# Phage-Host Interaction in Nature

Thesis by

Arbel D. Tadmor

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California
2011
(Defended May 17, 2011)

# Acknowledgments

To many people I owe my gratitude for making my stay at Caltech both fascinating as well as incredibly fun. First and foremost I would like to thank my advisor, Prof. Rob Phillips, who taught me how to be the most critical scientist I can be; Prof. Jared Leadbetter, from whom I learned about microbial diversity and the art of phylogenetics and who graciously agreed to head my committee, and my two other committee members, Prof. David Baltimore and Prof. Victoria J. Orphan, for kindly agreeing to be on my committee and whose questions and critique were critical to the success of my research. I wish to also thank Dr. Eric G. Matson and Dr. Elizabeth A. Ottesen for kindly answering my many questions, and Eric and Dr. Adam Z. Rosenthal for going termite collecting with me on several occasions. I would like to also thank Dr. Blake W. Axelrod, Prof. David Bensimon, Dr. Bertrand Ducos, Prof. Michael L. Roukes and Lijun Xu for fruitful and fascinating collaborations, and Prof. Grant J. Jensen for allowing me access to his imaging facility working closely with Dr. Alasdair McDowall and Dr. Bill Tivol. I am also indebted to the many people who kindly agreed to review our work along the way, including Prof. Sherwood Casjens, Prof. Daniel S. Fisher, Prof. Roger W. Hendrix, Prof. Ron Milo, Prof. Stephen R. Quake, Prof. Edward M. Rubin and Prof. Nathan D. Wolfe. I would also like to thank past and present members of the Phillips group and the Leadbetter group, and especially Dr. Heun Jin Lee, Dr. Martin Lindén, Damien Soghoian, Dr. David Wu and Dr. David Van Valen for the many stimulating conversations and valuable feedback I received from them over the years. Finally I wish to thank Dr. Paul Grayson for introducing me to the world of phages.

# Abstract

Though viruses may be the most abundant biological entities on the planet, very little is known about phage-host interaction in the wild due to the absence of proper experimental tools. In the present work we report of a method to pair environmental phages with their bacterial hosts at the single-cell level without having to culture either host or virus. The method utilizes microfluidic digital PCR in conjunction with a metagenome data mining tool that was developed to find a viral marker gene in an unknown environment. We implemented this technique on the microbial community residing in the hindgut of termites. Consequently, we discovered genus-wide infection patterns displaying remarkable intra-genus selectivity, with viral alleles displaying limited lateral gene transfer and/or host switching despite host proximity. To try and explain phage-host interactions from a theoretical perspective, we formulated a simple biophysical model describing the interaction of bacteria and viruses in aqueous environments. We predict that the radius $r$ of a bacterium is the most critical parameter determining its fixed point concentration, which scales as $r^{-4}$. Given the hypothesis that there is no selection pressure on bacterial radii, our model predicts that the size spectrum of marine bacteria follows a power law with slope -1, close to the observed average spectrum. Moreover, given the total concentration of bacteria in the ocean, our model enables us to estimate the total number of bacterial "species" per volume of water providing a lower and upper bound on the total number of species in the oceans. To elucidate the concept of a "species", we consider a bacterial-viral co-speciation model, which is consistent with the observed narrow host range of phages. Our model hints that the bacterial-viral "arms race" may be a critical component in the process of co-speciation. We suggest further experiments to test both models. Finally, we consider a recent high resolution measurement of the force as a function of time generated by stress fibers within a single fibroblast cell and suggest a stochastic model that is capable of accounting for the observed data.

# Table of Contents

## Chapter 1 Introduction

## Chapter 2 Probing Individual Environmental Bacteria for Viruses Using Microfluidic Digital PCR

## Chapter 3 MetaCAT—Metagenome Cluster Analysis Tool

## Chapter 4 The Biophysics of Prokaryotic and Viral Diversity in Aqueous Environments

## Chapter 5 An evolutionary model of phage-host interaction

## Chapter 6 A Kinetic Model for Stress Fiber Contraction and Relaxation

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.0 Preface

In the following introduction we begin with a brief overview of some basic facts known about phages and their interaction with bacterial hosts. Our purpose is not to exhaustively review the topic, but to introduce certain concepts that will be useful for the remaining chapters, especially Chapters 2, 4, and 5. We then highlight for each chapter the most interesting or promising findings. The remaining chapters of the thesis are organized as follows: Chapter 2 describes an experiment to co-localize phages with their hosts directly from the environment using single cell microfluidic technology. Chapter 3 describes a bioinformatic tool for metagenome analysis that was used in Chapter 2 to identify the must abundant viral genes in the metagenome of a higher Costa Rican termite. Chapter 4 analyzes the problem of phage-host interaction from a theoretical perspective. We first consider a biophysical model describing phage-host interaction of a single isolated phage-host system. We then make the leap to a distribution of phage-host systems in the environment, allowing us to calculate, for example, bounds on the total diversity in the ocean water column. Then, in Chapter 5, we consider the beginnings of an evolutionary model for phage-host co-speciation that we believe has much potential. The key feature of this model is that it is consistent with a "world" where phages have evolved to have a narrow host range. Presently, this model suggests how bacterial "species" and viral "species" are related (thus defining both terms), and hints that the arms race that bacteria and viruses are locked in is perhaps the engine driving bacterial and viral co-speciation, and thus perhaps bacterial evolution itself (a hypothesis). Finally in Chapter 6 we present an analysis of experimental data collected by Dr. Blake W. Axelrod, a research engineer in the Roukes lab, who measured with the highest resolution to date the force as a function of time of a stress fiber in a single fibroblast cell as this stress fiber is artificially disrupted and then allowed to naturally reassemble. Blake observed quantized steps in the force exhibiting exponential like temporal profiles that we can explain by a simple stochastic model, where each sarcomere perfectly obeys a law of exponentially distributed time delays.

## 1.1 Some facts phages in nature

### 1.1.1 Abundance and activity

Viruses may very well be the most abundant biological entities on the planet. In offshore surface waters viral concentrations are typically in the range of $10^5$-$10^6$ $ml^{-1}$, whereas in coastal environments, viral concentrations can reach $10^6$-$10^7$ $ml^{-1}$ [1]. High viral concentrations were also found in sea ice ($10^7$-$10^8$ $ml^{-1}$ [1]), marine sediments ($10^7$-$10^{10}$ $g^{-1}$) [1,2], in soil (~$10^8$ $g^{-1}$) [1] as well as in the rumen gut ($10^8$-$10^{10}$ $ml^{-1}$) [3,4]. Viral concentrations are typically correlated with bacterial concentrations. A variable often used by environmental virologists to gauge this correlation is the virus to bacterium ratio (VBR). The VBR for marine systems is consistently measured to be on the order of 10 [1,5,6,7], making viruses the most abundant life-forms in the oceans. The VBR can also reach as high as ~70 in sea ice [1] or could be as low as 0.04 in soil [1]. A VBR of ~6 is also observed when zooming in on a particular phage-host system. For example, *Synechococcus* cells from the Gulf of Mexico have been shown to be infected by an average of ~6 viral-like particles per bacterium [8].

Virions are also extremely active in the environment. It has been estimated that ~20% of all marine microbial cells, which constitute over 90% of the viable biomass in the Earth's oceans, are turned over daily by viral predation [6]. In the deep-sea, viruses are thought to be responsible for at least 80% of prokaryote mortality (calculated by taking the ratio of viral production and prokaryotic burst size) [9]. The same is true for low-oxygen lake waters (in which grazers do not thrive) where viruses are thought to be responsible for 50–100% of bacterial mortality [5]. Such high viral-induced mortality suggests that many if not most bacteria die from viral infection. For example, in environments where viral lysis accounts for 50–100% of the bacterial mortality, either every bacterial cell or every second cell will be lysed by a virus in order to maintain a steady state population of bacteria.

In terms of their life expectancy in the wild, marine viruses can survive only about 1 to 10 days without having to "feed". Viral decay rates of ~0.1 to ~2 $day^{-1}$ have been measured for inshore and offshore regions, respectively [8], with comparable decay rates in deep-sea sediments [2].

### 1.1.2 Lytic or lysogenic?

Viruses are also very abundant in the form of lysogenic viruses, with an estimated ~60% of sequenced bacterial genomes encoding at least one integrated viral element [10,11]. However, one might expect that with viruses being so abundant in nature and having such a major contribution to bacterial mortality, that the observed viral-like particles are lytic viruses. If these viruses were lysogenic, they would probably need to be continuously induced in order to reach the observed levels of abundance, obviating the need to integrate or to encode a genetic switch. Indeed, growth experiments with native bacterial communities in freshly filtered sea water indicated that under typical natural conditions, induction of lysogens was rare, with the vast majority of the observed viruses probably the result of successive lytic infection [5,12]. Furthermore, attempts to induce lysogens with bright continuous sunlight or pulsed sunlight did not result in increased viral concentrations [5,12] suggesting that lysogens are not easily inducible. Therefore, it appears that lysogenic induction may be occurring at low levels either continuously or sporadically [5] (possibly occasionally on larger scales [1]), with the vast majority of viruses in the sea probably the result of lytic infection [5].

## 1.2 Phage-host interaction

### 1.2.1 Predator-prey dynamics

Phages have effect on bacteria in many different levels and vice versa. Our intention here is not to give a comprehensive review of all mechanism of interaction between phages and bacteria, but to highlight a few important concepts used in later chapters. The most basic level that viruses affect bacteria is through concentration control. In a classic case of a predator-prey dynamical system (one predator-one prey), the fixed point concentration of the prey is determined by the predator. Therefore the fixed point concentration of the prey does not depend on its growth rate. As long as there is positive growth of the prey, its final concentration will be the same. Therefore, if a bacterial species has a very low growth rate, the concentration of viruses infecting it will be low. Conversely, if the bacterium grows very fast, the concentration of the viruses infecting it will be very high.

### 1.2.2 Population control versus species control

It is generally accepted that bacterial host mortality is primarily due to either grazing by protists or lysis by viruses [5,13,14,15]. The fundamental difference between these two predators is that protists, to a first-order approximation, are omnivores, i.e., they are not host specific [15,16], while phages display a species- or strain-level host range [1,17,18]. Protist therefore either control the *total* bacterial concentration (sum of all species), or — if they are themselves prey — do not exert control over bacteria [15] and simply reduce the bacterial production rate (with bacterial concentration being determined by competition for nutrients [15]). Viruses on the other hand exert control at the species level. Therefore, through predator-prey dynamics, viruses directly control the genetic diversity of bacteria in the environment.

### 1.2.3 Kill the winner hypothesis

In nature, every environment contains many species of bacteria. Given the narrow host range of phages, to a first-order approximation, we can think of this environment as comprised of a collection of non-interacting phage-hosts systems[1]. Given the individual predator-prey dynamics, based on our explanation above, we expect the concentration of each bacterial species to be controlled separately and be independent of the growth rate of the bacteria. By having viruses control the population in this way, fast growing cells will not be "allowed" to take over the population. If a bacterium's growth rate increases, the concentration of the viruses infecting it will also increase, thus keeping the (fixed point) concentration of that bacterium in check. Thus the equilibrium diversity in these networks is maintained by mechanisms that are selectively ''killing the winner.'' (i.e., a superior competitor) [16,19,20].

### 1.2.4 The bacterial-viral "arms race"

Recently it has been discovered that bacteria have a primitive immune system in the form of CRISPRs — clustered regularly interspaced short palindromic repeats — arrays found in nearly half of all sequenced bacterial genomes [21]. Short (26–72bps [21]) "spacer" sequences derived

---

[1] In Chapter 4 we show that the situation of more than one virus controlling the same bacterial species leads to extinction events.

from viral genes, present between the CRISPRs, are transcribed and interfere with viral gene expression in a mechanism thought to be similar to RNA silencing [21,22]. Bacteria continuously acquire CRISPR spacer sequences from viruses to evade these viruses. To evade new acquired spacers, the viruses rapidly evolve their genes though mutation, homologous recombination and deletion [23]. Conversely, CRISPR repeats and their associated proteins undergo evolution to escape a shut-down mechanism for the CRISPR system encoded by the phage [21]. Thus, bacteria and viruses are locked in an arms race [21]. This arms race may have long term evolutionary consequences on the bacterial population. From inspection of the history of spacers stored on the bacterial genomes of many individuals in a population it has been observed that all individuals can have essentially the same older spacers, with the new diverse set of spacers at the tip of the array, where new spacers are added [23]. One explanation for this observation could be a recent strong selection event caused by an unusually virulent virus to which potentially only one cell in the population was immune [23].

## 1.3 A coarse-grained view of phage-host interaction

### 1.3.1 The biophysics of a single phage-host system

#### Two perspectives: biophysical versus dynamical

From a dynamical point of view phage-host systems can be analyzed as a classic predator-prey problem. This type of problem has been studied extensively and is considered a textbook problem. From a biophysical perspective, the problem of phage-host interaction is that of viral transport. While this problem seems difficult to address in environments like soil or sediment, in aqueous environments the problem of viral transport can be reduced to solving the diffusion equation. This intuition did not escaped biophysicists who worked with viruses in the early days, and the first solution to this problem appeared in the book of Stent “Molecular Biology of

Bacterial Viruses" in 1963. Although both perspectives are known and are made use of, we have not seen in the literature an attempt to merge these two perspectives in one package, obtaining expressions for the concentrations of bacteria and viruses in terms of basic biophysical parameters such as temperature, viscosity, radii, and so on. We have also not seen any model attempting to exploit the empirical correlation between burst size and the volume of the bacterium and the inverse correlation between burst size and the volume of the virion (with empirical correlations measured up to 1 μm) [1,24,25]. Such correlations can have great implications on the scaling laws of these systems, and may be critical when attempting to draw conclusions on an entire community.

## Combining the two perspectives leads to new insight

In Sections 4.1–4.3 we construct a new biophysical model describing the interaction of a single isolated phage-host system and obtain interesting scaling laws for the steady-state concentration of bacteria and viruses. We find that the most critical parameter determining the fixed point concentration of a phage-host pair in the environment is the radius of the bacterium (Fig. 1.1). We found that the fixed point (i.e., steady state) concentration of bacteria scales as $r^{-4}$ with $r$ being the radius of the bacterium. Since in nature, the radii of bacteria vary by over three orders of magnitude, our model predicts that the concentration of bacteria can change by over 13 orders of magnitude! Furthermore, our model predicts that large bacteria will be exceedingly rare, with the largest known bacterium (*Thiomargarita namibiensis* having a diameter of 750 μm) predicted to have one cell in ~$10^3$ liters of water. On the other hand, we predict that the concentration of the viruses infecting large bacteria will be high enough so that, using molecular techniques, these viruses should be detectable in one ml of water.

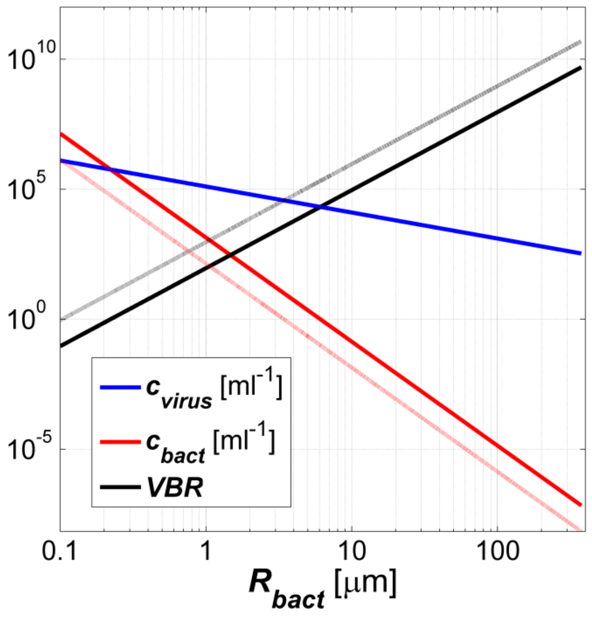

**Figure 1.1. Scaling of the virus concentration, the bacterium concentration and the VBR with the radius of the bacterium for a single phage-host system.** This figure shows that the radius of a bacterium is a critical parameter determining the fixed point concentration of the bacterium.

### 1.3.2 The biophysics of many phage-host systems

#### How many is many?

Thus far we have dealt with an artificial problem of a single phage-host system. In nature there are many such systems and in Section 4.4 we deal with the question of how to make the transition from a single phage-host system, to many such systems in the environment. How many is many? In the Venter expedition to the Sargasso Sea, every sample containing several hundreds of liters of ocean water was found to have at least 300 bacterial species (using a cutoff equivalent to a small subunit rRNA cutoff of 3%) [26]. Therefore, we expect that a natural environment will contain at least hundreds (probably thousands) of phage-host systems. Since the host range of phages is narrow, these phage-host systems can be treated, to a first-order approximation, as independent. Our realization from Section 1.3.1 that bacterial radii span such a wide range of values, and that the fixed point concentration of bacteria is extremely sensitive to this parameter, suggested to us that one cannot simply replace this parameter with an average value. One actually needs to calculate this average using the probability density function of bacterial radii for the given environment.

## A simple evolutionary scenario

The difficulty in making the transition between a single phage-host system and many phage-host systems in the environment, is figuring out what is the *a priori* probability density of radii in a given environment, i.e., the probability per radius that a given environment *a priori* would contain a bacterial species with this radius. This function (which we denoted by $f_R(r)$) has evolutionary significance and can be interpreted as the density of bacterial species, perhaps analogous to the density of states in statistical mechanics, and reflects the evolutionary history of bacteria in the given environment. If the radii of all bacterial species that have adapted to survive in the given environment were known, one could calculate this function. Since we cannot calculate this function from first principles, we considered the simplest evolutionary scenario where this function is a constant, which means there is no selection pressure on bacterial radii, i.e., all radii are *a priori* equally probable. Given this assumption we were able to obtain expressions for basic quantities, such as, the total concentration of bacteria in a given environment, the total concentration of viruses in a given environment, the VBR, and the total bacterial biomass in a given environment. These results are especially interesting given that they are calculated from basic physical parameters such as temperature, viscosity, radii, and so on.

### The size spectra of bacteria in the ocean

One additional quantity that we can calculate given $f_R(r)$ is the distribution of radii in a given environment, and from this function we can easily calculate the probability that a bacterium of random volume *V*, is greater than or equal to a given volume, *v,* or $\text{Prob}(V \geq v)$. This function is called the size spectra of radii and has been of interest to marine biologists for decades, with measurements dating back to the work of Sheldon in 1972 [27]. In 2001 the Chisholm lab from MIT measured the size spectra of microbes in the western north Atlantic Ocean. They found that the size spectra obeyed a power law with a slope between -1 and -1.4. The ensemble average of all environments was well described by a power law of slope -1.2. When expanding their dataset to include microzooplankton the slope was corrected to a value close to -1. Our calculation, given our simple evolutionary scenario, predicted a power law with a slope of -1, hopefully indicating that we are on the right track.

### Species richness

In section 1.2.2 we mentioned that the total concentration of bacteria is determined either by the protists or by the availability of nutrients [15]. Thus, given the total concentration of bacteria in an environment, one can in fact calculate the number of predicted species. By considering two extreme models of spatial distribution of diversity — complete homogeneity and maximal heterogeneity — we were able to calculate bounds on the total diversity in the ocean water column. In Section 4.4 we also explore how the number of species scales with basic parameters and found that, quite intuitively, warm, nutrient-rich environments where viruses have a long lifetime will sustain the greatest diversity of species. Finally we compared estimates of diversity with observations from metagenome studies.

**What is a species?**

When applying our model to data we realized that there is something strange about our biophysical model. No where did we define what a "species" is! How different do two genomes (bacterial or viral) need to be in order to be considered different "species"? It is this question that we tried to address in Chapter 5.

## 1.4 The evolutionary perspective

### 1.4.1 A model for co-speciation of viruses and bacteria

In order to answer what a bacterial or viral species are, one needs to go to a higher theory that takes into account the genetic aspect of these entities, and not just parameterize them with a radius, decay rate, and so on. We therefore sought to formulate an evolutionary model that could hopefully supply us with a definition of what is a species. This model needed to respect a few basic rules so that it would be equivalent to our biophysical model. These rules were basically that: (1) each bacterial species was associated with a single viral species and vice versa (i.e., there is no cross interaction between phage-host systems) and (2) each species (bacterial or viral) was unique and distinguishable from all other species. We then asked ourselves the following question: if we start from a state of a single bacterial species interacting with a single viral species, how would this state evolve so that after some time we obtained a state comprised of two bacterial species and two viral species, where the new species were independent of the old species. Such an evolutionary model would create a "world" with single viral species paired with single bacterial species, and vice versa, and where each bacteria-viral species pair was independent of all other pairs. We found that in order for these strains to evolve we needed to (1) define the concept of a "strain", which is like a species, only there is no restriction on whom this strain can or cannot interact with, and (2) assume that whenever a new bacterial strain emerges, a corresponding viral strain co-emerges so that the symmetry between bacteria and viruses is conserved. By considering how a species evolved through generation of strains, into a new species, a qualitative picture of what a species is, within the context of this model, emerged (Fig. 1.2; see also Fig. 5.1 that illustrates the process of co-speciation). How this complex structure was obtained is discussed in detail in Chapter 5. When this structure is viewed in a genetic coarse-grained way we recover the simple picture of our biophysical model: species interacting uniquely with species. Therefore we argue that this model can be used to interpret the results of the biophysical model. Such a situation is often encountered in physics, where one theory is the limiting case of another (such as nuclear physics' versus particle physics' description of a proton). Such limits are related to scale transformations in renormalization group theory, possibly suggesting a deeper connection between the two models.

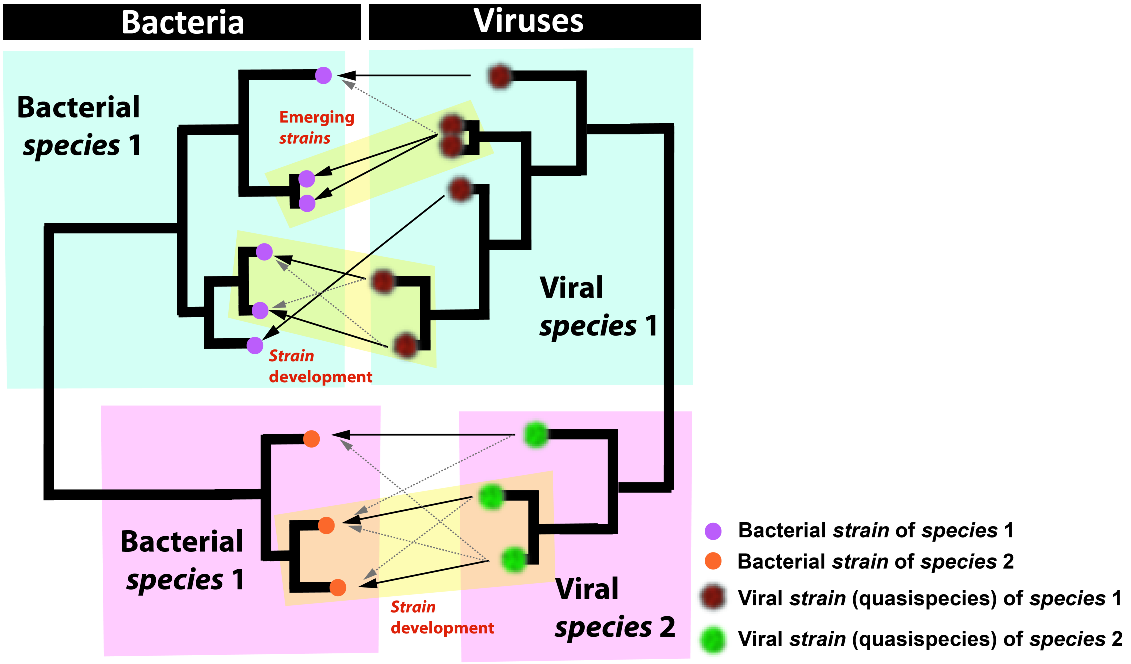

**Figure 1.2 Schematic depiction of bacterial and viral species and strains.** The relation between bacterial and viral species and strains according to a postulated evolutionary model considered in Chapter 5. Each bacterial species interacts with a single viral species. Bacterial strains on the other hand (that are simply emerging bacterial species) are initially part of a mesh of interactions with other strains. The interaction of the bacterial strain with the co-evolving viral strain is critical in order for both strains to evolve away from this state into a state of mutual independence (emerging as new species).

### 1.4.2 Is positive feedback driving co-speciation?

Perhaps the most interesting finding of this model was that in order for a new bacterial species and new viral species to co-emerge, the emerging bacterial and viral strains may be driving each other's evolution, through a positive feedback evolution mechanism. This positive feedback causes the strains to evolve as fast as possible from their initial state in order to sever the bonds with their parental strains and become independent (Fig. 1.3). This positive feedback evolution mechanism is the arms race between bacteria and viruses. The logic behind our suggestion is the following:

1. Phages for some reason have converged to an evolutionary solution where they have a narrow host range. This is not the most beneficial solution for a parasitic element, as a wide host range, such as that of a grazer, would be much more effective. Therefore, there appears to be some evolutionary advantage to this solution.
2. In the process of the arms race, viruses cause selective sweeps in the bacterial population. Such bottlenecks are known to accelerate evolution as traits in small populations can be fixed quickly. Thus, the phage is driving bacterial evolution, distancing the new bacterial strain as fast as possible from the bacterial strain from which is was born (Fig. 5.1). This evolution is necessary for the emerging bacterial species since this will lead the emerging viral species to lose its affinity to the parental bacterial strain and gain control over it. As it gains control, the concentration of the emerging bacterial species increases (Fig. 1.4). The concentration of the emerging bacterial species is maximal when the new viral species has total control over it. Thus, this process allows the emerging bacterial species in the end to "take up" its own concentration.
3. As the new viral species is emerging, it is controlling two populations, the parental bacterial species and the new bacterial strain (Fig. 5.1). In order for this phage to form a unique association with the new bacterial species (i.e., control only it) it must evolve away from its current state as far as possible until it can no longer infect the parental bacterial strain. This process appears to be achieved through the bacterial-viral arms race, since the new bacterial strain is forcing the virus to keep muting in order to track the new bacterial strain and the virus is causing the new bacterial strain to evolve.
4. Combining 2+3 we conclude that perhaps through positive feedback the bacterial and viral species are moving at the greatest possible pace from an initial state of one species to a final

state of two species (Fig. 5.1). Thus, viruses are the tool of evolution to generate species, and the narrow host range of phages is necessary to achieve this goal.

### Chaotic evolution?

Since a positive feedback mechanism amplifies noise exponentially (like the shrill of a microphone in front of a speaker), the process of bacterial speciation may be simply a process of “amplifying noise”. This may open the door to quantitative analysis via chaos theory. For example, it would be interesting to see if phylogenic trees of bacteria spanning many orders (strain, species, genus, family, order, etc.) display any features of self similarity, the hallmark of fractals generated by chaos theory. Further quantities that may be tractable are the rate speciation and the number of strains per species.

An experimental system to test the predictions of this theory would be Lenski-type evolution experiments with E. coli + a lytic phage. Specific experiments are suggested in section 5.4. In section 4.5 we suggest a series of experiments to test our biophysical model.

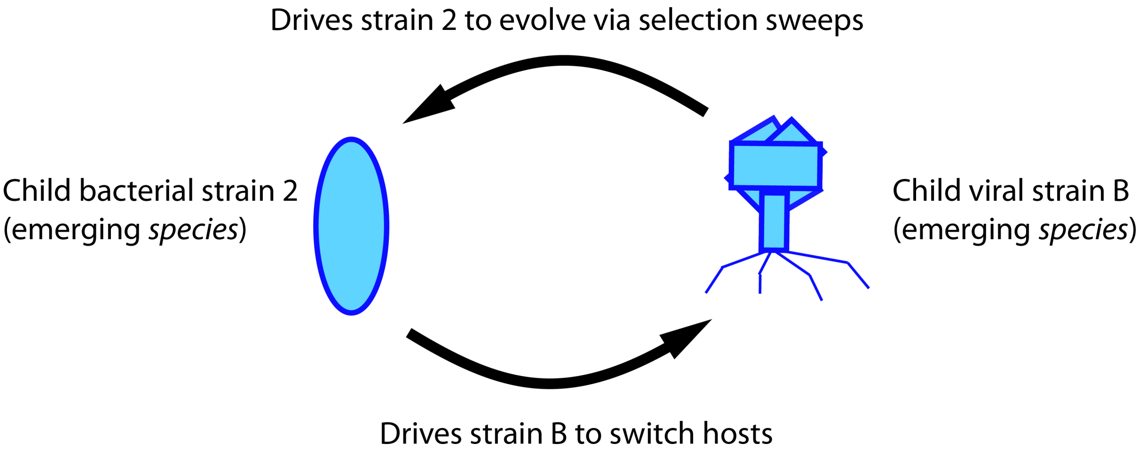

**Figure 1.3 Positive feedback evolution model for emerging bacterial and viral "species".** The arms race between bacteria and viruses may be a critical step in the formation of a new bacterial and viral "species". This process is critical in order to allow viral strain B to relinquish its control of its parental bacterial strain (strain 1) while at the same time gaining control over the new bacterial strain (strain 2). Therefore this "arms race" may allow the two emerging "species" to form a one-to-one association, leading to the result that vial species have a narrow ("species") host range. This process may also be critical for the bacterium, where by selective sweeps, the controlling child viral strain drives the bacterial strain to evolve away from its original parental strain. This positive feedback model may be initially amplifying "noise". Thus the process of co-speciation is perhaps equivalent to "amplification of noise" and thus may be a chaotic effect. Covering the genome space at such an exponential rate may be required in order to converge to a solution on a practical timescale, especially given the fact that bacteria are much less efficient and exploring this space than diploid organisms. Thus, the arms race may be an equivalent solution of bacteria to sexual reproduction (possibly a good enough solution for a smaller genome size).

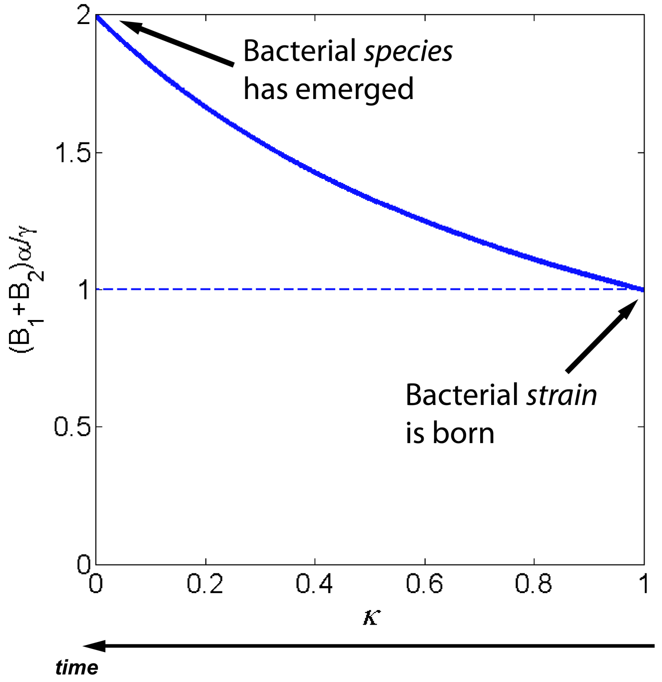

**Figure 1.4. Total concentration "taken up" by the evolving bacterial strain and its parental strain.** Initially, the total concentration of the parental bacterial species ($B_1$) and the just-emerging bacterial species ($B_2$), in normalized units, is 1 and is determined by the controlling viral species. As the new bacterial strain is emerging, it is driving the evolution of the emerging viral strain, causing its affinity to the parental bacterial strain to drop (i.e., κ, which is a measure of the affinity of the emerging viral species to the parental bacterial species, decreases). This causes the emerging viral species to gain more control over the emerging bacterial species, and so its concentration increases. When co-speciation is completed, the new viral species has total control of the new bacterial species and has lost its affinity to the parental bacterial species (i.e., it has a species host range). This allows the total concentration to double (i.e., $B_1$=$B_2$=1). See also Fig. 5.4.

## 1.5 The experimental frontier

### 1.5.1 Phage-host co-localization methodology

Thus far, phage-host interaction in the wild could only be investigated for certain systems such as cyanophages [28,29,30,31]. The challenge lays in the fact that traditional techniques in microbiology necessitate that hosts be culturable in order to isolate their phage. Yet when >99% of bacteria cannot be cultured [32] other methods need to be sought. In Chapter 2 we describe a method using digital microfluidic PCR array to pair phages with their bacterial host without having any prior assumptions regarding the host. The experimental scheme for a new environment is shown in Fig. 1.5.

The first stage involves obtaining a metagenome for the environment of interest. Once gene objects have been assembled and translated, one can run a bioinformatic tool called MetaCAT (Chapter 3) that was written for this purpose. MetaCAT (metagenome cluster analysis tool) is used to find the most abundant viral genes in a given metagenome by clustering together similar genes in the metagenome that are expected to be related (Fig. 1.6). This tool is used to find candidate viral marker genes. The idea behind using this tool for viral genes is that viral genes tend to have many mutations, and therefore are not collapsed by the assembler. Therefore we expect that the abundance of the genes in the metagenome reflects their abundance in the sample. Thus abundant viral genes found by MetaCAT would correspond to abundant and thus dominant genes in the sample. Furthermore, the more alleles one has for primer design, the more general the primers will be, and the more diversity they will recover.

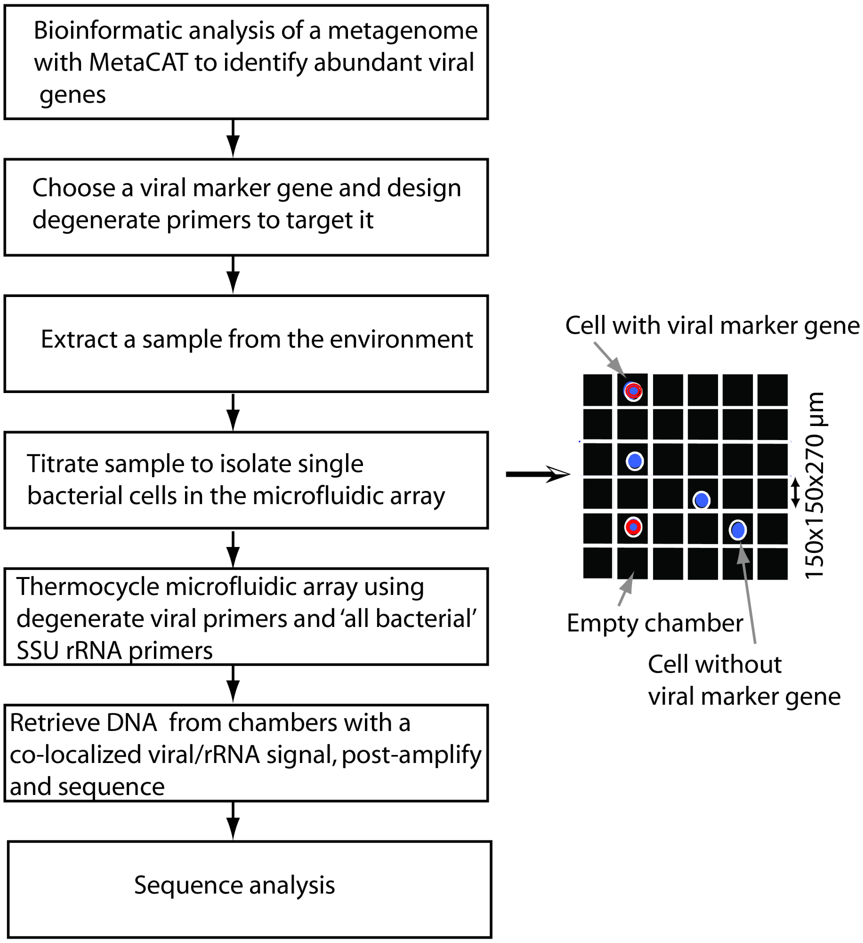

**Figure 1.5 Workflow using the microfluidic digital PCR array for host-virus co-localization in a novel environmental sample**

After degenerate primers have been designed, one can load an environmental sample onto a digital PCR microfluidic array panel, which distributes the sample evenly among 765 6 nl chambers. Samples are titered such that a small fraction of chambers contain a single cell that is probed for a universal small subunit rRNA gene and a viral marker gene. In Fig. 1.7 we show a typical digital PCR panel after PCR cycling. Each chamber that contains both colours (red for a viral marker gene and green for the small subunit rRNA gene) is a potential co-localization signal. Chambers displaying co-localization are retrieved and sequenced allowing later phylogenetic analysis. The great challenge with this experiment was that the phage gene displayed many mutations from cell to cell of the same host species. Therefore we needed to devise a statistical criterion and sampling strategy to separate repeated co-localizations due to chance from genuine co-localization.

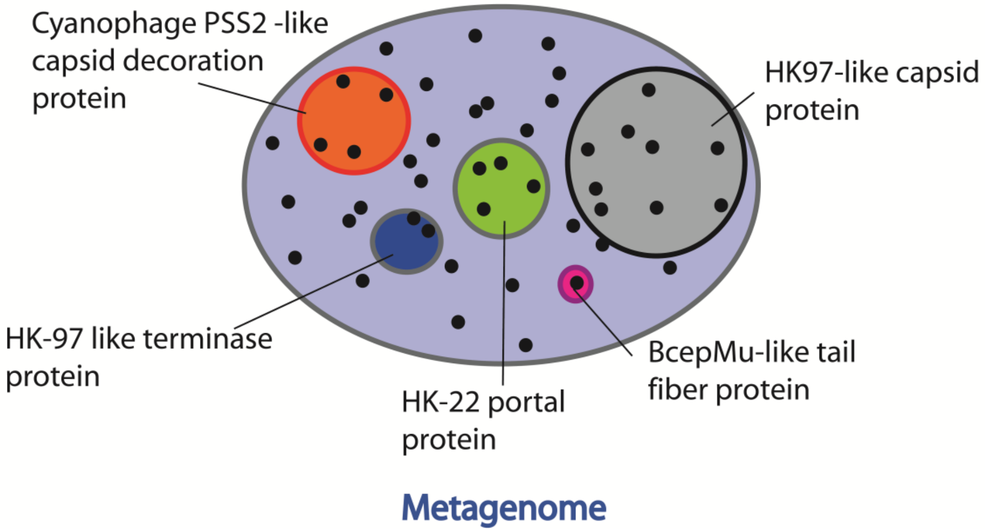

**Figure 1.6 Ideal clustering of gene objects in a metagenome.** Each dot represents a gene object in a metagenome, with the entire metagenome depicted by the blue oval. Similar genes are grouped into clusters (circles of different colors) and each cluster is represented by a single gene from a known reference database. In this schematic description, the distance between dots is interpreted in an abstract manner and does not correspond to a rigorous metric.

### 1.5.2 The case of the termite hindgut

The co-localization experiment described in Chapter 2 was performed for samples from the gut of a termite. Our analysis of the termite hindgut began by analyzing the metagenome of a higher termite collected from Costa Rica. This analysis detected several highly abundant unique viral genes. We then BLASTed these genes against the genomes of two spirochetes that were isolated from a lower termite collected from northern California. We found that two genes had very close homologs: a portal protein and a terminase protein. These two genes were part of larger prophage-like elements (two elements in each genome). We then proceeded to uncover the entire

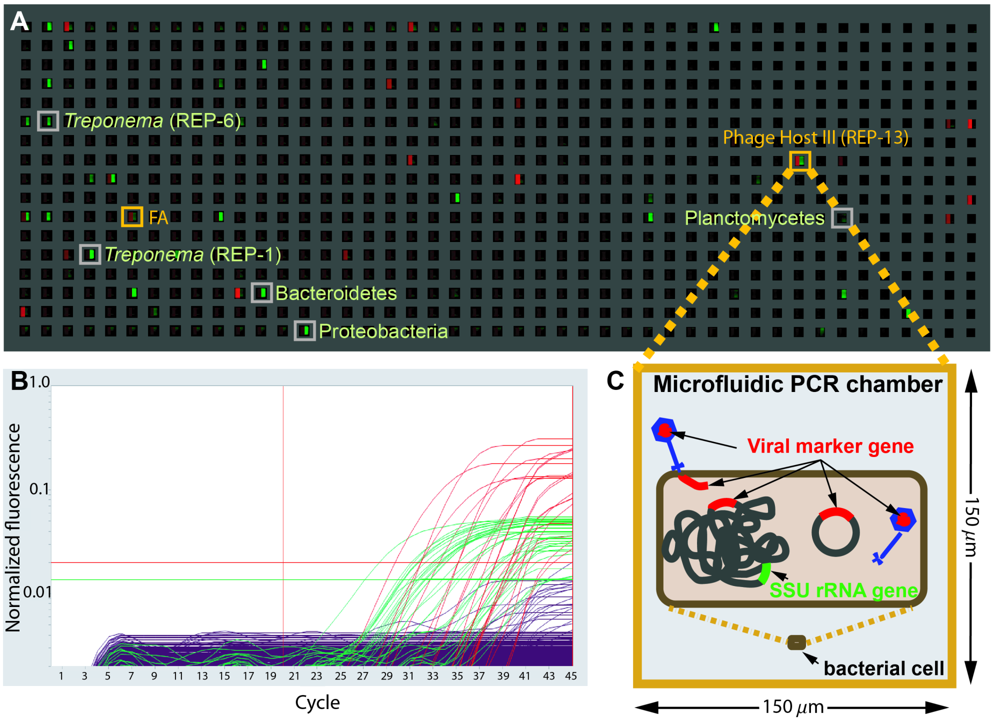

**Figure 1.7. End-point fluorescence measured in a panel of a microfluidic digital PCR array. A.** The measured end-point fluorescence from the rRNA channel (right half of each chamber) and the terminase channel (left half of each chamber) in a microfluidic array panel. **B.** Normalized amplification curves of all chambers (red/viral, green/rRNA). **C.** Specific physical associations between a bacterial cell and the viral marker gene resulting in co-localization include for example: an attached or assembling virion, injected DNA, an integrated prophage or a plasmid containing the viral marker gene.

prophage-like element in each genome (Fig. 1.8). To show that these prophage-like elements were also abundant in the metagenome we BLASTed each gene from in the prophage-like element against the metagenome. The result, shown in Fig. 1.8, indicated that these prophage-like elements were indeed abundant in both termites. We chose the terminase gene to be our viral marker gene and designed degenerate primers to amplify a large portion of this gene (~820 bp). To test the hypothesis that the prophage-like element that we found is ubiquitous in termites, we tested these primers against nine termite species belonging to seven families collected from five different geographical locations. Fig. 1.9 shows that indeed we obtained positive hits for all these termites confirming that this prophage-like element is ubiquitous to termites (at least of north and central America). We also obtained a positive hit for a wood feeding roach, raising the possibility that this prophage-like element has infected a common ancestor of termites and wood feeding roaches and has been transformed since. In Chapter 2 we describe the results of our co-localization experiment gut samples extracted from *Reticulitermes hesperus* using the same viral primers.

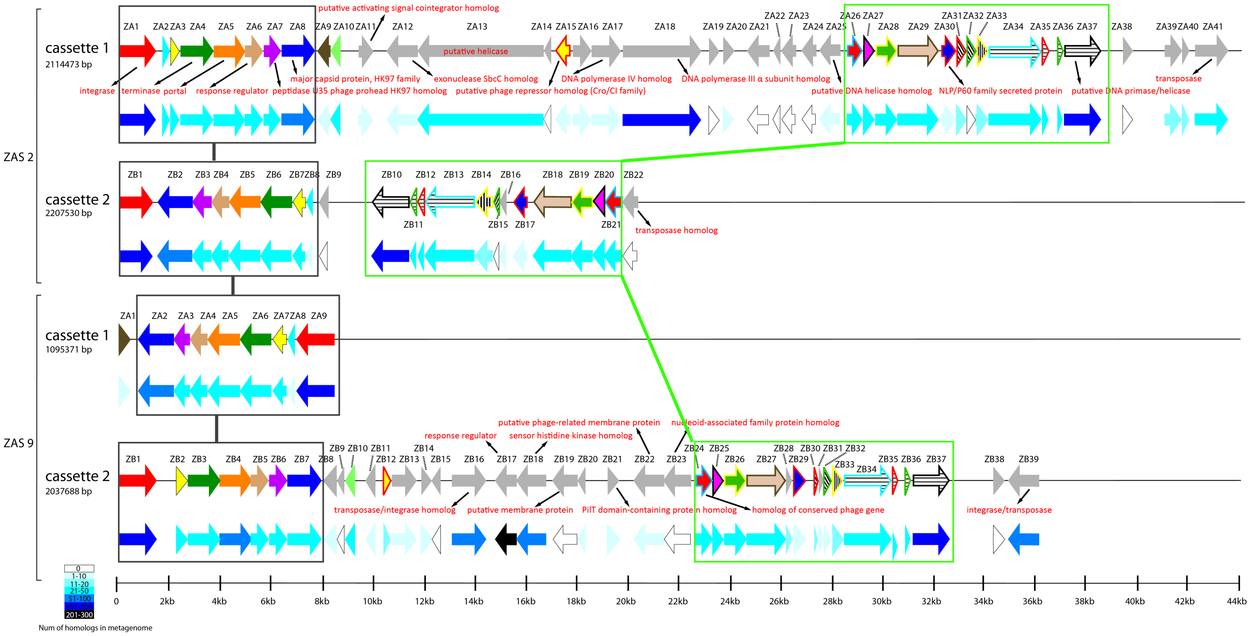

**Figure 1.8 Map of viral cassettes in ZAS2 and ZAS9 highlighting gene frequency in the higher termite metagenome.** Blue arrows represent abundance in the metagenome.

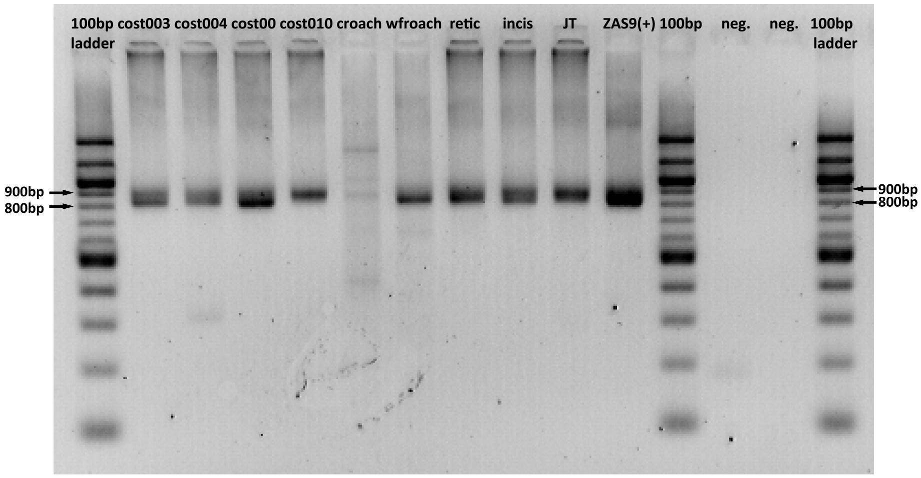

**Figure 1.9. Agarose gel electrophoresis analysis of terminase PCR product amplified from termite and related insect species.** PCR product using degenerate primers ter.7F and ter.5eR targeting the large terminase subunit gene. Specimens included were: *Nasutitermes sp.* (cost003), *Rhynchotermes sp.* (cost004), *Microcerotermes sp.* (cost008), *Amitermes sp.* (cost010), *Periplaneta americana* (croach), *Cryptocercus punctulatus* (wfroach), *Reticulitermes hesperus* (retic), *Incisitermes minor* (incis), Gnathiamitermes sp. JT5 (JT). ZAS9 was used as a positive control. Also shown are two negative PCR controls.

## 1.6 Stress fibers in single fibroblast cells

Dr. Blake W. Axelrod, a research engineer in the Roukes lab, built a microfluidic NEMS device allowing him to measure the force as a function of time of a stress fiber in a single fibroblast cell, performing the highest resolution measurement to date. In this experiment, a single fibroblast cell (Fig. 1.10A insert) contacts a NEMS force sensor. When the cell is placed in a recovery medium it exerts a force on the force sensor (Fig. 1.10A, blue region) corresponding to the force generated by an assembled stress fiber (about 20nN). Once a substance called cytochalasin D is flowed in, the stress fiber undergoes disassembly and consequently the force declines (Fig. 1.10A, red region). The process of disassembly is a reversible one, since when flowing the recovery medium back again, the stress fiber assembles again and the force is regenerated. When examined more closely, each assembly/disassembly profile appears to be comprised of steps that appear to, on average, increase in duration (Fig. 1.10B). These steps were remarkably uniform in the force amplitude and it was postulated that they are the result of individual sarcomeres failing or contracting.

At the time the data was presented it was not clear what the origin of the temporal dynamics is, what the mechanism leading to exponential-like “charging” and “discharging” curves is, and why the steps are increasing with time. In Chapter 6 we present our analysis of Blake’s dataset. We proposed a simple stochastic model for stress fiber assembly and disassembly, whereby individual sarcomeres assemble or dissemble (a) abruptly, (b) irreversibly, (c) independently, (d) with the time to the event of assembly or disassembly exponentially distributed with a fixed time constant (Fig. 1.11). With this model it is simple to explain why, for example, steps increase in time. According to this model, in the case of stress fiber disassembly for example, the time from the perturbation (t=0) until a sarcomere fails is exponentially distributed. If there are N sarcomeres, then at t=0 there are N independent sarcomeres that can fail. Thus one needs to wait a short period of time to observe a step. As more steps fail, one needs to wait longer until one sees a step because there are less remaining sarcomeres. In Chapter 6 we show that the inverse duration times of each step increase linearly with time. In Fig. 1.12 we show the inverse duration times versus the linear prediction of the model (where parameters were estimated from the data based on the model). The data appears to be behaving qualitatively as predicted. In Chapter 6 we present more rigorous tests to check our model. We show the (a) sarcomeres appear to be failing or assembling statistically as exponential variables. (b) When data was rescaled using model parameters estimated from the data, all profiles collapsed to the predicted ensemble average.

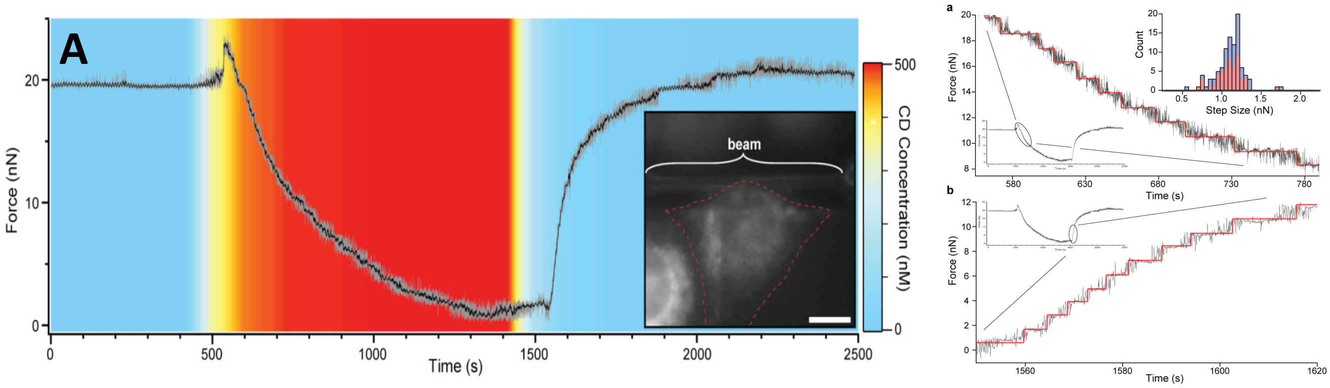

**Figure 1.10 Typical force-time response to Cytochalasin D perturbation. A.** Typical measured force response to the force disruptor Cytochalasin D (red region) and recovery medium (blue region). Cytochalasin D belongs to a class of substances called Cytochalasins that are fast acting and reversible disruptors of contractile force. When Cytochalasin D is flowed in, the force decays due to stress fiber disassembly. When recovery medium is flowed in, the stress fiber reassembles and the force increases with time. Inset shows fluorescent image of the cell attached to the beam taken immediately before data acquisition, scale bar is 10 µm. **B.** Steps during CD-induced force collapse (upper) or force recovery (lower). The average step size is 1.08 nN ± 0.18 nN (*n*=96). Figure and caption courtesy of Blake Axelrod (Roukes lab, Caltech).

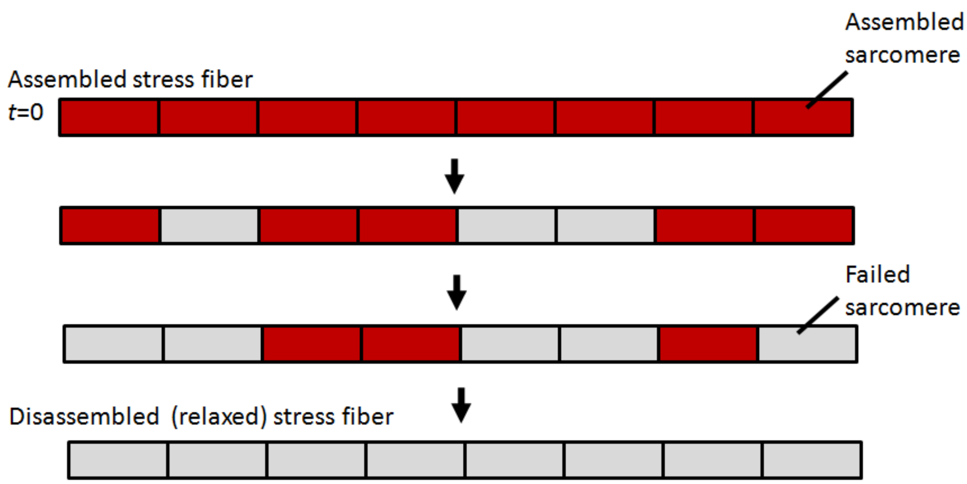

**Figure 1.11. Schematic model for stress fiber relaxation.** (1) Each sarcomere assembles or disassembles abruptly and irreversibly. (2) Sarcomeres assemble or disassemble independently of each other. (3) The time until a sarcomere assembles or disassembles is exponentially distributed (reflecting a certain constant probability rate for this event to occur). (4) The time constant for assembly is the same for all sarcomeres. Similarly, the time constant for disassembly is the same for all sarcomeres.

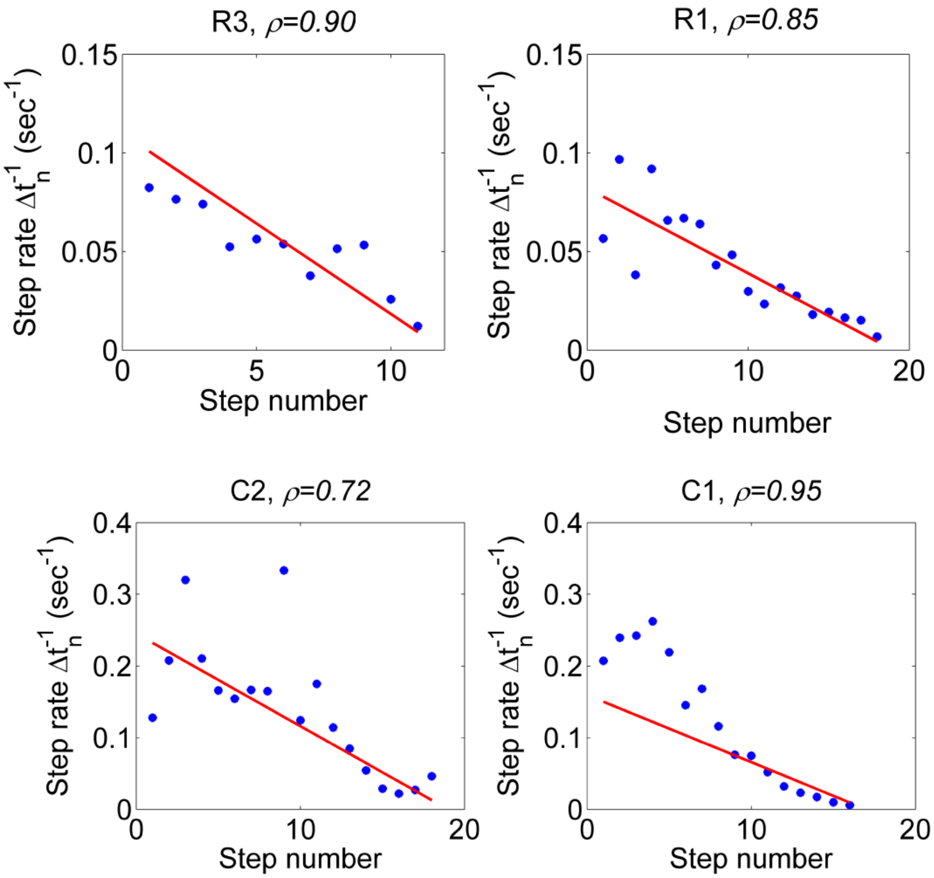

**Figure 1.12. Stochastic model prediction of force step durations versus experimental data.** The stochastic model prediction for the inverse of the force step durations (red curve) versus experimental data points (blue dots). $\rho$ is the Pearson correlation coefficient measuring the strength of the correlation. The red line is not a fit, as these lines were predicted based on parameters that were estimated from the data according to our stochastic model. Note that we anticipate a high level of noise since the standard deviation equals the predicted rates.

## 1.7 References

1. Weinbauer M (2004) Ecology of prokaryotic viruses. FEMS Microbiology Reviews 28: 127-181.
2. Danovaro R, Corinaldesi C, Luna GM, Magagnini M, Manini E, et al. (2009) Prokaryote diversity and viral production in deep-sea sediments and seamounts. Deep Sea Research Part II: Topical Studies in Oceanography 56: 738-747.
3. Klieve AV, Swain RA (1993) Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. Applied and Environmental Microbiology 59: 2299.
4. Kamra D (2005) Rumen microbial ecosystem. Current Science 89: 124-135.

5. Fuhrman J (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399: 541-548.
6. Suttle C (2007) Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5: 801-812.
7. Wommack K, Colwell R (2000) Virioplankton: viruses in aquatic ecosystems. Microbiology and Molecular Biology Reviews 64: 69.
8. Suttle CA, Chan AM (1994) Dynamics and distribution of cyanophages and their effect on marine Synechococcus spp. Applied and Environmental Microbiology 60: 3167.
9. Danovaro R, Dell'Anno A, Corinaldesi C, Magagnini M, Noble R, et al. (2008) Major viral impact on the functioning of benthic deep-sea ecosystems. Nature 454: 1084-1087.
10. Edwards R, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3: 504-510.
11. Casjens S (2003) Prophages and bacterial genomics: what have we learned so far? Molecular Microbiology 49: 277-300.
12. Wilcox R, Fuhrman J (1994) Bacterial viruses in coastal seawater: lytic rather than lysogenic production. Marine Ecology-Progress Series 114: 35-35.
13. Suttle C (2005) Viruses in the sea. Nature 437: 356-361.
14. Paul J, Kellogg C (2000) Ecology of bacteriophages in nature. Viral Ecology: 211–246.
15. Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. Nature Reviews Microbiology 3: 537-546.
16. Thingstad T, Lignell R (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquatic Microbial Ecology 13: 19-27.
17. Suttle C (2000) Ecological, evolutionary, and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae. Viral Ecology: Academic Press. pp. 247–296.
18. Kutter E, Sulakvelidze A (2005) Bacteriophages: biology and applications: CRC Press.
19. Thingstad T (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnology and Oceanography: 1320-1328.
20. Weinbauer MG, Rassoulzadegan F (2004) Are viruses driving microbial diversification and diversity? Environmental Microbiology 6: 1-11.
21. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Microbiology 6: 181-186.
22. Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D (2009) Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PLoS One 4: e4169.
23. Banfield J, Young M (2009) Variety--the Splice of Life--in Microbial Communities. Science 326: 1198.
24. Weinbauer M, Peduzzi P (1994) Frequency, size and distribution of bacteriophages in different marine bacterial morphotypes. Marine Ecology Progress Series 108: 11-20.
25. Weinbauer M, Hoefle M (1998) Size-specific mortality of lake bacterioplankton by natural virus communities. Aquatic Microbial Ecology 15: 103-113.
26. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66.
27. Sheldon R, Prakash A, Sutcliffe Jr W (1972) The size distribution of particles in the ocean. Limnology and Oceanography 17: 327-340.

28. Sullivan M, Huang K, Ignacio-Espinoza J, Berlin A, Kelly L, et al. (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol 12: 3035-3056.
29. Lindell D, Jaffe J, Coleman M, Futschik M, Axmann I, et al. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 449: 83-86.
30. Angly F, Felts B, Breitbart M, Salamon P, Edwards R, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368.
31. Williamson S, Rusch D, Yooseph S, Halpern A, Heidelberg K, et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3: 1456.
32. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: reviews0003.