Abstract

In this investigation, we examined the interaction of phages and bacteria in bacterial biofilm colonies, the evolution of prophages (viral genetic material inserted into the bacterial genome) and their genetic repertoire. To study the synergistic effects of lytic phages and antibiotics on bacterial biofilm colonies, we have developed a mathematical model of ordinary differential equations (ODEs). We have also presented a mathematical model consisting of a partial differential equation (PDEs), to study evolutionary forces acting on prophages. We fitted the PDE model to three publicly available databases and were able to show that induction is the prominent fate of intact prophages, with an average prophage loss of only 1% of its genome before induction. We also demonstrate that there is a tipping point at which the relation between prophage and bacteria transforms from being parasitic to mutualistic. Lastly, we investigated annotated prophages from two well-studied prophage databases. These genes were accessed using PHASTER, a bioinformatics tool to identify prophages in bacterial genomes. From this analysis, we observed that genes involved in phage lytic function are preferentially lost, while integrase and transposase are preferentially enriched in smaller prophages. We have also developed an ODE model and have carried out gene-level simulations to get more insight into the genetic repertoire of prophages. The results of our ODE model and gene-level simulations are in agreement with prophage repertoire data.

Keywords: Bacteriophage, Bacteria, Biofilm, Phage therapy, Mathematical model, Prophage & domestication, Genome evolution, Individual-based model, Genetic repertoire, Defective prophages.

Summary for lay audience

Bacteriophages are viral predators of bacteria. Upon infecting bacteria, bacteriophages either kill the host bacteria or enter a long-term genetic association with the bacterial host. The ability of bacteriophages to kill bacterial cells is used as a therapeutic strategy to cure bacterial infections, called phage therapy. If the bacteriophage enters a long-term association with the host, the viral genome integrated into the host bacterial genome is called a prophage. In this investigation, using mathematical modeling approaches, we study the synergistic effects of phage therapy and antibiotics on bacterial biofilm colonies, the evolution of prophages and the genetic composition of prophages.

Co-Authorship Statement

I, Amjad Khan, declare that this thesis titled, "Phage-bacteria interaction and prophage sequences in bacterial genomes" has been written by me under the supervision of Dr. Lindi M. Wahl.

Chapter 2: Phage therapy and antibiotics for biofilm eradication: a predictive model, has been published as a book chapter co-authored with L.M. Wahl and P. Yu in Recent Advances in Mathematical and Statistical Methods. AK developed the model with supervisory input from LMW, and AK performed the analysis. PY confirmed the analysis. AK drafted the paper. AK, PY and LMW finalized the paper.

Chapter 3: Quantifying the forces that maintain prophages in bacterial genomes, has been accepted as journal article co-authored with L.M. Wahl in Journal of Theoretical Population Biology. AK and LMW developed the model. AK and LMW contributed equally to the bioinformatics analysis. AK wrote the code, executed and analyzed all optimization results. AK drafted all results. AK and LMW jointly wrote the paper.

Chapter 4: The genetic repertoire of prophages, has been submitted for publication co-authored with L.M. Wahl and A.R. Burmeister in PLOS Computational Biology. AK, ARB and LMW designed the study. AK and LMW developed the mathematical and computational approaches. AK, ARB and LMW analysed the data. AK, ARB and LMW wrote the manuscript.

Acknowledgements

I am thankful to my supervisor Lindi Wahl for continued support and insights. My supervisor's guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D study. I am also indebted to Professor Pei Yu (Department of Applied Mathematics, UWO) and Alita Burmeister (Associate Research Scientist in Ecology and Evolutionary Biology, Yale University) for several insightful comments. Lastly, I would like to thank the professors and staff in the Department of Applied Mathematics, Western University, for their help during my graduate studies.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

1.1 Introduction

Bacteria were one of the first life forms to appear on earth and have a profound and diverse impact on our lives [1]. It has been shown that the ratio of human cells to bacterial cells, in the human body, is close to 1:1 [100], an update of the popular estimate that bacterial cells outnumber human cells in our body by the ratio 10:1 [75, 96]. In 1915, the British microbiologist Fredrick Twort [114] observed the phenomenon of clearing in a solution of bacteria and thought that it was caused by an enzyme responsible for killing bacteria [121]. In 1917, the French physician Felix d'Herelle observed the same clearing phenomenon and speculated that an organism, which he called bacteriophage (phage for short) or "bacteria eater", was responsible for killing the bacteria [37]. After the direct visualization of these bacteria eaters under the electron microscope, it was established that these bacteria eaters are organisms [17], the bacteriophages. Phages outnumber their bacterial hosts by a ratio of approximately 10:1 [11], making them the most abundant microorganisms in the biosphere [93, 32] and have been

critical players in the evolutionary history of bacteria [106].

1.2 Bacteriophages

1.2.1 Phage life cycle

Although some viruses do not exclusively depend on their host for replication and carry with them genes needed for the processes thought to be the distinguishing characteristic of life [92], most viruses require a host for their replication. A phage starts its life cycle by attaching to a specific receptor on the surface of the bacterial host. Once attached, the phage injects its viral genome into the host cell. After injection, these viral genomes can take several possible pathways, of which the most well-known are: (1) the viral genome takes over the host's cellular machinery, the viral genome is replicated and the structural components are produced, ultimately leading to the death of host cell and release of progeny virions into the environment, known as the lytic life cycle; and (2) the viral genome once inside the cell is integrated with the host bacterial genome, known as the lysogenic life cycle.

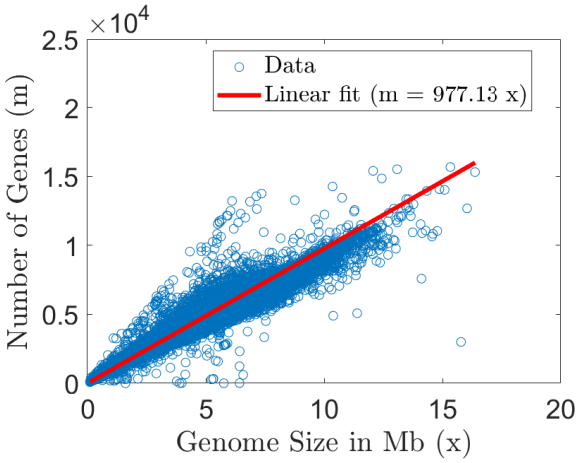
Phages that can follow only the lytic life cycle are called lytic phages whereas phages which follow either the lytic or lysogenic life cycle are called temperate phages. The majority of the identified temperate phages are double-stranded (ds) DNA viruses [109], with the exception of a few single-stranded (ss) DNA temperate phages [63]. Integrated viral DNA of a temperate phage is referred to as a prophage and the bacterial cell carrying prophage or prophages is called a lysogen [77]. These prophage sequences are transmitted vertically with the genome of the bacterial host as the host cell divides into daughter cells [43].

Since the discovery of phages, the ability of lytic phages to eradicate a bacterial population has been used to treat bacterial infectious diseases in humans, animals, and plants [2, 83]. After the advent of antibiotics to treat infectious diseases, phage therapy was confined to particular regions of the world [83]. Recently, due to the emergence of antibiotic-resistant bacteria, attention has been refocused on phage as a weapon against bacterial infectious diseases [2, 23].

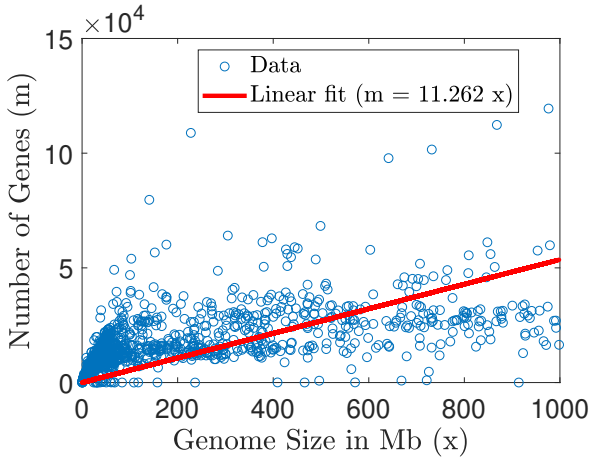
1.2.2 Bacterial genome sequencing

Bacterial genomes are compact as compared to eukaryotic genomes in a sense that bacterial genomes contain less noncoding DNA, see Figure 1.1. Bacterial genomes vary in size by at least an order of magnitude [12] and may be up to 30 Mb in size [20]. Bacterial genomes are dynamic and are exposed to various events dominated by insertions and deletions of genetic elements [99]. To understand the role of bacteria in shaping our environment and to eliminate various bacterial diseases, a complete understanding of the bacterial genome is important.

The year 1995 is marked as the year when the first two human pathogenic bacterial genomes, *Haemophilus influenzae* [44] and *Mycoplasma genetalium* [48], were sequenced. Since then bacterial genome sequencing has undergone substantial development. From 1995 – 2009 only 2010 sequenced bacterial genomes were submitted to NCBI, this number has increased with the advent of modern technologies and 213,581 bacterial genomes have been sequenced and submitted to NCBI (as of November 2019), see Figure 1.2. The process of extracting biological information from sequenced bacterial genome data and providing descriptive information to these features is called bacterial genome annotation.



(a) Number of genes versus bacterial genome size, showing a high correlation (correlation coefficient r =0.984) between genome size and number of genes contained.



(b) Number of genes versus eukaryotic genome size, showing a lower correlation (r =0.624) between eukaryotic genome size and number of genes.

Figure 1.1: Prokaryotic genomes are compact as compared to eukaryotic genomes. These data were downloaded from GenBank in November 2019 and consist of 190,618 bacterial genomes and 2,379 eukaryotic genomes.

1.2.3 Prophage abundance and detection in bacterial genomes

In the event of lysogeny, temperate phages can integrate their genome into different chromosomal sites of a bacterial genome. Phage λ DNA integrates at a unique site in the bacterial genome [110], phage P2 integrates its genome at at least 10 different sites in the host bacterial genome [8], whereas phage Mu can integrate randomly into the host bacterial genome [22]. Bacterial genome sequencing has revealed that prophages are not only frequently identified in pathogenic bacterial strains [26], but are abundant in many bacterial genomes. Prophages may constitute up to 20% of a bacterial genome [30]. The distribution of prophages in a bacterial

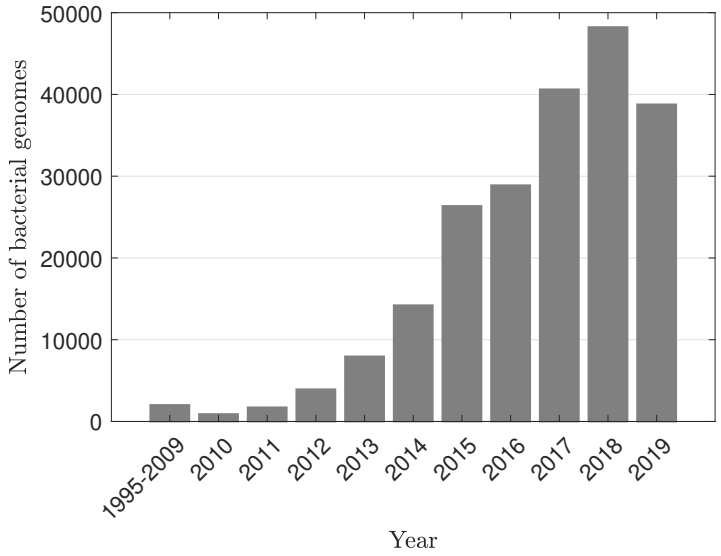


Figure 1.2: The number of bacterial genome sequences submitted to NCBI is growing rapidly. These data were downloaded from NCBI in November 2019.

genome is variable, ranging from no prophage to several prophages per bacterial genome [112]. For example, the sequenced *Escherichia coli* O157:H7 strain Sakai has been shown to contain 18 prophages which make up 16% of its total genome content.

Temperate phages, after integration with the bacterial genome, repress their lysis genes [66] and are subject to mutations that are biased towards deletions [30]. This mutational degradation may eliminate the ability of a prophage to enter into the lytic life cycle by deleting or damaging genes required for lysis and re-infection. Prophages lacking the ability to enter into the lytic life cycle are called defective or cryptic prophages [25]. It has been shown that defective prophages are abundant in bacterial genomes, for example, *Escherichia coli* K-12 contains nine cryptic prophage elements in its genome [120].

A prophage, after insertion into the bacterial genome, becomes a part of the bacterial genome. This relation between prophages and host bacterial cells is usually stable but intact

prophage may initiate the lytic life cycle spontaneously [46, 59], or in response to some environmental cues, or DNA damaging agents [7, 73], resulting in the killing of the host and release of progeny virions into the environment. This process is called the induction of a prophage and is very common in the bacterial world [4]. Some prophages, like λ, excise from the bacterial chromosome to initiate the lytic life cycle, while others, like Mu, produce viral particles before excision from the host bacterial genome [101].

Prophage sequences bring with them many genes, making prophage a prominent source of genetic diversity within bacterial populations [45]. Amongst the changes caused by prophages of particular interest has been the contribution of prophages to bacterial virulence and antibiotic resistance [119, 45, 52].

Prophages can be identified in a bacterial genome using experimental or computational approaches. In the experimental approach, bacteria are usually exposed to UV light or other DNA-damaging conditions to cause the induction of prophages present in the bacterial genome. This technique clearly overlooks the presence of defective prophages as well as other prophages that could not induce. Since a large number of sequenced bacterial genomes are publicly available, computational approaches to identify prophages are preferred.

Since the early 2000s many computational approaches have been developed to find prophages in bacterial genomes. Different computational programs used to identify prophages include Dinucleotide abundance [85, 105], Phage_Finder [47], Prophage Finder [18], Prophinder [71], PHAST (PHAge Search Tool) [123], PHASTER (PHAge Search Tool – Enhanced Release), an improved version of PHAST [6], PhiSpy [3], VirSorter [95], VRprofile [70] and others. Using these computational tools it has been shown that prophages are abundant in bacterial genomes. PhiSpy [3], written in Python and C++, was used to identify 36,488 prophages from the anal-

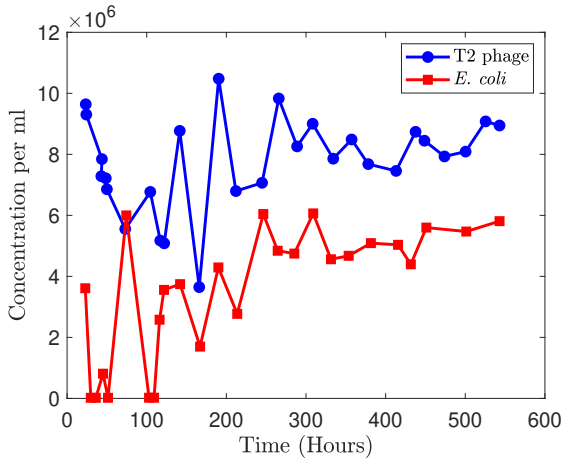
ysis of over 11,000 bacterial genomes; 83% of the bacterial genomes contained at least one prophage [60]. In another study, PHAST [123] was used to identify 4122 prophages in 795 *Acinetobacter baumannii* genomes, for an average of 5 prophages per bacterial genome [34]. Of these prophages 78% were identified as defective. Using PHASTER, Mottawea et al. were able to identify 11,297 prophages in 1760 *Salmonella enterica* genomes [82].

1.2.4 Previous studies of phage-bacteria interaction

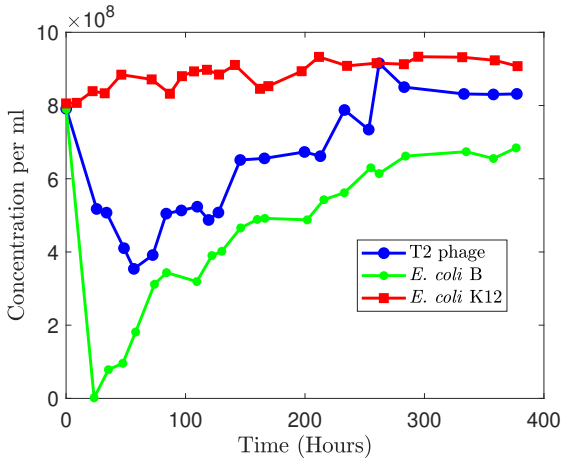
Phages contribute to maintaining bacterial diversity [21], alter competition between bacterial species [16] and mediate the exchange of genetic material among bacteria through horizontal gene transfer [104, 28]. Bacteria are constantly evolving to evade phage infection and phages are acquiring new strategies to infect their bacterial hosts; for details see [94, 61]. To elaborate the population dynamics of phage-bacteria interactions, both experimental and theoretical studies have been undertaken.

The population interaction between phages and bacteria in laboratory setting has been studied by many authors [56, 69, 67, 14]. These studies have considered the interaction between *E. coli* B with T2 phages and mixed species of bacteria with T2 and T3 phages. These investigations reported long-term coexistence of bacterial and viral populations in laboratory cultures. In 1977 Levin et al. investigated the interaction between phages and bacteria by considering a culture of *E. Coli* B (T2 sensitive) and K12 (T2 resistant) with virulent phage T2 [69]. Figure 1.3 illustrates the results of [69] and Figure 1.4 illustrates the results of a related, later study [15].

In the Lotka-Volterra model, introduced by Lotka in 1925 [74] and Volterra in 1926 [117],



(a) Density of glucose-limited populations with a T2-sensitive strain of *E. coli* B and the bacteriophage T2.



(b) Glucose-limited continuous culture populations with a T2-sensitive strain of *E. coli* B, a T2-resistant strain of *E. coli* K12, and the bacteriophage T2.

Figure 1.3: Population dynamics of sensitive, resistant bacteria and virulent phage. These data were extracted from Figure 5A and Figure 8 of Levin et al. (1977) [69], using PlotDigitizer.

the consumption of prey follows the law of mass action, i.e. the consumption of prey by a predator is proportional to the product of the population density of predator and prey. The prey population grows exponentially in the absence of a predator, and the predator population declines exponentially in the absence of the prey. The associated system of differential equations:

$$\begin{aligned} \frac{dx}{dt} &= rx - gxy \\ \frac{dy}{dt} &= \gamma gxy - \mu y \end{aligned} \tag{1.1}$$

gives the classical Lotka-Volterra model. Here, x corresponds to the population density of the prey, y corresponds to the population density of predator, r is the rate of increase of prey

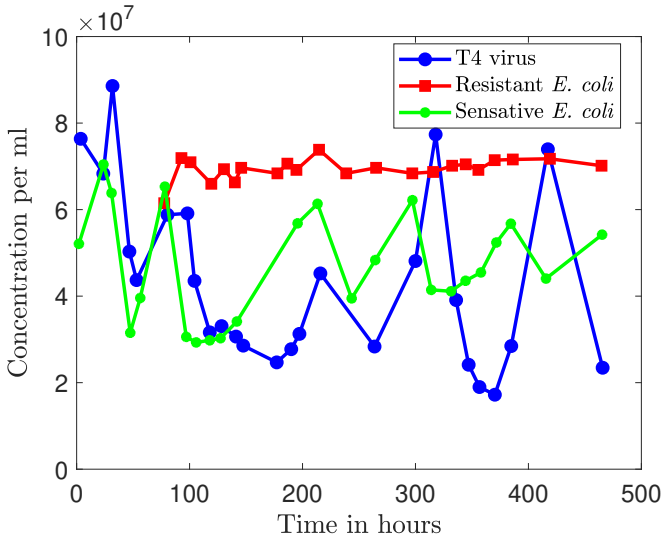


Figure 1.4: Dynamics of T4 sensitive *Escherichia coli* (green), T4 resistant *Eschericia coli* (red), and bacteriophage T4 (blue) in chemostats supplied with media containing glucose, showing oscillating dynamics. Resistant *E. coli* emerges approximately after 78 hours. The population of T4 phage and sensitive bacteria oscillate. The data was extracted from Figure 2B of Bohannan et al. (1999) [15] using PlotDigitizer.

population, g is the predation rate, γ is the reproduction rate of predators and μ is the mortality rate of predators. With appropriate parameter values, this model predicts the coexistence of predator and prey, see Figure 1.5.

This model has been further improved, for example, by introducing a logistic growth term $r(1 - \frac{x}{K})$, where K is the carrying capacity of the prey population, instead of the exponential growth term rx [86]. The linear functional response $\mathcal{F}(x) = g\,x$ in System 1.1, called the Holling Type I response, can also be replaced by more realistic functional responses (Holling Type II, Holling Type III and Holling Type IV) [36].

Based on the nature of the interaction between phages and bacteria and the qualitative

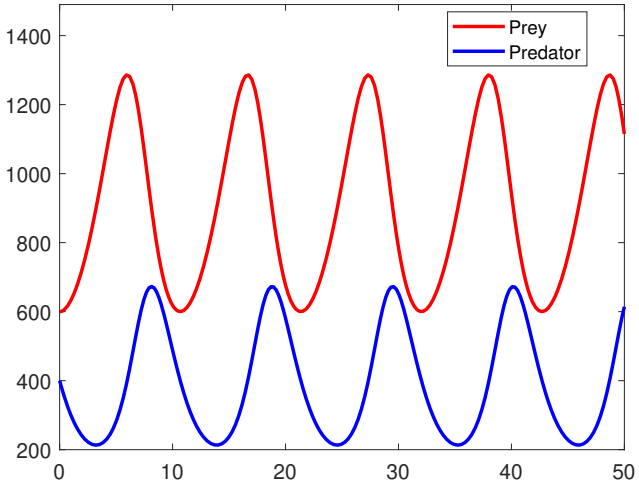


Figure 1.5: Lotka-Volterra model showing long-term existence of both predator and prey populations.

behaviour of such predator-prey models, predator-prey type models were the natural choice for modelling the interaction between phages and bacteria. In 1961 Campbell [24] presented the following model

$$\begin{aligned} \frac{dB}{dt} &= r_B B\left(1 - \frac{B}{K}\right) - aB - kPB \\ \frac{dP}{dt} &= kbB(t-l)P(t-l) - kBP - aP - dP, \end{aligned} \tag{1.2}$$

where B and P are population densities of bacteria and free phages, respectively. The parameters r_B, a, k, d represent the growth rate of bacteria, the flow rate constant, the absorption rate and the rate of spontaneous inactivation of phages, respectively. Each infected bacterial cell yields b phage particles (burst size) at a time l seconds after infection. The author performed a local steady-state analysis of this model, concluding that phages will maintain bacteria at a low but non-zero level if they grow rapidly; otherwise phages will die out. In System 1.2, the functional response is $\mathcal{F} = kB$.

Building on the Campbell model and to account for the relationship between prey growth and the availability of resources, a chemostat model with constant inflow of nutrient solution and outflow of culture was proposed by Levin et al. in 1977 [69]. Multiple resources and multiple species of bacteria and phages were considered. Therefore, phages compete in addition to bacterial competition. The behaviour of the model developed was compared with that of experimental populations of *E. coli* and its virulent virus T2, see Figure 1.3.

Campbell did not discuss the stability analysis at all, whereas Levin et al. investigated the model by integrating the equations numerically but did not carry out stability analysis analytically. Bremermann [19] presented a relatively simpler model by assuming that time delays were negligible on the timescale of consideration. He did not use delay differential equations. If S denotes the density of susceptible hosts, I denotes the density of infected hosts and P the density of phages, then the equations for this model are given as:

$$\begin{aligned} \frac{dS}{dt} &= r_s S\left(1-\frac{S}{K}\right) - \beta S P \\ \frac{dI}{dt} &= \beta S P - \lambda I \\ \frac{dP}{dt} &= bI - \mu P. \end{aligned} \tag{1.3}$$

Here r_s is the bacterial growth rate, K is its carrying capacity, β is the rate of adsorption, λ is the death rate of infected bacterial hosts, b is the rate at which new phages are produced (burst size) and μ is the death rate of phages. By carrying out stability analysis of system 1.3 the author was able to show that the existence of phages depends on the carrying capacity of the bacteria, K. If the carrying capacity falls below a certain threshold $\left(\frac{\lambda\mu}{b\beta}\right)$ the phage population and infected hosts will die out and they uninfected host population will approach its carrying capacity K.

Later, several authors modified the above models. For example Lenski et al. [67] included mutational events into the Levin et al. model [69]. These authors also compared model predictions with the results of experiments with *E. coli* and virulent phage and the evolutionary constraints for *E. coli* and virulent phage. In 1997, Bohannan and Lenski [14] model was a modification of the Levin model. This model ignored the dynamics of infected cells by considering them to instantaneously become dead. The authors solved the model analytically and examined the behaviour of the system numerically. The dynamics of the T4 and *E. coli* populations were also shown, see Figure 1.4. A model, similar to Bremermann's model, was proposed by Beretta and Kaung in 1998 [10] for marine bacteriophages. A rich literature about phage-bacteria interactions is now available, for detail see [87, 51, 103].

1.3 Mobile genetic elements

Mobile Genetic Elements (MGEs) are DNA segments that encode enzymes or other proteins that mediate DNA movements within genomes or between bacterial cells [50]. MGEs are involved in all aspects of genome organization, function, and evolution [35, 5]. Horizontal gene transfer (HGT) is the intercellular movement of DNA that allows transfer of genes from one bacterium (donor) to another bacterium (recipient) by means other than vertical transmission. This transfer of genetic material can take place through: (1) transformation, where bacteria take DNA from their environment; (2) conjugation, where DNA is exchanged between two bacteria; or (3) transduction, which is bacteriophage-mediated exchange of genetic material between two bacteria [104]. The role of HGT in bacterial evolution has long been recognised [57, 104].

1.3.1 Overview of mobile genetic elements

The discovery of MGEs is attributed to McClintock for her work on the maize chromosome, where the existence of jumping genes in maize chromosomes was reported [79]. The genomes of both prokaryotes and eukaryotes carry abundant MGEs [98, 78]. In this investigation, we will focus only on MGEs in prokaryotic genomes.

MGEs can be categorized into different classes; we outline four important classes here. (1) Bacteriophages (lytic/lysogenic/prophages) are one of the most important classes of MGEs which help bacterial cells to exchange genetic material with each other through HGT [27]. (2) Plasmids are extra-chromosomal genetic material and are very common in bacterial genomes [88]. Plasmids are transformed from donor bacterium to recipient bacterium through conjugation, a form of horizontal gene transfer in which bacteria exchange genetic material directly [88]. Plasmids usually carry genes that bring genotypic changes to its bacterial host, for example, antibiotic resistance genes [108]. (3) Transposable elements (TEs) are ubiquitous DNA sequences in bacterial genomes that move within their host bacterial genome [111]. (4) Insertion sequences (ISs), a type of TE [29], are the simplest of MGEs and can move around within genomes or horizontally as a part of other MGEs [116].

1.3.2 Evolutionary forces affecting mobile genetic elements

Mobile genetic elements are either intracellular (inserted from within the cell) or intercellular (inserted from another cell) and are typcically considered to be genetic parasites or junk DNA. The insertion of these new genetic materials comes with some cost to the bacterial host. The cost incurred by these MGEs to the host varies significantly and depends on the nature of the

inserted element. Bacteriophages kill their host to release progeny virions but other mobile genetic elements like plasmids or TEs do not kill their host cell. However, to maintain these elements in its genome the host bacteria must exert some extra energy, making them costly [38]. These MGEs can also disrupt important functions by disrupting a bacterial gene upon insertion [102, 84].

On the other hand, these inserted genetic materials can also endow some benefits to their bacterial host by carrying genes that are beneficial to the host [116]. Frost et al., in [50], called mobile genetic elements “the agents of open source evolution". Mobile genetic elements may also have a positive impact on the bacterial host’s neighbours by producing proteins that are beneficial for neighbours [72] or a negative impact by producing proteins that harm the host’s neighbours [41]. The relationship between mobile genetic elements and their host and host’s neighbours is further explained by Rankin et al. in [91].

Once inserted into the bacterial genome, MGEs are faced with several evolutionary forces:

(1) The rate of insertion of these MGEs into the bacterial genome is an important factor in determining their future distribution in the host population. For example, an important factor in determining the prophage distribution in bacterial genomes is the lysogeny [13], the integration of the phage genome with the bacterial genome after it enters into the bacterial cell.

(2) Mutation occurs randomly and is a change in the nucleotide sequence of a short region of a genome. Mutation is considered to be an important force in shaping bacterial genome evolution [54]. It has been shown that mutation in the bacterial genome is biased toward deletion [64]. Mutations have important and profound affects on these MGEs, for example, mutation can impair genes required for the excision of prophage from a bacterial genome, resulting in domestication of these prophages in bacterial genomes [30, 120].

(3) MGEs, inserted into bacterial genomes, are part of the bacterial host genome and are transmitted vertically from parent to offspring. If these MGEs contain genes that can help its host to proliferate this will, in turn, help these MGEs to proliferate in the host population.

(4) Many MGEs have the ability to invade host genomes horizontally. As described before, this horizontal transfer of genes between bacterial genomes is considered to be a very important factor in the evolution of bacterial genomes and may occur through conjugation, transformation or transduction [104].

(5) MGEs usually have a stable relationship with their host's genome and are transmitted vertically as a part of the bacterial genome to the daughter cells. However some MGEs, like prophages, can excise from the bacterial genome, in response to environmental signals or spontaneously, and kill the host bacterium [73].

1.3.3 Mathematical modelling of mobile genetic elements

Since the discovery of mobile genetic elements there have been many theoretical approaches to explain the nature of these DNA segments. Most of these approaches deal with a particular class of MGEs. The initial mathematical models of MGEs were mostly focused on understanding the mechanisms that prevent the unlimited expansion of MGEs in host populations, despite their tendency for proliferation. These studies also focused on obtaining the equilibrium copy number distribution under diverse evolutionary scenarios. Below we provide some overview of these models aimed at the evolution of MGEs in prokaryotic genomes.

Stewart and Levin, in 1977, developed an ODE model for the population dynamics of horizontally transmitted plasmids in bacterial genomes [107] and argued that plasmids could

not persist if there is a low rate of HGT. They also argued that if plasmids persist, then cells carrying them will maintain high frequencies in bacterial populations, even if the cells carrying them are less fit. In 1980, Levin and Stewart presented an ODE model for the population dynamics of nonconjugative plasmids [68]. Here they concluded that nonconjugative plasmids could be maintained even when the bacteria carrying them have a lower reproduction rate than other bacteria and it is highly unlikely that they will be maintained without conferring to the host some selective benefit. The above mentioned models assume that random encounters occur between members of the plasmid-bearing population and plasmid-free population, at a rate that is proportional to the densities of these populations, that is, the law of mass action. Several variants of these models were presented over the years, for details see [49, 76, 122, 113]. To capture the dynamics of the plasmid in spatially structured habitats other techniques have been applied, for example, computational models [62, 65, 89, 80, 33].

The evolution of TEs in prokaryotic genomes has been studied by many authors [53, 9, 81, 39, 118, 40, 90]. All these authors used the branching process method to study the evolution of TEs in prokaryotic genomes. In a Markov process the outcome of a state is independent of past states and depends only on the present state. A branching process is a special type of Markov process. Using these methods, Sawyer and Hartl, in 1986, developed a model for the distribution of TEs in prokaryotic genomes [97]. The model assumed that TEs are entering prokaryotic genomes at a constant rate and can reduce the fitness of the host in proportion to their numbers in the host genome. A model in which the TEs can convey a selective advantage to the host was also considered. The equilibrium distributions of copy numbers for these models were determined. Relevant parameters were estimated using data regarding the distribution of insertion sequences in natural isolates of *Escherichia coli*. In another study, Hartl

and Sawyer (1988) used a branching process to model the insertion sequences in *E. coli* and concluded that horizontal gene transfer is essential in maintaining bacterial insertion sequences [53].

Basten and Moody formulated a model to analyze the spread of transposable genetic elements in prokaryotic genomes, in 1991 [9]. The authors incorporated selection, transposition and deletion in their model. They concluded that TEs can spread through a population despite selection against them. In [42], the effect of a fluctuating environment on the spread and persistence of TEs was studied.

Van Passel et al. developed a birth–death–diversification model for mobile genetic elements subject to sequence diversification [115]. They applied the model to putative mobile promoters, a type of MGE, and quantified the relative importance of duplication, loss, horizontal gene transfer (HGT), and diversification to the maintenance of the PMP reservoir.

Finally, in 2018, Iranzo and Koonin developed an ODE model and carried out comparative genomic study to explain the roles of selection, horizontal gene transfer, gene duplication and gene loss in the spread and persistence of MGEs [58]. By quantifying the fitness of MGEs to the bacterial hosts they showed that these MGEs are deleterious at evolutionary timescales and characterized them as parasites.

1.4 Motivation and outline of the thesis

Due to the alarming spread of antibiotic resistance and its consequences to public health, the evolution of antibiotic resistance genes has been a topic of interest for many scientists [55]. Experimental results have shown that the synergistic use of antibiotics and phages has a promising

effect against antibiotic resistance bacteria, especially those in biofilms [31]. In **Chapter 2**, we have developed an ODE model to study the effect of antibiotics and phages on the bacterial population in a biofilm. We have exploited the idea of a group defense mechanism by assuming that as the biofilm becomes more and more mature, the harder it becomes for phages to kill bacteria in the biofilm colony. Here we show that the synergistic use of phages and antibiotics, especially using phages first and then antibiotics, can incur maximum damage to the biofilm bacteria. We also show that neither phages nor antibiotics, alone, can eliminate the biofilm completely. Complete elimination of biofilm bacteria is possible only if we could stop the further attachment of planktonic bacteria to the biofilm colony.

The advent of modern technologies has resulted in huge databases about MGEs and has opened new ways to investigate the evolution of MGEs and their role in the evolution of bacteria. Although a substantial literature is available about the evolution of MGEs, these studies are lacking the evolution of an important player responsible for the evolution of bacterial genomes and hence antibiotic resistance genes, the prophages. In **Chapter 3** we attempted fill this void. Here, we have developed a PDE model to mimic the prophage size distribution in bacterial genomes. The basic question we investigate here is *"Why is the prophage size distribution in bacterial genomes bimodal?"* We fitted our PDE model to the three publicly available data sets and were able to quantify various evolutionary forces acting on prophages.

The next question we investigated is *"Which genes are enriched in defective prophages?"*. In **Chapter 4**, we study the genetic repertoire of prophages, especially, which genes are enriched in defective prophages. We downloaded data describing prophages in two well-studied data sets from GenBank and examine their genetic repertoires. Here we developed an ODE model to investigate possible steady states of prophages genes. We also developed a gene-level

model to get more insight into the genetic repertoire of prophages.

In **Chapter 5**, we present conclusions and possible future work.

Bibliography

[1] S. T. Abedon, editor. *Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses*. Cambridge University Press, Cambridge, June 2008.

[2] S. T. Abedon, P. García, P. Mullany, and R. Aminov. Editorial: Phage therapy: past, present and future. *Frontiers in Microbiology*, 8:981, June 2017.

[3] S. Akhter, R. K. Aziz, and R. A. Edwards. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, 40(16):e126–e126, Sept. 2012.

[4] S. Alexeeva, J. A. Guerra Martínez, M. Spus, and E. J. Smid. Spontaneously induced prophages are abundant in a naturally evolved bacterial starter culture and deliver competitive advantage to the host. *BMC Microbiology*, 18(1):120, Dec. 2018.

[5] I. R. Arkhipova and P. A. Rice. Mobile genetic elements: *in silico, in vitro, in vivo*. *Molecular Ecology*, 25(5):1027–1031, Mar. 2016.

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

[7] B. J. Barnhart, S. H. Cox, and J. H. Jett. Prophage induction and inactivation by UV light. *Journal of Virology*, 18(3):950–955, June 1976.

[8] V. Barreiro and E. Haggård-Ljungquist. Attachment sites for bacteriophage P2 on the *Escherichia coli* chromosome: DNA sequences, localization on the physical map, and detection of a P2-like remnant in *E. coli* K-12 derivatives. *Journal of Bacteriology*, 174(12):4086–4093, June 1992.

[9] C. J. Basten and M. E. Moody. A branching-process model for the evolution of transposable elements incorporating selection. *Journal of Mathematical Biology*, 29(8):743–761, Aug. 1991.

[10] E. Beretta and Y. Kuang. Modeling and analysis of a marine bacteriophage infection. *Mathematical Biosciences*, 149(1):57–76, Apr. 1998.

[11] Ø. Bergh, K. Y. Børsheim, G. Bratbak, and M. Heldal. High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–468, Aug. 1989.

[12] L.-M. Bobay and H. Ochman. The evolution of bacterial genome architecture. *Frontiers in Genetics*, 8:72, May 2017.

[13] L.-M. Bobay, M. Touchon, and E. P. C. Rocha. Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132, Aug. 2014.

[14] B. J. M. Bohannan and R. E. Lenski. Effect of resource enrichment on a chemostat community of bacteria and bacteriophage. *Ecology*, 78(8):2303–2315, Dec. 1997.

[15] B. J. M. Bohannan and R. E. Lenski. Effect of prey heterogeneity on the response of a

model food chain to resource enrichment. *The American Naturalist*, 153(1):73–82, Jan. 1999.

[16] B. J. M. Bohannan and R. E. Lenski. The relative importance of competition and predation varies with productivity in a model community. *The American Naturalist*, 156(4):329–340, Oct. 2000.

[17] B. Borries, E. Ruska, and H. Ruska. Bakterien und virus in "Ubermikroskopischer aufnahme: Mit einer einführung in die technik des "Ubermikroskops. *Klinische Wochenschrift*, 17(27):921–925, July 1938.

[18] M. Bose and R. D. Barber. Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biology*, 6(3):223–227, 2006.

[19] H. J. Bremermann. Parasites at the origin of life. *Journal of Mathematical Biology*, 16(2):165–180, Jan. 1983.

[20] T. A. Brown. *Genomes 3*. Garland Science Pub, New York, 3rd ed edition, 2007. OCLC: ocm65195299.

[21] A. Buckling and P. B. Rainey. The role of parasites in sympatric and allopatric host diversification. *Nature*, 420(6915):496–499, Dec. 2002.

[22] A. I. Bukhari and D. Zipser. Random insertion of Mu-1 DNA within a single gene. *Nature New Biology*, 236(69):240–243, Apr. 1972.

[23] J. J. Bull, B. R. Levin, T. DeRouin, N. Walker, and C. A. Bloch. Dynamics of success and

failure in phage and antibiotic therapy in experimental infections. *BMC microbiology*, 2:35, Nov. 2002.

[24] A. Campbell. Conditions for the existence of bacteriophage. *Evolution*, 15(2):153, June 1961.

[25] A. M. Campbell. Prophages and cryptic prophages. In F. J. de Bruijn, J. R. Lupski, and G. M. Weinstock, editors, *Bacterial Genomes*, pages 23–29. Springer US, Boston, MA, 1998.

[26] C. Canchaya, G. Fournous, and H. Brüssow. The impact of prophages on bacterial chromosomes. *Molecular Microbiology*, 53(1):9–18, July 2004.

[27] C. Canchaya, G. Fournous, S. Chibani-Chennoufi, M.-L. Dillmann, and H. Brüssow. Phage as agents of lateral gene transfer. *Current Opinion in Microbiology*, 6(4):417–424, Aug. 2003.

[28] C. Canchaya, C. Proux, G. Fournous, A. Bruttin, and H. Brüssow. Prophage genomics. *Microbiol Mol Biol Rev*, 67(2):238–276, June 2003.

[29] P. Capy, editor. *Dynamics and evolution of transposable elements*. Molecular biology intelligence unit. Springer [u.a.], New York, 1998. OCLC: 845253921.

[30] S. Casjens. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.*, 49(2):277–300, July 2003.

[31] W. N. Chaudhry, J. Concepción-Acevedo, T. Park, S. Andleeb, J. J. Bull, and B. R.

Levin. Synergy and order effects of antibiotics and phages in killing *Pseudomonas aeruginosa* biofilms. *PLoS One*, 12(1):e0168615, Jan. 2017.

[32] M. R. Clokie, A. D. Millard, A. V. Letarov, and S. Heaphy. Phages in nature. *Bacteriophage*, 1(1), 2011.

[33] B. D. Connelly, L. Zaman, P. K. McKinley, and C. Ofria. Modeling the evolutionary dynamics of plasmids in spatial populations. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation - GECCO '11*, page 227, Dublin, Ireland, 2011. ACM Press.

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

[35] N. L. Craig, M. Chandler, M. Gellert, A. Lambowitz, P. A. Rice, and S. Sandmeyer, editors. *Mobile DNA III*. ASM Press, Washington, DC, 2015.

[36] J. Dawes and M. Souza. A derivation of Holling's type I, II and III functional responses in predator–prey systems. *Journal of Theoretical Biology*, 327:11–22, June 2013.

[37] F. d'Herelle. Sur un microbe invisible antagoniste des bacilles dysentériques. *Comptes Rendus de l'Académie des Sciences—Series D*, 165(27):373–375, 1917.

[38] J. C. Diaz Ricci and M. E. Hernández. Plasmid effects on *Escherichia coli* metabolism. *Critical Reviews in Biotechnology*, 20(2):79–108, Jan. 2000.

[39] E. S. Dolgin and B. Charlesworth. The fate of transposable elements in asexual populations. *Genetics*, 174(2):817–827, Oct. 2006.

[40] N. E. Drakos and L. M. Wahl. Extinction probabilities and stationary distributions of mobile genetic elements in prokaryotes: the birth–death-diversification model. *Theoretical Population Biology*, 106:22–31, Dec. 2015.

[41] G. A. Dykes and J. W. Hastings. Selection and fitness in bacteriocin–producing bacteria. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1382):683–687, May 1997.

[42] R. J. Edwards and J. F. Y. Brookfield. Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Molecular Biology and Evolution*, 20(1):30–37, Jan. 2003.

[43] R. Feiner, T. Argov, L. Rabinovich, N. Sigal, I. Borovok, and A. A. Herskovits. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Micro*, 13(10):641–650, Oct. 2015.

[44] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, July 1995.

[45] L.-C. Fortier and O. Sekulovic. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, 4(5):354–365, July 2013.

[46] J. L. Fothergill, E. Mowat, M. J. Walshaw, M. J. Ledson, C. E. James, and C. Winstanley. Effect of antibiotic treatment on bacteriophage production by a cystic fibrosis

epidemic strain of *Pseudomonas aeruginosa*. *Antimicrobial Agents and Chemotherapy*, 55(1):426–428, Jan. 2011.

[47] D. E. Fouts. Phage_finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, 34(20):5839–5851, Nov. 2006.

[48] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, and T. S. Lucier. The minimal gene complement of mycoplasma genitalium. *Science*, 270(5235):397–404, Oct. 1995.

[49] R. Freter, R. R. Freter, and H. Brickner. Experimental and mathematical models of *Escherichia coli* plasmid transfer in vitro and in vivo. *Infection and Immunity*, 39(1):60–84, Jan. 1983.

[50] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, Sept. 2005.

[51] S. A. Gourley and Y. Kuang. A delay reaction-diffusion model of the spread of bacteriophage infection. *SIAM Journal on Applied Mathematics*, 65(2):550–566, Jan. 2004.

[52] J. Haaber, J. J. Leisner, M. T. Cohn, A. Catalan-Moreno, J. B. Nielsen, H. Westh, J. R. Penadés, and H. Ingmer. Bacterial viruses enable their host to acquire antibiotic resis-

tance genes from neighbouring cells. *Nature Communications*, 7:ncomms13333, Nov. 2016.

[53] D. L. Hartl and S. A. Sawyer. Why do unrelated insertion sequences occur together in the genome of *Escherichia coli? Genetics*, 118(3):537–541, Mar. 1988.

[54] R. Hershberg. Mutation—the engine of evolution: studying mutation and its role in the evolution of bacteria. *Cold Spring Harbor Perspectives in Biology*, 7(9):a018077, Sept. 2015.

[55] B. Hong, Y. Ba, L. Niu, F. Lou, Z. Zhang, H. Liu, Y. Pan, and Y. Zhao. A Comprehensive research on antibiotic resistance genes in microbiota of aquatic animals. *Frontiers in Microbiology*, 9:1617, July 2018.

[56] M. T. Horne. Coevolution of *Escherichia coli* and bacteriophages in chemostat culture. *Science*, 168(3934):992–993, May 1970.

[57] J. R. Huddleston. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. *Infection and Drug Resistance*, page 167, June 2014.

[58] J. Iranzo and E. V. Koonin. How genetic parasites persist despite the purge of natural selection. *EPL (Europhysics Letters)*, 122(5):58001, July 2018.

[59] C. E. James, J. L. Fothergill, H. Kalwij, A. J. Hall, J. Cottell, M. A. Brockhurst, and C. Winstanley. Differential infection properties of three inducible prophages from an epidemic strain of *Pseudomonas aeruginosa*. *BMC Microbiology*, 12(1):216, 2012.

[60] H. S. Kang, K. McNair, D. Cuevas, B. Bailey, A. Segall, and R. A. Edwards. Prophage genomics reveals patterns in phage genome organization and replication. *bioRxiv*, page 114819, Mar. 2017.

[61] B. Koskella and M. A. Brockhurst. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiology Reviews*, 38(5):916–931, Sept. 2014.

[62] S. M. Krone, R. Lu, R. Fox, H. Suzuki, and E. M. Top. Modelling the spatial dynamics of plasmid transfer and persistence. *Microbiology (Reading, England)*, 153(Pt 8):2803–2816, Aug. 2007.

[63] M. Krupovic and P. Forterre. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes: Integration of ssDNA viruses into cellular genomes. *Annals of the New York Academy of Sciences*, 1341(1):41–53, Apr. 2015.

[64] C.-H. Kuo and H. Ochman. Deletional bias across the three domains of life. *Genome Biology and Evolution*, 1:145–152, Jan. 2009.

[65] C. Lagido, I. J. Wilson, L. A. Glover, and J. I. Prosser. A model for bacterial conjugal gene transfer on solid surfaces. *FEMS microbiology ecology*, 44(1):67–78, May 2003.

[66] J. G. Lawrence, R. W. Hendrix, and S. Casjens. Where are the pseudogenes in bacterial genomes? *Trends in Microbiology*, 9(11):535–540, Nov. 2001.

[67] R. E. Lenski and B. R. Levin. Constraints on the coevolution of bacteria and virulent phage: A model, some experiments, and predictions for natural communities. *The American Naturalist*, 125(4):585–602, Apr. 1985.

[68] B. R. Levin and F. M. Stewart. The population biology of bacterial plasmids: a priori conditions for the existence of mobilizable nonconjugative factors. *Genetics*, 94(2):425–443, Feb. 1980.

[69] B. R. Levin, F. M. Stewart, and L. Chao. Resource-limited growth, competition, and predation: a model and experimental studies with bacteria and bacteriophage. *The American Naturalist*, 111(977):3–24, Jan. 1977.

[70] J. Li, C. Tai, Z. Deng, W. Zhong, Y. He, and H.-Y. Ou. VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria. *Briefings in Bioinformatics*, page bbw141, Jan. 2017.

[71] G. Lima-Mendez, J. Van Helden, A. Toussaint, and R. Leplae. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, 24(6):863–865, Mar. 2008.

[72] D. M. Livermore. β-Lactamases in laboratory and clinical resistance. *Clinical Microbiology Reviews*, 8(4):557–584, Oct. 1995.

[73] E. López, A. Domenech, M.-J. Ferrándiz, M. J. Frias, C. Ardanuy, M. Ramirez, E. García, J. Liñares, and A. G. de la Campa. Induction of prophages by fluoroquinolones in *Streptococcus pneumoniae*: implications for emergence of resistance in genetically-related clones. *PloS One*, 9(4):e94358, 2014.

[74] A. J. Lotka. Elements of physical biology. *Nature*, 116(2917):461–461, Sept. 1925.

[75] T. D. Luckey. Introduction to the ecology of the intestinal flora. *The American Journal of Clinical Nutrition*, 23(11):1430–1432, Nov. 1970.

[76] P. D. Lundquist and B. R. Levin. Transitory derepression and the maintenance of conjugative plasmids. *Genetics*, 113(3):483–497, July 1986.

[77] A. Lwoff. Lysogeny. *Bacteriological Reviews*, 17(4):269–337, Dec. 1953.

[78] M. Lynch. *The origins of genome architecture*. Sinauer Associates, Sunderland, Mass, 2007. OCLC: ocm77574049.

[79] B. McClintock. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355, June 1950.

[80] B. V. Merkey, L. A. Lardon, J. M. Seoane, J.-U. Kreft, and B. F. Smets. Growth dependence of conjugation explains limited plasmid invasion in biofilms: an individual-based modelling study. *Environmental Microbiology*, 13(9):2435–2452, Sept. 2011.

[81] M. E. Moody. A branching process model for the evolution of transposable elements. *Journal of Mathematical Biology*, 26(3):347–357, June 1988.

[82] W. Mottawea, M.-O. Duceppe, A. A. Dupras, V. Usongo, J. Jeukens, L. Freschi, J.-G. Emond-Rheault, J. Hamel, I. Kukavica-Ibrulj, B. Boyle, A. Gill, E. Burnett, E. Franz, G. Arya, J. T. Weadge, S. Gruenheid, M. Wiedmann, H. Huang, F. Daigle, S. Moineau, S. Bekal, R. C. Levesque, L. D. Goodridge, and D. Ogunremi. *Salmonella enterica* prophage sequence profiles reflect genome diversity and can be used for high discrimination subtyping. *Frontiers in Microbiology*, 9, May 2018.

[83] D. Myelnikov. An alternative cure: The adoption and survival of bacteriophage therapy in the USSR, 1922-1955. *Journal of the History of Medicine and Allied Sciences*, 73(4):385–411, Oct. 2018.

[84] M. Naito and T. E. Pawlowska. The role of mobile genetic elements in evolutionary longevity of heritable endobacteria. *Mobile Genetic Elements*, 6(1):e1136375, Jan. 2016.

[85] P. Nicolas. Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Research*, 30(6):1418–1426, Mar. 2002.

[86] R. M. Nisbet and W. S. C. Gurney. *Modelling fluctuating populations*. Wiley, Chichester ; New York, 1982.

[87] R. J. Payne and V. A. Jansen. Understanding bacteriophage therapy as a density-dependent kinetic process. *Journal of Theoretical Biology*, 208(1):37–48, Jan. 2001.

[88] G. Phillips and B. E. Funnell, editors. *Plasmid biology*. ASM Press, Washington, D.C, 2004.

[89] J. M. Ponciano, L. De Gelder, E. M. Top, and P. Joyce. The population biology of bacterial plasmids: a hidden Markov model approach. *Genetics*, 176(2):957–968, June 2007.

[90] M. Rabbani and L. M. Wahl. The dynamics of mobile promoters: Enhanced stability in promoter regions. *Journal of Theoretical Biology*, 407:401–408, Oct. 2016.

[91] D. J. Rankin, E. P. C. Rocha, and S. P. Brown. What traits are carried on mobile genetic elements, and why? *Heredity*, 106(1):1–10, Jan. 2011.

[92] D. Raoult. The 1.2-megabase genome sequence of mimivirus. *Science*, 306(5700):1344–1350, Nov. 2004.

[93] F. Rohwer. Global phage diversity. *Cell*, 113(2):141, Apr. 2003.

[94] J. T. Rostøl and L. Marraffini. (Ph)ighting phages: How bacteria resist their parasites. *Cell Host & Microbe*, 25(2):184–194, Feb. 2019.

[95] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, 2015.

[96] D. C. Savage. Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology*, 31(1):107–133, Oct. 1977.

[97] S. Sawyer and D. Hartl. Distribution of transposable elements in prokaryotes. *Theoretical Population Biology*, 30(1):1–16, Aug. 1986.

[98] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, ..., and R. K. Wilson. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, Nov. 2009.

[99] I. Sela, Y. I. Wolf, and E. V. Koonin. Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences*, 113(41):11399–11407, Oct. 2016.

[100] R. Sender, S. Fuchs, and R. Milo. Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology*, 14(8):e1002533, Aug. 2016.

[101] J. A. Shapiro. Molecular model for the transposition and replication of bacteriophage *Mu* and other transposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, 76(4):1933–1937, Apr. 1979.

[102] J. A. Simser, M. S. Rahman, S. M. Dreher-Lesnick, and A. F. Azad. A novel and naturally occurring transposon, ISRpe1 in the Rickettsia peacockii genome disrupting the rickA gene involved in actin-based motility: Rickettsia peacockii ISRpe1 Disruption of rickA. *Molecular Microbiology*, 58(1):71–79, Aug. 2005.

[103] H. L. Smith. Models of virulent phage growth with application to phage therapy. *SIAM Journal on Applied Mathematics*, 68(6):1717–1737, Jan. 2008.

[104] S. M. Soucy, J. Huang, and J. P. Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, Aug. 2015.

[105] K. V. Srividhya, V. Alaguraj, G. Poornima, D. Kumar, G. P. Singh, L. Raghavenderan, A. V. S. K. M. Katta, P. Mehta, and S. Krishnaswamy. Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS ONE*, 2(11):e1193, Nov. 2007.

[106] A. Stern and R. Sorek. The phage-host arms-race: shaping the evolution of microbes. *Bioessays*, 33(1):43–51, Jan. 2011.

[107] F. M. Stewart and B. R. Levin. The population biology of bacterial plasmids: a priori

conditions for the existence of conjugationally transmitted factors. *Genetics*, 87(2):209–228, Oct. 1977.

[108] F. Svara and D. J. Rankin. The evolution of plasmid-carried antibiotic resistance. *BMC Evolutionary Biology*, 11(1):130, Dec. 2011.

[109] A. J. Székely and M. Breitbart. Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiology Letters*, 363(6):fnw027, Mar. 2016.

[110] A. Tal, R. Arbel-Goren, N. Costantino, D. L. Court, and J. Stavans. Location of the unique integration site on an *Escherichia coli* chromosome by bacteriophage lambda DNA in vivo. *Proceedings of the National Academy of Sciences*, 111(20):7308–7312, May 2014.

[111] M. Tollis and S. Boissinot. The evolutionary dynamics of transposable elements in eukaryote genomes. In M. Garrido-Ramos, editor, *Genome Dynamics*, volume 7, pages 68–91. S. KARGER AG, Basel, 2012.

[112] M. Touchon, A. Bernheim, and E. P. Rocha. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J*, 10(11):2744–2754, Nov. 2016.

[113] P. E. Turner, V. S. Cooper, and R. E. Lenski. Tradeoff between horizontal and vertical modes of transmission in bacterial plasmids. *Evolution*, 52(2):315, Apr. 1998.

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

[115] M. W. van Passel, H. Nijveen, and L. M. Wahl. Birth, death, and diversification of mobile promoters in prokaryotes. *Genetics*, 197(1):291–299, May 2014.

[116] J. Vandecraen, M. Chandler, A. Aertsen, and R. Van Houdt. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*, 43(6):709–730, Nov. 2017.

[117] V. Volterra. Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science*, 3(1):3–51, Apr. 1928.

[118] A. Wagner. Cooperation is fleeting in the world of transposable elements. *PLoS Computational Biology*, 2(12):e162, 2006.

[119] P. L. Wagner and M. K. Waldor. Bacteriophage control of bacterial virulence. *Infection and Immunity*, 70(8):3985–3993, Aug. 2002.

[120] X. Wang, Y. Kim, Q. Ma, S. H. Hong, K. Pokusaeva, J. M. Sturino, and T. K. Wood. Cryptic prophages help bacteria cope with adverse environments. *Nat Commun*, 1:147, Dec. 2010.

[121] J. Weitz. *Quantitative viral ecology: dynamics of viruses and their microbial hosts*. Princeton University Press, Princeton, NJ, 2017.

[122] X. Zhong, J. E. Krol, E. M. Top, and S. M. Krone. Accounting for mating pair formation in plasmid population dynamics. *Journal of Theoretical Biology*, 262(4):711–719, Feb. 2010.

[123] Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart. PHAST: a fast phage search tool. *Nucleic Acids Research*, 39(suppl):W347–W352, July 2011.

Chapter 2

Phage therapy and antibiotics for biofilm eradication: a predictive model

Bacteria that make up the complex physical structures known as biofilms can be 10-1000 fold more resistant to antibiotics than planktonic (free-living) bacteria. In this chapter we develop a mathematical model to analyze therapeutic techniques that have been proposed to reduce and/or eradicate biofilms, specifically, antibiotics and phage therapy. In this context, the biofilm can be understood as a group defense mechanism, such that the functional response of phages to the biofilm bacterial density is reduced as the biofilm approaches carrying capacity. To capture this mechanism we introduce the function $f(x) = \left(\kappa - \frac{x}{K}\right) x$, where x is biofilm density, K is biofilm carrying capacity and $1 < \kappa < 2$ is the group defense parameter. The model predicts that two therapeutic strategies of recent experimental interest (phage therapy followed by antibiotics, or antibiotics followed by phage therapy) can reduce but not eradicate the biofilm. In contrast, we predict that complete elimination of biofilm bacteria can be achieved by mechanisms that block the attachment of planktonic bacteria to the biofilm.

2.1 Introduction

Bacteria are ubiquitous unicellular organisms, with critical importance in both human health and disease [1]. Bacteria can exist as planktonic (free-living) cells, or in complex communities known as biofilms. In the biofilm state, the bacterial colony is attached to a surface; within the biofilm each cell is sessile and surrounded by extracellular polymeric substances (EPS), substances produced by bacteria in the colony that determine the physical and chemical properties of the biofilm [18]. Biofilms are responsible for a variety of problems in water distribution systems [11], the food industry [27], and medical treatment [25, 8]. Most importantly, biofilms have been implicated as a key factor in two-thirds of human infections [17].

Bacteria are able to rapidly develop resistance against agents employed to eradicate them. In particular, bacteria in a biofilm have been shown to increase resistance to antibiotics by factors of ten to 1000 [9]. Amongst the reasons for enhanced resistance in the biofilm state is the EPS structure surrounding the biofilm colony, which can completely block the infiltration of antibiotics, and the presence of persister cells in the biofilm colony, which are in a metabolically inactive state and thus protected from antibiotic action [9].

The goal of reducing or eradicating biofilm populations has been the focus of research over many years, and there has been much experimental work in this regard [14, 8, 12]. Many agents have been employed for this purpose, which include but are not limited to natural inhibitors of biofilm, for example honey [21], drugs (antibiotics, biofilm-degrading components) [23, 24], bacteriophages and phage-derived enzymes [5, 13, 2] or combinations of some of these [7].

While phage therapy has been proposed as possibly the most effective of these agents, phages alone may not be sufficient to completely eradicate a biofilm[2]. Most recently, experimental work demonstrated that using phage therapy first, followed by antibiotics, maximized the killing of bacteria in an established biofilm.

In this chapter, we develop a mathematical model to study these therapeutic strategies in detail. In section 2.2, we develop the model, tracking biofilm and planktonic bacteria in two linked compartments. In section 2.3, we explore therapeutic strategies including: phage followed by antibiotics; antibiotics followed by phage; and a novel strategy we propose which may have the potential to eradicate the biofilm. In section 2.4, we derive some conclusions from our analysis.

2.2 Mathematical model

We model the interaction between bacteria and bacteriophages (viruses that infect bacteria) using an established predator-prey approach [22]. Our model considers cells of a single bacterial species in either a biofilm or planktonic compartment. The model studies the population dynamics of biofilm cells, B, planktonic cells, P and phage, V_B and V_P, in the biofilm and planktonic compartments respectively. The parameters of the model are described as follows.

The bacterial populations (biofilm or planktonic) are modeled as cell densities per unit volume, cells/cm^3. The biofilm population can increase logistically with a maximum growth rate r, but is limited by a fixed number of available attachment sites in the biofilm matrix,

given by carrying capacity K_B cells/cm^3. Similarly, planktonic bacteria can grow logistically with maximum growth rate r but are limited by carrying capacity K_P. The planktonic bacteria join the biofilm at rate $\mathcal{A}(B,P)$ and biofilm bacteria leave the biofilm with detachment rate $\mathcal{D}(B,P)$. It has been shown that T4 can diffuse fairly through biofilm channels [10]; in the model, phages enter the biofilm compartment at rate p and leave at rate q. In addition, as described above, bacteria in a mature biofilm present substantial resistance to bacteriophages. The expression $f(B)\,V_B$ gives the number of adsorption events per unit time in the biofilm, where $f(B)$, the phage response function, will model this group defense mechanism. The number of adsorption events per unit time in the planktonic compartment is given by $g(P)\,V_P$, where $g(P)$ is the phage response function in the absence of group defense. We neglect the time delay between infection and lysis and assume that each adsorption event instantaneously produces b daughter phages, resulting in new $bf(B)\,V_B$ and $b\,g(P)\,V_P$ bacteriophages in the biofilm and planktonic compartments respectively. Bacteriophage are cleared or denatured at rate c. These assumptions yield the following system:

$$\begin{aligned}
\frac{dB}{dt} &= r\left(1-\frac{B}{K_B}\right)B - f(B)V_B + \mathcal{A}(P,B) - \mathcal{D}(B,P)\\
\frac{dP}{dt} &= r\left(1-\frac{P}{K_P}\right)P - g(P)V_P - \mathcal{A}(P,B) + \mathcal{D}(P,B)\\
\frac{dV_B}{dt} &= bf(B)V_B - cV_B + pV_P - qV_B\\
\frac{dV_P}{dt} &= bg(P)V_P - cV_P + qV_B - pV_P. \qquad (2.1)
\end{aligned}$$

We note that the attachment and detachment rates, $\mathcal{A}(B,P)$ and $\mathcal{D}(B,P)$, satisfy $\mathcal{A}(B,0) = 0$ and $\mathcal{D}(0,P) = 0$. More generally, system 2.1 can also be considered as a two-patch predator-prey model, with group defense acting in one patch only, as illustrated in Figure 2.1.

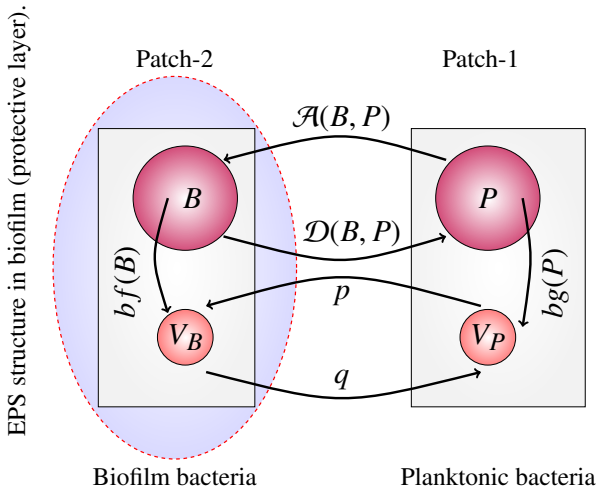


Figure 2.1: Diagram of the model. In patch-1 there is no group defence mechanism but prey (bacteria) can take refuge in patch-2, where they can have group defence mechanism and can protect themselves against predator (phage).

2.3 Therapeutic strategies

Recent experimental work has addressed approaches for minimizing or eradicating bacterial biofilms [7]. In particular, Chaudhry *et al.* compared two therapeutic strategies: applying antibiotics and then phages, or applying the same two agents in the reverse order. Treatment with phages first followed by antibiotics resulted in maximum killing of biofilm bacteria. Here we predict that although these strategies can indeed reduce the biofilm, neither strategy can eradicate the biofilm completely.

2.3.1 Using antibiotics first and then phages

Given that planktonic bacteria are many-fold more sensitive to antibiotics than biofilm bacteria, we assume that an appropriate antibiotic is administered such that planktonic bacteria can be effectively eliminated before phage therapy. We also assume that $\mathcal{A}(B,P) = \mathcal{D}(B,P)$. After the

application of antibiotic we will arrive at the following system

$$\begin{aligned} \frac{dB}{dt} &= r\left(1-\frac{B}{K_B}\right)B - f(B)V_B \\ \frac{dV_B}{dt} &= bf(B)V_B - cV_B. \end{aligned} \tag{2.2}$$

This is a standard predator-prey system with group defense, as studied in [28, 15, 30, 6]. In particular, $f(B)$ must satisfy $f(0) = 0$, $f(B) > 0$ for all $B > 0$, and if there exists a constant $M > 0$, such that $f'(B) > 0$ if $B < M$ and $f'(B) < 0$ if $B > M$, then the system models group defence [15]. The function $f(B) = \frac{m\,B}{\alpha\,B^2+\beta\,B+1}$, called the Holling Type-IV or the Monod-Haldane function, was introduced in [4] and satisfies these properties. System (2.2) has been previously studied with the above functional response for $\beta > -2\sqrt{\alpha}$ [30, 19], and with $f(B) = \alpha e^{-\beta B}$ [29].

In this study, we consider biofilm bacteria that cannot exceed their carrying capacity, such that $B \leq K_B$ at all times. Hence we replace the property $f(B) > 0$ for all $B > 0$ by $f(B) > 0$ for all $0 < B \leq K_B$. To model this phenomenon, we propose a relatively simple functional response $f(B) = \alpha\left(\kappa - \frac{B}{K_B}\right)B$. The rationale for this function is similar to the rationale underpinning logistic growth: we assume that as the biofilm population approaches carrying capacity, the ability of phage to penetrate the biofilm is reduced, linearly. The resulting functional response has the same properties as that of $f(B)$ defined in [30, 29] for $0 < B \leq K_B$. Here α is proportional to the adsorption rate of phages to bacteria, $1 < \kappa < 2$ is the group defense parameter, and K_B is the carrying capacity of the biofilm bacteria. A convenient feature of this model is that the group defense mechanism can be controlled through the parameter κ; $\kappa = 1$ corresponds to a perfect group defense mechanism (no phage adsorption when the biofilm is at carrying capacity) and $\kappa = 2$ corresponds to the absence of effective group defense ($f(B)$

increasing on $0 < B \leq K_B$).

System (2.2) has a maximum of four equilibria. Two boundary equilibria are: $E_0 = (0, 0)$, which represents the complete extinction of biofilm bacteria and phages; and $E_{K_B} = (K_B, 0)$, which represents the extinction of phages while the biofilm bacteria reaches carrying capacity. In addition, two positive equilibria are: $E_{\mu_1} = (\mu_1, \mathcal{F}(\mu_1))$ and $E_{\mu_2} = (\mu_2, \mathcal{F}(\mu_2))$ subject to some conditions of existence. Here $\mathcal{F}(B) = \frac{r\left(1-\frac{B}{K_B}\right)}{\alpha\left(\kappa-\frac{B}{K_B}\right)}$ and μ_1 and μ_2 are solutions to the equation $\hat{f}(B) = \hat{c}$, where $\hat{f}(B) = \left(\kappa - \frac{B}{K_B}\right) B$ and $\hat{c} = \frac{c}{b\,\alpha}$ and $\mu_1 < \frac{\kappa K_B}{2} < \mu_2 < K_B$. The existence of the two positive equilibria E_{μ_1} and E_{μ_2} depend on the positioning of the prey isocline $V_B = \mathcal{F}(B)$ and predator isoclines $B = \mu_1$ and $B = \mu_2$. As we increase $\hat{c}$ in the interval $(0, \hat{c}_M)$, μ_1 and μ_2 become closer to each other; when $\hat{c} = \hat{c}_M$ the two equilibria coincide and we get $E_{\mu_1} = E_{\mu_2} = \left(\frac{\kappa K_B}{2}, \frac{r\left(1-\frac{\kappa}{2}\right)}{\alpha\left(\kappa-\frac{\kappa}{2}\right)}\right)$. Equilibria and their existence can be summarized in the following theorem.

Theorem 2.3.1. *System (2.2) has four equilibria E_0, E_{K_B}, E_{μ_1} and E_{μ_2} if $\hat{c} \in (\hat{c}_m, \hat{c}_M)$, three equilibria E_0, E_{K_B} and E_{μ_1} if $\hat{c} \in (0, \hat{c}_m)$, three equilibria E_0, E_{K_B} and $E_{\mu_1} = E_{\mu_2} = E_\mu = \left(\frac{\kappa K_B}{2}, \frac{r\left(1-\frac{\kappa}{2}\right)}{\alpha\left(\kappa-\frac{\kappa}{2}\right)}\right)$ if $\hat{c} = \hat{c}_M$, only two equilibria E_0 and E_{K_B} if $\hat{c} > \hat{c}_M$, where $\hat{c} = \frac{c}{b\,\alpha}$, $\hat{c}_M = \frac{\kappa^2}{4} K_B$ and $\hat{c}_m = (\kappa - 1)K_B$.*

Stability analysis

It can be easily shown that $E_0 = (0, 0)$ has eigenvalues $\lambda_1 = r > 0$, $\lambda_2 = -c < 0$ showing that $E_0 = (0, 0)$ is a saddle point. The equilibria $E_{K_B} = (K_B, 0)$ has $\lambda_1 = -r$, $\lambda_2 = b\alpha(\hat{c}_m - \hat{c})$, as eigenvalues, showing that E_{K_B} is an attractive node, if $\hat{c} > \hat{c}_m$, and is a saddle point if $\hat{c} < \hat{c}_m$.

To study the stability of the other two equilibria, if they exist, we write the model (2.2) as

$$\begin{aligned} \frac{dB}{dt} &= f(B)\left(\mathcal{F}(B) - V_B\right) \\ \frac{dV_B}{dt} &= bf(B)V_B - cV_B. \end{aligned} \tag{2.3}$$

The eigenvalues for E_{μ_1} are $\lambda_{1,2} = \frac{\xi_1 \pm \sqrt{\xi_1^2 - 4\Delta_1}}{2}$, where $\xi_1 = f(\mu_1)\mathcal{F}'(\mu_1)$ is the trace of Jacobian matrix of (2.3) at E_{μ_1}. Since $\mathcal{F}'(B) < 0$, hence $\xi_1 < 0$ and $\Delta_1 = b\,\alpha^2\,\mu_1\left(\kappa - \frac{\mu_1}{K_B}\right)\left(\kappa - \frac{2\mu_1}{K_B}\right)$ is the determinant of the Jacobian matrix at E_{μ_1}. As $\mu_1 < \frac{\kappa K_B}{2}$, hence $\Delta_1 > 0$. This demonstrates that E_{μ_1} is an attracting point. Similarly, the eigenvalues corresponding to E_{μ_2} are $\lambda_{1,2} = \frac{\xi_2 \pm \sqrt{\xi_2^2 - 4\Delta_2}}{2}$, where $\xi_2 = f(\mu_2)\mathcal{F}'(\mu_2) < 0$ is the trace of Jacobian matrix at E_{μ_2} and $\Delta_2 = b\,\alpha^2\,\mu_2\left(\kappa - \frac{\mu_2}{K_B}\right)\left(\kappa - \frac{2\mu_2}{K_B}\right) < 0$ is the determinant of the Jacobian matrix at E_{μ_2}. We conclude that E_{μ_2} is a saddle point. Since only one equilibrium corresponds to the extinction of biofilm bacteria, and it is a saddle point for all feasible parameter values, we conclude that complete eradication of the biofilm is not possible using this therapeutic strategy. This conclusion is consistent with the view, as discussed in an extensive review [3], that phage action is not sufficient for complete eradication of biofilms.

2.3.2 Using phages first and then antibiotics

In order to understand phage therapy, we return to model (2.1), approximating the complicated processes of attachment and detachment by simpler functions to gain tractability. Specifically, we assume biofilm bacteria detach at constant per capita rate n; this assumption has a long history in the literature, extending back to Freter's influential research on bacterial colonization of the intestinal tract [16, 20]. We further assume that planktonic bacteria attach at constant per capita rate m. In Freter's original biofilm model, attachment is also proportional to the number

of planktonic bacteria, but is further restricted by the number of available "wall attachment sites" [16]. In our model, we restrict biofilm *growth* by the number of attachment sites, K_B, but take a linear attachment rate. Since B and P are densities (cells per unit volume), the net transfer of cells between compartments must be scaled, yielding $\mathcal{A}(B,P) = \left(\frac{vol_P}{vol_B}\right) m\,P$ and $\mathcal{D}(B,P) = \left(\frac{vol_B}{vol_P}\right) n\,B$, where vol_B and vol_P are the volumes of the biofilm and planktonic compartments respectively. After the substitution of these function into system (2.1), it can be shown by direct calculation that the resulting system has three equilibrium solutions: the trivial equilibrium, an equilibrium with both classes of bacteria only, and the all-existing equilibrium (exact expressions omitted for brevity). Out of these equilibria the only equilibrium which corresponds to the complete eradication of biofilm bacteria is E_0. It can be shown by a direct calculation that this equilibrium E_0 is a saddle point for all feasible parameter values. This demonstrates that phage therapy will not eradicate the biofilm. Since the biofilm bacteria are resistant to antibiotics, we can conclude that even phage therapy followed by antibiotics will not remove the biofilm.

2.3.3 A novel therapeutic strategy: blocking attachment

The model developed here allows us to address the following question: is there a therapeutic strategy, *in principle*, that could eradicate the biofilm? Since attachment of planktonic bacteria is critical to biofilm maintenance, we investigated the model assuming this attachment is negligible, and phage therapy is also applied. In this case it can be shown by direct calculations that system (2.1), with the substitutions $\mathcal{A}(B,P) = 0$ and $\mathcal{D}(B,P) = \left(\frac{vol_B}{vol_P}\right) n\,B$, has five equilibrium

solutions:

$$\begin{aligned}
&\mathrm{E}_0: \quad (B,P,V_B,V_P) = (0,0,0,0)\\
&\mathrm{E}_1: \quad (B,P,V_B,V_P) = (0,K_P,0,0)\\
&\mathrm{E}_2: \quad (B,P,V_B,V_P) = (0,\tfrac{c(c+p+q)}{\alpha\, b(c+q)},\tfrac{M\,p\,r}{b\alpha^2(c+q)^2K_P}(\tfrac{vol_P}{vol_B}),\tfrac{M\,r}{b\alpha^2K_P}),\\
&\mathrm{E}_3: \quad (B,P,V_B,V_P) = (B^*,P^*,0,0),\\
&\mathrm{E}_4: \quad (B,P,V_B,V_P) = (B^{**},P^{**},V_B^{**},V_P^{**}),
\end{aligned} \tag{2.4}$$

where

$$M = bq\alpha K_P - c(c+p+q-b\alpha K_P). \tag{2.5}$$

Three of these equilibria, $\mathrm{E}_0, \mathrm{E}_1$ and E_2, represent complete eradication of the biofilm. The equilibria E_0 and E_1 exist for any positive parameter values, while E_2 exists only for $M \geq 0$, i.e. $\alpha \geq \frac{c(c+p+q)}{bK_P(c+q)}$. The equilibrium E_0 is a saddle point for all feasible values of parameters. The equilibrium E_1 is asymptotically stable if $n > r$ and $\alpha < \frac{c(c+p+q)}{b\,K_P(c+q)}$. This implies that if the detachment rate is greater than the birth rate of bacteria in the biofilm and adsorption rate is less than $\frac{c(c+p+q)}{b\,K_P(c+q)}$, then elimination of biofilm bacteria is possible; in particular, planktonic bacteria will reach their carrying capacity and there will be no biofilm or planktonic viruses. Using the Hurwitz criterion, it can be shown that the equilibrium E_2 is stable if $M > 0$, i.e. $\alpha > \frac{c(c+p+q)}{b\,K_P(c+q)}$ (which guarantees its existence), and $n > \max(0, \bar{N}_1)$, where

$$\bar{N}_1 = r - \frac{r\kappa p(\frac{vol_P}{vol_B})\,M}{b\alpha(c+q)^2K_P}.$$

If $\bar{N}_1$ is negative, the conditions for elimination of the biofilm become $n > 0$ and $\alpha > \frac{c(c+p+q)}{b\,K_P(c+q)}$. Thus, the model predicts that biofilm eradication is possible if the attachment of planktonic bacteria to the biofilm, $\mathcal{A}(B,P)$ can be blocked. Although an analysis of realistic numerical parameter values is outside the scope of this contribution, we note that the rate at which the

biofilm could be eliminated depends on the difference between the logistic growth rate, r, and the loss rate of biofilm $(f(B)V_B - \mathcal{D}(B,P))/B$.

2.4 Summary and Conclusions

Biofilm formation starts with the attachment of planktonic (free-living) bacteria to a surface. As these bacteria become sessile and start producing the extracellular matrix (EPS) which defines the biofilm, other bacteria from the planktonic state continue to attach. In this way the bacteria develop a colony that can minimize the infiltration of antibacterial agents. In particular, antibiotics are often ineffective against biofilms, both due to the extracellular structure and the presence of persister cells, which are metabolically inactive. Phages (viruses that infect bacteria) offer the most promising alternative strategy for removing biofilms. Some phages such as T4 can easily infiltrate the EPS structure and can also infect and kill persister cells [18].

A range of experimental studies have shown that phages, antibiotics or other agents alone are not enough to eradicate a biofilm completely, hence a combination of these agents is typically recommended [7, 26]. In this study we derive a mathematical model which predicts that a combination of antibiotics and phage therapy cannot eradicate a biofilm, whether applied as antibiotics followed by phage, or in the reverse order, as studied in [7].

In subsection 2.3.3, we investigate a novel, hypothetical therapeutic strategy. In particular, we demonstrate that if further attachment of planktonic bacteria to the biofilm can be blocked (even if the biofilm is already mature), complete elimination of the biofilm is possible using phages. After eliminating the biofilm, antibiotics can be used to eliminate any remaining

planktonic bacteria. This result suggests that blocking attachment, perhaps by blocking EPS production, is a promising avenue for biofilm eradication. Interestingly, the genetic pathways associated with quorum sensing may in fact be the targets of several natural biofilm inhibitors [21].

Mathematically, the model we derive is a two-patch predator-prey system with group defense by the prey in one patch. Our analysis was made tractable by proposing a simple, novel functional response describing group defense. While this function is invalid (become negative) for biofilm densities that exceed an upper bound, in reality physical constraints limit the density of cells in biofilms, and this limitation did not impede analysis. We expect that this functional form may have further uses in the study of group defense mechanisms, particularly when other aspects of the model become more complex.

Bibliography

[1] S. T. Abedon, editor. *Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses*. Cambridge University Press, Cambridge, June 2008.

[2] S. T. Abedon. Ecology of anti-biofilm agents I: antibiotics versus bacteriophages. *Pharmaceuticals*, 8(3):525–558, Sept. 2015.

[3] S. T. Abedon. Ecology of anti-biofilm agents II: bacteriophage exploitation and biocontrol of biofilm bacteria. *Pharmaceuticals (Basel)*, 8(3):559–589, Sept. 2015.

[4] J. F. Andrews. A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrates. *Biotechnol. Bioeng.*, 10(6):707–723, Nov. 1968.

[5] J. Azeredo and I. W. Sutherland. The use of phages for the removal of infectious biofilms. *Current Pharmaceutical Biotechnology*, 9(4):261–266, Aug. 2008.

[6] H. W. Broer, V. Naudot, R. Roussarie, and K. Saleh. A Predator-prey model with non-monotonic response function. In *Regular & chaotic dynamics*, volume 11, pages 155–165, 2006.

[7] W. N. Chaudhry, J. Concepción-Acevedo, T. Park, S. Andleeb, J. J. Bull, and B. R. Levin.

Synergy and order effects of antibiotics and phages in killing *Pseudomonas aeruginosa* biofilms. *PLoS One*, 12(1):e0168615, Jan. 2017.

[8] O. Ciofu, E. Rojo-Molinero, M. D. Macià, and A. Oliver. Antibiotic treatment of biofilm infections. *APMIS*, 125(4):304–319, Apr. 2017.

[9] D. Davies. Understanding biofilm resistance to antibacterial agents. *Nature Reviews Drug Discovery*, 2(2):114–122, Feb. 2003.

[10] M. M. Doolittle, J. J. Cooney, and D. E. Caldwell. Tracing the interaction of bacteriophage with bacterial biofilms using fluorescent and chromogenic probes. *Journal of Industrial Microbiology*, 16(6):331–341, June 1996.

[11] I. Douterelo, S. Husband, V. Loza, and J. Boxall. Dynamics of biofilm regrowth in drinking water distribution systems. *Applied and Environmental Microbiology*, 82(14):4155–4168, July 2016.

[12] G. Feng, Y. Cheng, S.-Y. Wang, D. A. Borca-Tasciuc, R. W. Worobo, and C. I. Moraru. Bacterial attachment and biofilm formation on surfaces are reduced by small-diameter nanoscale pores: how small is small enough? *NPJ Biofilms and Microbiomes*, 1:201522, Dec. 2015.

[13] L. Fernández, S. González, A. B. Campelo, B. Martínez, A. Rodríguez, and P. García. Low-level predation by lytic phage phiIPLA-RODI promotes biofilm formation and triggers the stringent response in *Staphylococcus aureus*. *Scientific Reports*, 7:srep40965, Jan. 2017.

[14] D. Fleming and K. P. Rumbaugh. Approaches to dispersing medical biofilms. *Microorganisms*, 5(2), Apr. 2017.

[15] H. I. Freedman and G. S. K. Wolkowicz. Predator-prey systems with group defence: The paradox of enrichment revisited. *Bulletin of Mathematical Biology*, 48(5-6):493–508, Sept. 1986.

[16] R. Freter, H. Brickner, J. Fekete, M. M. Vickerman, and K. E. Carey. Survival and implantation of *Escherichia coli* in the intestinal tract. *Infect. Immun.*, 39(2):686–703, Feb. 1983.

[17] C. A. Fux, P. Stoodley, L. Hall-Stoodley, and J. W. Costerton. Bacterial biofilms: a diagnostic and therapeutic challenge. *Expert Review of Anti-Infective Therapy*, 1(4):667–683, Dec. 2003.

[18] D. R. Harper, H. M. R. T. Parracho, J. Walker, R. Sharp, G. Hughes, M. Werthén, S. Lehman, and S. Morales. Bacteriophages and biofilms. *Antibiotics (Basel)*, 3(3):270–284, June 2014.

[19] J. Jiang and P. Yu. Multistable phenomena involving equilibria and periodic motions in predator–prey systems. *International Journal of Bifurcation and Chaos*, 27(03):1750043, Mar. 2017.

[20] D. Jones, H. V. Kojouharov, D. Le, and H. Smith. The Freter model: a simple model of biofilm formation. *J Math Biol*, 47(2):137–152, Aug. 2003.

[21] J.-H. Lee, J.-H. Park, J.-A. Kim, G. P. Neupane, M. H. Cho, C.-S. Lee, and J. Lee. Low

concentrations of honey reduce biofilm formation, quorum sensing, and virulence in *Escherichia coli* O157:H7. *Biofouling*, 27(10):1095–1104, Nov. 2011.

[22] R. E. Lenski. Dynamics of interactions between bacteria and virulent bacteriophage. In *Advances in Microbial Ecology*, Advances in microbial ecology, pages 1–44. Springer, Boston, MA, 1988. DOI: 10.1007/978-1-4684-5409-3_1.

[23] A. S. Lynch and D. Abbanat. New antibiotic agents and approaches to treat biofilm-associated infections. *Expert Opinion on Therapeutic Patents*, 20(10):1373–1387, Oct. 2010.

[24] H. Mu, J. Tang, Q. Liu, C. Sun, T. Wang, and J. Duan. Potent antibacterial nanoparticles against biofilm and intracellular bacteria. *Scientific Reports*, 6:srep18877, Jan. 2016.

[25] A. Omar, J. B. Wright, G. Schultz, R. Burrell, and P. Nadworny. Microbial biofilms and chronic wounds. *Microorganisms*, 5(1), Mar. 2017.

[26] E. M. Ryan, M. Y. Alkawareek, R. F. Donnelly, and B. F. Gilmore. Synergistic phage-antibiotic combinations for the control of *Escherichia coli* biofilms in vitro. *FEMS Immunology & Medical Microbiology*, 65(2):395–398, July 2012.

[27] R. Van Houdt and C. Michiels. Biofilm formation and the food industry, a focus on the bacterial outer surface. *Journal of Applied Microbiology*, 109(4):1117–1131, Oct. 2010.

[28] G. Wolkowicz. Bifurcation analysis of a predator-prey system involving group defence. *SIAM J. Appl. Math.*, 48(3):592–606, June 1988.

[29] D. Xiao and S. Ruan. Global analysis in a predator-prey system with nonmonotonic functional response. *SIAM J. Appl. Math.*, 61(4):1445–1472, Jan. 2001.

[30] H. Zhu, S. A. Campbell, and G. S. K. Wolkowicz. Bifurcation analysis of a predator-prey system with nonmonotonic functional response. *SIAM Journal on Applied Mathematics*, 63(2):636–682, 2002.

Chapter 3

Quantifying the Forces that Maintain Prophages in Bacterial Genomes

Genome sequencing has revealed that prophages, viral sequences integrated in a bacterial chromosome, are abundant, accounting for as much as 20% of the bacterial genome. These sequences can confer fitness benefits to the bacterial host, but may also instigate cell death through induction. Several recent investigations have revealed that the distribution of prophage lengths is bimodal, with a clear distinction between small and large prophages. In this chapter we develop a mathematical model of the evolutionary forces affecting the prophage size distribution, and fit this model to three recent data sets. This approach offers quantitative estimates for the relative rates of lysogeny, induction, mutational degradation and selection acting on a wide class of prophage sequences. The model predicts that large prophages are predominantly maintained by the introduction of new prophage sequences through lysogeny, whereas shorter prophages can be enriched when they no longer encode the genes necessary for induction, but still offer selective benefits to their hosts.

3.1 Introduction

Bacteriophages (phages), the viral predators of bacteria, are the most abundant microorganisms in the biosphere [13] and have been critical players in the evolutionary history of bacteria [53]. Many phages reproduce exclusively through the lytic life cycle: after attachment to a bacterial cell surface, the phage infects the bacterium, uses bacterial machinery to produce progeny virions, and then kills the cell, through lysis, to release these viral particles into the environment. In contrast, temperate phages are defined by their ability to switch between the lytic and lysogenic life cycles. In the lysogenic life cycle, after infecting the bacterial cell, the phage DNA is integrated into the bacterial chromosome, and does not produce progeny virions [30]. Prophage refers to phage DNA which has been integrated into the bacterial chromosome in this way, and bacterial host cells containing prophages are referred to as lysogens [61]. Prophage sequences are then transmitted vertically with the host bacterial genome as the host cell divides into daughter cells.

Prophages are frequently identified in sequenced bacterial genomes and contribute up to 20% of a bacterial DNA sequence [11]. The number of prophages in a bacterial genome is extremely variable, ranging from zero to more than a dozen prophages per genome [55]. The identity of these prophages also varies both within and among species [45], with prophage content being particularly high in bacterial pathogenic strains [10].

While integrated in the bacterial genome, many temperate virus genes are not expressed and are thus not under selection for function [36]. Prophages are thus subject to loss-of-function mutations or deletions [11] which are hidden from purifying selection. This sequence degradation may affect genes required for lysis and may render the prophage defective, that is, unable

to enter the lytic life cycle. Defective or "cryptic" prophages are abundant in bacterial genomes, for example *Escherichia coli* K-12 contains nine cryptic prophage elements [60]. Some defective prophages are able to re-enter lysis with the help of co-infecting phages [41].

Recently, using *PhiSpy*, a bioinformatics tool for identifying prophages [2], 36,488 prophages were identified from the analysis of over 11,000 bacterial genomes; 83% of the bacterial genomes contained at least one prophage [32]. In a similar study [14], 4,122 prophages were identified in 795 genomes of *Acinetobacter baumannii*, for an average of 5 prophages per bacterial genome; PHAge Search Tool (PHAST) [63] was used for the identification of these prophages. Of these prophages, 78% were identified as defective [14]. Using PHASTER (PHAge Search Tool - Enhanced Release) [3], 11,297 prophages were identified in 1,760 *Salmonella enterica* genomes, for an average of 6.4 prophages per bacterial genome [45]. Due to this abundance of lysogeny in the bacterial world, it has been suggested that for temperate phages, the amount of viral genetic material encoded in prophages likely surpasses the total amount of viral DNA in free phage particles [59].

The relationship between bacteria and prophage is complex and multifaceted. Integration of the phage genome into a bacterial genome may be a survival strategy for the phage [49], at the risk of abandoning an independent, predatory existence. The acquisition of viral genomes comes with obvious costs for the bacterial host cell, most importantly the risk of future lysis [47], as well as the energy costs of maintaining extra genetic material in the bacterial genome [43]. But prophage often carry beneficial genes and may confer novel adaptive traits to their bacterial hosts, which in turn helps the prophages themselves to proliferate [6].

Table 3.1 lists several such benefits, conferred by prophages to their bacterial hosts. These traits are often interrelated, for example, prophages can enhance the host's capacity to form

Beneficial trait	Reference
protection from infection by the same phage (super infection exclusion)	[10, 29]
protection from phagocytosis	[42]
increase in cell growth rate	[60]
increase in antibiotic resistance	[26, 60]
increase in tolerance to environmental stress	[19, 60]
enhanced ability to form biofilm	[60, 24]
virulence factors	[27, 21, 4]
suppression of metabolic activity to increase survival in harsh environments	[46]
activation/deactivation of regulatory switches	[20]
adaptation to new host	[18]

Table 3.1: Beneficial traits that bacterial hosts may acquire from integrated prophages.

biofilm and biofilm can enhance antibiotic resistance. Biofilm bacteria are $500 - 1000$ times more resistant to antibiotics as compared to planktonic (free-living) bacteria [28].

Although prophages are quite stable within bacterial genomes, intact prophages are able to initiate the lytic life cycle. Induction results in the death of the bacterial host cell and release of progeny virions. Some prophages, like λ, excise from the bacterial chromosome to initiate the lytic life cycle, while for others, like Mu, the original prophage sequence remains in the bacterial host genome while viral particles are produced [50]. Induction can be triggered spon-

taneously [22, 31] or by DNA damaging agents, including external stimuli such as exposure to UV light or antibiotics [5, 40].

Prophage sequences are the single most prominent source of genetic diversity within bacterial populations [21], and prophage-encoded genes contribute to many aspects of bacterial physiology. The contribution of prophages to bacterial virulence, as well as to antibiotic resistance, have been particularly well-studied [58, 21, 26]. Understanding the spread and maintenance of prophage sequences in bacterial genomes is thus an important first step in estimating the impact of phage populations on these critical public health issues. Our aim is to make use of the wealth of recent data regarding the distribution of prophages in bacterial genomes to shed light on the evolutionary forces responsible for maintaining prophage sequences. We seek answers to basic questions about prophage evolution, such as: how does the rate at which prophage enter bacterial genomes through lysogeny compare with the induction rate; how does loss through induction compare with loss through mutational degradation; can we quantify the magnitude of the selective benefit conferred by prophages to their hosts?

3.1.1 The prophage size distribution

When a prophage is first integrated into a bacterial genome, its length is determined by the sequence length of the corresponding viral genome. Random mutation in the bacterial genome, however, is strongly biased toward deletions [35, 44, 16]. In addition, intact prophages may be lethal to the host cell, thus there should be strong selection for inactivation. Hence in a random sample of prophage sequences, one might expect a few large prophages, corresponding to fully inducible sequences, followed by a gradient of smaller and smaller prophages that have been

subject to degradation over evolutionary timescales [6]. In other words, assuming prophage sequences enter bacterial genomes at a relatively constant rate, we might expect a unimodal distribution with a peak on the right and a long tail to the left (negative skewness).

In contrast with this expectation, three recent datasets – available in public databases and discussed in greater detail below – suggest that the distribution of prophage lengths, across a wide variety of bacterial species, is multimodal [6, 15, 8, 37]. In the sections to follow, we will use each of these datasets to determine the simplest evolutionary model that is consistent with these data; in other words, which forces or processes are, at a minimum, necessary to recover these distributions?

The data sets we analyze further are:

Data Set 1 [6]: Bobay et al. identified 624 prophages (474 from *E. coli* and 150 from *S. enterica*) and recovered a bimodal distribution of prophage lengths; see Figure 1A. Note that prophages that had no resemblance to core phage genomes from this study were discarded.

Data Set 2 [15]: 128 prophages in *Desulfovibrio*, the sulphate-reducing bacteria, were shown to have a bimodal or possibly trimodal distribution; see Figure 1B.

Data Set 3 [37]: The ACLAME database (http://aclame.ulb.ac.be) contains 760 prophage sequences. These data were retrieved using Prophinder [38], a prophage detection algorithm. Note that this data set includes a sparse tail of prophages with a length greater than 60 kb; in the data fitting described below, we neglect these outliers, reducing the data set from 760 to 737 prophages; see Figure 1C.

A further summary of these data sets is provided in Table 3.2.

In addition to these three publicly-available datasets, further evidence regarding the distribution of prophage lengths has recently appeared in a study of the molecular epidemiology of

Pneumcocci [8]; 482 prophages were collected from clinical isolates that spanned 36 countries and nearly a century (1916-2008). Although the lengths of individual prophages in these data are not publicly available, overall the lengths exhibit a bimodal distribution, as shown in Figure 1D.

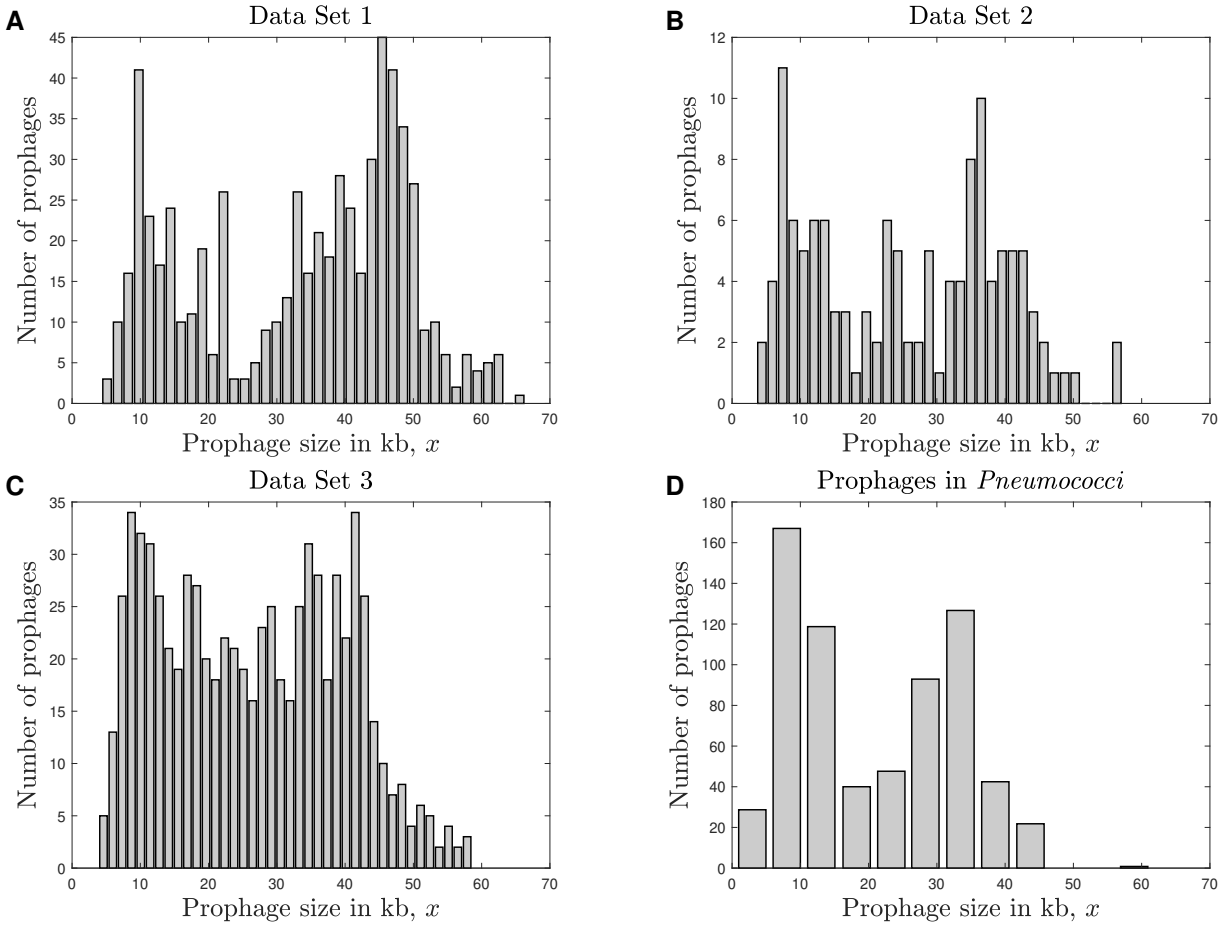


Figure 3.1: Prophage size distributions. (A) Prophages identified in *E. coli* and *S. enterica* [6]. (B) Prophages identified in *Desulfovibrio* [15]. (C) Prophages across a range of sequenced bacterial genomes [37]. (D) Prophages identified in *Pneumococci* [8].

Data Set	Prophage number	Min (kb)	Max (kb)	Average (kb)	Bacterial Species	Reference
1	624	4.348	66.345	33.268	*E. coli*, *S. enterica*	[6]
2	128	3.603	57.140	25.580	*Desulfovibrio*	[15]
3	737	3.918	58.560	26.450	Diverse	[37]

Table 3.2: Summary of the three data sets analyzed in this study.

3.1.2 Our approach

The natural question that arises is: why are prophage size distriubutions so often bimodal? In a preliminary investigation, [6] arrived at the conclusion that the bimodal distribution in their data is neither due to taxonomic biases nor due to large neutral deletions of genetic material. In addition, if there were substantial heterogeneity in the prophage integration rate over evolutionary time, we might expect greater diversity in the multimodal distributions observed across a range of bacterial species; in other words, it seems unlikely that the peak at shorter lengths in four datasets is due to a single historical “burst" of prophage integration. The aim of our study is to investigate these prophage distributions in more detail and, using the available data, determine which evolutionary processes are necessary to explain these observations.

The organization of this article is as follows: in Section 3.2, we derive a partial differential equation that models the time evolution of the prophage distribution, including expressions for the effects of mutation, selection, horizontal gene transfer and induction. In Section 3.3, we describe our analysis of the model, including model selection and fitting the model to the

available data sets. In Section 3.4, we present the results of data fitting. Finally, in Section 3.5, we discuss the conclusions derived from this analysis and suggest further directions.

3.2 Model Derivation

To better understand the observed distributions of prophage lengths, we seek an expression for the expected frequency of prophages of length x, $P(x)$, in a population of bacterial genomes. Here x is prophage length in kb, with $x_0 < x < x_M$ (x_0 and x_M are the minimum and maximum prophage length, respectively). We begin by considering the processes that change the distribution of prophages over time, and then solve for the expected steady state (long-term behavior) of the model. Let $Q(x,t)$ be the frequency of prophages of length x at time t. After time step δt, this distribution may change due to: (1) new temperate viral genomes entering the bacterial genome; (2) horizontal gene transfer (HGT) adding new prophages or partial prophages to the existing prophage pool; (3) mutational degradation reducing prophage lengths; (4) selection promoting the proliferation of the bacterial population, and therefore the prophage population; and (5) induction removing prophages from the population .

Taking into account these five possible processes, we arrive at the following partial differential equation describing the time evolution of the prophage size distribution:

$$\frac{\partial Q(x,t)}{\partial t} = \alpha f(x) + \beta\, g(x) + \frac{\partial}{\partial x}\left[D(x)Q(x,t)\right] + r_S\, S(x)\, Q(x,t) - r_I\, I(x)\, Q(x,t)\ . \qquad (3.1)$$

The five terms on the right describe the influx of new prophage via lysogeny, influx via HGT, mutational degradation, selection, and induction respectively. The distribution of interest, $P(x)$, if it exists, is the steady state solution of $Q(x,t)$, i.e., $P(x) = \lim_{t\to\infty} Q(x,t)$. In the following

subsections we explain each of the terms in Equation 3.1 in greater detail, and derive mathematical expressions for the underlying functions. For this purpose, in subsection (3.2.1), we show that there is a linear relationship between the length of a prophage and the number of genes it contains.

$Q(x,t)$	frequency of prophages of length x (kb) at time t
$P(x)$	steady state solution of $Q(x,t)$
$f(x)$	length distribution of prophage sequences entering via lysogeny
$g(x)$	length distribution of phages transferred by HGT
$D(x)$	mutational degradation rate
$S(x)$	expected fraction of r_S conferred by prophage of length x
$I(x)$	probability that prophage carries genes required for induction
α	rate of lysogeny
β	rate of horizontal gene transfer (HGT)
r_S	selection coefficient (intact prophage)
r_I	rate of induction

Table 3.3: Model functions and parameters.

3.2.1 Number of genes and length of prophage

Genome size and number of genes are strongly correlated in prokaryotes [25]. We used data available in the ACLAME database (http://aclame.ulb.ac.be), to confirm that this relationship between sequence length and the number of genes extends to prophage sequences in bacterial genomes, as illustrated in Figure 3.2. We find that the length of a prophage, x, and the number

of genes it carries, m, are highly correlated (correlation coefficient $r = 0.912$). We model this correspondence as the simple linear relation $m = \kappa\, x$, with x in kb and $\kappa = 0.808$ genes per kb, corresponding to 1.2 kb per gene.

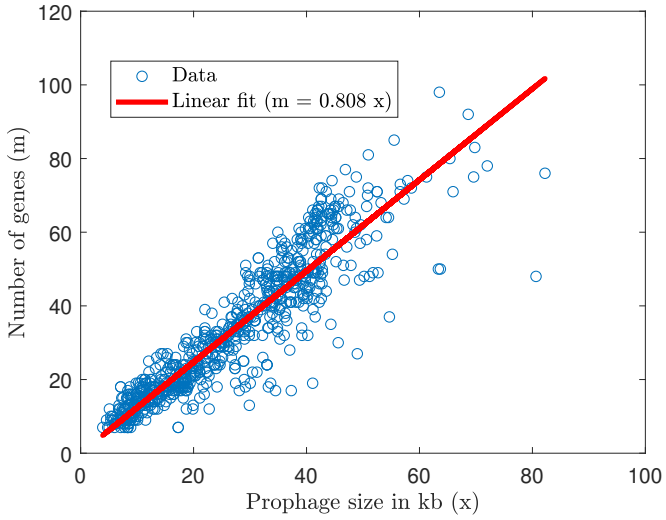


Figure 3.2: Data from ACLAME database (http://aclame.ulb.ac.be), showing the strong correlation between length of prophage and the number of genes on that prophage. All prophages in the database of length up to 85 kb were considered.

3.2.2 Lysogeny

The function $f(x)$ in Equation 3.1 represents the length distribution of prophages which enter bacterial genomes through lysogeny. In other words, in the event that a phage genome will be integrated into a bacterial genome through lysogeny, $f(x)$ gives the probability density for the length of the phage genome to be integrated. Thus, $f(x)$ captures not only the distribution of autonomous temperate phage genome lengths, but also any potential differences in lysogeny probabilities among these phages (see Appendix C for further details).

Most temperate phages are double-stranded (ds) DNA viruses [54], although several single-

stranded (ss) DNA phages have also been identified as temperate [34]. We therefore expect that $f(x)$ might resemble the distribution of dsDNA phage lengths in nature. Consistent with results reported for dsDNA phage (see Figure 1 in [6]), we assume that the length distribution of these active phages may be multimodal, and that the minimum length for an autonomous phage is longer than the minimum length for a (possibly degraded and cryptic) prophage. For convenience, we describe the multimodal distribution of active phages as the sum of $g = 1$ to 3 Gaussian probability density functions (i.e. a mixed distribution with up to three components):

$$f(x) = \begin{cases} \sum_{i=1}^{g} p_i e^{-\frac{(x-(\theta+\mu_i))^2}{\sigma_i^2}} & x \geq \theta \\ 0, & x < \theta. \end{cases} \tag{3.2}$$

where $p_i > 0$, $\mu_i > 0$, $\sigma_i > 0$ for $i = 1$ to g represent relative weights (in the convex combination), means and standard deviations of the component distributions, respectively. Note that θ in this expression gives the length of the smallest autonomous temperate phage. Temperate dsDNA phages have diverse genomes, with the smallest reported size of about 15 kb (*Bacillus phage Bam35c*) [48, 23]; [6] report that the smallest autonomous dsDNA phage that can infect enterobacteria is 30 kb. Here we assume that the smallest autonomous dsDNA phage that may successfully lysogenize a host has a genome size of $\theta = 20$ kb, however our results were not sensitive to this choice of threshold parameter (see Appendix B).

3.2.3 Horizontal gene transfer (HGT)

HGT describes several processes by which genetic material is transferred from a donor bacterium to a recipient bacterium. The transfer of genetic material can be accomplished through *conjugation*, *transformation* and *transduction*. Although it has been inferred that transduc-

tion is 1000 times less likely than conjugation to transfer antibiotic resistance genes [57], we note that transduction might be especially relevant to the HGT of chromosomal prophages. In specialized transduction, a prophage erroneously excises from the bacterial genome, taking a neighboring piece of the host chromosome and possibly integrating an incomplete prophage sequence into the new bacterial host; in generalized transduction, DNA from elsewhere in the host chromosome (not necessarily adjacent to the prophage) is packaged into the viral capsid and transmitted to a new host [56]. Since, for example, prophages can encode packaging-recognition sites (*pac* signals) [12], transduction could play a significant role in the prophage size distribution.

The function $g(x)$ in equation 3.1 represents the transfer of prophage sequences, or partial prophage sequences, into host genomes by any of these processes of HGT. We reasoned that shorter DNA sequences should be transferred with higher probability than larger ones, therefore, we assume that $g(x)$ must be a non-negative decreasing function, i.e, $g(x) \geq 0$ and $g^{'}(x) \leq 0$ for all $x \in [0, x_M]$ (recalling that x_M is the size of the largest prophage in the dataset). The simplest function which satisfies these properties is a decreasing linear function, $g(x) = -x + x_M$. The slope of this line is scaled by the free parameter β representing the maximum rate of prophage integration via HGT. We also investigated more complex functions describing HGT (data not shown) but these were not justified in data fitting.

3.2.4 Degradation

Mutational processes in bacterial genomes exhibit a strong bias toward deletion [35], and mutational degradation may render intact prophages cryptic, i.e. incapable of induction or unable

to form plaques. Under the assumption that the probability of deletion is constant along the genome, we recover a linear relation between prophage sequence length and the rate of loss, $D(x) = r_D\, x$, where again x is the length of the prophage sequence in kb, and r_D is the rate of degradation (kb of prophage sequence lost, per bacterial generation, per kb of prophage sequence). Degradation from larger to smaller prophages depends on the gradient of the prophage length distribution, thus this process introduces an advective term.

3.2.5 Selective advantage

Integrated prophage often confer fitness benefits to their bacterial hosts [7]. Longer prophage sequences may encode a greater number of beneficial genes, conferring greater advantage to host cells and increasing the prophage population in turn. Let n_b be the total number of potentially beneficial genes carried by phage. If L is the number of genes in a prophage as it attaches to the bacterial genome, suppose after degradation m intact genes remain. Then the probability, $\mathcal{P}(i)$, where $0 < i \leq m$, that the prophage of length m carries i beneficial genes is

$$\mathcal{P}(i) = \frac{\binom{n_b}{i}\binom{L-n_b}{m-i}}{\binom{L}{m}},$$

which is a Hypergeometric distribution, with expected value $\sum_{i=1}^{m} i\,\mathcal{P}(i) = \frac{n_b\, m}{L}$. Converting this expectation to units of sequence length, a prophage degraded to length $x = m/\kappa$ from initial length $x_L = L/\kappa$ is expected to carry fraction x/x_L of all possible beneficial genes (a result we will use in Equation 3.4).

The probability that a prophage of length x is a degraded version of an active phage of length x_L is given by the proportion of active phages that have length x_L, as a fraction of all the

active phages that might have given rise to this prophage:

$$\mathcal{A}(x, x_L) = \begin{cases} 0, & x > x_L \\ \dfrac{f(x_L)}{\sum_{y=x}^{x_M} f(y)}, & x \leq x_L. \end{cases} \tag{3.3}$$

We assume that an active prophage confers an overall selective advantage r_S per bacterial generation. For mathematical tractability, we assume that the magnitude of this maximum selective effect does not vary with the initial length of the active phage sequence; this is a simplification that should be relaxed in future work. We can then deduce $S(x)$, the expected fraction of r_S that will be conferred by a prophage of length x, by conditioning and summing over all possible active phages that may have produced this prophage:

$$S(x) = \sum_{x_L=x}^{x_M} \mathcal{A}(x, x_L)\frac{x}{x_L}. \tag{3.4}$$

Thus, the term $r_S S(x)$ gives the change in the intrinsic growth rate conferred on average by a prophage of length x.

Finally, we note that along with potential fitness benefits, prophage genes may also impose fitness costs on their hosts [33]. The parameter r_S, which could be positive or negative, reflects the sum total of these costs and benefits. Thus for example if the best-fit value of r_S were negative, the model would predict that independent of induction, the carriage of prophage genes comes at a net cost to the host cell.

3.2.6 Induction

Prophage may re-instigate the lytic life cycle, either spontaneously or in response to some stress, resulting in the death of the host cell and release of progeny phages. As before, we let

L denote the number of genes in an active phage, while m is the number of genes retained on a (possibly degraded) prophage. Assume n_l is the number of phage genes required for the loss of prophage from the bacterial genome through induction. (Since the model tracks only the frequency of bacterial hosts carrying prophages of a given length, this loss could reflect either the excision of prophage from the bacterial genome, or the death of the bacterial host through lysis.) The probability that a prophage carries the genes required for induction, given that after degradation it carries m out of L genes is:

$$\begin{cases} 0, & m < n_l \\ \dfrac{\binom{L-n_l}{m-n_l}}{\binom{L}{m}}, & n_l \leq m \leq L. \end{cases}$$

We again use the parameter κ to convert between gene number and sequence length, and make use of Stirling's approximation to simplify factorial terms. We thus approximate the probability that a prophage, initially of length x_L but degraded to length x, contains the genes required for induction as:

$$\mathcal{R}(x, x_L) \approx \begin{cases} 0, & x < x_n \\ \dfrac{x^{\kappa x}\,(x_L - x_n)^{\kappa(x_L - x_n)}}{x_L^{\kappa x_L}(x - x_n)^{\kappa(x - x_n)}}, & x_n < x \leq x_L, \end{cases} \tag{3.5}$$

where $x_n = n_l/\kappa$ is the length of the genes required. The probability that a prophage of length x is a degraded version of prophage of length x_L is given by (3.3). Thus, the overall probability that the prophage contains the genes required for induction (excision from the host genome), after conditioning and summing over all possible active phages that may have produced this prophage is given as:

$$I(x) \approx \sum_{x_L = x}^{x_M} \mathcal{A}(x, x_L)\,\mathcal{R}(x, x_L). \tag{3.6}$$

The parameter r_I gives the induction rate, that is, the rate per prophage per bacterial generation at which prophages are lost from bacterial genomes due to induction. Finally, we note that because prophages lost to induction may or may not contain the genes required for re-infection, we do not expect that the product $I(x)P(x)$ will directly yield the distribution of lysogenizing phages, $f(x)$ (but see Appendix C).

Typical geometries of these five functions – lysogeny, HGT, degradation, selection and induction – are illustrated in Figure 3.3.

3.2.7 Closed-form solution

If we consider $\lim_{t\to\infty} Q(x,t) = P(x)$ and $D(x) = r_D\, x$ then the differential equation generating the steady state solution of equation 3.1 is given by

$$\frac{dP(x)}{dx} + \left(\frac{1}{x} + \mathcal{F}(x)\right) P(x) + \frac{\alpha}{r_D\, x}\, f(x) + \frac{\beta}{r_D\, x}\, g(x) = 0 \tag{3.7}$$

where $\mathcal{F}(x) = (r_S\, S(x) - r_I\, I(x))/(r_D\, x)$. Equation (3.7) is first order linear ordinary differential equation and its solution is given by

$$P(x) \;=\; \frac{-\,e^{-\int \mathcal{F}(x)dx}}{r_D\, x} \int (\alpha\, f(x) \,+\, \beta\, g(x))\, e^{\int \mathcal{F}(x)dx} dx + \frac{C}{x}\, e^{-\int \mathcal{F}(x)dx}, \tag{3.8}$$

where C is a constant of integration. Although in general a numerical approach is required to evaluate $P(x)$ due to the complexity of the functions $S(x)$, $I(x)$ and $f(x)$, the form of this solution will prove valuable in eliminating some solutions during model selection (see Table 3.4 below).

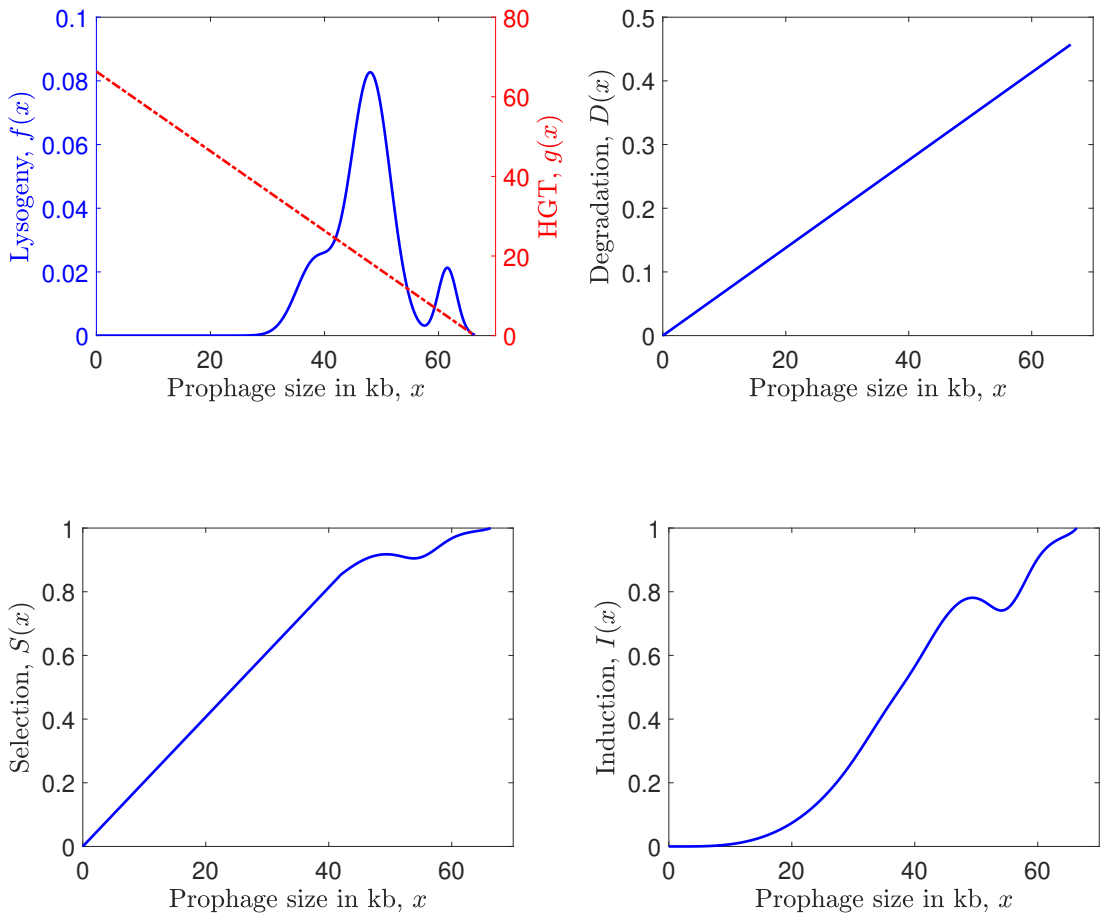


Figure 3.3: Example geometries of the influx distributions via lysogeny and HGT (top panel, left and right axes respectively), as well as the degradation, selection and induction functions, plotted against prophage size. To illustrate the shapes of these functions, we have used parameters corresponding to the best fit to Data Set 1 (see Results to follow). Parameter values are $p_1 = 17.95$, $\mu_1 = 18.37$, $\sigma_1 = 4.73$, $p_2 = 64.19$, $\mu_2 = 28.06$, $\sigma_2 = 4.93$, $p_3 = 16.57$, $\mu_3 = 41.54$, $\sigma_3 = 2.32$, $r_D = 0.0069$.

3.3 Model selection and data fitting

Although lysogeny, HGT, degradation, selection and induction may all contribute to the maintenance of the prophage population in nature, we performed rigorous model selection to determine which of these processes are statistically justified in modeling the data. This approach allows us to identify the key evolutionary processes underlying the prophage distribution, and to estimate the relative magnitude of their effects.

Since lysogeny is a necessary prerequisite for the prophage distribution, we considered models that included incoming prophage ($f(x)$) but included or excluded HGT, degradation, induction and selection in all possible combinations. Thus in total, we tested all $2^4 = 16$ possible models. For brevity, Table 3.4 lists the full model, as well as all possible models that exclude HGT; analogous models including HGT were also tested.

The first step in analyzing these models was to exclude models that are qualitatively unable to capture the prophage size distribution data. From the analytical solutions of models 3 (with both induction and selection present), 7 (with only induction present), and 8 (with only selection present), we see that in these cases, the steady-state solution $P(x) = 0$ wherever the incoming phage distribution $f(x) = 0$. Thus these models predict the absence of prophage with lengths smaller than $\theta = 20$ kb. These models are clearly unable to capture the distributions illustrated in Figure 3.1 and were excluded from further analysis. This result makes intuitive sense: the three excluded models do not include degradation, and therefore cannot explain prophage with lengths shorter than the lengths of autonomous temperate phage.

We proceeded with model selection using the remaining five models (models 1, 2, 4, 5 and 6), fitted to each data set. For each model, we also allowed the function $f(x)$ (the incoming

Model	Processes $f\ D\ I\ S\ g$	Steady-state solution, $P(x)$
1	✓✓✓✓✓	$\frac{-e^{-\int \mathcal{F}(x)dx}}{r_D\, x} \int (\alpha\, f(x) + \beta\, g(x))\, e^{\int \mathcal{F}(x)dx} dx + \frac{C}{x} e^{-\int \mathcal{F}(x)dx}$
2	✓✓✓✓✗	$\frac{-\alpha\ e^{-\int \mathcal{F}(x)dx}}{r_D\, x} \int f(x)\, e^{\int \mathcal{F}(x)dx} dx + \frac{C}{x} e^{-\int \mathcal{F}(x)dx}$
3	✓✗ ✓✓✗	$\frac{-\alpha f(x)}{S(x)-I(x)}$ where $S(x) \neq I(x)$
4	✓✓✗ ✓✗	$-\frac{\alpha}{r_D}\frac{1}{x} e^{-\int \left(\frac{r_S}{r_D}\frac{S(x)}{x}\right)dx} \int f(x) e^{\int \left(\frac{r_S}{r_D}\frac{S(x)}{x}\right)dx} dx + \frac{C}{x} e^{-\int \left(\frac{r_S}{r_D}\frac{S(x)}{x}\right)dx}$
5	✓✓✓✗ ✗	$-\frac{\alpha}{r_D}\frac{1}{x} e^{\int \left(\frac{r_I}{r_D}\frac{I(x)}{x}\right)dx} \int f(x) e^{-\int \left(\frac{r_I}{r_D}\frac{I(x)}{x}\right)dx} dx + \frac{C}{x} e^{\int \left(\frac{r_I}{r_D}\frac{I(x)}{x}\right)dx}$
6	✓✓✗ ✗ ✗	$-\frac{\alpha}{r_D\, x} \int f(x)dx + \frac{C}{x}.$
7	✓✗ ✓✗ ✗	$\frac{\alpha f(x)}{I(x)}$, where $I(x) \neq 0.$
8	✓✗ ✗ ✓✗	$\frac{-\alpha f(x)}{S(x)}$, where $S(x) \neq 0.$
9	✓✗ ✗ ✗ ✗	no steady-state solution

Table 3.4: A detailed description of the models considered. Each model includes or excludes terms on the right-hand side of Equation 3.1 as indicated. The analytical forms for the steady-state solutions, as shown in the right-most column, allow us to eliminate several models from further analysis (see text for details). Here $\mathcal{F}(x) = \frac{r_S\, S(x)}{r_D\, x} - \frac{r_I\, I(x)}{r_D\, x}$ and C is an arbitrary constant.

phage distribution) to be described by a mixed distribution incorporating one to three Gaussian distributions. While the data sets included between $n = 128$ and $n = 737$ data points, the tested models included between $k = 4$ and $k = 15$ free parameters. We used a finite difference scheme to obtain, numerically, the steady-state solution to the model, and compared this steady-state solution to the data, optimizing the log-likelihood to identify the best fit parameter values. The log-likelihood is defined as $\log(L) = \sum \log P(x_i)$, where x_i are the n observed lengths of prophage sequences in the data set, $P(x)$ is the numerically obtained steady-state solution, and the sum is taken for $i = 1$ to n. To select the best model among the candidate models, we used the Akaike Information Criterion (AIC) [1], defined as:

$$\text{AIC} = 2k - 2\ \log\left(\hat{L}\right) \tag{3.9}$$

where k is the number of free parameters, and $\log\left(\hat{L}\right)$ is the maximum log-likelihood.

While the lowest AIC value corresponds to the best fit, it is possible that several candidate models may offer equivalently good fits; these correspond to models that cannot be rejected, statistically. To address this issue, we compute the relative probability. If AIC_{min} is the lowest AIC value obtained for one of the candidate models, the relative probability [9] is defined for each candidate model as $R = \exp\left((AIC_{min} - AIC)/2\right)$. The best fit model will thus have relative probability 1. If we imagine adding a single “dummy" variable to the best fit model, that is, we add an additional parameter that has no effect on the fit, the AIC will increase by 2 and the log-likelihood will not change. Thus the relative probability of the best fit model including an extra dummy parameter will be $\exp(-1) = 0.368$. We therefore reject candidate models with $R \leq \exp(-1)$. If candidate models have relative probability values that exceed $\exp(-1)$, we are unable to reject them and consider them as possible “best fits” to the data.

3.4 Results

We fit models 4, 5 and 6 to each of the three data sets, including in each model the possibility of up to three components in the incoming phage distribution, $f(x)$, and including or excluding HGT, $g(x)$. A full summary of the model-fitting results is provided in Appendix A.2 (Table A.1, Table A.2, Table A.3 for Data Sets 1, 2 and 3, respectively).

Despite the reduced number of free parameters in the simpler models (an attribute rewarded by the AIC criterion), in no case did models 4 (degradation and selection), 5 (degradation and induction) or 6 (degradation) achieve the best fit to any of the data sets. To illustrate, we have reproduced the fits obtained to Data Set 1 with these three models in Figure 3.4.

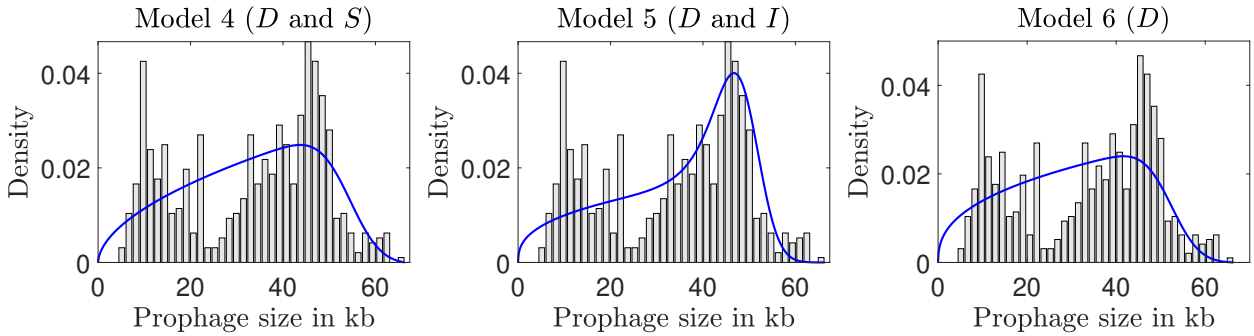


Figure 3.4: Lines of best fit ($P(x)$, shown in blue) obtained to Data Set 1 (histograms) by models 4, 5 and 6 (see Table 3.4); these models did not provide adequate fits to the data.

For all three data sets, then, model 2, as described by Equation 3.1 with $\beta = 0$, provided the best fit, with varying degrees of complexity in the function describing the incoming phage distribution ($f(x)$). We describe and illustrate these results below.

3.4.1 Data Set 1

The best fit to these data from *E. coli* and *S. enterica* was obtained using the full model without HGT (model 2), with the distribution of incoming phages described by a mixture of three

underlying Gaussian distributions ($g = 3$). This model has 14 free parameters (see Figure 5A). The second-best fit is the same model with HGT; this fit has relative probability 0.372 and thus cannot be rejected. We note however that the contribution of HGT in this second-best fit is very small (see Table 3.5). The best fits are illustrated in Figure 5B and 5C. Detailed results of data fitting are provided in Table A.1, Appendix A.2.

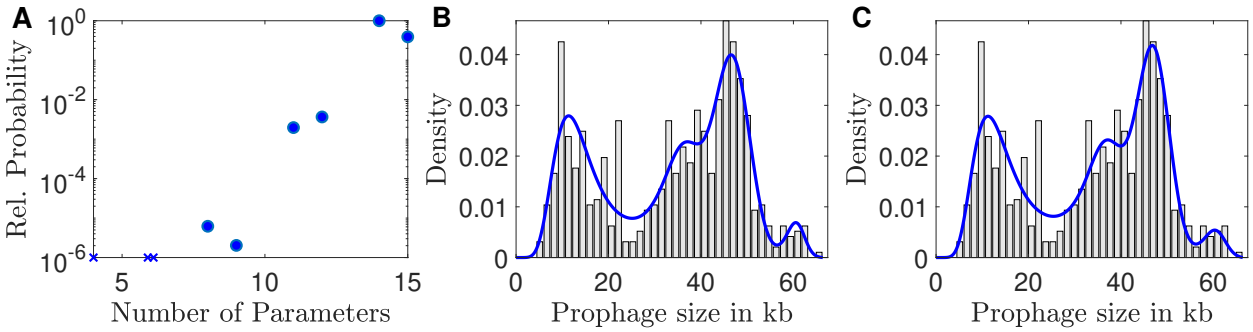


Figure 3.5: Model fitting results for Data Set 1. (A) The relative probability of candidate models for Data Set 1, plotted as a function of the number of parameters in that model; crosses indicate relative probabilities $\leq 10^{-6}$. (B) The best fit predicted by the model ($P(x)$, blue curve) to Data Set 1 (histogram). The best fit included 14 free parameters. (C) The second-best fit model (blue curve) to Data Set 1 (histogram). The second-best fit included 15 free parameters.

3.4.2 Data Set 2

The prophage length distribution from *Desulfovibrio* was best described by the full model without HGT, and a single Gaussian describing the incoming phage lengths ($g = 1$, 8 parameter model), see Figure 6A. This fit is illustrated in Figure 6B. Details are provided in Table A.2 in Appendix A.2.

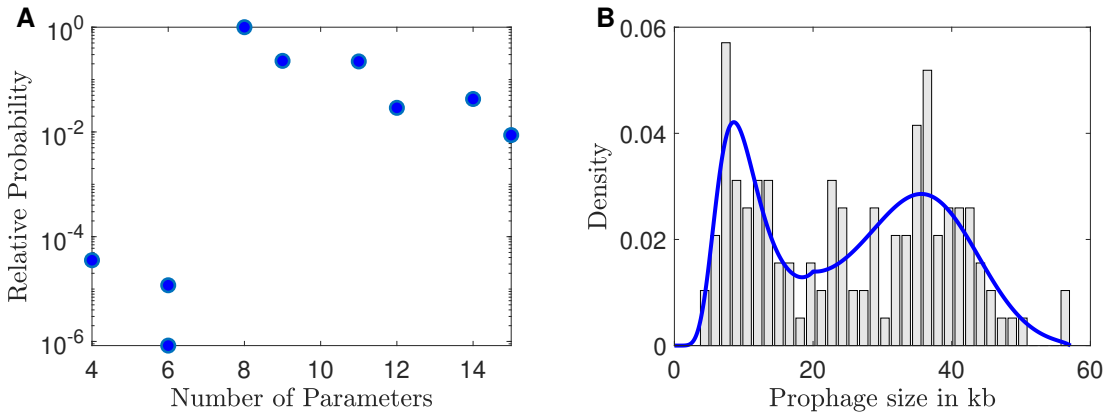


Figure 3.6: Model fitting results for Data Set 2. (A) The relative probability of candidate models for Data Set 2, plotted as a function of the number of parameters in that model. (B) The best fit predicted by the model ($P(x)$, blue curve), to Data Set 2 (histogram). The best model includes 8 free parameters and has relative probability 1.

3.4.3 Data Set 3

The prophage length distribution from the ACLAME database was also best described by the full model with a single Gaussian describing incoming phage ($g = 1$, 8 parameter model), see Figure 7A. The best fit is illustrated in Figure 7B. Model fitting details are provided in Table A.3 in Appendix A.2.

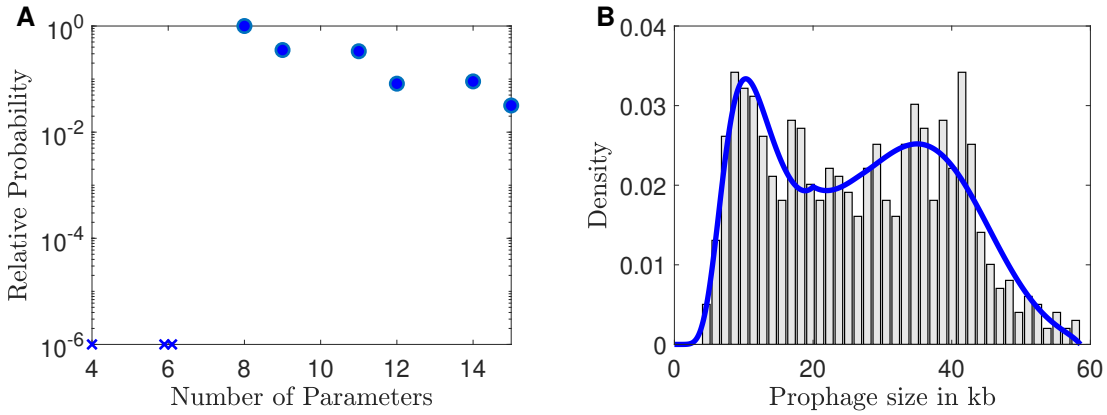


Figure 3.7: Model fitting results for Data Set 3. (A) The relative probability of candidate models for Data Set 3, plotted as a function of the number of parameters in that model; crosses indicate relative probabilities $\leq 10^{-6}$. (B) The best fit predicted by the model ($P(x)$, blue curve), to Data Set 3 (histogram). The best model includes 8 free parameters and has relative probability 1.

Table 3.5 provides a summary of the best-fit parameter values obtained for the three data sets. A sensitivity analysis, using Data Set 1, indicates a high degree of confidence in the parameter values, that is, all parameters are well-constrained by the data (see Appendix B). However some care must be taken in interpreting these numerical values. We present these rates in comparable units and address the implications of the quantitative results further in the Discussion.

		Data Set 1		Data Set 2	Data Set 3	
	Parameter	best	2nd best			**Mean**†
α	Relative rate of lysogeny	0.1301	0.1982	0.1191	0.0734	0.1175
r_D	Relative rate of degradation	0.0069	0.01361	0.0051	0.0052	0.0066
r_S	Relative selection coeff. (intact prophage)	0.3137	0.7276	0.2397	0.2249	0.3110
r_I	Relative rate of induction	0.6169	1.1139	0.5291	0.4713	0.6025
n_l	Number of genes required for induction	2.440	1.9512	2.440	2.5856	2.4198
β	Relative rate of horizontal gene transfer	—-	8.76×10^{-13}	—-	—-	8.76×10^{-13}

Table 3.5: Parameter values for the best fits. †Mean across all data sets, weighted by relative probability for Data Set 1.

3.5 Discussion

Because we can only fit the steady-state solution of Equation 3.1 to the data, the resulting rates are only meaningful relative to other rates in the model. Thus, although the time units of the best-fit rates are an arbitrary number of generations, we can express each of these rates relative to the induction rate. This allows us to compare the evolutionary forces at play in terms of what

we will call the “expected prophage lifetime", that is, the average time between lysogeny and induction, for prophages that retain all the genes necessary for induction. We find that the time between lysogeny events (new prophages entering the genome) is about 5 prophage lifetimes, while the selection coefficient, for an intact prophage, is approximately 0.5 per prophage lifetime. If a prophage remains in the host genome for 100 bacterial generations before induction, for example, this selection coefficient would correspond to a selection coefficient $s = 0.004$ per bacterial generation. Finally, we predict that degradation of the prophage genome occurs at a rate of about 0.01 kb per kb in the prophage genome, per prophage lifetime. Thus on average the model predicts that prophages have lost only 1% of their genome to degradation at the time of induction. These normalized rates are presented in summary in Table 3.6.

Rates expressed per expected prophage lifetime	
Lysogeny	0.20
rate at which new prophage enters genome	
Degradation	0.01
kb lost per kb of prophage genome	
Selection	0.52
overall selection coefficient per prophage lifetime	
Induction	1.00
rate at which fully competent prophage induces	

Table 3.6: Rates of the processes in the model, normalized by the induction rate. Induction rate and selection coefficient are provided for fully intact (non-degraded) phage. See text for details.

From these normalized rates, a picture emerges in which induction is the dominant fate for active prophages, occurring at a much higher rate than any other process. New prophages enter the bacterial genome, on average, at a rate that is about one fifth of the induction rate. These new sequences degrade very slowly relative to their induction rate, an observation that seems reasonable given that prophages would be unable to induce if degradation were rapid.

Despite the slow degradation rate, over evolutionary time smaller and smaller prophages accrue in host genomes. These are maintained due to the balance between two effects: induction and selection. In particular, short prophage sequences typically lack the genes required for excision or induction, but may still confer some benefit to their host.

Thus, our model predicts that the peak on the right of the prophage size distribution is due to the contribution of autonomous free phage, entering bacterial genomes via lysogeny, the term $\alpha f(x)$ in our model. In contrast, the peak on the left is maintained in the region for which $r_s S(x) > r_I I(x)$, that is, where the benefits of selection outweigh the costs of induction, as illustrated in Figure 3.8.

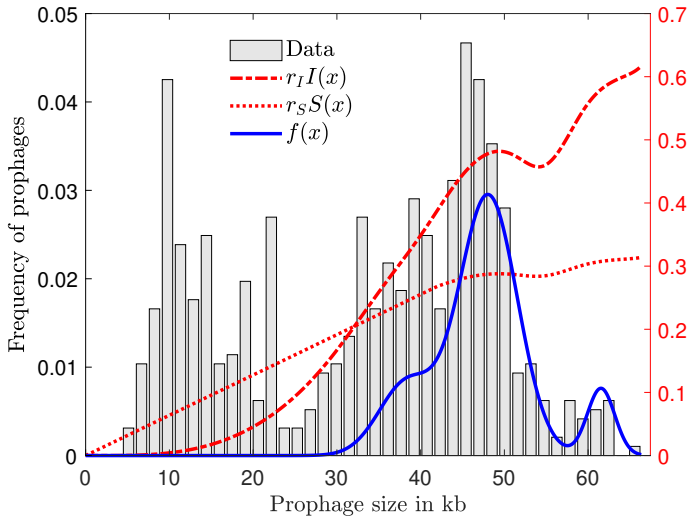


Figure 3.8: Components of the best-fit model prediction for Data Set 1 (histogram, left axis). The distribution of autonomous temperate phages ($f(x)$, solid curve) is plotted along with the induction curve ($r_I I(x)$, dash-dotted, right axis) and the selection curve ($r_S S(x)$, dotted, right axis); induction and selection intersect near 30 kb.

An unexpected result of our analysis is that a process that preferentially adds prophages of shorter lengths to bacterial genomes (the process we describe as HGT in the model derivation) was not required to provide a fit to any of the three data sets. For Data Set 1, HGT was included

in the second-best fit, but the overall rate of HGT was extremely small relative to other rates in the model. These results of course do not preclude a role for HGT in the maintenance of prophages in bacterial genomes, but indicate that HGT is not required to explain the empirical data currently available. As mentioned previously, it seems likely that transduction may be the most important HGT process for prophages, and transduction rates are inferred to be low relative to other modes of HGT [57].

Our findings suggest that a minimum of two to three prophage genes are required to enable prophage excision from the bacterial chromosome. The excision of phage λ, for example, requires at least two enzymes, an integrase and exonuclease, which are produced from a single transcript encoding the *xis* and *nit* genes [62]. We would of course expect wide variability in these steps across phage-host systems. For example, phage Mu replicates first and then excises from the host chromosome [39], and would presumably require further intact genes for excision.

The relation between prophage and their bacterial hosts may either be parasitic or mutualistic depending on the balance between the cost the bacterial host incurs due to the integration of foreign DNA into its genome, and benefits conferred by the foreign DNA to the bacterial host [51]. The biggest cost is incurred by induction, although due to the compactness of bacterial genomes, small insertions of foreign DNA may result in significant energy costs as well [33]. Our results predict a tipping point between parasitism and mutualism, the point at which $r_I I(x) = r_S S(x)$. Shorter prophage sequences are unlikely to maintain all the genes required for induction, and our model predicts that they persist at high frequencies in host genomes because of the selective benefits they still confer. Thus the peak on the left of the bimodal prophage distributions may correspond to predominantly mutualistic prophage, consistent with high levels

of purifying selection inferred for prophage genes using comparative genomics [6]. Metagenomic data, that is, prophage distributions obtained from environmental samples of bacterial populations, could help clarify the role of positive selection in maintaining prophage sequences.

Although an active body of research addresses the comparison and refinement of algorithms for prophage identification [52, 17], short prophage remnants can be difficult to detect and are likely underrepresented in the available data. For example, prophage remnants with sequence similarity to short mobile genetic elements were excluded from Data Set 1 [6], and a minimum of six phage-like genes are required to identify phage gene clusters in the PHAST search tool [63]. Most algorithms to date rely on homology-based techniques to identify prophages, making it difficult to detect prophages that are not similar to known phages (but see [2]), and making the identification of shorter sequences more challenging. This detection bias likely shifts the position of the lower peak in bimodal prophage distributions – the true peak in the data might occur at even shorter prophage lengths – but would not affect our conclusions regarding the underlying mechanisms in play.

We note that for one of the three data sets in this analysis, the best fit to the data was obtained using 14 of 15 possible parameters, that is, the best fit supported a relatively complex model. This implies that as richer data sets become available, further features could, and should, be added to the model to better describe the prophage distribution. As mentioned previously, two assumptions that could clearly be relaxed are that all phage genomes, irrespective of their length, offer the same average selective benefit to their host, and require the same number of genes for induction. In reality, longer active phages presumably have the capacity to encode further beneficial functions and more complex excision mechanisms. Another assumption, inherent in our approach, is that degraded prophage have lost the genes required for induction,

or the genes conferring benefit to the host, in proportion to their total gene loss. Thus the selection or induction rates depend only on prophage length. A more nuanced (but less tractable) approach will be to follow the loss and enrichment of specific classes of genes in degraded prophage sequences.

Acknowledgements

The authors are indebted to Alita Burmeister for several insightful comments that strengthened the work. The Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged for funding.

Bibliography

[1] H. Akaike. Likelihood of a model and information criteria. *Journal of Econometrics*, 16(1):3–14, May 1981.

[2] S. Akhter, R. K. Aziz, and R. A. Edwards. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, 40(16):e126–e126, Sept. 2012.

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

[4] D. J. Banks, S. B. Beres, and J. M. Musser. The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends in Microbiology*, 10(11):515–521, Nov. 2002.

[5] B. J. Barnhart, S. H. Cox, and J. H. Jett. Prophage induction and inactivation by UV light. *Journal of Virology*, 18(3):950–955, June 1976.

[6] L.-M. Bobay, M. Touchon, and E. P. C. Rocha. Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132, Aug. 2014.

[7] J. Bondy-Denomy and A. R. Davidson. When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. *J. Microbiol.*, 52(3):235–242, Mar. 2014.

[8] A. B. Brueggemann, C. L. Harrold, R. R. Javan, A. J. van Tonder, A. J. McDonnell, and B. A. Edwards. Pneumococcal prophages are diverse, but not without structure or history. *Scientific Reports*, 7:srep42976, Feb. 2017.

[9] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York, 3rd ed. edition, Dec. 2003.

[10] C. Canchaya, G. Fournous, and H. Brüssow. The impact of prophages on bacterial chromosomes. *Molecular Microbiology*, 53(1):9–18, July 2004.

[11] S. Casjens. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.*, 49(2):277–300, July 2003.

[12] S. R. Casjens and E. B. Gilcrease. Bacteriophages, methods and protocols, volume 2 molecular and applied aspects. *Methods in molecular biology*, 502:91–111, 2009.

[13] M. R. Clokie, A. D. Millard, A. V. Letarov, and S. Heaphy. Phages in nature. *Bacteriophage*, 1(1), 2011.

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

[15] J. S. Crispim, R. S. Dias, P. M. P. Vidigal, M. P. de Sousa, C. C. da Silva, M. F. San-

tana, and S. O. de Paula. Screening and characterization of prophages in *Desulfovibrio* genomes. *Scientific Reports*, 8(1):9273, June 2018.

[16] B. Danneels, M. Pinto-Carbó, and A. Carlier. Patterns of nucleotide deletion and insertion inferred from bacterial *Pseudogenes*. *Genome Biology and Evolution*, 10(7):1792–1802, July 2018.

[17] A. L. de Sousa, D. Maués, A. Lobato, E. F. Franco, K. Pinheiro, F. Araújo, Y. Pantoja, A. L. da Costa da Silva, J. Morais, and R. T. J. Ramos. PhageWeb web interface for rapid identification and characterization of prophages in bacterial genomes. *Frontiers in Genetics*, 9:644, Dec. 2018.

[18] S. M. Diene, A. R. Corvaglia, P. François, and N. van der Mee-Marquet. Prophages and adaptation of *Staphylococcus aureus* ST398 to the human clinic. *BMC Genomics*, 18:133, Feb. 2017.

[19] G. Edlin, L. Lin, and R. Bitner. Reproductive fitness of P1, P2, and Mu lysogens of *Escherichia coli*. *J. Virol.*, 21(2):560–564, Feb. 1977.

[20] R. Feiner, T. Argov, L. Rabinovich, N. Sigal, I. Borovok, and A. A. Herskovits. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Micro*, 13(10):641–650, Oct. 2015.

[21] L.-C. Fortier and O. Sekulovic. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, 4(5):354–365, July 2013.

[22] J. L. Fothergill, E. Mowat, M. J. Walshaw, M. J. Ledson, C. E. James, and C. Winstanley. Effect of antibiotic treatment on bacteriophage production by a cystic fibrosis epidemic

strain of *Pseudomonas aeruginosa*. *Antimicrobial Agents and Chemotherapy*, 55(1):426–428, Jan. 2011.

[23] A. Gaidelyte, S. T. Jaatinen, R. Daugelavicius, J. K. H. Bamford, and D. H. Bamford. The linear double-stranded DNA of phage Bam35 enters lysogenic host cells, but the late phage functions are suppressed. *Journal of Bacteriology*, 187(10):3521–3527, May 2005.

[24] J. Gödeke, K. Paul, J. Lassak, and K. M. Thormann. Phage-induced lysis enhances biofilm formation in *Shewanella oneidensis* MR-1. *ISME J*, 5(4):613–626, Apr. 2011.

[25] T. R. Gregory. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics*, 6(9):699–708, Sept. 2005.

[26] J. Haaber, J. J. Leisner, M. T. Cohn, A. Catalan-Moreno, J. B. Nielsen, H. Westh, J. R. Penadés, and H. Ingmer. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nature Communications*, 7:ncomms13333, Nov. 2016.

[27] J. Hacker and E. Carniel. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO reports*, 2(5):376–381, May 2001.

[28] D. R. Harper, H. M. R. T. Parracho, J. Walker, R. Sharp, G. Hughes, M. Werthén, S. Lehman, and S. Morales. Bacteriophages and biofilms. *Antibiotics (Basel)*, 3(3):270–284, June 2014.

[29] B. Hofer, M. Ruge, and B. Dreiseikelmann. The superinfection exclusion gene (sieA) of bacteriophage P22: identification and overexpression of the gene and localization of the gene product. *Journal of Bacteriology*, 177(11):3080–3086, June 1995.

[30] C. Howard-Varona, K. R. Hargreaves, S. T. Abedon, and M. B. Sullivan. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME journal*, 11(7):1511–1520, 2017.

[31] C. E. James, J. L. Fothergill, H. Kalwij, A. J. Hall, J. Cottell, M. A. Brockhurst, and C. Winstanley. Differential infection properties of three inducible prophages from an epidemic strain of *Pseudomonas aeruginosa*. *BMC Microbiology*, 12(1):216, 2012.

[32] H. S. Kang, K. McNair, D. Cuevas, B. Bailey, A. Segall, and R. A. Edwards. Prophage genomics reveals patterns in phage genome organization and replication. *bioRxiv*, page 114819, Mar. 2017.

[33] E. V. Koonin. Evolution of genome architecture. *The International Journal of Biochemistry & Cell Biology*, 41(2):298–306, Feb. 2009.

[34] M. Krupovic and P. Forterre. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes: Integration of ssDNA viruses into cellular genomes. *Annals of the New York Academy of Sciences*, 1341(1):41–53, Apr. 2015.

[35] C.-H. Kuo and H. Ochman. Deletional bias across the three domains of life. *Genome Biology and Evolution*, 1:145–152, Jan. 2009.

[36] J. G. Lawrence, R. W. Hendrix, and S. Casjens. Where are the pseudogenes in bacterial genomes? *Trends in Microbiology*, 9(11):535–540, Nov. 2001.

[37] R. Leplae, G. Lima-Mendez, and A. Toussaint. ACLAME: A CLAssification of mobile genetic elements, update 2010. *Nucleic Acids Research*, 38(suppl_1):D57–D61, Jan. 2010.

[38] G. Lima-Mendez, J. Van Helden, A. Toussaint, and R. Leplae. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, 24(6):863–865, Mar. 2008.

[39] E. Ljungquist and A. I. Bukhari. State of prophage Mu DNA upon induction. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8):3143–3147, 1977.

[40] E. López, A. Domenech, M.-J. Ferrándiz, M. J. Frias, C. Ardanuy, M. Ramirez, E. García, J. Liñares, and A. G. de la Campa. Induction of prophages by fluoroquinolones in *Streptococcus pneumoniae*: implications for emergence of resistance in genetically-related clones. *PloS One*, 9(4):e94358, 2014.

[41] R. C. Matos, N. Lapaque, L. Rigottier-Gois, L. Debarbieux, T. Meylheuc, B. Gonzalez-Zorn, F. Repoila, M. d. F. Lopes, and P. Serror. *Enterococcus faecalis* prophage dynamics and contributions to pathogenic traits. *PLoS Genetics*, 9(6):e1003539, June 2013.

[42] K. Meltz Steinberg and B. R. Levin. Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage. *Proceedings of the Royal Society B: Biological Sciences*, 274(1621):1921–1929, Aug. 2007.

[43] A. S. Millan, M. Toll-Riera, Q. Qi, and R. C. MacLean. Interactions between horizontally acquired genes create a fitness cost in *Pseudomonas aeruginosa*. *Nature Communications*, 6:6845, Apr. 2015.

[44] A. Mira, H. Ochman, and N. A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10):589–596, Oct. 2001.

[45] W. Mottawea, M.-O. Duceppe, A. A. Dupras, V. Usongo, J. Jeukens, L. Freschi, J.-G. Emond-Rheault, J. Hamel, I. Kukavica-Ibrulj, B. Boyle, A. Gill, E. Burnett, E. Franz, G. Arya, J. T. Weadge, S. Gruenheid, M. Wiedmann, H. Huang, F. Daigle, S. Moineau, S. Bekal, R. C. Levesque, L. D. Goodridge, and D. Ogunremi. *Salmonella enterica* prophage sequence profiles reflect genome diversity and can be used for high discrimination subtyping. *Frontiers in Microbiology*, 9, May 2018.

[46] J. H. Paul. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *The ISME Journal*, 2(6):579–589, June 2008.

[47] M. Ptashne. *A genetic switch: ghage lambda revisited*. CSHL Press, 2004.

[48] J. J. Ravantti, A. Gaidelyte, D. H. Bamford, and J. K. Bamford. Comparative analysis of bacterial viruses Bam35, infecting a gram-positive host, and PRD1, infecting gram-negative hosts, demonstrates a viral lineage. *Virology*, 313(2):401–414, Sept. 2003.

[49] D. Refardt and P. B. Rainey. Tuning a genetic switch: experimental evolution and natural variation of prophage induction. *Evolution*, 64(4):1086–1097, Apr. 2010.

[50] J. A. Shapiro. Molecular model for the transposition and replication of bacteriophage *Mu* and other transposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, 76(4):1933–1937, Apr. 1979.

[51] J. W. Shapiro and P. E. Turner. Evolution of mutualism from parasitism in experimental virus populations. *Evolution*, 72(3):707–712, Mar. 2018.

[52] W. Song, H.-X. Sun, C. Zhang, L. Cheng, Y. Peng, Z. Deng, D. Wang, Y. Wang, M. Hu, W. Liu, H. Yang, Y. Shen, J. Li, L. You, and M. Xiao. Prophage Hunter: an integrative

hunting tool for active prophages. *Nucleic Acids Research*, 47(W1):W74–W80, July 2019.

[53] A. Stern and R. Sorek. The phage-host arms-race: shaping the evolution of microbes. *Bioessays*, 33(1):43–51, Jan. 2011.

[54] A. J. Székely and M. Breitbart. Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiology Letters*, 363(6):fnw027, Mar. 2016.

[55] M. Touchon, A. Bernheim, and E. P. Rocha. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J*, 10(11):2744–2754, Nov. 2016.

[56] M. Touchon, J. A. Moura de Sousa, and E. P. Rocha. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Current Opinion in Microbiology*, 38:66–73, Aug. 2017.

[57] V. V. Volkova, Z. Lu, T. Besser, and Y. T. Gröhn. Modeling the infection dynamics of bacteriophages in enteric *Escherichia coli*: estimating the contribution of transduction to antimicrobial gene spread. *Applied and Environmental Microbiology*, 80(14):4350–4362, July 2014.

[58] P. L. Wagner and M. K. Waldor. Bacteriophage control of bacterial virulence. *Infection and Immunity*, 70(8):3985–3993, Aug. 2002.

[59] L. M. Wahl and T. Pattenden. Prophage provide a safe haven for adaptive exploration in temperate viruses. *Genetics*, 206(1):407–416, May 2017.

[60] X. Wang, Y. Kim, Q. Ma, S. H. Hong, K. Pokusaeva, J. M. Sturino, and T. K. Wood. Cryptic prophages help bacteria cope with adverse environments. *Nat Commun*, 1:147, Dec. 2010.

[61] M. G. Weinbauer. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.*, 28(2):127–181, May 2004.

[62] J. Weitz. *Quantitative viral ecology: dynamics of viruses and their microbial hosts*. Princeton University Press, Princeton, NJ, 2017.

[63] Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart. PHAST: a fast phage search tool. *Nucleic Acids Research*, 39(suppl):W347–W352, July 2011.

Chapter 4

The genetic repertoire of prophages

Bacterial genome sequencing has revealed that prophages – the functional or cryptic genome sequences of temperate bacteriophages – are far more numerous than previously recognized. Prophages are subject to mutational degradation, but they may also be maintained by selection if they confer benefits to their bacterial hosts; these evolutionary forces will have different effects on prophage genes of different function. In this chapter, we examine the distributions of 53,356 annotated prophage genes identified in 1384 prophage sequences, comparing the gene repertoires of intact and incomplete prophages. These data indicate that genes involved in the replication, packaging, and release of phage particles have been preferentially lost in incomplete prophages, while transposase and integrase genes are significantly enriched. In this chapter, we also developed mathematical and computational models to test how evolutionary forces affect prophage gene repertoires. These approaches demonstrate that genes involved in phage lytic function are preferentially lost, resulting in shorter prophages that often retain genes that benefit the host. Meanwhile, the model suggests that the enrichment of transposase sequences in shorter prophages is likely due to their role in disruption of phage lysis genes and generation of cryptic prophages that cannot harm their hosts. Overall, we show that variation in positive and negative selection on different prophage gene classes explains the diversity of

prophage genome structures, including the evolution and maintenance of cryptic and domesticated prophage sequences.

4.1 Introduction

Bacteriophages, viruses that infect bacteria, are the most prevalent life form on the planet, vastly outnumbering both their bacterial hosts and all other life forms combined [3, 26, 8]. As lethal pathogens, lytic bacteriophages typically reproduce in large numbers, causing the death of their hosts in the process. Temperate phages, however, are so-named because they also have the ability to integrate their genetic code into the host cell DNA, leaving the host cell unharmed. Once integrated, these viral sequences can persist as *prophages* for many bacterial generations, being replicated as part of the host cell genome during cellular fission.

While integrated in the bacterial genome, prophage sequences are subject to selection, mutation, and horizontal gene transfer (HGT). A wealth of recent evidence argues for the role of positive selection in the maintenance of prophages which confer benefits such as immunity against other infecting phages, antibiotic resistance, resistance to environmental stress and numerous virulence factors [17]. Mutation in bacterial genomes is biased toward deletion [19, 22, 12], and thus prophage sequences are subject to mutational degradation over long time scales. In addition, some families of prophages carry transposase genes, enabling replicative (copy-and-paste) transposition of the prophage sequence to other locations in the bacterial genome.

If a prophage retains the functional genes required for replication, packaging and cellular lysis, the prophage sequence can initiate *induction*, a process in which the prophage resumes its role as a lethal pathogen, produces a large number of daughter phage, and kills the host.

Spontaneous induction rates are in the range of 10^{-5} to 10^{-3} per bacterial generation [21, 31], but can be substantially increased if the bacterial host cell is in stress [1].

Genome sequencing and comparative genomics have recently revealed that prophages are far more numerous and more widely-shared across bacterial genomes than previously recognized [10, 23]. In addition, four recent studies have independently reported that the distribution of prophage lengths is bimodal [4, 5, 11, 20], a phenomenon that may be explained by the balance between selection for prophage maintenance (via beneficial effects of prophage genes) and selection against prophage (via harmful effects of induction and cell lysis) [18]. These fundamental evolutionary forces will differentially affect prophage genes of different function. For example, while short deletions might affect all prophage genes, positive selection will affect only those prophages that carry genes of benefit to their host; negative selection (via induction) will affect only prophages that carry functional induction genes. Thus, prophages of differing lengths, or differing degrees of integrity when compared to the ancestor phage genome, may carry different genetic repertoires – signatures of the evolutionary forces in play. Indeed, tail fiber and integrase coding sequences are significantly enriched in small prophages (Figure S1, [4]), but little else is understood about the gene repertoires of intact or degraded (incomplete) prophage sequences.

In this contribution, we examine the distributions of prophage genes identified in publicly available genome sequences, comparing the gene repertoires of intact and incomplete prophages. To better understand these results, we also develop both a mathematical and computational model describing the fates of distinct gene classes in prophages. Our results support the roles of both positive and negative selection in maintaining prophage sequences with diverse genetic repertoires, and offer explanations for both the enrichment and loss of specific

gene functions in cryptic prophages.

4.2 Gene repertoire of sequenced prophages

We investigated bacterial genomes studied in two previously published data sets [4, 20], using the PHASTER interface [2] for rapid prophage identification and gene annotation. Data Set 1, originally studied by Bobay et al. [4], includes 624 prophages from 85 bacterial genomes; these prophage sequences contain 24,877 annotated genes. Data Set 2, as studied by Leplae et al. [20], includes 760 prophages from 306 bacterial genomes, with 28,479 annotated genes. For the 13 phage gene functions listed in Table 4.1, we tracked the number of prophages identified as containing at least one gene of that class. We further partitioned these data based on whether the prophage sequence was classified as “intact", “questionable" or “incomplete" by the PHASTER algorithm.

Figure 4.1 plots the frequency of each gene class in intact (f_{int}), questionable and incomplete (f_{inc}) prophages. The genes are ordered left to right according to their degree of enrichment in incomplete prophages; due to small numbers, gene types that constituted less than 1% of the data have been excluded. These results are summarized in the lower panels of Figure 4.1, which show the percent change in gene frequency between incomplete and intact prophages, that is:

$$\%\ \text{change} = \frac{100(f_{\text{inc}} - f_{\text{int}})}{f_{\text{int}}} \ .$$

Positive values of % change thus indicate genes that are relatively enriched in incomplete prophages, while negative values indicate genes that are preferentially lost.

To evaluate the statistical significance of these results, for each gene type we use the same number of identified genes (e.g. 317+25+14 = 356 for terminase in Data Set 1), and ran-

Gene	Number of prophages containing a gene of this type					
	Count in Data Set 1			Count in Data Set 2		
	Intact	**Questionable**	**Incomplete**	**Intact**	**Questionable**	**Incomplete**
terminase	317	25	14	292	53	58
portal	277	9	3	283	67	48
head	299	16	25	281	79	86
injection	14	0	2	4	0	0
tail	413	46	86	419	116	141
protease	82	3	5	72	12	22
transposase	195	32	75	190	173	144
integrase	346	54	85	312	94	165
lysis	226	19	11	52	6	6
plate	121	5	0	143	20	28
capsid	225	14	18	233	50	40
lysin	235	17	19	165	32	22
flippase	2	0	20	0	1	4
Total	2752	240	363	2446	603	764

Table 4.1: The genetic repertoire of prophages in Data Set 1 and Data Set 2.

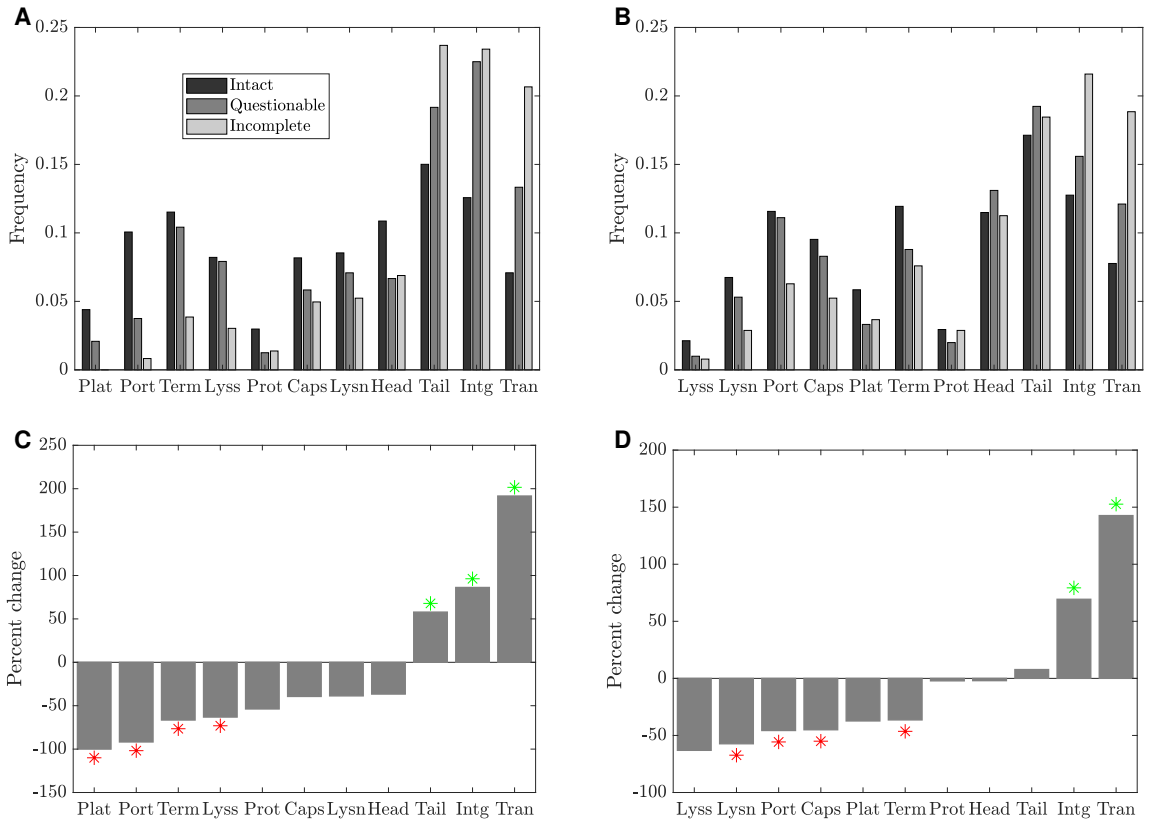


Figure 4.1: Changes in prophage gene frequencies, for intact, questionable and incomplete prophages. (A) The frequency of each gene class identified in Table 4.1 in prophages from Data Set 1 [4], for prophages identified as intact, questionable or incomplete. Gene classes that constituted less than 1% of the data have been excluded. Gene classes are ordered by the percent change in frequency (degree to which they are enriched in incomplete prophages, see panel C.) (B) Gene frequencies as in panel A, but for Data Set 2 [20]. (C) Percent change in gene frequency; the frequency of each gene class in incomplete prophages is compared to the baseline frequency of that class in intact prophages. Frequencies that were significantly lower (red) or higher (green) than expected by chance with a two-sided 5% significance threshold are indicated by stars. Thus red stars indicate gene classes that are preferentially lost, while green stars indicate classes that are enriched in short prophages. (D) Percent change in gene frequency for Data Set 2.

	Number of transposase genes identified					
	Count in Data Set 1			**Count in Data Set 2**		
	Intact	**Questionable**	**Incomplete**	**Intact**	**Questionable**	**Incomplete**
IS transposase	174	34	76	459	90	109
non-IS transposase	278	37	88	464	101	99
All phage proteins	21054	2271	1552	19250	5097	4132

Table 4.2: The distribution of transposase genes identified in Data Set 1 and Data Set 2.

domly assign the genes to one of the three prophage classes. Because intact prophages in the dataset contain more genes than incomplete prophages, we also preserve the proportion of genes assigned to each class. Thus in randomly assigning genes to prophage classes, we assign 2752/(2752+240+363) = 82% of genes to intact prophages in Data Set 1, for example, while only 363/(2752+240+363) = 11% are assigned to incomplete prophages.

We computed the percent change in gene frequency for these bootstrapped data as described above, and repeated this procedure 10,000 times. Stars in the lower panels of Figure 4.1 indicate % change values in the data that were lower than the 2.5 percentile or higher than the 97.5 percentile in the bootstrapped distributions.

These results reveal several features that are conserved between data sets. We note that lysis or lysin genes, as well as portal and terminase proteins, are preferentially lost in incomplete prophages. In contrast, transposase and integrase genes are substantially enriched. We explore these results further in both the computational and mathematical models described in the sections to follow.

In light of the striking enrichment of transposase genes in incomplete prophages, we ex-

amined the transposase annotations in greater detail. For each prophage in the dataset, we downloaded the coding sequences and the BLAST hits identified for each coding sequence by PHASTER [2]. We counted the number of coding sequences with a BLAST hit annotated as an insertion sequence (IS) transposase (e.g. "IS3 transposase B"), as well as those annotated as a transposase but without a BLAST hit to an IS. As a control, we also counted the total number of proteins identified as a "phage hit protein" by PHASTER in each data set.

As shown in Table 4.2, IS transposases account for 41.4% of the transposase sequences identified in Data Set 1, and 49.8% of those in Data Set 2. In Data Set 1, the frequency of IS transposases (calculated as the fraction of all phage proteins identified) is enriched 4.9-fold in incomplete prophages as compared to intact prophages; the frequency of non-IS transposases also increased but to a lesser degree (3.3-fold). In Data Set 2, the frequency of IS transposases increased by 10% in incomplete prophages, while non-IS transposases were reduced by 0.6%. Thus in both data sets, the frequency of IS transposases is enriched in incomplete prophages, and enriched to a greater degree than non-IS transposases. As discussed further below, this suggests that the enrichment of transposase sequences in incomplete prophages may be due to the disruption of essential prophage functions due to IS insertion; in other words, the presence of the IS transposase has rendered the prophage cryptic.

4.3 Analytical model of prophage gene content

To investigate the preferential loss or maintenance of specific classes of phage genes in prophage sequences over time, we developed a mathematical model. The model, although simplified, allows us to predict the effects of key parameters on the longterm genetic repertoire of prophages.

The model tracks a population of bacterial genomes, which contain prophages with genes

in three possible types – beneficial genes, excision genes and re-infection genes. Beneficial genes are genes that confer a selective advantage to the host, thus increasing the prophage population through vertical transmission. Biological examples of beneficial genes include host virulence factors that help the bacterial cell during colonization of its host (for example phage lambda's *lom* gene). In contrast, excision genes are the genes involved in prophage induction into the lytic cycle, which leads to the death of the host cell. Examples of excision genes include lambda's *O* and *P* genes, which switch on the lytic cycle by commandeering the host's DNA polymerase. Phage induction will typically lead to bacterial cell death regardless of the quantity or quality of phage progeny. Phage progeny success is determined by the phage's re-infection genes, comprising the genes required for phage genome replication, packaging, lysis, transmission to a new host, and reestablishment of lysogeny. This re-infection class, in particular, includes a large number of genes of different function; yet their net effect, taken together, is to increase the prophage population through horizontal transmission.

In the simplified model, we consider a full prophage as one containing just three 'genes', one of each class. Here, we can think of a 'gene' as a full functional complement of the underlying sequences required for each function. We denote the frequency of full prophages in the population at time t as $P_{111}(t)$. More generally, we use the notation P_{ber} to represent the frequency of prophages with (1) or without (0) the beneficial, excision or re-infection genes respectively. For completeness, the model also includes a population P_{000} corresponding to bacterial genomes in which the prophage has been completely lost. Note that in the computational model to follow, we will both expand these gene classes and include multiple genes per class.

The analytical model includes the following processes:

Degradation: Each gene in each prophage in the population is lost (gene deletion) at rate r_D. For example, the frequency of P_{111} is lost at overall rate $3r_D$, contributing at rate r_D to each of the populations P_{011}, P_{101} and P_{110}.

Induction: If a bacterial genome contains a prophage which carries the excision gene, the prophage induces at rate r_I and the bacterium is lost from the population.

Re-infection: Prophages that carry both the excision and re-infection genes (P_{111} and P_{011}) reproduce (create copies of themselves in new bacterial genomes), through lysis, re-infection and lysogeny, at rate r_L.

Selection: To model the potential selective benefit conferred by the prophage, we assume that bacterial populations that carry the beneficial prophage gene grow at per capita rate r_S.

These assumptions yield the following system of ordinary differential equations, illustrated as a schematic in Figure 4.2:

$$\begin{aligned}
\frac{dP_{111}}{dt} &= (r_L + r_S - 3r_D - r_I)P_{111} - \phi P_{111} \\
\frac{dP_{011}}{dt} &= (r_L - 2r_D - r_I)P_{011} + r_D P_{111} - \phi P_{011} \\
\frac{dP_{101}}{dt} &= (r_S - 2r_D)P_{101} + r_D P_{111} - \phi P_{101} \\
\frac{dP_{110}}{dt} &= (r_S - 2r_D - r_I)P_{110} + r_D P_{111} - \phi P_{110} \\
\frac{dP_{001}}{dt} &= (-r_D)P_{001} + r_D P_{011} + r_D P_{101} - \phi P_{001} \\
\frac{dP_{010}}{dt} &= (-r_D - r_I)P_{010} + r_D P_{011} + r_D P_{110} - \phi P_{010} \\
\frac{dP_{100}}{dt} &= (r_S - r_D)P_{100} + r_D P_{101} + r_D P_{110} - \phi P_{100} \\
\frac{dP_{000}}{dt} &= r_D(P_{100} + P_{010} + P_{001}) - \phi P_{000} \ . \qquad (4.1)
\end{aligned}$$

Here, terms involving ϕ simply ensure that the frequencies sum to unity at all times, with ϕ defined as:

$$\phi = (r_L + r_S - r_I)P_{111} + (r_L - r_I)P_{011} + r_S P_{101} + (r_S - r_I)P_{110} - r_I P_{010} + r_S P_{100} \, .$$

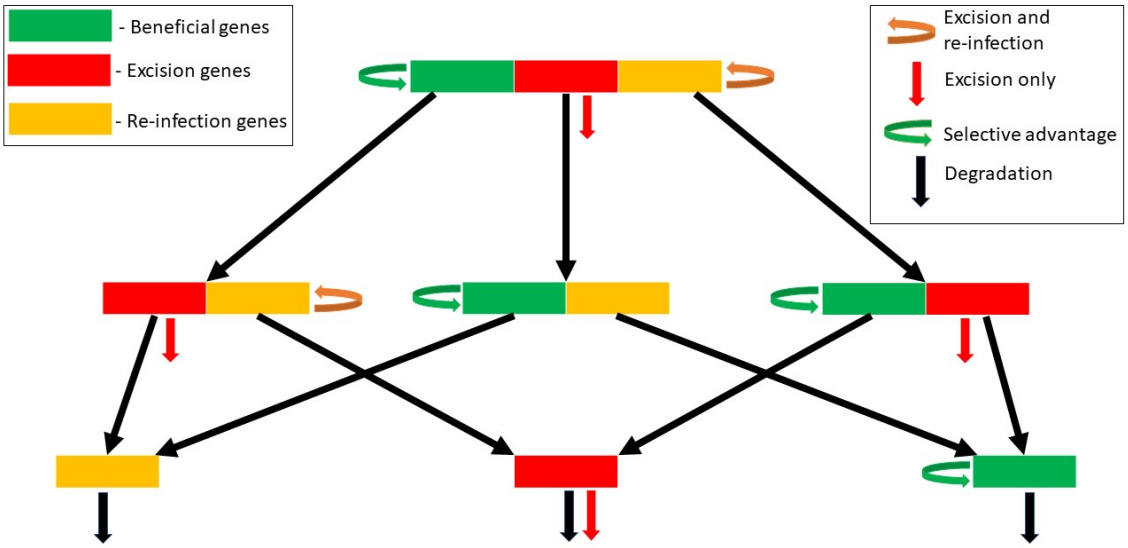


Figure 4.2: Schematic diagram of the mathematical model.

A detailed analysis of the equilibria and stability of this 8-dimensional model is provided in the Supplementary Material. From this analysis, we find four possible longterm outcomes: (1) equilibrium E_0, in which all prophage genes are lost; (2) equilibrium E_B, in which both excision and re-infection genes are lost, but beneficial genes persist. This equilibrium reflects complete domestication of the prophage; (3) equilibrium E_{ER}, in which beneficial genes are lost but both excision and re-infection genes persist. This corresponds to a virulent prophage that does not contribute to host fitness; (4) equilibrium E_A, in which all three types of genes persist.

As described in the Supplementary Material, two critical conditions are sufficient to deter-

Longterm prediction Condition	B genes do not persist $r_S < r_D$	B genes persist $r_S > r_D$
ER genes do not persist $r_L < 2r_D + r_I$	**extinction** E_0	**domestication** E_B
ER genes persist $r_L > 2r_D + r_I$	**virulence** E_{ER}	**persistence** E_A

Table 4.3: Conditions determining which classes of prophage genes persist longterm.

mine the long-term behaviour of the prophage gene distribution.

Condition 1: $r_S > r_D$. Note that r_S is the rate at which a beneficial gene produces a new copy of itself, while r_D is the rate at which a beneficial gene is lost, which occurs through mutational degradation. Thus $r_S > r_D$ implies that on average, a beneficial gene makes more than one copy of itself before it is lost: beneficial genes persist.

Condition 2: $r_L > 2r_D + r_I$. Similarly, the combination of an excision and a re-infection gene, co-occuring on a prophage, is able to produce a new copy of itself at rate r_L. These genes may be lost through induction, but also lost if either gene is degraded by mutation, so the total rate of loss is $2r_D + r_I$. Thus this gene combination can persist if $r_L > 2r_D + r_I$.

The predicted behaviour of the mathematical model can therefore be summarized as shown in Table 4.3.

In Figure 4.3, we illustrate the approach of System 4.1 to each of these four equilibrium states, for appropriate parameter values. To simplify the presentation, we plot the the average number of genes of each type carried per prophage, where for example the average number of beneficial genes per prophage is given by $P_{111} + P_{110} + P_{101} + P_{100}$. Equivalently, this is the fraction of prophages that carry the beneficial gene.

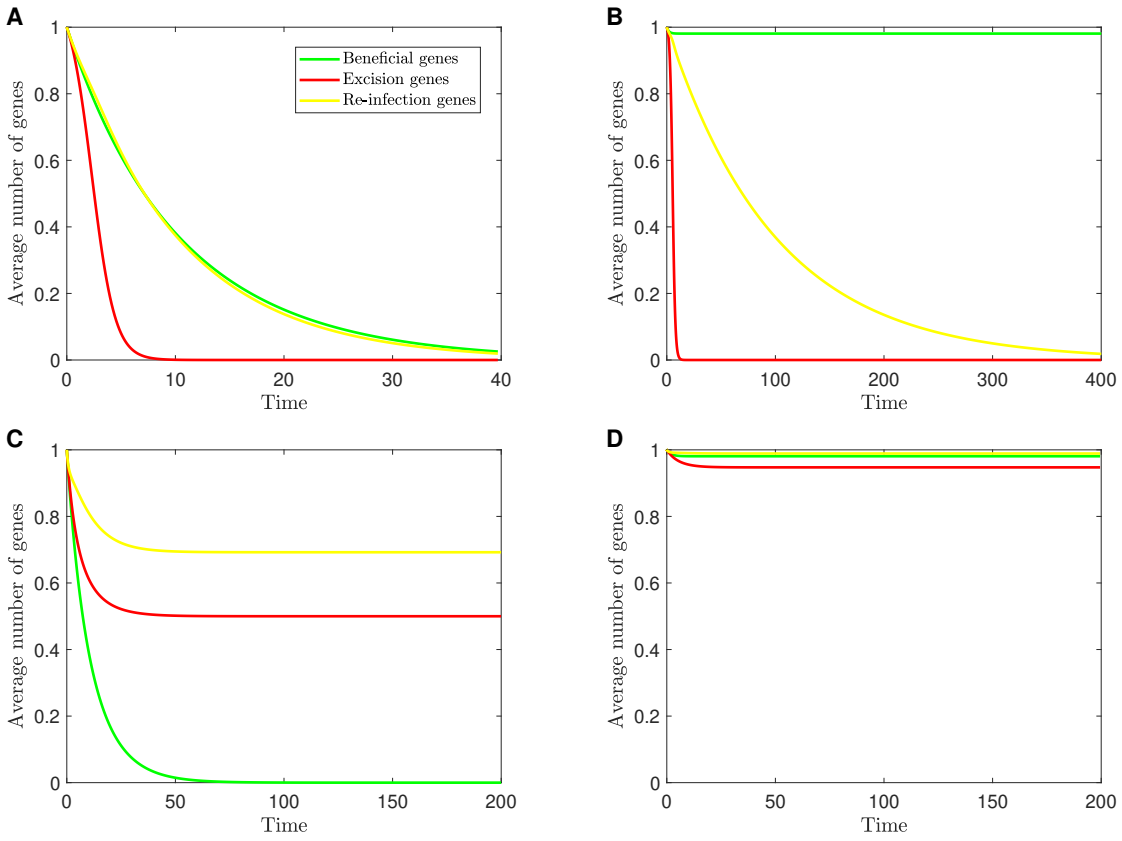


Figure 4.3: Numerical integration of the analytical model, showing System 4.1 converging toward four possible equilibria: (A) **Extinction**, $E_0(r_S = 0.01, r_D = 0.1, r_L = 0.2)$; (B) **Domestication**, E_B $(r_S = 0.52, r_D = 0.01, r_L = 0.2)$; (C) **Virulence**, E_{ER} $(r_S = 0.02, r_D = 0.1, r_L = 1.3)$; (D) **Persistence**, E_A $(r_S = 0.52, r_D = 0.01, r_L = 1.2)$. In all cases, $r_I = 1$.

4.4 Gene Repertoire Simulations

We also developed a computational model which is able to describe the gene content of prophage sequences in greater detail. Here, we assume that prophages exist in a population of bacterial genomes that are linked by both cellular reproduction (vertical transmission of the prophage) and horizontal gene transfer (horizontal transmission). We can vary the initial conditions such

that all or only some of the bacterial genomes initially carry prophages. Bacterial genomes that carry inducible prophage sequences may be lost through induction and lysis, whereas bacterial genomes that carry beneficial prophage sequences may be preferentially copied to the next generation. We describe these steps in detail below.

Each prophage sequence may contain genes of the following four types: excision genes, re-infection genes, beneficial genes and neutral genes. A full prophage carries n_E excision genes, n_R re-infection genes, n_B beneficial and n_N neutral genes. We include mutational degradation and also include the possibility that an insertion sequence (or other transposable element) could disrupt the prophage genome.

We track the presence or absence of each gene in each prophage sequence. A discrete timestep in the simulation corresponds to a bacterial generation time. The rates of the underlying processes, however, are expressed in units of the "prophage generation time", that is, the average time that a single prophage sequence is maintained in a bacterial genome before induction [18]. Since the bacterial generation time, Δt, is much shorter than the prophage generation time, if a process occurs at rate r per prophage generation, the probability that it occurs in timestep Δt is small and well-approximated by $r\Delta t$.

The following processes are included in the model, with parameters as described in Table 4.4:

Degradation: In each time step, each gene in each prophage in the population is removed (gene deletion) with probability $r_D\Delta t$.

Induction: If a prophage carries all n_E excision genes, it may induce with probability $r_I\Delta t$. When a prophage induces, it is removed from the population. We thus assume that all n_E excision genes are required for excision and death of the host cell. Note that we ignore

polylysogeny, that is, we make the simplifying assumption that excision and cell death affect only the excising prophage.

Re-Infection: To simulate the process of lysis followed by re-infection and lysogeny, a copy of any prophages that carry all n_E excision genes *and* all n_R re-infection genes may be added to the prophage pool with probability $r_L \Delta t$. Thus, full complements of both the excision and re-infection genes are required to reinfect. In addition, in some simulations new (full length) prophages are added to the prophage pool with probability $r_F \Delta t$. This might occur for example if there is an influx of prophages to the local population from an external pool.

Selection: Copies of existing prophages are also added to the population at rate $r_S \Delta t \, n_b / n_B$. Here n_b is the number of beneficial genes carried by the prophage, and r_S is the maximum selective benefit provided to the host cell if the prophage contains all n_B beneficial genes.

Population regulation: We consider a pool of prophages that exists within a bacterial population that cannot grow unbounded. To regulate the population size, if the current population size N, is greater than the bacterial carrying capacity, K, each bacterial genome is copied into the subsequent generation with probability K/N.

While all of our simulation studies include the processes described above, in some simulations we also explored the impact of disruption by transposable elements (TEs, such as bacterial insertion sequences) as follows.

TE disruption: Motivated by the observed frequencies of IS transposase sequences in incomplete prophages, we include the possibility of TE disruptions in prophage genes. For each gene in each timestep, a TE disruption may occur with probability $r_T \Delta t$. When this occurs, we assume that gene function has been disrupted: if a beneficial gene has been disrupted, the gene confers no benefit to the host thereafter; if a gene required for excision or re-infection

is disrupted, the prophage is no longer able to kill the host or re-infect respectively. Thus TE disruptions have the same effect as gene deletions, but leave a signature of TE sequences in the prophage genome.

We wondered whether it was reasonable to include TE disruptions in the model, since their rates might be negligible relative to mutational degradation. Rates of base pair substitutions in *E. coli* K12 have been estimated to be on the order of 2×10^{-10} per nucleotide, per generation [14]. Multiplying by 1.2 kb per prophage gene [18] yields an estimate of 2.4×10^{-7} base pair substitutions, per prophage gene, per bacterial generation. Presumably only a fraction of base pair substitutions result in loss of function. In addition, small indels are estimated to occur at about one tenth of this rate [14]. Thus taking in *E. coli* as a model organism, rates of prophage gene degradation through mutation (base pair substitution and short indels) might occur on the order of 10^{-7} or 10^{-8} per gene per generation.

In comparison, rates of transposition, for 5 insertion sequences in *E. coli* K12, have been estimated to be about 1×10^{-5} per element per generation [28]; this includes both copy-and-paste and cut-and-paste transpositions. The ancestral genome in this mutation accumulation study carried a total of 33 copies of these ISs, yielding an overall transposition rate of 3.3×10^{-4} transpositions per generation. Given that a typical prophage gene comprises 1.2 kb [18] of a 4.6 Mb *E. coli* genome, we arrive at an estimated transposition rate of 8.6×10^{-8} per prophage gene, per bacterial generation, similar to our estimate for gene loss through mutational degradation.

Parameter	Description
n_B	number of beneficial genes
n_E	number of genes necessary for excision
n_R	number of genes necessary for re-infection
n_N	number of neutral genes
r_D	rate of loss through mutational degradation
r_I	rate of loss through induction, excision and host death
r_L	rate of increase through lysis, reinfection and lysogeny
r_T	rate of loss through TE disruption
r_S	selective advantage to host cell if prophage carries all beneficial genes

Table 4.4: Parameters of the computational model.

4.4.1 Gene content of active temperate phage

We used phage lambda's genome architecture as a model for the number of excision, beneficial, and reinfection genes in a temperate phage genome (see Figure 1 in [25]). Lambda has long been a model system for the study of lytic-lysogeny cycles, phage genome arrangement, and phage evolution [6].

Excision genes: Lambda's excision genes are those that switch phage gene expression to the lytic cycle. Corresponding to the early right operon (6.5 kb), these excision genes make up approximately 13.3 percent of the phage genome.

Beneficial genes: Lambda carries several genes thought to confer benefit to the bacterial host during lysogeny. These include *cI*, *rexA*, *rexB*, *sieB*, *lom*, and *bor*, comprising 3.7kb total, about 7.6 percent of the genome.

Reinfection genes: The rest of the lambda genome contains genes that allow a phage to form viable progeny capable of reinfecting other cells: phage particle production, packaging, lysis, and lysogeny. These genes include about 38.4 kb, 79 percent of the genome. Most of these genes are contained in the late Operon (27 kb, phage particle production) and the early left operon (13 kb, lysogeny). The host-beneficial genes encoded in those operons (*sieB*, *lom*, *bor*, 1.6 kb total) are included instead in the beneficial genes category discussed above.

We note that not all lambda genes have been fully characterized. For example, lom and bor are thought to be host-beneficial during lysogeny, but more work is needed to establish the host fitness components. We also note that not all excision and reinfection phage genes are likely essential.

Taken together, these gene frequencies motivated the choice to model a full prophage

genome in the ratio 1:1:8 for benefical:excision:re-infection genes. In addition to these genes, in some simulations we also included neutral genes as a control.

4.4.2 Computational Model Results

Figure 4.4 illustrates that like the analytical model, the long-term behaviour of the simulation predicts four possible outcomes for the prophage: extinction, domestication, virulence (loss of genes that benefit the host but retention of genes necessary for virulent function) or persistence (of all gene types). Although omitted for brevity, it is straightforward to derive conditions similar to those provided in Table 4.3 which predict the loss or retention of each gene type. For example with 1 excision gene and 8 re-infection genes, the 'ER' function can be lost by a mutation in any of these 9 genes, so the overall rate of loss is $9r_D + r_I$, while the rate of gain is r_L.

We further examined the qualitative features of the prophage population at the persistence equilibrium. To do this, we simulated the prophage population with parameter values as described in panel D of Figure 4.4 for 200 generations, and then compared the gene content of prophages of different lengths. We define all prophages as either "intact" or "incomplete": an intact prophage carries all the genes necessary for excision and re-infection, whereas if any of these genes is missing, the prophage is incomplete. Figure 4.5A shows the frequency of each type of gene in intact and incomplete prophages; the percent change in incomplete prophages, as compared to the baseline of an intact prophage, is shown in panel B. We find that genes involved in excision and re-infection are preferentially lost in incomplete prophages.

These results are clarified in Figure 4.5C, which shows a histogram of prophage lengths

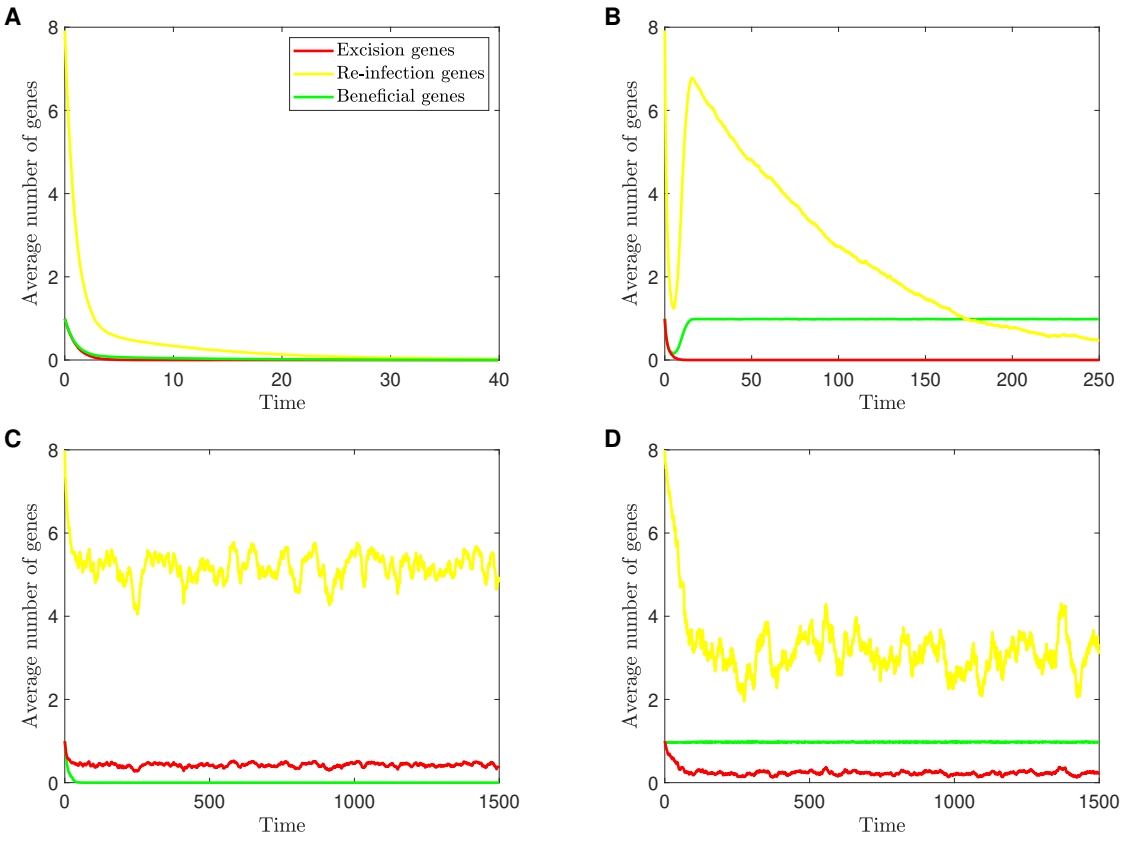


Figure 4.4: Simulations results showing the approach to four possible long-term outcomes: (A) **Extinction** ($r_S = 0.01$, $r_D = 0.1$, $r_L = 0.2$); (B) **Domestication** ($r_S = 0.52$, $r_D = 0.01$, $r_L = 0.2$); (C) **Virulence** ($r_S = 0.02$, $r_D = 0.1$, $r_L = 2.0$); (D) **Persistence** ($r_S = 1.5$, $r_D = 0.05$, $r_L = 1.5$). In all cases, $r_I = 1$, $r_T = 0$, $n_B = 1$, $n_E = 1$ and $n_R = 8$. The average number of genes of each type, per prophage, is plotted against time.

(grey bars), along with the gene frequency for each gene type, for prophages of each length. Thus for example full prophages have 10 genes and have 80% re-infection genes, 10% excision genes and 10% beneficial genes. We see a bimodal distribution of prophage lengths, with the smallest prophages becoming domesticated, that is, retaining only the gene that benefits the host.

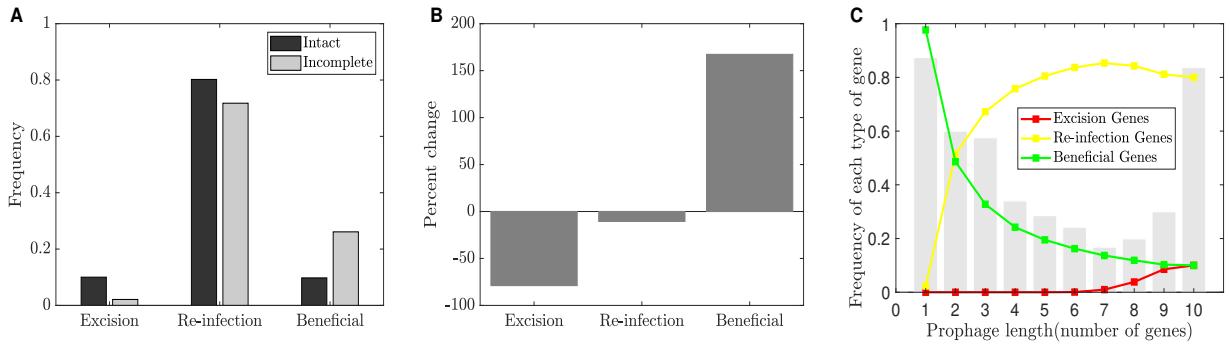


Figure 4.5: Gene frequencies in intact and incomplete prophages. (A) Frequency of genes of each type in intact and incomplete prophages, for the computational model simulated at the persistence equilibrium (see text for details); (B) Percent change in gene frequency from intact to incomplete; (C) A histogram of prophage lengths (grey bars), as well as the frequency of gene classes at each length. We find a bimodal distribution of prophage sizes, with smaller prophages losing the excision and re-infection genes but retaining the beneficial gene.

Fig. 4.6 illustrates the effect of adding transposable element disruptions to the computational model. In panel A, despite TE disruptions, the prophage population persists and retains all genes. Here we have also added a single neutral gene, which has no effect on fitness, for comparison (grey line). Panel D shows the average number of TE disruptions sustained in each type of gene; TEs accumulate in neutral genes but their presence in functional genes is minimized by purifying selection. Panels B through F show similar results, except that the rate of TE disruption, r_T, and the selective advantage, r_S, are altered. Increasing the transposition rate has the same qualitative effect as increasing the mutation rate, r_D, in Table 4.3; the long-term outcome can change from persistence (panel A) to either virulence (panel B) or domestication, depending on the value of r_S, and then ultimately to extinction (panel C) as r_T increases.

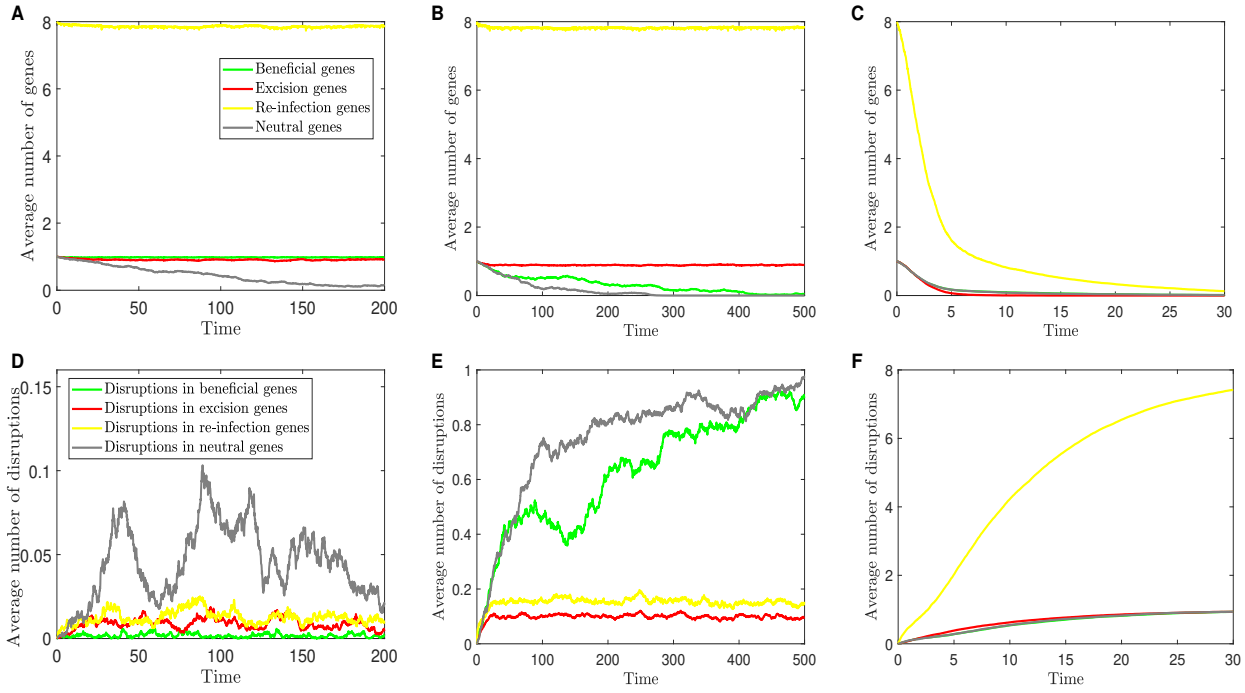


Figure 4.6: The effect of TE disruptions on the long-term outcome for prophage sequences. In panels A through C, the average number of genes of each type per prophage is plotted against time. As the transposition rate is increased, the long-term prediction for the prophage changes from persistence (panel A) to virulence (panel B) and finally to extinction (panel C). Panels D through F show the average number of TE disruptions sustained in genes of each type versus time. TEs accumulate in neutral genes but are limited in functional genes due to purifying selection. Parameter values are: (A and D) $r_S = 0.52$, $r_T = 0.009$; (B and E) $r_S = 0.01$, $r_T = 0.01$; (C and F) $r_S = 0.002$, $r_T = 0.1$. In all cases, $r_I = 1$, $r_L = 1.2$, $r_D = 0.001$, $n_B = 1$, $n_E = 1$, $n_R = 8$ and $n_N = 1$.

Again, we simulated the prophage population with parameter values as described in panel D of Figure 4.4 for 5000 generations, but including TE disruptions ($r_T = 0.002$), comparing the gene content of intact and incomplete prophages. Using the strict definition of “intact" or

“incomplete" described above, transposase genes were enriched nearly 400-fold in incomplete prophages (see Supplementary Material).

This result may be artificially inflated by the fact that only the single beneficial gene can sustain a TE disruption in an intact prophage in our simulations. In reality, algorithms such as PHASTER are not able to classify prophages as intact based on the certainty that they contain a full complement of functional phage genes. Instead, approximate metrics are used, based for example on the number of identified phage genes in close proximity in the sequence [2]. For a better comparison with the data shown in Figure 4.1, we therefore classified prophages as “intact" if they contained 80% of more of the possible prophage genes; prophages with less than 80% were classified as incomplete.

Figure 4.7A shows the frequency of each type of gene in intact and incomplete prophages classified in this way; the percent change in incomplete prophages, as compared to the baseline of an intact prophage, is shown in panel B. Again we see that genes involved in excision and re-infection are preferentially lost, beneficial genes are preferentially maintained, and transposase genes are substantially enriched in shorter prophages.

Figure 4.7C shows a histogram of prophage lengths (grey bars), along with the gene frequency for each gene type, for prophages of each length. A bimodal distribution of prophage lengths is again demonstrated, with the smallest prophages becoming domesticated, that is, retaining only the gene that benefits the host. We note that transposase genes accumulate in prophages of intermediate length, but are absent from the smallest prophages.

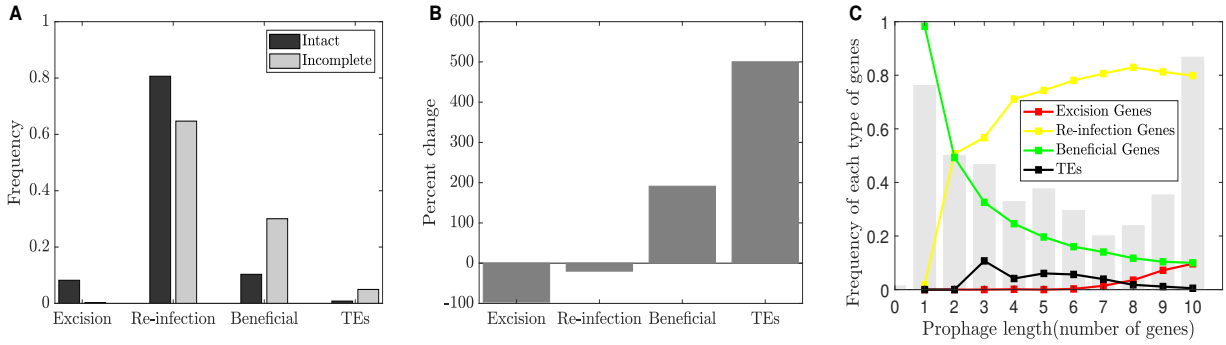


Figure 4.7: Gene frequencies in intact and incomplete prophages, when TEs are included (r_S = 1.5, r_L = 1.5, r_D = 0.048, r_T = 0.002). (A) Frequency of genes of each type in intact and incomplete prophages, for the computational model simulated at the persistence equilibrium with TE disruptions (see text for details); (B) Percent change in gene frequency from intact to incomplete; (C) A histogram of prophage lengths (grey bars), as well as the frequency of gene classes at each length. We find a bimodal distribution of prophage sizes, with TEs accumulating in prophages of intermediate lengths.

4.5 Summary and Discussion

We bring three lines of evidence to bear on the diverse genetic repertoire of active and cryptic prophages. First, we examine over 50,000 gene annotations from sequenced prophages to demonstrate that genes involved in lytic function – structural genes such as plate, capsid and portal genes, as well as lysis, lysin and terminase genes – are preferentially lost in incomplete (presumably cryptic) prophages. In constrast, three gene classes are enriched: tail fiber, integrase and transposase genes (Fig. 1c, d).

Secondly, a simplified mathematical model predicts that depending on the balance among dynamic processes such as the rates of lysis and infection, selection and mutational degra-

dation, four longterm outcomes are predicted for prophage sequences: the maintenance of an active prophage that also carries host-beneficial genes, the maintenance of an active but virulent prophage, domestication, or extinction (Fig. 3 and Table 3).

Thirdly, a more complex computational approach examines the genetic repertoires of prophages of differing lengths. The computational model predicts a bimodal distribution of prophage lengths, as observed in a number of recent studies [4, 20, 5, 11], and consistent with our recent predictions regarding the interplay of selection and mutation on the prophage length distribution [18]. The computational model also demonstrates that genes involved in excision and re-infection are preferentially lost in shorter prophages (Fig. 5, 7), consistent with the loss of lytic-cycle specific genes observed our bioinformatic analyis (Fig. 1). This result is intuitively appealing since selection at the level of the host favors loss of intact lytic-cycle alleles, which contribute to greater rates of host lysis and death.

As summarized above, our bioinformatic analysis supported a previous finding [4] that some genes are significantly enriched in shorter prophages; in our data these enriched genes included transposases, phage tail protein-encoding genes, and integrases. Consistent with these data, transposases preferentially accumulated in shorter prophages in our simulation studies, which assumed that transposable elements disrupted gene function and left an identifiable transposase gene sequence as a signature.

Tail fibre genes, in contrast, would be classified as re-infection genes, and thus were not predicted to be enriched in cryptic prophages in our simulations. Bobay et al. [4] hypothesize that tail genes may be domesticated by bacterial hosts through the longer-term processes of co-option and evolution of novel function, for example the evolution of bacterial tailocin toxins from phage tail ancestors. For domestication of tail genes in this way, such processes

would presumably require a specific combination of multiple accumulated mutations and the appropriate selective environment for a novel function to emerge. These conditions were well beyond the scope of our models here, but modeling the additional complexity of domestication via the accumulation of *de novo* adaptive mutations is an interesting idea for future work.

The enrichment of integrase genes in short prophages was an interesting and unexpected result of our bioinformatic analysis. Prophages typically possess integrase genes, facilitating chromosomal integration, and excision in conjugation with excisionase [16, 7, 15, 24]. Prophage integrase genes have been used as recognized diagnostic markers for prophages within bacterial genomes [30, 29], as markers to identify temperate phage genomes [27], and as signature genes to measure prophage diversity, and hence, host genome diversity [9]. Based on these facts we were expecting that integrase genes may be missing in incomplete prophages (as suggested for example in [9]), yet the opposite trend was observed.

We suggest that integrase genes may be maintained through the evolution of phage-like genetic selfish elements, such as satellite phages and molecular parasites that don't require the full complement of phage lytic cycle genes but benefit from horizontal transfer among hosts. Integrase genes are commonly found on mobile genetic elements [13]; when foreign DNA enters a host cell through horizontal transfer, integrase acts as a catalyst to mediate the process of recombination, thus integrases may facilitate the horizontal transfer of prophage-derived elements. The horizontal transmission of integrase genes along with their neighboring prophage genes would then result in significant enrichment of integrase genes in incomplete prophages. Although our computational model has not yet incorporated horizontal transmission of full or partial prophages, this hypothesis could be explored more fully in an expanded model that includes horizontal gene transfer.

Bibliography

[1] S. Alexeeva, J. A. Guerra Martínez, M. Spus, and E. J. Smid. Spontaneously induced prophages are abundant in a naturally evolved bacterial starter culture and deliver competitive advantage to the host. *BMC Microbiology*, 18(1):120, Dec. 2018.

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

[3] Ø. Bergh, K. Y. Børsheim, G. Bratbak, and M. Heldal. High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–468, Aug. 1989.

[4] L.-M. Bobay, M. Touchon, and E. P. C. Rocha. Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132, Aug. 2014.

[5] A. B. Brueggemann, C. L. Harrold, R. R. Javan, A. J. van Tonder, A. J. McDonnell, and B. A. Edwards. Pneumococcal prophages are diverse, but not without structure or history. *Scientific Reports*, 7:srep42976, Feb. 2017.

[6] R. Calendar, editor. *The bacteriophages*. Oxford University Press, Oxford ; New York, 2nd ed edition, 2006. OCLC: ocm56103967.

[7] S. Casjens. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.*, 49(2):277–300, July 2003.

[8] M. R. Clokie, A. D. Millard, A. V. Letarov, and S. Heaphy. Phages in nature. *Bacteriophage*, 1(1), 2011.

[9] A. Colavecchio, Y. D'Souza, E. Tompkins, J. Jeukens, L. Freschi, J.-G. Emond-Rheault, I. Kukavica-Ibrulj, B. Boyle, S. Bekal, S. Tamber, R. C. Levesque, and L. D. Goodridge. Prophage integrase typing ts a useful indicator of genomic diversity in Salmonella enterica. *Frontiers in Microbiology*, 8:1283, July 2017.

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

[11] J. S. Crispim, R. S. Dias, P. M. P. Vidigal, M. P. de Sousa, C. C. da Silva, M. F. Santana, and S. O. de Paula. Screening and characterization of prophages in *Desulfovibrio* genomes. *Scientific Reports*, 8(1):9273, June 2018.

[12] B. Danneels, M. Pinto-Carbó, and A. Carlier. Patterns of nucleotide deletion and insertion inferred from bacterial *Pseudogenes*. *Genome Biology and Evolution*, 10(7):1792–1802, July 2018.

[13] S. Domingues, G. J. da Silva, and K. M. Nielsen. Integrons: Vehicles and pathways for horizontal dissemination in bacteria. *Mobile Genetic Elements*, 2(5):211–223, Sept. 2012.

[14] P. L. Foster, H. Lee, E. Popodi, J. P. Townes, and H. Tang. Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 112(44):E5990–E5999, Nov. 2015.

[15] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, Sept. 2005.

[16] A. C. Groth and M. P. Calos. Phage integrases: biology and applications. *Journal of Molecular Biology*, 335(3):667–678, Jan. 2004.

[17] E. Harrison and M. A. Brockhurst. Ecological and evolutionary benefits of temperate phage: what does or doesn't kill Yyou makes you stronger. *BioEssays*, 39(12):1700112, Dec. 2017.

[18] A. Khan and L. M. Wahl. Quantifying the forces that maintain prophages in bacterial genomes. *Theoretical Population Biology*, page S0040580919301844, Nov. 2019.

[19] C.-H. Kuo and H. Ochman. Deletional bias across the three domains of life. *Genome Biology and Evolution*, 1:145–152, Jan. 2009.

[20] R. Leplae, G. Lima-Mendez, and A. Toussaint. ACLAME: A CLAssification of mobile genetic elements, update 2010. *Nucleic Acids Research*, 38(suppl_1):D57–D61, Jan. 2010.

[21] J. W. Little. Robustness of a gene regulatory circuit. *The EMBO Journal*, 18(15):4299–4307, Aug. 1999.

[22] A. Mira, H. Ochman, and N. A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10):589–596, Oct. 2001.

[23] W. Mottawea, M.-O. Duceppe, A. A. Dupras, V. Usongo, J. Jeukens, L. Freschi, J.-G. Emond-Rheault, J. Hamel, I. Kukavica-Ibrulj, B. Boyle, A. Gill, E. Burnett, E. Franz, G. Arya, J. T. Weadge, S. Gruenheid, M. Wiedmann, H. Huang, F. Daigle, S. Moineau, S. Bekal, R. C. Levesque, L. D. Goodridge, and D. Ogunremi. *Salmonella enterica* prophage sequence profiles reflect genome diversity and can be used for high discrimination subtyping. *Frontiers in Microbiology*, 9, May 2018.

[24] M. Ptashne. *A genetic switch: ghage lambda revisited*. CSHL Press, 2004.

[25] S. V. Rajagopala, S. Casjens, and P. Uetz. The protein interaction map of bacteriophage lambda. *BMC Microbiology*, 11(1):213, 2011.

[26] F. Rohwer. Global phage diversity. *Cell*, 113(2):141, Apr. 2003.

[27] W. Song, H.-X. Sun, C. Zhang, L. Cheng, Y. Peng, Z. Deng, D. Wang, Y. Wang, M. Hu, W. Liu, H. Yang, Y. Shen, J. Li, L. You, and M. Xiao. Prophage Hunter: an integrative hunting tool for active prophages. *Nucleic Acids Research*, 47(W1):W74–W80, July 2019.

[28] A. Sousa, C. Bourgard, L. M. Wahl, and I. Gordo. Rates of transposition in *Escherichia coli*. *Biology Letters*, 9(6):20130838, Dec. 2013.

[29] P. M. Tran, M. Feiss, K. J. Kinney, and W. Salgado-Pabón. ϕSa3mw prophage as a molecular regulatory switch of *staphylococcus aureus* β-toxin production. *Journal of Bacteriology*, 201(14):e00766–18, /jb/201/14/JB.00766–18.atom, Apr. 2019.

[30] M. Ventura, C. Canchaya, M. Kleerebezem, W. M. de Vos, R. J. Siezen, and H. Brüssow. The prophage sequences of Lactobacillus plantarum strain WCFS1. *Virology*, 316(2):245–255, Nov. 2003.

[31] C. Zong, L. So, L. A. Sepúlveda, S. O. Skinner, and I. Golding. Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. *Molecular Systems Biology*, 6(1):440, Jan. 2010.

Chapter 5

Discussion & Conclusions

The study of phage-bacteria interactions involves gigantic numbers [1]: if all the viruses on earth were laid end to end they would stretch for 100 million light years; there are 13×10^{28} bacteria in the oceans (100 million times more than the number of stars in the known universe); 1×10^{31} viral infections occur per second in the ocean environment which results in the removal of 20% to 40% of the bacterial mass in the oceans per day; one gram dental plaque contains approximately 1×10^{11} bacteria; a teaspoon of soil contains 1×10^{9} microorganisms; bacteria present in the human gut weigh about 1 kg; 8% of human DNA is of viral origin; we have only sequenced 1×10^{-22}% of the total DNA on earth.

These quantities, along with the importance of the microorganisms to our ecosystem [12, 13], public health [24], food, and the possible discovery of new life on distant planets [18], have transformed our views about phage-bacteria interaction from being a simple system to a complex and important set of interactions. The intriguing dynamics of phage-bacteria interactions and the urgency of these global issues have invited us to dive deep into this ocean of knowledge. Once immersed in this study, we realized that there are more surprising facts, such

as: the amount of temperate phage DNA in bacterial genomes surpasses the amount of DNA in free-living phage [26]; and pathogenic bacteria, responsible for the death and miseries of millions of people, are domesticating these viral genomes with high frequency [11]. In this thesis, we contributed to this vast field. Our contribution is in the form of three projects, called **Chapter 2**, **Chapter 3** and **Chapter 4**, in this investigation. Below we provide details of these projects and the conclusions derived from them.

Here, we started with a topic directly related to public health, "Phage therapy and antibiotics for biofilm eradication: a predictive model" (**Chapter 2**). In that chapter, we developed a simple predictive model to capture the effect of the synergistic use of phages and antibiotics in biofilms. For this study, we assumed that bacteria are offering structural resistance (by grouping together, constructing biofilm and developing an EPS structure around the biofilm) to the antibiotic. In this model, we also used the idea of a group defense mechanism, that is, a phage functional response to bacteria resembling a Holling type IV functional response. We were able to show that neither antibiotic nor phages alone can eliminate the biofilm completely, and the synergistic effect of applying phages first and then antibiotics works better.

Prophages, being the main source of bacterial genome diversity and important factors in bacterial genome evolution, are more abundant in bacterial genomes than previously thought [6, 20]. In **Chapter 3**, we developed a mathematical model to study the effect of various evolutionary forces acting on prophages. We investigated the model in detail, fitting against some publicly available datasets, and were able to quantify the relative rates of these evolutionary forces in time units expressed in terms of the "expected prophage lifetime", that is, the average time between lysogeny and induction (see Table 3.6). From these rates, we conclude that: (1) the time between lysogeny events is about 5 prophage lifetimes, that is, new prophages enter

into bacterial genomes at a rate one-fifth of the induction rate; (2) the selection coefficient is 0.5 per prophage lifetime; (3) on average a prophage has lost only 1% of its genome at the time of induction; (4) a minimum of two to three prophage genes are required to excise prophage from the bacterial genome. The relation between prophages and their bacterial hosts is defined to be parasitic or mutualistic depending on the balance between the cost and benefits of the integration of foreign DNA [23]. The biggest cost due to prophage integration is the possibility of induction, which results in the killing of the bacterial host, although there are other small costs as well, such as energy costs [15]. Our results predict a tipping point between parasitism and mutualism, the point at which cost equals benefit. Our model predicted that the bimodal prophage size distribution is due to the balance between selective advantage (benefit) and induction (cost). The peak on right is due to the lysogeny of new prophages and the peak on left is due to the accumulation of smaller prophages, conferring more benefits to the bacterial hosts than their cost, as shown in Figure 3.8.

The domestication of defective prophages, the retention of defective prophages that confer some benefit to their hosts, is a common phenomenon in bacterial populations [3]. We believe that these defective prophages have a prominent role in shaping bacterial genome evolution. The genetic material of domesticated defective prophages may also serve as a tool-box for other prophages to use for repair through recombination [7]. In **Chapter 4**, "The genetic repertoire of prophages" we focus on genes enriched in smaller prophages and the role of evolutionary forces. First, we downloaded data regarding the genetic repertoire for two well-studied prophage databases [3, 17], using PHASTER [2]. The distributions of 53,356 annotated prophage genes identified in 1384 prophage sequences were examined, showing that: (1) genes involved in the lytic life cycle were preferentially lost in smaller prophages; (2) transposes

and integrases are significantly enriched in smaller prophages. We also developed an ODE model and computational model to study the effect of these evolutionary forces on the genetic repertoire of prophages. While the models were able to explain many interesting features of the data, we were unable to explain the enrichment of integrase genes in smaller prophages.

5.1 Future Work

We believe that the ODE model 2.1, representing phage-bacteria interaction in bacterial biofilm colonies may have some rich dynamics. We are planning to extend this work further and carry out further detailed bifurcation analysis of the system. In our model, phages and bacteria interact with each other according to the law of mass action and we ignore spatial structure. Several studies have concluded that ignoring the spatial structure of a biological system may lead to inaccuracies [8], and bacterial biofilms have a complex and interesting structure [10]. The inclusion of spatial structure in this model is needed to better understand the dynamics of interaction between phages and bacteria in bacterial biofilm colonies.

Prophages are abundant in bacterial genomes and are particularly prominent in pathogenic bacterial genomes [5, 4]. Prophages from pathogenic bacteria have been shown to encode virulence factors [14]. The insertion of these extra genes in bacterial genomes may increase the bacterial genome size. But studies have shown that pathogenic bacteria have smaller genomes and fewer genes than their closest non-pathogenic relatives [19, 25]. Gene acquisition and deletion may be the events underlying the emergence and evolution of bacterial pathogens [22]. The fact that despite having more prophages in their genomes, pathogen genomes are smaller than the closest non-pathogen relative gives rise to a question: what is the relation between

prophages and rates of gene gain and loss, acting on the whole host genome?. The availability of a huge amount of data regarding bacterial pathogens and the usefulness of mathematical modeling can give insights into this question.

Similarly, mutational deletion is a prominent evolutionary force acting on prophages. Such mutations make these prophages shorter, eventually resulting in domestication or deletion from the bacterial genome [3]. Prophages can excise from the bacterial genome randomly or due to some DNA damaging agent, resulting in free life as a temperate phage [21]. Does this mutational deletion cause a reduction, over evolutionary time, in the genome size of temperate phages? If not, how can temperate phage keep their genomes intact in the presence of mutational deletions as an important evolutionary force acting on prophages? In other words, what is the role of mutational deletions in the evolution of temperate phages?

One of the strongest signals obtained from the genetic repertoire data of prophages, in **Chapter 4**, was the significant enrichment of integrase genes in incomplete prophages (see Table 4.1 and Figure 3.1). Full and partial prophage sequences are frequently transferred horizontally through transduction [9] and related processes such as molecular parasitism by GTAs [16]. Once foreign DNA enters a host cell through HGT, it needs to recombine with the host genome; integrase acts as a catalyst to mediate the process of recombination, resulting in an increased rate of recombination. Therefore, integrases likely help prophage-derived elements with recombination into the host genome during horizontal transmission. The horizontal transmission of integrase genes along with their neighboring prophage genes may have caused significant enrichment of integrase in incomplete prophages. From this, we hypothesize that in the event of horizontal transmission, a mobile genetic element containing the integrase gene may have a selective advantage over a mobile genetic element lacking the integrase gene. We are

planning to explore this hypothesis in much more detail, through an individual-based model like the one developed in Chapter 4, but including more detail such as horizontal gene transfer.

Bibliography

[1] Microbiology by numbers. *Nature Reviews Microbiology*, 9(9):628–628, Sept. 2011.

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

[3] L.-M. Bobay, M. Touchon, and E. P. C. Rocha. Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132, Aug. 2014.

[4] H. Brussow, C. Canchaya, and W.-D. Hardt. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews*, 68(3):560–602, Sept. 2004.

[5] C. Canchaya, G. Fournous, and H. Brüssow. The impact of prophages on bacterial chromosomes. *Molecular Microbiology*, 53(1):9–18, July 2004.

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

[7] M. De Paepe, G. Hutinet, O. Son, J. Amarir-Bouhram, S. Schbath, and M.-A. Petit. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-Like recombinases. *PLoS Genetics*, 10(3):e1004181, Mar. 2014.

[8] U. Dieckmann, R. Law, J. A. J. Metz, and I. I. for Applied Systems Analysis, editors. *The geometry of ecological interactions: simplifying spatial complexity*. Cambridge studies in adaptive dynamics. Cambridge University Press, Cambridge, U.K. ; New York, NY, USA, 2000.

[9] A. Fillol-Salom, A. Alsaadi, J. A. M. d. Sousa, L. Zhong, K. R. Foster, E. P. C. Rocha, J. R. Penadés, H. Ingmer, and J. Haaber. Bacteriophages benefit from generalized transduction. *PLOS Pathogens*, 15(7):e1007888, July 2019.

[10] H.-C. Flemming, J. Wingender, U. Szewzyk, P. Steinberg, S. A. Rice, and S. Kjelleberg. Biofilms: an emergent form of bacterial life. *Nature Reviews Microbiology*, 14(9):563–575, Sept. 2016.

[11] L.-C. Fortier and O. Sekulovic. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, 4(5):354–365, July 2013.

[12] S. M. Gibbons and J. A. Gilbert. Microbial diversity — exploration of natural ecosystems and microbiomes. *Current Opinion in Genetics & Development*, 35:66–72, Dec. 2015.

[13] E. B. Graham, J. E. Knelman, A. Schindlbacher, S. Siciliano, M. Breulmann, A. Yannarell, J. M. Beman, G. Abell, L. Philippot, J. Prosser, A. Foulquier, J. C. Yuste,

H. C. Glanville, D. L. Jones, R. Angel, J. Salminen, R. J. Newton, H. Bürgmann, L. J. Ingram, U. Hamer, H. M. P. Siljanen, K. Peltoniemi, K. Potthast, L. Bañeras, M. Hartmann, S. Banerjee, R.-Q. Yu, G. Nogaro, A. Richter, M. Koranda, S. C. Castle, M. Goberna, B. Song, A. Chatterjee, O. C. Nunes, A. R. Lopes, Y. Cao, A. Kaisermann, S. Hallin, M. S. Strickland, J. Garcia-Pausas, J. Barba, H. Kang, K. Isobe, S. Papaspyrou, R. Pastorelli, A. Lagomarsino, E. S. Lindström, N. Basiliko, and D. R. Nemergut. Microbes as engines of ecosystem function: when does community structure enhance predictions of ecosystem processes? *Frontiers in Microbiology*, 7, Feb. 2016.

[14] P. Hyman, S. T. Abedon, and C. International, editors. *Bacteriophages in health and disease*. Number 24 in Advances in molecular and cellular microbiology. CABI, Wallingford, Oxfordshire, 2012.

[15] E. V. Koonin. Evolution of genome architecture. *The International Journal of Biochemistry & Cell Biology*, 41(2):298–306, Feb. 2009.

[16] A. S. Lang, O. Zhaxybayeva, and J. T. Beatty. Gene transfer agents: phage-like elements of genetic exchange. *Nature Reviews Microbiology*, 10(7):472–482, July 2012.

[17] R. Leplae, G. Lima-Mendez, and A. Toussaint. ACLAME: A CLAssification of mobile genetic elements, update 2010. *Nucleic Acids Research*, 38(suppl_1):D57–D61, Jan. 2010.

[18] J. V. Lopez, R. S. Peixoto, and A. S. Rosado. Inevitable future: space colonization beyond Earth with microbes first. *FEMS Microbiology Ecology*, 95(10):fiz127, Oct. 2019.

[19] N. A. Moran. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108(5):583–586, Mar. 2002.

[20] W. Mottawea, M.-O. Duceppe, A. A. Dupras, V. Usongo, J. Jeukens, L. Freschi, J.-G. Emond-Rheault, J. Hamel, I. Kukavica-Ibrulj, B. Boyle, A. Gill, E. Burnett, E. Franz, G. Arya, J. T. Weadge, S. Gruenheid, M. Wiedmann, H. Huang, F. Daigle, S. Moineau, S. Bekal, R. C. Levesque, L. D. Goodridge, and D. Ogunremi. *Salmonella enterica* prophage sequence profiles reflect genome diversity and can be used for high discrimination subtyping. *Frontiers in Microbiology*, 9, May 2018.

[21] A. M. Nanda, K. Thormann, and J. Frunzke. Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. *Journal of Bacteriology*, 197(3):410–419, Feb. 2015.

[22] H. Ochman. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science*, 292(5519):1096–1099, May 2001.

[23] J. W. Shapiro and P. E. Turner. Evolution of mutualism from parasitism in experimental virus populations. *Evolution*, 72(3):707–712, Mar. 2018.

[24] V. H. Smith, R. J. Rubinstein, S. Park, L. Kelly, and V. Klepac-Ceraj. Microbiology and ecology are vitally important to premedical curricula. *Evolution, Medicine, and Public Health*, page eov014, July 2015.

[25] C. Toft and S. G. E. Andersson. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews. Genetics*, 11(7):465–475, July 2010.

[26] L. M. Wahl and T. Pattenden. Prophage provide a safe haven for adaptive exploration in temperate viruses. *Genetics*, 206(1):407–416, May 2017.

Appendix A

Appendix for Chapter 3

A.1 Derivation of the PDE 3.1 and its steady state solution

Let $Q(x,t)$ be the length distribution of prophages of length x, at a time t. Then after time step, δt, we have

$$\begin{aligned} Q(x,t+\delta t) &= Q(x,t) + D(x+\delta x)\,P(x+\delta x,t)\,\frac{\delta t}{\delta x} - D(x)\,Q(x,t)\,\frac{\delta t}{\delta x} + r_S S(x)\,Q(x,t)\,\delta t \\ &- r_I I(x)\,Q(x,t)\,\delta t + \alpha\, f(x)\,\delta t + \beta\, g(x)\,\delta t. \end{aligned}$$

Using Tylor's series expansion and after simplification we arrive at the following

$$\frac{Q(x,t+\delta t) - Q(x,t)}{\delta t} = \left(\frac{D(x+\delta x) - D(x)}{\delta x}\right) Q(x,t) + D(x+\delta x)\,\frac{\partial Q(x,t)}{\partial x} + O(\delta x) \tag{A.1}$$

Now taking $\lim_{\delta t \to 0}$ and $\lim_{\delta x \to 0}$, we arrive at

$$\begin{aligned} \frac{\partial Q(x,t)}{\partial t} &= \frac{\partial D(x)}{\partial x}\,Q(x,t) + D(x)\,\frac{\partial Q(x,t)}{\partial x} + r_S S(x)\,Q(x,t) - r_I I(x)\,Q(x,t) + \alpha\, f(x) + \beta\, g(x) \\ &= \frac{\partial}{\partial x}[D(x)\,Q(x,t)] + [r_S S(x) - r_I I(x)]\,Q(x,t) + \alpha\, f(x) + \beta\, g(x). \end{aligned} \tag{A.2}$$

If we consider $\lim_{t\to\infty} Q(x,t) = P(x)$ and $D(x) = r_D\, x$ then the differential equation generating steady state solution, of the PDE 3.1, is given by

$$\frac{dP(x)}{dx} + \left(\frac{1}{x} + \mathcal{F}(x)\right) P(x) + \frac{\alpha}{r_D\, x}\, f(x) + \frac{\beta}{r_D\, x}\, g(x) = 0 \tag{A.3}$$

where $\mathcal{F}(x) = \frac{r_S\, S(x)}{r_D\, x} - \frac{r_I\, I(x)}{r_D\, x}$. Equation (3.7) is first order linear ODE and its solution is given by

$$P(x) \;=\; \frac{-\,e^{-\int \mathcal{F}(x)dx}}{r_D\, x} \int (\alpha\, f(x) \,+\, \beta\, g(x))\, e^{\int \mathcal{F}(x)dx} dx + \frac{C}{x}\, e^{-\int \mathcal{F}(x)dx}, \tag{A.4}$$

where C is a constant of integration.

A.2 Results from model selection and data fitting

The AIC value is the measure of loss of information for the model under consideration and is an ordinal number, used for ranking models. The lowest AIC value corresponds to the best fit. If the number of data points are small enough compared to the number of parameters then the AIC value is not penalized enough. To remedy this problem a second order Akaike Information criteria, the corrected Akaike Information Criteria (AICc), is defined. The corrected Akaike Information Criteria (AICc) is given as [2]:

$$AIC_c = AIC + \frac{2k(k+1)}{n-(k+1)}. \tag{A.5}$$

As the number of data points becomes large enough, AICc values converge to AIC values and either of these criteria can be used to determine the best fit model amongst the candidate models [2]. In the tables to follow, we provide both AIC and AICc values, and compute relative probabilities using the AICc values.

A.2.1 Data Set 1

#	Parameters	AIC	AICc	Log-likelihood	Relative probability (AICc)
1	15	4884.1734	4884.9642	-2427.0867	0.3719
2	14	4882.2952	4882.9860	-2427.1476	1
3	12	4893.5207	4894.0322	-2434.7603	0.0039
4	11	4894.7613	4895.1920	-2436.3807	0.0022
5	9	4908.5087	4908.8014	-2445.2544	2.4788e-06
6	8	4906.2759	4906.5097	-2445.1379	7.7964e-06
7	6	5044.8136	5044.9499	-2515.9068	6.7604e-36
8	6	5069.6683	5069.8042	-2528.8341	2.7098e-41
9	4	5063.9143	5063.9788	-2527.9571	4.9878e-40

Table A.1: Number of parameters, AIC, AICc values, log-liklihood and the corresponding relative probabilities for Data Set 1 [1]. The best fit model includes a mixed distribution to describe autonomous temperate phages (g=3), degradation, induction and selection. The second best fit model is the same model with HGT and has relative probability 0.3791.

A.2.2 Data Set 2

#	Number of Parameters	AIC	AICc	Log-likelihood	Relative probability
1	15	993.726	998.583	-480.863	0.0016
2	14	990.549	994.797	-480.275	0.0108
3	12	991.327	994.492	-482.663	0.0126
4	11	987.247	989.937	-481.624	0.123
5	9	987.197	989.062	-483.599	0.191
6	8	984.237	985.749	-483.119	1
7	6	1006.929	1007.855	-496.464	1.585e-05
8	6	1012.234	1013.159	-499.117	1.117e-06
9	4	1004.729	1005.216	-497.364	5.927e-05

Table A.2: Number of parameters, AIC, AICc values and the corresponding relative probabilities for Data Set 2 [3]. The best fit model includes degradation, induction and selection as well as one Gaussian distribution to describe autonomous temperate phages (g=1).

A.2.3 Data Set 3

#	Number of Parameters	AIC	AICc	Log-likelihood	Relative probability
1	15	5671.819	5672.579	-2819.909	0.0318
2	14	5669.819	5670.489	-2819.909	0.0904
3	12	5670.179	5670.685	-2822.089	0.0819
4	11	5667.438	5667.872	-2821.719	0.3347
5	9	5667.461	5667.766	-2823.731	0.3528
6	8	5665.434	5665.683	-2823.717	1
7	6	5731.084	5731.239	-2858.542	5.8167e-15
8	6	5757.726	5757.880	-2871.863	9.5404e-21
9	4	5753.680	5753.762	-2871.840	7.4776e-20

Table A.3: Number of parameters, AIC, AICc values and the corresponding relative probabilities for Data Set 3 [4]. The best fit model includes degradation, induction and selection as well as one Gaussian distribution to describe autonomous temperate phages (g=1).

Bibliography

[1] Bobay, L.-M., Touchon, M., and Rocha, E. P. C. (2014). Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132.

[2] Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York, 3rd ed. edition.

[3] Crispim, J. S., Dias, R. S., Vidigal, P. M. P., de Sousa, M. P., da Silva, C. C., Santana, M. F., and de Paula, S. O. (2018). Screening and characterization of prophages in *Desulfovibrio* genomes. *Scientific Reports*, 8(1):9273.

[4] Leplae, R., Lima-Mendez, G., and Toussaint, A. (2010). ACLAME: A CLAssification of mobile genetic elements, update 2010. *Nucleic Acids Research*, 38(suppl_1):D57–D61.

Appendix B

Appendix for Chapter 3

B.1 Sensitivity Analysis

B.1.1 Sensitivity to the smallest autonomous phage length.

We tested fitting model (3.1) to Data Set 1, but assuming that the smallest autonomous phage to infect *E. Coli* and *S. Enterica* has length θ = 30 kb, as suggested in [1]. We compared these results to results obtained with θ = 20 kb, as described in Section 2.2 of the main text. Figure B.1 demonstrates that our results are insensitive to the choice of this parameter.

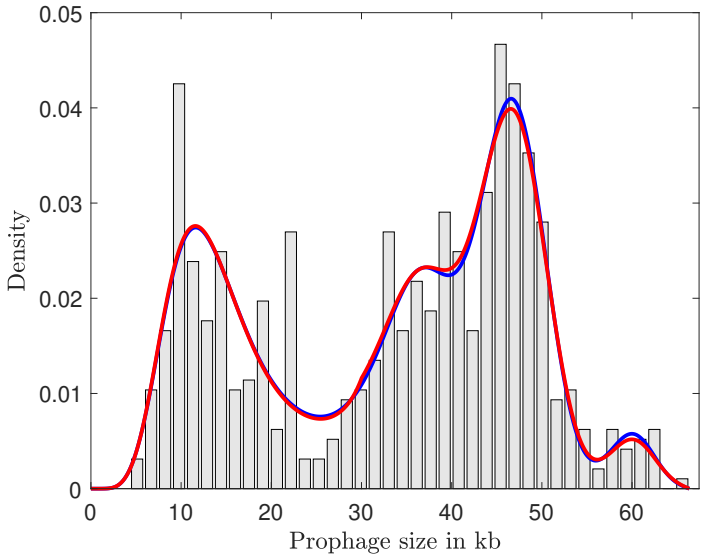


Figure B.1: Results of data fitting are not sensitive to the choice of the parameter θ representing the genome size of the smallest autonomous temperate phage in kb. Best fits obtained to Data Set 1 (histogram) for $\theta = 20$ (blue, solid) and $\theta = 30$ (red, solid) are indistinguishable.

B.1.2 Rate parameters

We performed a bootstrap sensitivity analysis for all parameters of the model using Data Set 1. In brief, we assumed that the best fit model for Data Set 1 represented the true distribution, and resampled this true distribution 335 times, each time creating a simulated data set of 624 observed prophage lengths. We then subjected each of these data sets to the model fitting exercise described in Section 3 of the main text. Table B.1 shows the mean and standard deviations for the relative rate parameters of the model (each rate normalized by the induction rate, r_I), after the analysis of 335 simulated data sets. These results indicate that the quantitative conclusions of our work are relatively insensitive to variations in data sampling; the coefficient of variation (standard deviation/mean) of the degradation rate is largest at 16%.

Parameter	Description	Mean	Standard deviation	Coefficient of Variation
α	Relative rate of lysogeny	0.2078	0.0118	0.0569
r_D	Relative rate of degradation	0.0125	0.0021	0.1644
r_S	Relative selection coefficient	0.5012	0.0483	0.0964

Table B.1: Sensitivity analysis of rate parameters.

B.1.3 Influx of active phage, $f(x)$

In addition, this process produced 335 estimates of the influx distribution $f(x)$. In Figure B.2, we plot the mean of these functions at every value of x (blue line), plus/minus one standard deviation (grey area). The best fit $f(x)$ from Data Set 1 is also shown for comparison (red line). These results indicate that the form of $f(x)$ is very tightly constrained by the data, a result that is perhaps not surprising given the large number of data points.

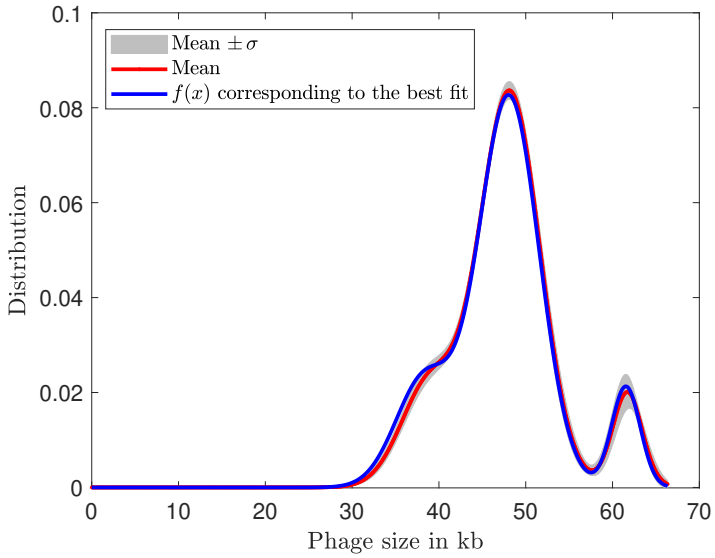


Figure B.2: Sensitivity analysis of the prophage influx function. The mean (red) and standard deviation (σ) of best-fit $f(x)$ curves for all simulated data sets are shown, along with the best-fit $f(x)$ function from the true data (blue). See text for details.

Bibliography

[1] Bobay, L.-M., Touchon, M., and Rocha, E. P. C. (2014). Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132.

Appendix C

Appendix for Chapter 3

C.1 The influx distribution

As described in Section 2.2 of the main text, the function $f(x)$ gives the length distribution for prophages that are newly integrating into bacterial genomes. Here, we note that this is neither the length distribution of active temperate phages, nor is it the length distribution of inducing phage.

To clarify, suppose $A(x)$ is the length distribution of active temperate phages. Let $L(x)$ be the average lysogeny probability for a temperate phage of length x. Since $A(x)$ consists of phages of different classes (lambdoid, mu-like, etc.), we expect that $L(x)$ is not constant in x. In this case, the influx distribution $f(x)$ is given by the product $f(x) = A(x)L(x)$. Thus, unfortunately, we cannot use empirical data describing $A(x)$ to infer $f(x)$.

Similarly, from the model at steady state, the product $P(x)I(x)$ gives the length distribution of excising prophage. Suppose $R(x)$ gives the probability that a prophage of length x retains the genes required for re-infection (genes involved in replication, packaging, and adsorption, for example). If re-infection competent phage enter the lysogenic life cycle with probability $L(x)$, we could also express the influx distribution as $f(x) = P(x)I(x)R(x)L(x)$. Again, we are

unable to use $P(x)I(x)$ to directly infer $f(x)$.

Despite these limitations, some qualitative features of $f(x)$ and $A(x)$ appear surprisingly robust. Along with prophage sequences, the length distribution of 68 dsDNA temperate phages infecting enterobacteria are reported in [1]. While the weight of the peaks in this multimodal distribution vary, the number and position of the peaks is strikingly similar with our best fit estimate for $f(x)$ for Data Set 1, as shown in Table C.1.

Feature	Empirical Data	Model Prediction
Number of main peaks	3	3
Position of first peak	$\approx$ 40 kb	$\approx$ 38 kb
Position of second peak	$\approx$ 45 kb	$\approx$ 48 kb
Position of third peak	$\approx$ 59 kb	$\approx$ 61 kb

Table C.1: Comparison of the main features of empirical data describing the length distribution of autonomous dsDNA phages [1], and the best-fit model predictions for the phage influx distribution, $f(x)$.

Similarly, we find that $I(x)P(x)$ yields a surprisingly good approximation for $f(x)$, as illustrated in Figure C.1, again for Data Set 1. This suggests that most prophage sequences that retain the genes for necessary for excision also retain the genes necessary for re-infection.

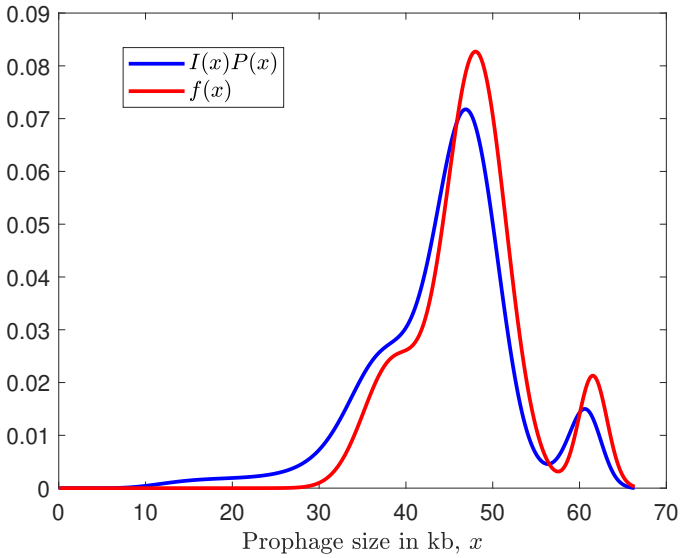


Figure C.1: Comparison of best-fit $f(x)$ with the product $P(x)I(x)$; results shown for Data Set 1.

Bibliography

[1] Bobay, L.-M., Touchon, M., and Rocha, E. P. C. (2014). Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132.

Appendix D

Appendix for Chapter 3

D.1 MATLAB code

In this appendix we are presenting MATLAB [1] code for solving PDE model (3.1). The MATLAB routine `fminsearch` was used to maximize the log-likelihood. The AIC criteria was then used to select the best fit model from all the candidate models. Here we present code for the full model used to fit the steady state solution of 3.1 to Data Set 1 [2]. This code calculates selection (3.2.5) and induction (3.2.6) before fitting the steady state solution to the data. The code then maximizes the log-likelihood, which is calculated by the function `sol_error`, also provided.

```
clear all
clc
global pls
load('prophage_sizes.csv')
pro = prophage_sizes;
pls = pro;
pls=pls/1000; % prophage length in kb
Pinit=  [0.204686704151821   4.099811476589647  15.724017308458798  ...
    11.536956361960861  65.822440602842576  27.198192070426217  ...
    25.986837564236168   10.382028701765515  41.551463526177145...
         6.882397692568709  0.007780330712747   0.013293852127858   ...
            2.799975078907655   0.949368876463170  ...
            44.738263580612710   0.020163712415658]; % inatial guess
 options = optimset('MaxFunEvals',2000000);
 Pbest = fminsearch(@sol_error,Pinit);
```

```
Pbest;
P = Pbest;
```

```
function err = sol_error(P)
global pls xs pfinal
ts = 0:0.01:200;
xs = linspace(0, max(pls), 664);
f= @ (x) abs(P(2))*exp(-(x-(20+abs(P(3)))).^2./abs(P(4)))+ ...
    abs(P(5))*exp(-(x-(20+abs(P(6)))).^2./abs(P(7)))+ ...
    abs(P(8))*exp(-(x-(20+abs(P(9)))).^2./abs(P(10))); % gaussian of ...
       incoming phages
g = @(x) abs(P(11))*(-x+max(pls)); % HGT
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
D= @(x) abs(P(12))*x;
dD= @(x) abs(P(12));
%%
% Calculating induction curve
M1min = 20;  % smallest dsDNA phage
MM= max(pls);
n = abs(P(13)); % minimum number of genes required for induction
M1max = max(pls);
M1 = M1min :0.1: M1max;
b = zeros(length(M1), length(xs));
bsum = zeros (1, length(xs));
i =1;
while i ≤ length(M1)
L = M1(i);
m = n+10^-13:0.1: L;
N= 0:0.1:n;
b(i, 1:length(N))=0;
j=1;
while  j≤ length(m)
       b(i, length(N)+j) = ((m(j).^m(j)).*((L-n).^(L-n))./( ...
           (L.^L).*(m(j)-n).^(m(j)-n));  % continuous approximation ...
           to probability of induction
       j= j+1;
end
bsum = bsum + f(L).*b(i, :);
 i = i+1;
end
k=1;
while k≤length(bsum)
    p =0:0.1:M1min;
    if k > length(p)
    l= k+1 -length(p);
    L= M1(l);
    end
    L1 = L:0.1:M1max;
    if k≤length(p)
        bsum(k)= bsum(k)./sum(f(M1));
```

```
    else
        bsum(k)= bsum(k)./sum(f(L1));
    end
    k=k+1;
end
I = abs(P(14)).*(bsum./max(bsum));
%%
% Calculating selection curve
%%
Max = max(pls);
m = 0:0.1:MM;
M1min = floor(abs(P(15)));
M1max = max(cdfx);
M1 = M1min:0.1:M1max;
%Max = 66;
s=  abs(P(15));
s1 = 0:0.1:s;
expec = zeros(1, length(m));
j=length(s1);
i=1;
while j ≤length(m)
    M = M1(j-(length(s1)-1));
    M2 = M:0.1:M1max;
    if j ≤ length(s1)
    while i ≤ length(s1)
    S = (m(i).*s).*((sum(f(M2)./M2))./sum(f(M2)));
    expec(i)=S;
    i = i+1;
    end
    else
        S = (m(j).*s).*((sum(f(M2)./M2))./sum(f(M2)));
   expec(j)=S;
    end
j = j+1;
end
Sel = abs(P(16)).*expec;
% solving PDE model
∆t = ts(2)-ts(1);
∆x = xs(2)-xs(1);
b = ∆t / ∆x;
Q=zeros(length(xs), length(ts));
f= abs(P(2))*exp(-(xs-(20+abs(P(3)))).^2./abs(P(4)))+ ...
    abs(P(5))*exp(-(xs-(20+abs(P(6)))).^2./abs(P(7)))+ ...
    abs(P(8))*exp(-(xs-(20+abs(P(9)))).^2./abs(P(10)));
for i = 1:length(xs)
    if xs(i)≤20
        f(i) = 0;
    end
end
f = f./trapz(xs, f);
D= abs(P(12))*xs;
dD= abs(P(12));
 for k=1:length(ts)-1
  for i=2:length(xs)-1
```

```
99
100           Q(i,k+1) = Q(i,k)+Δt.*Q(i,k).*dD+b.*D(i)*(Q(i+1,k)-Q(i,k))+ ...
101           (Sel(i)-I(i)).*Q(i,k).*Δt+P(1).*(f(i))*Δt+g(xs(i)).*Q(i, ...
                  k).*Δt;
102
103    end
104      Q(:,k+1) = Q(:,k+1)./trapz(xs,Q(:,k+1));
105   end
106  pfinal = cumtrapz(xs,Q(:,length(ts)));
107  pguess = interp1(xs,Q(:,length(ts)),cdfx);
108  err = -sum(log(pguess(2:end-1)));
```

Bibliography

[1] (2018). *MATLAB version 9.5.0.944444 (R2018b)*. The Mathworks, Inc., Natick, Massachusetts.

[2] Bobay, L.-M., Touchon, M., and Rocha, E. P. C. (2014). Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, 111(33):12127–12132.

Appendix E

Appendix for Chapter 4

E.1 Fixed point and stability analysis of system 4.1

System 4.1 has six equilibrium points, four of which are biologically meaningful (non-negative). We will use the notation $E_i = \left(\bar{P}_{111}, \bar{P}_{011}, \bar{P}_{101}, \bar{P}_{110}, \bar{P}_{001}, \bar{P}_{010}, \bar{P}_{100}, \bar{P}_{000}\right)$, where $\bar{P}_{ber}$ denotes the equilibrium value of $P_{ber}(t)$, and describe these equilibria below.

1) The fixed point $E_0 = (0, 0, 0, 0, 0, 0, 0, 1)$ corresponds to the complete elimination of prophages from bacterial genomes. This fixed point always exists. The eigenvalues of the corresponding linearized Jacobian are: 0, $-r_D$, $r_S - r_D$, $r_S - 2r_D$, $-r_D - r_I$, $r_S - 2r_D - r_I$, $r_L - 2r_D - r_I$, and $r_L + r_S - 3r_D - r_I$. This fixed point is stable if $r_S < r_D$ and $r_L < 2\,r_D + r_I$.

2) The fixed point $E_B = (0, 0, 0, 0, 0, 0, \frac{r_S - r_D}{r_S}, \frac{r_D}{r_S})$, corresponds to the existence of beneficial prophage genes only. This fixed point exists only if $r_S > r_D$. The eigenvalues of the corresponding Jacobian matrix are: $-r_D$, $-r_S$, $r_D - r_S$, $r_D - r_S$, $-r_I - r_S$, $-r_D - r_I$, $r_L - 2r_D - r_I$, and $r_L - r_D - r_I - r_S$. Thus the conditions for stability are $r_S > r_D$ and $r_L < 2\,r_D + r_I$.

3) The fixed point $E_{LI} = \left(0, \frac{\alpha\gamma}{r_L\eta}, 0, 0, \frac{r_D\alpha\gamma}{r_L\eta^2}, \frac{r_D\gamma}{r_L\eta}, 0, \frac{r_D^2\xi}{r_L\eta^2}\right)$, where $\alpha = r_L - r_D$, $\gamma = r_L - 2r_D - r_I$, $\eta = r_L - r_D - r_I$ and $\xi = 2r_L - 2r_D - r_I$. E_{LI} corresponds to the coexistence of lysis and infectious genes, and exists if $r_L > 2r_D + r_I$. Eigenvalues of the corresponding linearized Jacobian are:

$r_S - r_D$, $r_S - r_L$, $r_D - r_L$, $r_I + r_S - r_L$, $r_I + r_D - r_L$, $r_S + r_D + r_I - r_L$, $r_I + 2\,r_D - r_L$, and $r_I + 2\,r_D - r_L$. These eigenvalues are all negative under the two conditions $r_L > 2r_D + r_I$ and $r_S < r_D$.

4) $E_A = \left(\frac{\alpha\beta\gamma}{r_L r_S \eta}, \frac{r_D\alpha\gamma}{r_L r_S \eta}, \frac{r_D\alpha\beta\gamma}{r_L r_S \eta^2}, \frac{r_D\beta\gamma}{r_L r_S \eta} \frac{r_D^2\alpha\gamma}{r_L r_S \eta^2}, \frac{r_D^2\gamma}{r_L r_S \eta}, \frac{r_D^2\beta\xi}{r_L r_S \eta^2}, \frac{r_D^2\beta\gamma}{r_L r_S \eta^2}, \frac{r_D^3\xi}{r_L r_S \eta^2}\right)$, where $\beta = r_S - r_D$. The eigenvalues of the Jacobian are: $r_D - r_S$, $r_D - r_L$, $2\,r_D - r_L - r_S$, $r_I + 2\,r_D - r_L$, $r_D + r_I - r_L$, $r_I + 2\,r_D - r_L - r_S$, $r_I + 3\,r_D - r_L - r_S$, and $r_I + 3\,r_D - r_L - r_S$. These eigenvalues are all negative under the conditions $r_S > r_D$ and $r_L > 2r_D + r_I$.

Appendix F

Appendix for Chapter 4

F.1 Transposase enrichment in incomplete prophages.

As described in the main text, we simulated the prophage population with parameter values $r_S = 1.5$, $r_D = 0.048$, $r_L = 1.5$, $r_T = 0.002$ for 5000 generations to compare the gene content of intact and incomplete prophages. Using a strict definition for “intact" prophages, that is, only prophages containing all the genes required for excision and re-infection were considered intact, transposase genes were enriched nearly 400-fold in incomplete prophages (see Figure F.1) (A) and (B). When the classification of “intact” prophages was relaxed to prophages that contain 90% or more of the possible prophage genes, the results showed a 5.6-fold increase in transposase genes (see Figure F.1) (C) and (D).

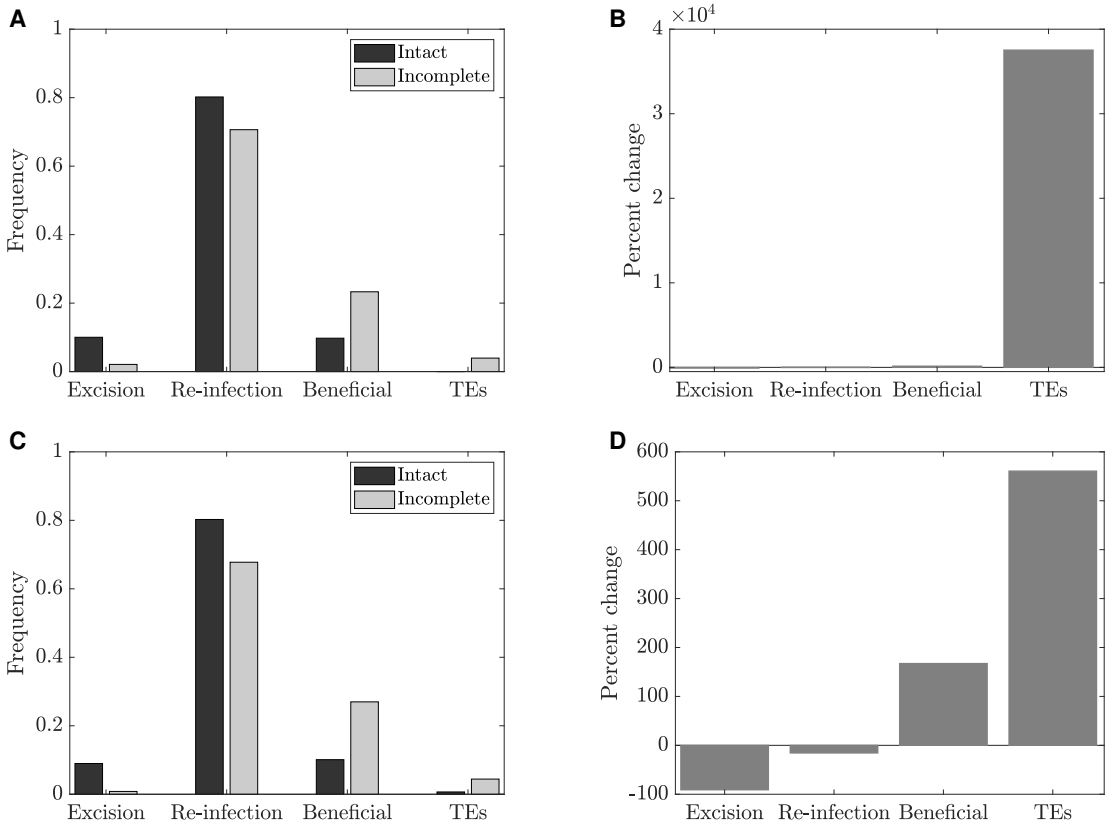


Figure F.1: Gene frequencies in intact and incomplete prophages, when TEs are included (r_S = 1.5, r_L = 1.5, r_D = 0.048, r_T = 0.002). (A) Frequency of genes of each type in intact and incomplete prophages, for the computational model simulated at the persistence equilibrium with TE disruptions; (B) Percent change in gene frequency from intact to incomplete; for (A) and (B) intact prophagea are defined as sequences containing all the genes required for excision and reinfection; (C) Frequency of genes of each type in intact and incomplete prophages, for the computational model simulated at the persistence equilibrium with TE; (D) Percent change in gene frequency from intact to incomplete; for (C) and (D) intact prophages are defined as sequences containing 90% or more of the possible prophage genes.

Appendix G

Appendix for Chapter 4

G.1 C++ code for computational model.

Here we present the C++ code, used to carry out all the calculations of the computational model in chapter 4. This code uses the routine "random.h", adopted from [1], to generate a random number between 0 and 1.

```
#include<stdio.h>
#include<math.h>
#include<stdlib.h>
#include<strings.h>
#include"random.h"
#include "getdata.h"

#define MAXNS 31000
#define MAXGENES 40
#define MAXPRINT 2000

long seed=-1;

int main(int argc, char **argv)
{

  float tend;
  float delt;
  int nsteps;
  int nlys;
  int ninf;
  int nneut;
  int nben;
  int ngenes;
  float rs;
```

```
26   float rl;
27   float rd;
28   float ri;
29   float rt;
30   int Ninit;
31
32   short (*prophages)[MAXGENES], (*tmpptr)[MAXGENES];
33   short (*newprophages)[MAXGENES];
34   short pro1[MAXNS][MAXGENES];
35   short pro2[MAXNS][MAXGENES];
36   float genemeans[MAXGENES];
37   float ismeans[MAXGENES];
38   float tempsum[MAXGENES];
39   float tempsumt[MAXGENES];
40   int occupied[MAXNS];
41   int induce[MAXNS];
42   int inducible, induciblesum, numben, noinduceflag = 0;
43   float ss, css[MAXNS], r, fractiontolose;
44   int maxind, i, newi, j, k, ii, jj, lyscapable;
45   float ran3(long *);
46   void getdata(FILE *,float *, float *,int *, int *, int *, int *, ...
         float *, float *, float *, float *, float *, int *);
47   int ntoprint, sizeflag=1, sizes[MAXNS], sizehist[MAXGENES+1], ksum=0;
48
49   FILE *fpin, *fpout, *fpout2, *fpout3, *fpout4, *fpout5;
50
51  if (argc>1) seed = -((long)(atof(argv[1])));
52  if (argc>2) sizeflag = (int)(atof(argv[2]));
53
54   if ((fpout=fopen("prosim.out","w"))==NULL) {
55     fprintf(stderr,"Error opening prosim.out\n");
56     printf("\a");
57     exit(1);
58   }
59   if ((fpout2=fopen("genemeans.out","w"))==NULL) {
60     fprintf(stderr,"Error opening genemeans.out\n");
61     printf("\a");
62     exit(1);
63   }
64   if ((fpin=fopen("prosim.in","r"))==NULL) {
65     fprintf(stderr,"Error opening prosim.in\n");
66     printf("\a");
67     exit(1);
68   }
69   if (sizeflag)
70   if ((fpout3=fopen("sizes.out","w"))==NULL) {
71     fprintf(stderr,"Error opening sizes.out\n");
72     printf("\a");
73     exit(1);
74   }
75   if ((fpout4=fopen("Ns.out","w"))==NULL) {
76     fprintf(stderr,"Error opening Ns.out\n");
77     printf("\a");
78     exit(1);
```

```
79    }
80    if ((fpout5=fopen("ismeans.out","w"))==NULL) {
81      fprintf(stderr,"Error opening ismeans.out\n");
82      printf("\a");
83      exit(1);
84    }
85
86    getdata(fpin,&tend,&delt,&nlys,&ninf,&nneut,&nben,&rs,&rl,&rd,...
87    &ri,&rt,&Ninit);
88    if (Ninit≥MAXNS) {
89      fprintf(stderr,"Error, Ninit too large\n");
90      exit(1);
91    }
92    nsteps = tend/delt;
93    ngenes = nlys+ninf+nneut+nben;
94    if (ngenes≥MAXGENES) {
95      fprintf(stderr,"Error, too many genes\n");
96      exit(1);
97    }
98    rs = rs*delt;
99    rl = rl*delt;
100   rd = rd*delt;
101   ri = ri*delt;
102   rt = rt*delt;
103   float big = 0.02;
104   if ((rs>big) || (rl>big) || (rd>big) || (rs*nben>big))
105     fprintf(stdout,"Error: big changes in one timestep\n");
106   prophages = pro1;
107   newprophages = pro2;
108   ntoprint = (int)((float)nsteps/(float)MAXPRINT);
109   if (ntoprint == 0) ntoprint = 1;
110
111   //
112   for (int i = 0; i < Ninit; i++) {
113     occupied[i] = 1;
114     for (int j = 0; j < ngenes; j++) prophages[i][j] = 1;
115     }
116   maxind = Ninit;    // maxind is the maximum possible occupied row ...
        in prophages
117 //----------TIME LOOP
118     for (int istep =0; istep < nsteps; istep++){
119
120 //-----------------Induction
121 //
122 // scan through the prophages, ineducable only if all genes required ...
       for induction are present
123 // also keep track that at least some prophages are ineducable
124     inducible sum = 0;
125     for (i = 0; i < maxind; i++) {
126       inducible = prophages[i][0];
127       for (j = 1; j < nlys; j++)
128     if (prophages[i][j] != -1) inducible = inducible*prophages[i][j];
129       if ( (ran3(&seed) < ri) && (inducible >0)) {
130        occupied[i] = 0;
```

```
131             for (j=0;j<ngenes;j++) prophages[i][j]=0; }
132       induciblesum += inducible;
133     }
134
135     if ((noinduceflag ==0) && (induciblesum == 0)) {
136       noinduceflag =1 ;
137       fprintf(stdout,"There is no more inducible phage in the ...
              population at timestep %d of %d.\n",istep,nsteps);
138     }
139
140 //knockout the genes that have been degraded
141     for (i = 0; i < maxind; i++)
142       for (j = 0; j < ngenes; j++)
143         if (ran3(&seed) < rd)  prophages [i][j] = 0;
144
145     //if all genes from a given prophage have been knocked out ...
          replace the corresponding entry at "occupied" by 0.
146     for (i = 0; i < maxind; i++) {
147       int sumpro = 0;
148       for (j = 0; j < ngenes; j++) sumpro += prophages[i][j];
149       if (sumpro == 0) occupied[i] = 0;
150     }
151
152 //IS insertions change the sequence to -1
153     for (i = 0; i < maxind; i++)
154       for (j = 0; j < ngenes; j++)
155         if (ran3(&seed) < rt)  prophages [i][j] = -1;
156
157
158   j = 0;   // first possible place to put the new prophage
159   for (i=0; i<maxind; i++) {
160     lyscapable = 1;
161     for (k=0; k<nlys+ninf; k++)
162       if ((prophages[i][k]==-1)||(prophages[i][k]==0)) lyscapable = 0;
163     if (lyscapable==1)  // ran3 is expensive.  don't call unless ...
          lyscapable
164       if (ran3(&seed) < rl) {   // make a new copy of prophage[i]
165         while (occupied[j]==1) j++; //find the next empty spot
166         occupied[j] = 1;
167         for (k=0; k<ngenes; k++) prophages[j][k] = prophages[i][k];
168       }
169    }
170    if (j≥maxind) maxind = j+1;
171    if (j>MAXNS) { fprintf(stderr,"Error: MAXNS exceeded\n"); istep = ...
         nsteps;}
172
173 // selection:
174 // put a copy of each prophage into the next generation with ...
      probability rs*(num ben genes)
175 //  newi will count the number of prophages in the next generation
176    newi = -1;
177    for (i=0;i<maxind;i++) {
178      numben = 0;
179      for (j=ngenes-nben;j<ngenes;j++)
```

```
180         if (prophages[i][j]==1) numben++;
181       if (ran3(&seed)<(float)(rs*numben)) {
182         newi++;
183         for (k=0;k<ngenes;k++) {
184           newprophages[newi][k] = prophages[i][k];
185         }
186         occupied[newi]=1;
187         }
188     }
189  /* population size regulation:  Every prophage is copied to the next
190     generation with high probability.  If the current population < Ninit,
191     every prophage is copied.  If the current population > Ninit, the
192     probability is reduced so that on average Ninit are maintained  */
193
194     fractiontolose = 1.0-(float)Ninit/(maxind+newi);
195     for (i=0;i<maxind;i++) {
196       if (ran3(&seed)>fractiontolose) {
197         newi++;
198         ksum = 0;
199         for (k=0;k<ngenes;k++) {
200           newprophages[newi][k] = prophages[i][k];
201           ksum += prophages[i][k];
202         }
203         if (ksum>0) occupied[newi]=1;
204       }
205     }
206     for (i=newi+1;i<maxind+1;i++) occupied[i] = 0;   // after newi, ...
          unoccupied
207     tmpptr = prophages;
208     prophages = newprophages;
209     newprophages = tmpptr;
210     maxind = newi;   // newi is the population size of the next ...
          population
211
212
213     // be sure to use maxind, not Ninit, now that the popn size is ...
          not constant
214     if ((float)istep/ntoprint == (int)istep/ntoprint) {
215       for (jj = 0; jj < maxind; jj++) sizes[jj] = 0;  //initialize
216       for (ii = 0; ii < ngenes; ii++) {
217        tempsum[ii] = 0;
218        tempsumt[ii] = 0;   // for transposase genes
219        sizehist[ii]=0;  // initialize for later
220        for (jj = 0; jj <maxind; jj++) {
221          if (prophages[jj][ii] == -1) {
222        tempsumt[ii]++;
223        sizes[jj]++;
224      }
225      else  {
226        tempsum[ii]=tempsum[ii]+ prophages[jj][ii];
227        sizes[jj] += prophages[jj][ii];
228      }
229        }
230        tempsum[ii] = (float)(tempsum[ii]/((float)maxind));
```

```
231         tempsumt[ii] = (float)(tempsumt[ii]/((float)maxind));
232       }
233       sizehist[ngenes]=0;  // last entry in array didn't get ...
             initialized yet
234       for (jj=0;jj<maxind;jj++)  sizehist[sizes[jj]]++;
235       genemeans[0] = 0; genemeans[1] = 0; genemeans[2] = 0 ; ...
             genemeans[3] = 0;
236       for (i = 0; i < nlys; i++) genemeans[0] = genemeans[0] + tempsum[i];
237       for (i = nlys; i < nlys+ninf; i++) genemeans[1] = genemeans[1] + ...
             tempsum[i];
238       for (i = nlys+ninf; i < nlys+ninf+nneut; i++) genemeans[2] = ...
             genemeans[2] + tempsum[i];
239       for (i = nlys+ninf+nneut; i < ngenes; i++) genemeans[3] = ...
             genemeans[3] + tempsum[i];
240       fprintf(fpout2,"%f %f %f %f ...
             %f\n",delt*istep,genemeans[0],genemeans[1],genemeans[2], ...
241                 genemeans[3]);
242       ismeans[0] = 0; ismeans[1] = 0; ismeans[2] = 0 ; ismeans[3] = 0;
243       for (i = 0; i < nlys; i++) ismeans[0] = ismeans[0] + tempsumt[i];
244       for (i = nlys; i < nlys+ninf; i++) ismeans[1] = ismeans[1] + ...
             tempsumt[i];
245       for (i = nlys+ninf; i < nlys+ninf+nneut; i++) ismeans[2] = ...
             ismeans[2] + tempsumt[i];
246       for (i = nlys+ninf+nneut; i < ngenes; i++) ismeans[3] = ...
             ismeans[3] + tempsumt[i];
247       fprintf(fpout5,"%f %f %f %f ...
             %f\n",delt*istep,ismeans[0],ismeans[1],ismeans[2],ismeans[3]);
248
249       if (sizeflag) {
250         for (i=0;i≤ngenes;i++) fprintf(fpout3,"%d ",sizehist[i]);
251         fprintf(fpout3,"\n");
252       }
253       fprintf(fpout4,"%d\n",maxind);
254     } //end toprint if statement
255     if (sizehist[0] == maxind) {
256         fprintf(stdout,"Prophage population is extinct\n");
257         istep = nsteps;
258     }
259   }  //end of loop on istep
260   for (i = 0; i < maxind; i++){
261       //fprintf(fpout,"%d ",occupied[i]);
262       for (j  =0; j < ngenes; j++) fprintf(fpout," %d ...
              ",prophages[i][j]);
263       fprintf(fpout,"\n");
264   }
265   fclose(fpout2);
266   fclose(fpout);
267   printf("\a");
268 }  // end of main
```

Bibliography

[1] Press, W. H., editor (1992). *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, Cambridge ; New York, 2nd ed edition.

Curriculum Vitae

Name: Amjad Khan

Education:

- 2015–2020 PhD candidate in Applied Mathematics, Western University, London, ON, Canada
- 2013–2015 MSc in Mathematics, McMaster University, Hamilton, ON, Canada
- 2006–2009 MPhil in Mathematics, NUST, Islamabad, Pakistan
- 2000–2004 BSc in Mathematics & Computer Sciences, University of Peshawar, Pakistan

Research Interests:

- Differential Equations
- Mathematical Biology
- Dynamical Systems & Bifurcation Analysis

Publications:

- Khan A., Wahl L. M., Burmeister A. R. The genetic repertoire of prophages. (submitted for publication in PLOS Computational Biology)
- Khan A., Wahl L. M. Quantifying the Forces that Maintain Prophages in Bacterial Genomes. (accepted Theoretical Population Biology)
- Khan A., Wahl L.M., Yu P. (2018) Phage Therapy and Antibiotics for Biofilm Eradication: A Predictive Model. In: Kilgour D., Kunze H., Makarov R., Melnik R., Wang X. (eds) Recent Advances in Mathematical and Statistical Methods. AMMCS 2017. Springer Proceedings in Mathematics & Statistics, vol 259. Springer.
- Khan A., Pelinovsky D., Long-time stability of small FPU solitary waves. Discrete Continuous Dynamical Systems Series A, April 2017, 37(4): 2065-2075. doi: 10.3934/dcds.2017088

Conferences and Poster Presentations

- Systems Modeling in the Pharmaceutical Industry - Problem Solving Workshop. August 12 - 16, 2019, The Fields Institute, Toronto, ON, Canada
- Khan A., Wahl L. M., The Evolutionary Forces Acting on Prophages: A Mathematical Study. Annual Meeting and Conference of the Society for Mathematical Biology (SMB 2019), July 21-26, SMB 2019 Annual Meeting at Montral, Qubec, Canada
- Wahl L. M., Khan A., Blurring the Lines between Predator and Prey: The Evolution of Temperate Viruses. Pokhara, Nepal June 28, 2019
- Khan A., Wahl L. M., Mathematical Model of the Prophage Size Distribution in Bacterial Genomes. "Canadian Society of Applied

and Industrial Mathematics (CAIMS 2018)" June 4 to 7, 2018 at Ryerson University in Toronto, ON

- Khan A., Wahl L. M., Population dynamics of phages and biofilm bacteria. "The IV AMMCS International Conference" Waterloo, Ontario, Canada, August 20-25, 2017
- Khan A., Pelinovsky D., Approximations of the lattice dynamics. April 21, 2015, Department of Mathematics and Statistics, McMaster University, Hamilton, ON.

Teaching

- Teaching Assistant, *Differential Equations, Probability for Life Sciences*, Department of Applied Mathematics, Western University, London, ON, Canada– 2019
- Teaching Assistant, *Calculus with Analysis for Statistics*, Department of Applied Mathematics, Western University, London, ON, Canada – 2018
- Instructor, *Calculus 2*, School of Applied Science and Technology, Fanshawe College, London, ON, Canada – 2018
- Instructor, *Business Mathematics*, Lawrence Kinlin School of Business, Fanshawe College, London, ON, Canada – 2017
- Teaching Assistant, *Applied Mathematics for Engineers*, Department of Applied Mathematics, Western University, London, ON, Canada – 2015, 2016 & 2017
- Teaching Assistant, *Introduction to Differential Equations*, Department of Mathematics & Statistics, McMaster University, Hamilton, On, Canada – 2015
- Teaching Assistant, *Engineering Mathematics*, Department of Mathematics & Statistics, McMaster University, Hamilton, On, Canada – 2014
- Teaching Assistant, *Linear Algebra*, Department of Mathematics & Statistics, McMaster University, Hamilton, On, Canada – 2014
- Teaching Assistant, *Linear Algebra*, Department of Mathematics & Statistics, McMaster University, Hamilton, On, Canada – 2013
- Instructor, *Differential Equations & Transforms*, NUST Institute of Civil Engineering, NUST, Islamabad, Pakistan – 2013
- Instructor, *Numerical Methods*, NUST Institute of Civil Engineering, NUST, Islamabad, Pakistan – 2012
- Instructor, *Calculus and Analytical Geometry*, NUST Institute of Civil Engineering, NUST, Islamabad, Pakistan – 2012
- Instructor, *Probability & Statistics*, NUST Institute of Civil Engineering, NUST, Islamabad, Pakistan – 2011
- Instructor, *Calculus and Analytical Geometry*, NUST Institute of Civil Engineering, NUST, Islamabad, Pakistan – 2011
- Instructor, *Calculus and Analytical Geometry*, NUST Institute of Civil Engineering, NUST, Islamabad, Pakistan – 2010
- Instructor, *Numerical Methods*, NUST Institute of Civil Engineering, NUST, Islamabad, Pakistan – 2009

Awards and Scholarships

- Student paper prize, AMMCS International Conference, Waterloo, Ontario, Canada -August 20-25, 2017.
- Graduate Research Scholarship (2015-2019), Western University, London, Ontario, Canada
- Graduate Research Scholarship (2013- 2015), McMaster University, Hamilton, Ontario, Canada
- Scholarship for M.Phil. studies (2007- 2009), Higher Education Commission (HEC), Islamabad, Pakistan

Technical Skills

Experience with computers and programming languages on Linux and windows operating systems:

- C++
- MATLAB
- TEX (LATEX)
- Maple