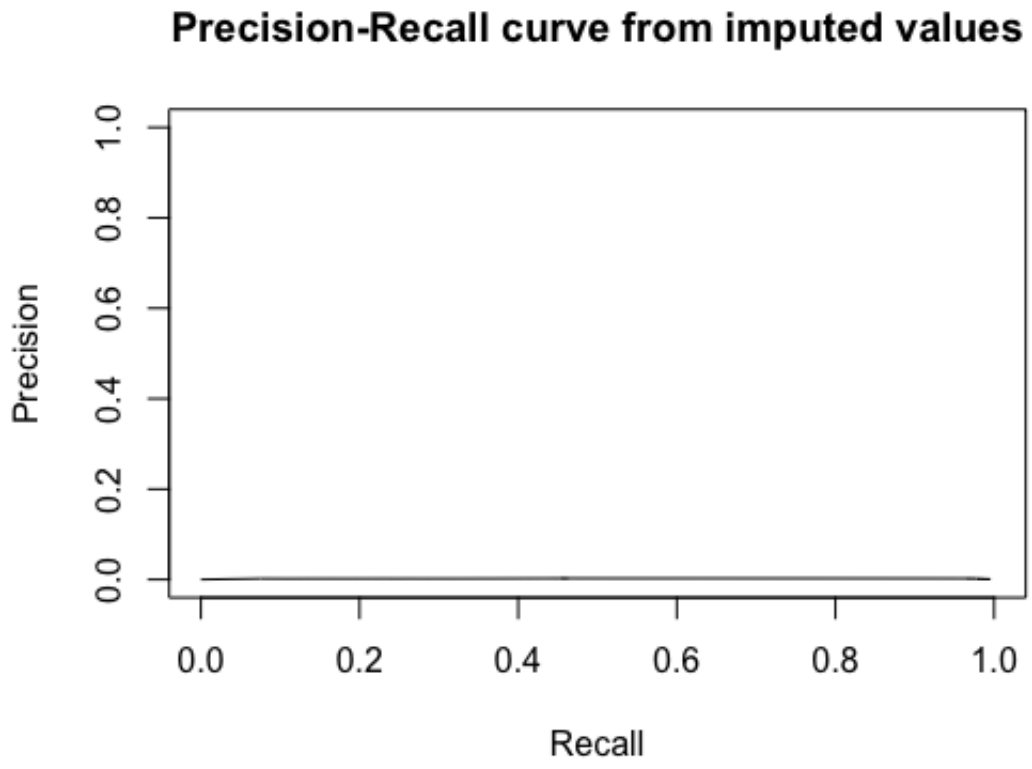

Figure 4.4: Precision-recall curve for the imputed interactions with hyperparameters k = 0.01 and g = 0.01.

Figure 4.5 can be seen that this model was not very successful in predicting the correct label either. More precisely, the model could not predict positive interactions. Since we could not see any significant differences at the intraspecies level in the heatmaps, we fitted a model on the largest species, i.e. *Staphyloccous aureus*. Hyperparameters were also tuned with a 2D grid-search (k = 0.001, g = 0.000359) and a precision-recall plot was plotted (Figure 4.6) with the predicted interactions from LOOCV.

As can be seen in Figures 4.5 and 4.6, these models are not exactly optimal for predicting positive interactions. The precision is in both cases lower than expected. These values are low because of the low value of the numerator (the true positives). The area under the curve is in both models approximately zero. Looking at the MSA (Figure 3.9), it is comprehensible that predictions at the strain level potentially are hard. The prophages in the subsample are really similar to each other. In addition, predictions at the strain level require alternative kernels, for example, kernels that allow for more differentiating between bacterial strains in a species. This will be further discussed in

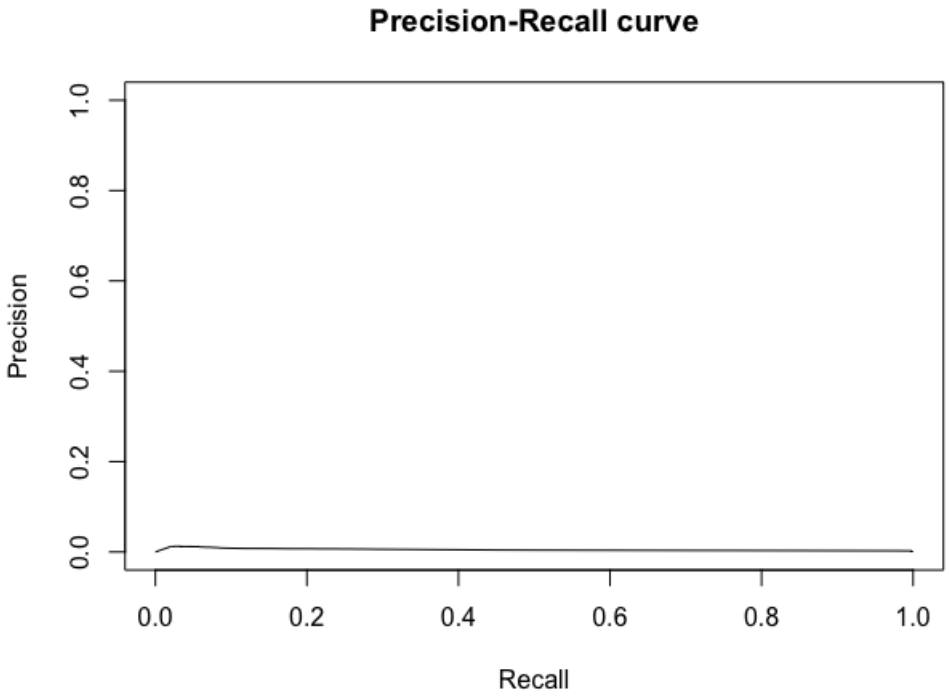

Figure 4.5: Precision-recall curve with hyperparameters k = 0.001 and g = 0.000359 for the first model.

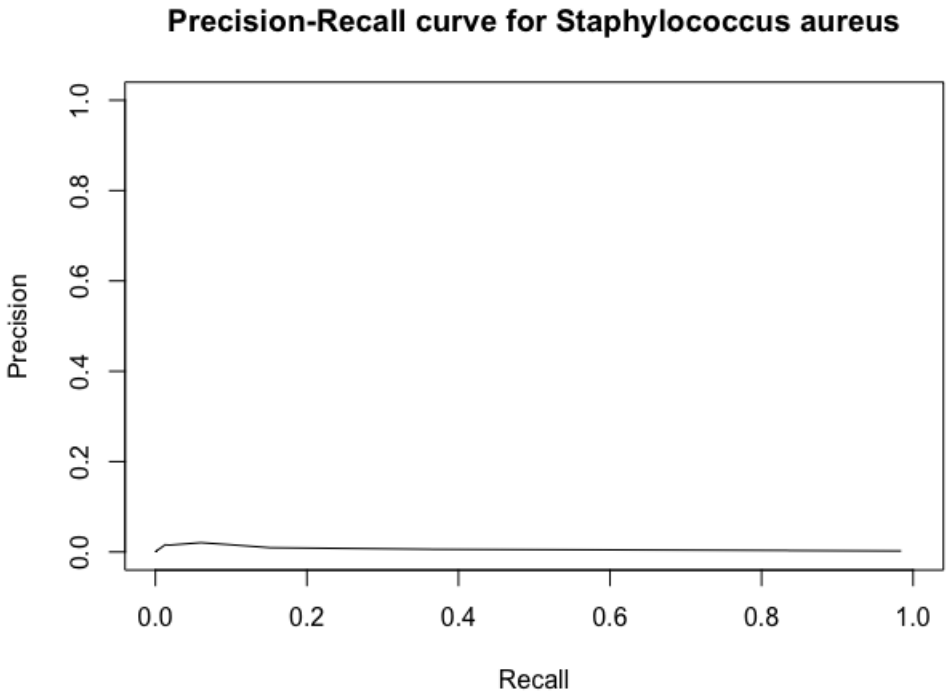

Figure 4.6: Precision-recall curve with hyperparameters k = 0.001 and g = 0.000359 for the *Staphylococcus aureus* model.

the last chapter.

As seen on the heatmaps before, the difference within bacterial strains is not noticeable. The difference in bacterial species, on the other hand, can be seen clearly. Therefore, we have additionally constructed a TSKRR model that predicts interactions at the species level, by setting the intraspecies interactions to one instead of zero. This causes the TSKRR to recognize interaction at the species level instead of the strain level. With this new interaction matrix, the hyperparameters were tuned using a 2D grid-search and a TSKRR model was fitted (k = 0.00069 and g = 0.00001). Again, the interactions were predicted with LOOCV.

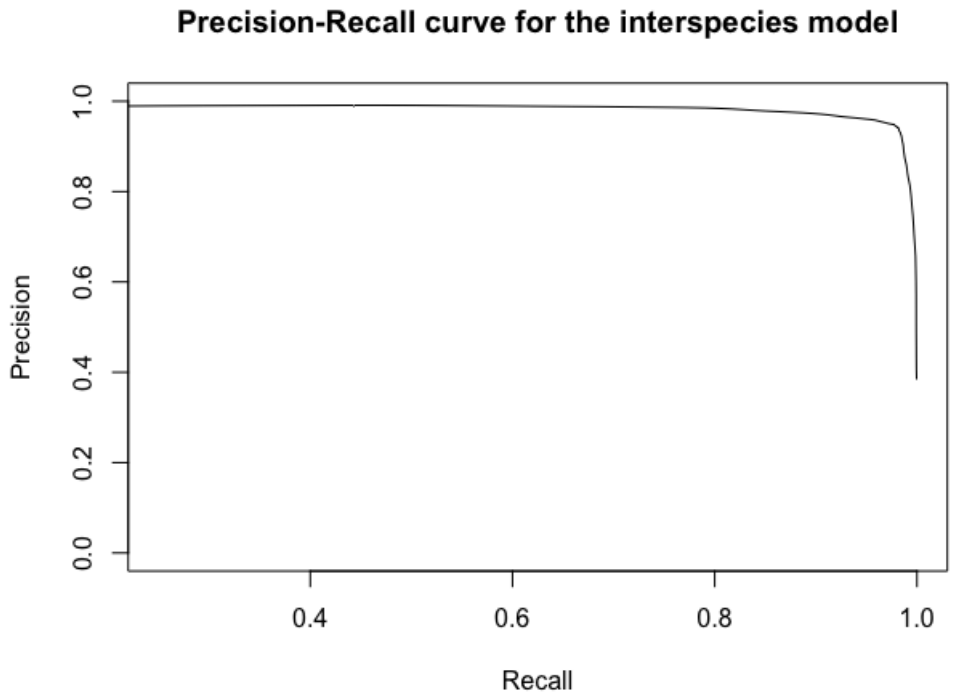

Figure 4.7: Precision-recall curve with hyperparameters k = 0.00069 and g = 0.00001 for the interspecies model.

Setting the intraspecies interactions to one instead of zero made the dataset more balanced. Instead of only 0.07% positive interactions, the dataset now contains 25.29% positive interactions. Furthermore, the heatmaps (Figures 4.2 and 4.3) showed a significant difference between bacterial species in both kernels, which is expected to lead to better predictions in this new setting. The precision-recall curve (Figure 4.7) shows a better prediction for positive interactions and has a bigger area under the curve.

The F1-score plot reaches values close to one near the middle which points to a good

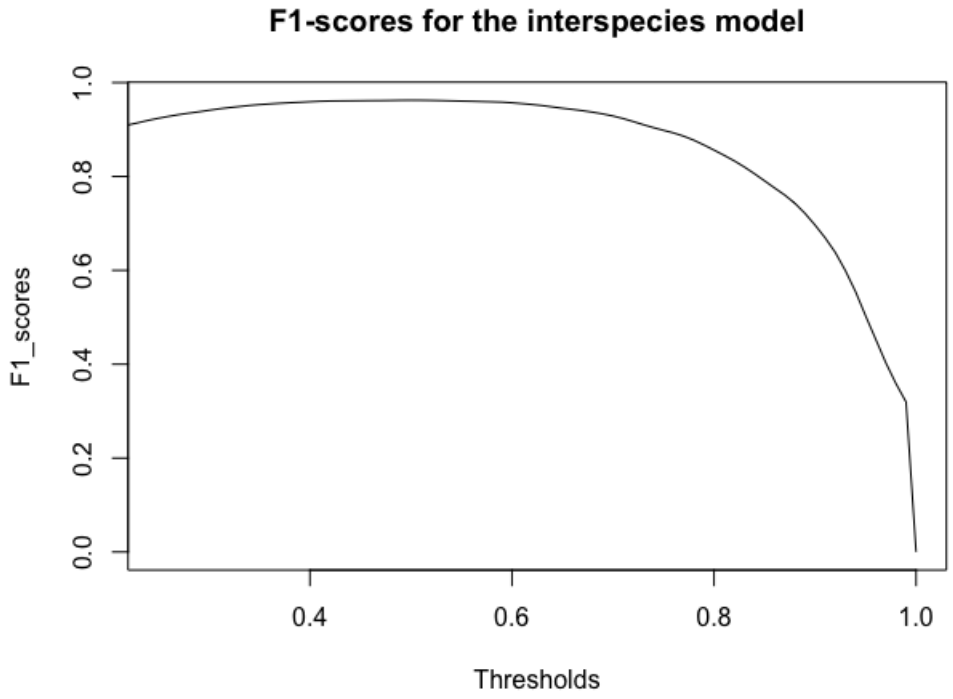

Figure 4.8: F1-score curve with hyperparameters k = 0.00069 and g = 0.00001 for the interspecies model.

performance of the model (Figure 4.8). On the edges, the F1-score is going down which is expected since these correspond to the most extreme thresholds. If the threshold is set too low, there will be many false positives and thus a lower F1-score. In contrast, if the threshold is too high, there will be more false negatives so a lower F1-score. Generally, the performance of this model is decent for most thresholds. At a threshold of 0.5, the accuracy of the model predictions was 98.1%, the error rate was 1.9%, the precision 94.9% and the specificity 98.2%.

4.3.4 Other widely used machine learning models

After the data exploration and fitting of the TSKRR models, a selection of other widely used machine learning models was fitted to the data. First, the dataset was split in a test set and a training set. Secondly, the training set was again split into a training and tuning set. The training and tuning sets were used to optimize the hyperparameters, while the test set was used to validate the model. Table 4.1 summarizes the various performance metrics for the weighted selected negative interactions for the prophage sequences were the reverse complement is added. The LDA model has the lowest accuracy (48.7%), while the Support Vector Classification (SVC) model has the highest accuracy (85.2%).

Table 4.1: The different performance meassurements for the weighted selected negative interactions from the dataset where the reversed complements were added.

Method	Accuracy	Precision	Recall	F1-score
Linear Discriminant Analysis	48.7	48.4%	48.2%	48.3%
Random Forests	84.0%	86.8%	84.1%	84.1%
K-Nearest Neighbors	84.6%	87.7%	84.7%	84.3%
Support Vector Classification	85.2%	88.1%	85.2%	84.9%

The precision-recall curves for the RF (82 estimators), KNN (21 neighbors), SVM (C = 46.41 and gamma = 100) and LDA models are plotted in Figure 4.9. The average precision (AP) is always given in the legend. Based on this plot, one can say that the RF model performs best regarding the precision, with an average precision of 85% but the other models, except the LDA model, are also performing quite well. The KNN model has an average precision of 81% and the Support Vector Classification 82%. The lowest average precision is 47% for the LDA model.

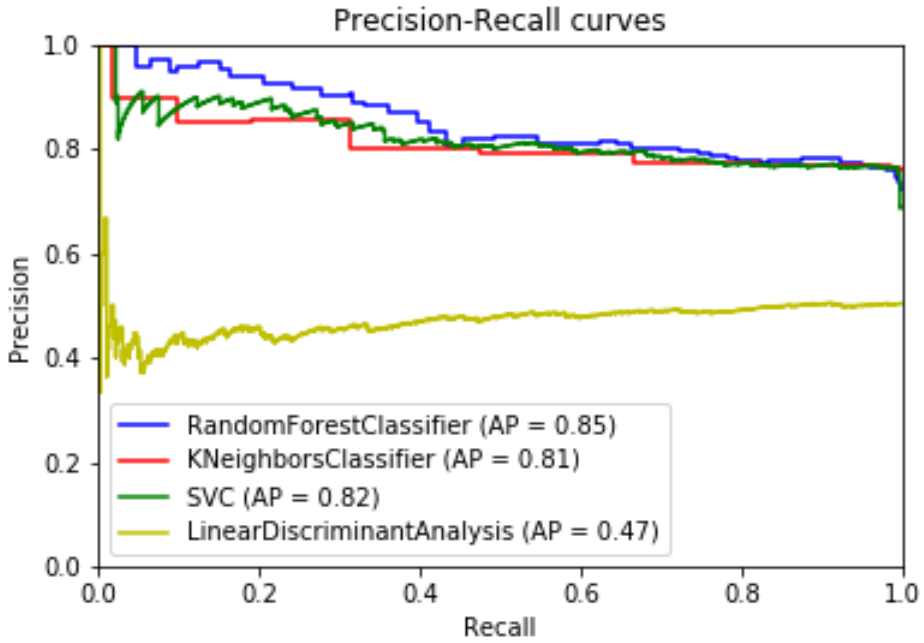

Figure 4.9: Precision-recall curves for different models.

In addition to the precision-recall curves, the F1-score was computed. As mentioned before, the F1-score is the product of the precision and the recall divided by the sum of the precision and the recall multiplied by two. The F1-score plots can be seen in Figure 4.10. Ideally, an F1-score is as close to one as possible. For the LDA model, the F1-score is lower than the other models. Also, it needs a higher threshold to make positive predictions than the other models. Since LDA is a simple linear method this result was expected. For the other three different models (RF, KNN and SVC) the F1-score is between 0.8 and 1.0 for thresholds in the interval of 0.2 and 1.0. One can see that if the threshold is too low, the F1-score is going down. If the threshold is

too high the F1-scores are decreasing as well. If the threshold is too low, more false positives will occur, leading to a lower precision and F1-score. If the threshold is too high, there will be more false negatives. The decrease in F1-score is more noticeable for lower thresholds than for higher thresholds. This points to the fact that our models are better in predicting negative interactions than in positive interactions.

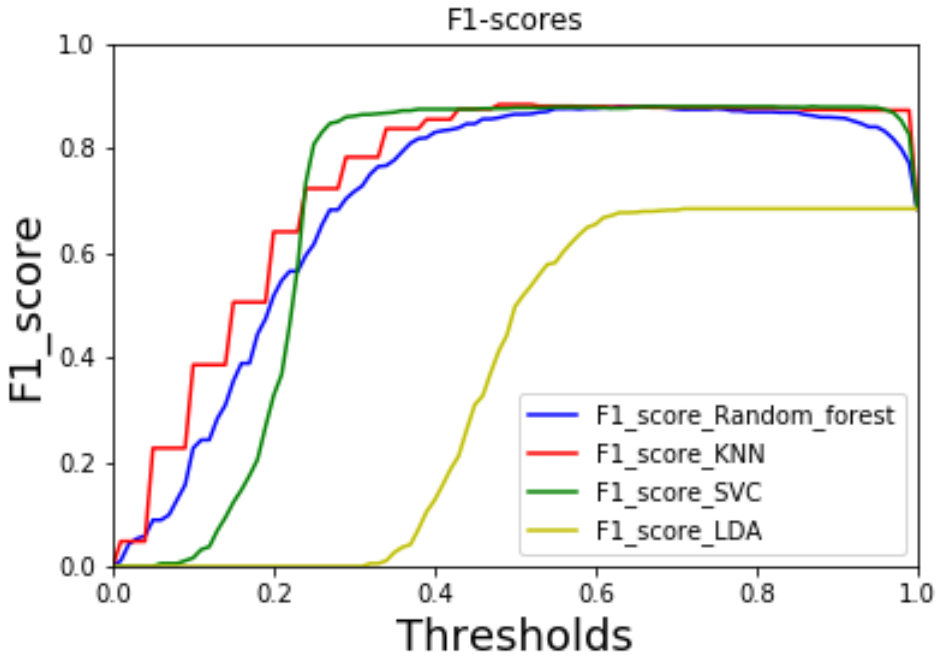

Figure 4.10: F1-scores for Random Forest (82 estimators), *K*-Nearest Neighbors (21 neighbors) and Support Vector Classification (C = 46.41 and gamma = 100).

To evaluate the model performance and avoid being biased for the negative interaction set, we repeated the weighted negative selection procedure 200 times while computing the accuracy of the different models (Figure 4.11). The average accuracy for the RF model is around 84% as already computed in Table 4.1, together with accuracy for the KNN model of 84% and 85% for the SVM model. For the shuffled labels, the models have an accuracy of on average 50%, which is the equivalent of a model that predicts randomly. Therefore, in the case of the shuffled labels, the models do not perform better than randomly guessing which object pair has a positive or negative interaction. If random patterns would occur in the data, accuracies of the models trained after shuffling the labels would be remarkably higher than 50%. Since our models have an accuracy of on average 50% with shuffled labels, this means no random patterns occur in our data.

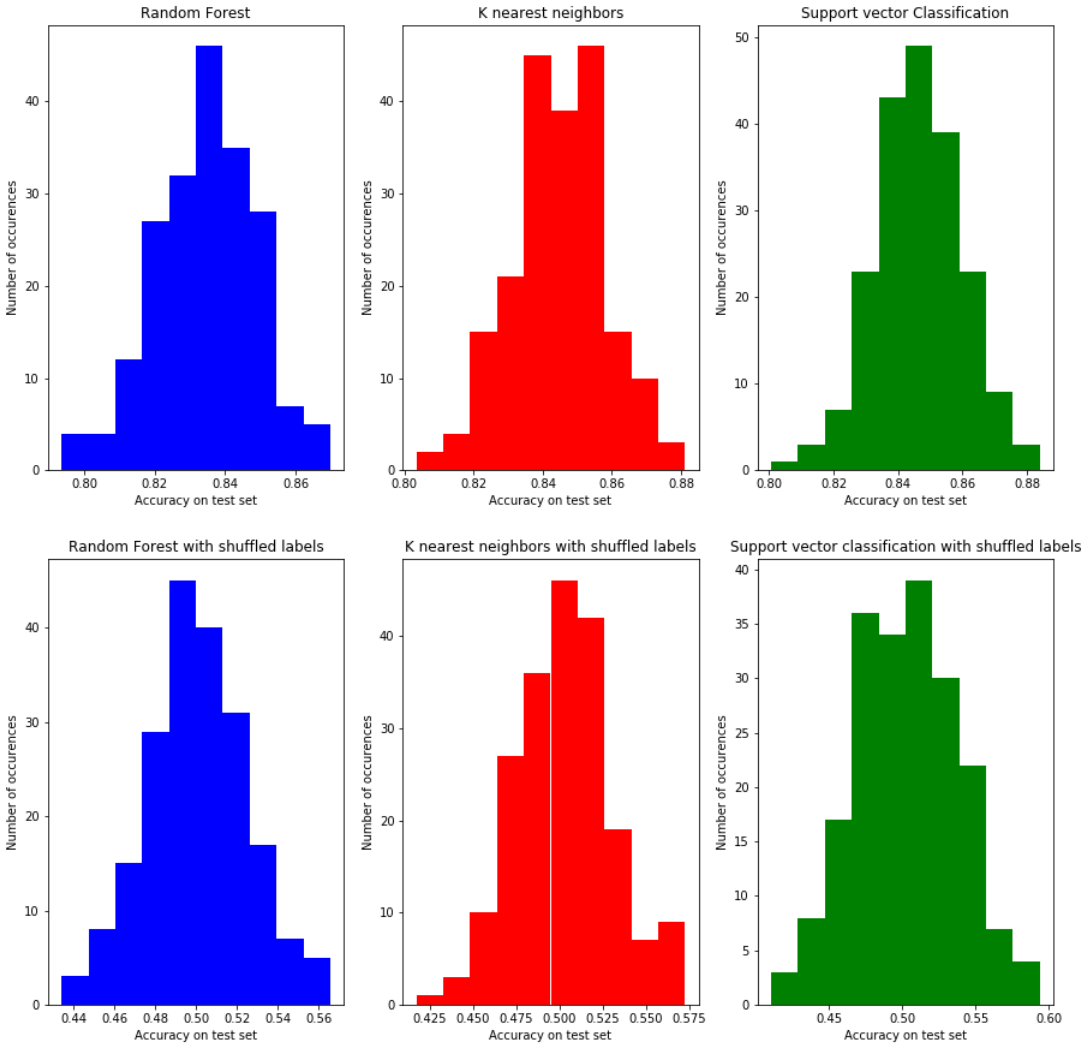

Figure 4.11: The accuracy for a Random Forest, *K*-Nearest Neighbors and Support Vector Classification model. The first row of plots are with the true labels and the second row are with shuffled labels.

CHAPTER 5

CONCLUSION AND FUTURE PERSPECTIVES

5.1 Conclusion

In chapter 1 and 2, we introduced the current problems with antibiotic-resistant bacteria, leading to bacterial infections that are almost impossible to treat. As one of the promising alternatives to antibiotics, phage therapy is beginning to re-emerge into the medical world. One of the remaining bottlenecks with phage therapy is the strain-level specificity of the phages, which makes it a long and costly procedure to find the matching phage(s) against a particular bacterial infection.
A practical hurdle in applying machine learning models to predict bacterium-phage interactions at the strain level is the need for strain-level annotation. To circumvent this lack of annotation, we have constructed a first-in-class database containing interactions between bacteria and their prophages annotating at the strain level. By employing the state-of-the-art tool PHASTER, we managed to collect active prophages that were integrated into the bacterial genomes. By doing so, the interaction between the bacteria and the prophages are by definition confirmed at the strain level. We found a total of 3,532 active prophages in 1,457 bacterial genomes.

Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) were used to visualize the sequences of the prophages. Some interesting results came out of this PCA and t-SNE. Using only two components, one can appropriately distinguish the prophages related to the different bacterial species. Initially, the second principal component had a symmetrical element for all the prophages. This made us think that there were reverse complements of the prophages present in our database. We selected a subset of 50 prophages that infect *Staphylococcus aureus* and computed for 25 of them the reverse complement sequence. After performing a PCA to the reverse complements we saw that the symmetry was gone. Phages do not have specific orientation in the bacterial genome. Prophage genes

can occur on both DNA strands, so taking only one DNA strands leads unfortunately to collecting reverse complements as well. We added to each prophage sequence its reverse complement to handle this problem. This was done to eliminate the variability that was explained by the difference between the prophages and the reverse complements. A possible alternative is to find the genes in the prophage sequences, collect them in the forward orientation and compute the 3-mers on the genes instead of the entire prophage sequence. This method solves the problem regarding reverse complements. An interesting fact is that there seems to be a negative correlation between GC% and the frequencies of both AAA and TTT 3-mers. Since *P. aeruginosa*, *K. pneumoniae* and *Enterobacter* all have high GC% in their sequences and are clearly different from the other prophages in the PCA (Figure 3.6) according to the first principal component. Since both AAA and TTT 3-mers have negative loadings for the first principal component, the prophages that have a positive first principal component have less AAA and TTT 3-mers in their genomes.

To deal with the unknown negative interactions we assumed that all non-observed interactions are negative interactions. This reasoning was based on the hypothesis that most prophages infect bacteria with strain level specificity. One could argue that this approach is not perfect but due to time constraints, other approaches were not tested. We used the normalized 3-mer frequencies as features for the prophages and the bacterial sequences.

In the Multiple Sequence Alignment, one can see that the prophages are similar to each other. This can carefully be extrapolated to other prophages that infect the same bacterial species. Furthermore, in the heatmaps, PCA and t-SNE it became clear that the differences between bacterial species and prophages that infect the same bacterial species are clearly noticeable. In contrast, the differences within bacterial species and the prophages that infect the same bacterial host are almost not visible. The PCA, t-SNE and MSA hinted to possible difficulties to predict interactions at the strain level. In addition, more features that characterize the strain-level differences are needed to predict accurately at the strain level. The 3-mer frequencies did not distinguish enough within intraspecies to predict at the strain level. An alternative would be to compute other k-mer frequencies and do the same analysis. Computing the k-mer frequencies of longer k-mers will result in more specific features for both the prophage and bacterial sequences, leading to more differences within a bacterial species and the prophages that infect the same bacterial species. Another option is to focus on specific regions in the genome instead of computing the k-mers for the

entire genome, this will be discussed in the next section.

In the heart of this thesis, namely chapter four, we introduced pairwise learning and proposed some alternative approaches to predict bacterium-phage interactions at the strain level. We used two different approaches to predict the interactions. Firstly, a TSKRR method was used. We created two pairwise kernels by taking the dot product of the feature matrix and the transpose of the feature matrix. This resulted in a bacteria kernel matrix of 1,293 by 1,293 and a prophage kernel matrix of 3,532 by 3,532. A TSKRR model was used to predict bacterium-phage interactions at the strain level and species level. Therefore, we needed both positive as negative labelled interactions. We have 3,532 positive interactions confirmed by PHASTER and used several methods to handle the negative interactions. A naive approach is to set all the unlabelled interactions to zero. Considering the strain specificity of bacterium-phage interactions, this simple approach does make sense from a biological perspective. First, we imputed with a TSKRR model the negative interactions. Secondly, the intraspecies interactions were set to one instead of zero. Thirdly, for the other widely used machine learning methods, we selected 3,532 weighted negative interactions based on the number of occurrences of each bacterial species. A possible alternative for the negative interactions would be to look at the genetic similarity between the prophages and determine a threshold above or below which a specific bacterium-phage pair has a positive or negative interaction. If a prophage is similar to a prophage that already infects the bacterium, the bacterium-phage pair would have a positive interaction. In contrast, if a prophage is dissimilar to a prophage that infects the bacterium, the bacterium-phage pair would have a negative interaction.

Predictions at the strain level were not so accurately predicted by the TSKRR model. This could already be seen in the heatmaps of the kernels, in which no intraspecies differences were noticeable. Conversely, the predictions at the species level showed excellent results but this was already expected since interspecies differences are clearly visible in the heatmaps. This resulted in our optimal model for predictions at the species level with an accuracy of 98.1%, an error rate of 1.9%, a precision of 94.9% and a specificity of 98.2%. Which is not so surprising since prophages adjust their 3-mer profiles to their host throughout evolution and our predictions at the species level use the 3-mer frequencies as features.
Secondly, widely used methods like Random Forest, K-Nearest Neighbors and Support Vector Machine were applied to predict the interactions. We selected 7,064 bacterium-phage interactions to store in a new database. Overall, the widely used machine learning models showed high accuracies and good precision-recall curves.

Our best-performing SVM model (C = 46.41, gamma = 100 and kernel = Radial basis function) got an accuracy of 85.2%, a precision of 88.1%, a recall of 85.2% and an F1-score of 84.9% for predictions at the strain level. The model with the lowest accuracy was the LDA with an accuracy of 49.5% and a precision of 48.3%, recall of 48.3% and an F1-score of 48.3%. We believe that the big discrepancy between the TSKRR and the other widely used machine learning methods is caused by the extreme sparseness of the interaction matrix used to train the TSKRR model. With only 0.077% of the interactions matrix being positively labelled, the TSKRR has to handle a very sparse interaction matrix, while the other widely used machine learning methods are using a more balanced dataset.

To conclude, this thesis aimed to construct machine learning models that can accurately predict whether or not a specific bacterium-phage pair will interact and therefore, shorten the time that is needed to find the right phage against a bacterial infection. Ideally, these models can shorten the time that is needed to find a specific phage to treat infection by a particular bacterium in, for example, a hospital. Therefore, a first-in-class database, containing 1,293 bacteria and 3,532 prophages, was constructed that links prophages with their bacterial host at the strain level. This database is focused on the ESKAPE organisms because of their clinical significance and their increasing resistance to commonly used antibiotics such as vancomycin and penicillin. Despite bottlenecks like the need for both positive and negative samples, we were able to accurately predict interactions at the species level with a Two-Step Kernel Ridge Regression model. With other widely used machine learning methods like Random Forests, *K*-Nearest Neighbors and Support Vector Classification, we were able to predict whether there is a positive or negative interaction at the strain level with an accuracy between 84-86%. Better feature representation is necessary to predict with a TSKRR model at the strain level and to improve our accuracy. This being said, we only tested the models on our self-constructed database. Never were new bacterial strains or new prophages being introduced from the ESKAPE group. We would expect that prophages from other bacterial hosts i.e., a bacterium that is not from the ESKAPE group, have no positive interactions predicted by our models. In practice, lytic phages are used in phage therapy. Our models are trained on lysogenic phages, nevertheless, lysogenic phages can be converted into the lytic variant and the infection machinery is for both the same. One can test this by introducing some lytic phages into our dataset and evaluated the predictions. Again, due to limited time, this analysis could not be performed.

5.2 Future perspectives

These constructed modelling approaches are far from perfect. One could explore many more alternatives for predicting bacterium-phage interactions. A couple of alternatives and future works are given below.

5.2.1 The use of other kernels

As already seen in the heatmaps (Figures 4.2 and 4.3) our kernels did not distinguish between prophages that infect the same bacterial species or within bacterial species. Since the TSKRR method needs two pairwise kernels, one can find many alternative ways to compute these similarities. Further efforts should be undertaken to efficiently compute kernels for sequences at this scale. Having kernels that can distinguish between prophages that infect the same bacterial species is a must for predicting interactions at the strain level. Also, multiple sequence alignments, for larger subsets, could not be computed in time because of limited computational resources. A possible alternative to mitigate the limitation in computational resources is to work on the most diverse regions in the bacterial DNA and the prophage sequence instead of the entire genome. Since these regions are smaller, the computational time will be significantly shorter. Also, by taking the most diverse regions for phages e.g., the early genes (the proteins that take over the host metabolism and molecular machinery) and the RBPs, there will be an increase in differences between prophages that infect the same bacterial species. This same idea can be applied to the bacteria. If one can compute the 3-mers or other k-mers for the most diverse regions in the bacterial genome e.g., the restriction-modification systems and the CRISPR regions, the differences within a bacterial species will be significantly larger and easier to distinguish.

For example, one can extract the genes from the prophage sequences with Phanotate (McNair et al., 2019). Phanotate can be used to get the correct orientation of the genes i.e., forward or reverse, which would be a more elegant solution compared to adding reverse complements to each of the sequences. Subsequently, multiple sequence alignment can be performed for clusters in our dataset, since a multiple sequence alignment on the entire dataset would be too complex and computationally intensive. Finally, on the new database, one can compute the 3-mers or other k-mers again and use the machine learning methods discussed in this thesis to predict the bacterium-phage interactions at the strain level.

5.2.2 Alternative feature matrices

We computed the 3-mer frequencies for each bacterial or prophage sequence but this is a rather simple representation for an entire genome, presumably masking the diversity in our database. This could be avoided by looking, for example, specific to the early genes or the receptor binding proteins of a phage as explained above. These genes are typically less conserved and thus more diverse than other genes. This diversity originates from the fact that phages and bacteria are constantly in an evolutionary arms-race in which they have to adapt to each other to survive. Also, other *k*-mers could be computed, since the decision of taking 3-mers instead of any other *k*-mer was more of an arbitrary choice. According to Edwards et al. (2015) the 4-mers could be promising to predict the bacterium-phage interaction at the species level. But in fact, many *k*-mers should be computed. There is no simple evidence that points to particular 3-mers or 4-mers being the best choice.

5.2.3 Alternative approaches to handle the negative interactions

As already said, the need for both negatively and positively labelled interactions is needed for most machine learning methods. One could use one-class learning methods such as one-class SVM. These methods try to identify objects of a specific class amongst all objects, by learning from a training set containing only objects of that class (Oliveri, 2017). These methods are typically used to detect outliers in a dataset. One could also incentivise laboratories to evaluate unknown interactions, starting with the unknown sequences that contain the most useful information. In this way, confirmed negative interactions could be added to the constructed database. In another approach, followed by Wang et al. (2006), they computed the distances from every unknown instance to all labelled and unlabelled instances and chose the negative interactions that are maximally separated from the known positive and maximally separated from each other to be evaluated in the lab. This way, more information is gathered by confirming negative interactions. Liu et al. (2015) followed a slightly different approach. Instead of selecting the unknown interactions as negative by comparing them to all other positive pairs, Liu et al. (2015) looked at objects that have known interactions with each member of the pair. Thus computing distances for every member of the pair separately and only to those instances that are linked to the pair. For example, one can look at the genetic similarity between the prophages and determine with a threshold whether a specific bacterium-phage pair consists of a positive or negative interaction. If a prophage is similar to another prophage that infects a specific bacterium, this prophage is more likely to infect this specific bacterium as

well. If a prophage is different from another prophage that already infects a specific bacterium, this prophage is less likely to infect this specific bacterium. This same idea can be applied to specific regions of the genome such as the RBPs or the early genes.

BIBLIOGRAPHY

Abedon, S. T. (1990). Selection for lysis inhibition in bacteriophage. *Journal of Theoretical Biology*, 146(4):501–511.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D. S. (2016). Phaster: a better, faster version of the phast phage search tool. *Nucleic Acids Research*, 44(W1):W16–W21.

Association, W. M. (2013). World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, 310(20):2191–2194.

Atamer, Z., Samtlebe, M., Neve, H., Heller, K., and Hinrichs, J. (2013). Review: elimination of bacteriophages in whey and whey products. *Frontiers in Microbiology*, 4:191.

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M. F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K. L., and Davidson, A. R. (2015). Multiple mechanisms for crispr–cas inhibition by anti-crispr proteins. *Nature*, 526(7571):136–139.

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L., and Davidson, A. R. (2013). Bacteriophage genes that inactivate the crispr/cas bacterial immune system. *Nature*, 493(7432):429–432.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brock, T. D. (1998). *Milestones in Microbiology*. ASM Press.

Brüssow, H. (2019). Hurdles for phage therapy to become a teality—an editorial comment. *Viruses*, 11:557.

Carlton, R. (1999). Phage therapy: past history and future prospects. *Archivum Immunologiae et Therapiae Experimentalis*, 47(5):267–274.

Carvalho Leite, D. M., Lopez, J. F., Brochet, X., Barreto-Sanz, M., Que, Y.-A., Resch, G., and Peña-Reyes, C. A. (2019). Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. *Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine*, page 8 p.

Cenens, W., Makumi, A., Mebrhatu, M. T., Lavigne, R., and Aertsen, A. (2013). Phage–host interactions during pseudolysogeny. *Bacteriophage*, 3(1):e25029.

Clokie, M. R., Millard, A. D., Letarov, A. V., and Heaphy, S. (2011). Phages in nature. *Bacteriophage*, 1(1):31–45.

Costa, A. R., Monteiro, R., and Azeredo, J. (2018). Genomic analysis of acinetobacter baumannii prophages reveals remarkable diversity and suggests profound impact on bacterial virulence and fitness. *Scientific Reports*, 8(1):15346.

Cunningham, P. and Delany, S. (2007). K-nearest neighbour classifiers. *Multi-Classification System*.

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1403.

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

Dedrick, R. M., Guerrero-Bustamante, C. A., Garlena, R. A., Russell, D. A., Ford, K., Harris, K., Gilmour, K. C., Soothill, J., Jacobs-Sera, D., Schooley, R. T., Hatfull, G. F., and Spencer, H. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant mycobacterium abscessus. *Nature Medicine*, 25(5):730–733.

Edwards, R. A., McNair, K., Faust, K., Raes, J., and Dutilh, B. E. (2015). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews*, 40(2):258–272.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Garretto, A., Miller-Ensminger, T., Wolfe, A. J., and Putonti, C. (2019). Bacteriophages of the lower urinary tract. *Nature Reviews Urology*, 16(7):422–432.

Gordillo Altamirano, F. and Barr, J. (2019). Phage therapy in the postantibiotic era. *Clinical Microbiology Reviews*, 32.

Griffiths, A. J., Miller, J., and Suzuki, D. (2000). *An Introduction to Genetic Analysis*. W.H. Freeman, New York.

Hesse, S. and Adhya, S. (2019). Phage therapy in the twenty-first century: facing the decline of the antibiotic era; is it finally time for the age of the phage? *Annual Review of Microbiology*, 73:155–174.

Khan, H. A., Baig, F. K., and Mehboob, R. (2017). Nosocomial infections: Epidemiology, prevention, control and surveillance. *Asian Pacific Journal of Tropical Biomedicine*, 7(5):478–482.

Kutter, E. M. and Sulakvelidze, A. (2005). *Bacteriophages: Biology and Applications*. Boca Raton (Fla.): CRC Press.

Leite, D. M. C., Brochet, X., Resch, G., Que, Y.-A., Neves, A., and Pena-Reyes, C. (2018). Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics*, 19(Suppl 14).

Liu, H., Sun, J., Guan, J., Zheng, J., and Zhou, S. (2015). Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12):i221–i229.

Magill, S. S., Edwards, J. R., Bamberg, W., Beldavs, Z. G., Dumyati, G., Kainer, M. A., Lynfield, R., Maloney, M., McAllister-Hollod, L., Nadle, J., Ray, S. M., Thompson, D. L., Wilson, L. E., Fridkin, S. K., Infections, E. I. P. H.-A., and Team, A. U. P. S. (2014). Multistate point-prevalence survey of health care-associated infections. *The New England Journal of Medicine*, 370(13):1198–1208.

Mansour, N. (2017). Bacteriophages are natural gift, could we pay further attention! *Journal of Food Microbiology*, 1:22.

McNair, K., Zhou, C., Dinsdale, E. A., Souza, B., and Edwards, R. A. (2019). PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics*, 35(22):4537–4542.

Miller, M. B. and Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Review of Microbiology*, 55:165–199.

Monteiro, R., Pires, D. P., Costa, A. R., and Azeredo, J. (2019). Phage therapy: going temperate? *Trends in Microbiology*, 27(4):368–378.

Natarajan, S. and Usha, B. (2018). Eskape pathogens: Trends in antibiotic resistance pattern.

NCBI Resource, C. (2018). Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 46(D1):D8–D13.

Nobrega, F. L., Vlot, M., de Jonge, P. A., Dreesens, L. L., Beaumont, H. J. E., Lavigne, R., Dutilh, B. E., and Brouns, S. J. J. (2018). Targeting mechanisms of tailed bacteriophages. *Nature Reviews Microbiology*, 16(12):760–773.

Oliveri, P. (2017). Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues –a tutorial. *Analytica Chimica Acta*, 982:9 – 19.

O'Neill, J. (2016). Tackling drug-resistant infections globally: final report and recommendations. Technical report.

Owen, S. V., Canals, R., Wenner, N., Hammarlöf, D. L., Kröger, C., and Hinton, J. C. D. (2020). A window into lysogeny: revealing temperate phage biology with transcriptomics. *Microbial Genomics*, 6(2).

Park, J. Y., Moon, B. Y., Park, J. W., Thornton, J. A., Park, Y. H., and Seo, K. S. (2017). Genetic engineering of a temperate phage-based delivery system for crispr/cas9 antimicrobials against staphylococcus aureus. *Scientific Reports*, 7(1):44929.

Paul, L., Suman, A., and Sultan, N. (2013). Methodological analysis of principal component analysis (pca) method. *International Journal of Computational Engineering and Management*, 16:32–38.

Pendleton, J., Gorman, S., and Gilmore, B. (2013). Clinical relevance of the eskape pathogens. *Expert Review of Anti-infective Therapy*, 11:297–308.

Pirnay, J.-P. (2014). Faagtherapie: de medische toepassing van de evolutionaire wapenwedloop tussen bacteriën en fagen. *Belgisch Militair Tijdschrift*, (8):107–120.

Pirnay, J.-P., De Vos, D., Verbeken, G., Merabishvili, M., Chanishvili, N., Vaneechoutte, M., Zizi, M., Laire, G., Lavigne, R., Huys, I., Van den Mooter, G., Buckling, A., Debarbieux, L., Pouillot, F., Azeredo, J., Kutter, E., Dublanchet, A., Górski, A., and Adamia, R. (2010). The phage therapy paradigm: Prêt-à-porter or sur-mesure? *Pharmaceutical Research*, 28:934–7.

Pirnay, J.-P., Verbeken, G., Ceyssens, P.-J., Huys, I., De Vos, D., Ameloot, C., and Fauconnier, A. (2018). The magistral phage. *Viruses*, 10(2).

Prescott, L. M. (1993). *Microbiology*. Wm. C. Brown Publishers.

Rayens, W. (2012). Discriminant analysis and statistical pattern recognition. *Technometrics*, 35:324–326.

Rice, L. B. (2008). Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no eskape. *The Journal of Infectious Diseases*, 197(8):1079–1081.

Ross, A., Ward, S., and Hyman, P. (2016). More is better: Selecting for broad host range bacteriophages. *Frontiers in Microbiology*, 7:1352.

Roux, S., Hallam, S. J., Woyke, T., and Sullivan, M. B. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife*, 4:e08490.

Russell, D. A. and Hatfull, G. F. (2016). Phagesdb: the actinobacteriophage database. *Bioinformatics*, 33(5):784–786.

Samson, J. E., Magadán, A. H., Sabri, M., and Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nature Reviews Microbiology*, 11(10):675–687.

Santajit, S. and Indrawattana, N. (2016). Mechanisms of antimicrobial resistance in eskape pathogens. *BioMed Research International*, 2016:2475067–2475067.

Sarker, S. A., McCallin, S., Barretto, C., Berger, B., Pittet, A.-C., Sultana, S., Krause, L., Huq, S., Bibiloni, R., Bruttin, A., Reuteler, G., and Brussow, H. (2012). Oral t4-like phage cocktail application to healthy adult volunteers from bangladesh. *Virology*, 434(2):222–232.

Schmidt, C. (2019). Phage therapy's latest makeover. *Nature Biotechnology*, 37(6):581–586.

Shen, J., Zhou, J., Xu, Y., and Xiu, Z. (2020). Prophages contribute to genome plasticity of klebsiella pneumoniae and may involve the chromosomal integration of args in cg258. *Genomics*, 112(1):998–1010.

Srividhya, K., Rao, G., Raghavenderan, L., Mehta, P., Prilusky, J., Manicka, S., Sussman, J., and Krishnaswamy, S. (2006). *Database and comparative identification of prophages*, volume 344, pages 863–868.

Stern, A. and Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 33(1):43–51.

Stock, M. (2017). *Exact and Efficient Algorithms for Pairwise Learning*. PhD thesis, Ghent University.

Stock, M., Pahikkala, T., Airola, A., De Baets, B., and Waegeman, W. (2016). Efficient pairwise learning using kernel ridge regression: an exact two-step method.

Stock, M., Pahikkala, T., Airola, A., Waegeman, W., and De Baets, B. (2020). Algebraic shortcuts for leave-one-out cross-validation in supervised network inference. *Briefings in Bioinformatics*, 21(1):262–271.

Sulakvelidze, A., Alavidze, Z., and Morris, J G, J. (2001a). Bacteriophage therapy. *Antimicrobial agents and chemotherapy*, 45(3):649–659.

Sulakvelidze, A., Alavidze, Z., and Morris, J. G. J. (2001b). Bacteriophage therapy. *Antimicrob Agents Chemother*, 45(3):649–659.

Tang, L. (2019). Anti-anti-crispr. *Nature Methods*, 16(11):1080–1080.

Trevors, J. T. (1999). Evolution of gene transfer in bacteria. *World Journal of Microbiology and Biotechnology*, 15(1):1–6.

Twort, F. (1915). An investigation on the nature of ultra-microscopic viruses. *The Lancet*, 186(4814):1241 – 1243.

van der Maaten, L. and Hinton, G. (2008). Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *P T*, 40(4):277–283.

Waegeman, W. (2019). Predictive modelling.

Wang, C., Ding, C., Meraz, R. F., and Holbrook, S. R. (2006). PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, 22(21):2590–2596.

Wang, I.-N., Deaton, J., and Young, R. (2003). Sizing the holin lesion with an endolysin-galactosidase fusion. *Journal of Bacteriology*, 185:779–87.

Wang, L., editor (2005). *Support Vector Machines: Theory and applications*. Springer Berlin Heidelberg.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *Federation of European Microbiological Societies Microbiology Reviews*, 28(2):127–181.

Yosef, I., Manor, M., Kiro, R., and Qimron, U. (2015). Temperate and lytic bacteriophages programmed to sensitize and kill antibiotic-resistant bacteria. *Proceedings of the National Academy of Sciences*, 112(23):7267–7272.

Zeldis, J., Bukhari, A., and Zipser, D. (1973). Orientation of prophage mu. *Virology*, 55(1):289 – 294.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). Phast: A fast phage search tool. *Nucleic Acids Research*, 39(2):W347–W352.

Örmälä, A.-M. and Jalasvuori, M. (2013). Phage therapy. *Bacteriophage*, 3(1):e24219.