Chapter 5

An evolutionary model of phage-host interaction

5.1 Introduction

In Section 4.4.2.6 we predicted the total number of "species" in a given environment. It is therefore of interest to define what a "species" is. In the biophysical model a "species" of bacterium or virus is defined by a set of random variables drawn from some distribution. However, as explained previously, two "species" with the same parameters can be totally different organisms and should be counted separately. Therefore the biophysical model described in Section 4.4 cannot provide us with an adequate definition of a "species" that would be useful for testing the predictions of our model. The problem lies in the fact that at that level of abstraction, bacteria and viruses are the equivalent of "point particles" without internal structure. In the present section we will attempt to go one step further and define an evolutionary model, which when viewed at a coarse-grained level, would be equivalent to the description of the biophysical model in Section 4.4.

In order to understand what a species is in the context of our biophysical model, we propose definitions for both bacterial and viral species that ensure that the assumptions of the biophysical model are respected. These assumptions include: (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. Based on these definitions, we will construct an evolutionary model for the emergence

of new bacterial and viral "species" in nature. While this model is equivalent to our biophysical model when viewed in a genetic coarse-grained way, the evolutionary model leads to the prediction that bacterial "strains" are part of interaction networks with viral "strains", whereas bacterial "species" form a unique association with a single viral "species" and vice versa. Furthermore, in order for new bacterial and viral species to emerge as independent elements, the emerging viral species needs to abandon the parental bacterial strain that it previously controlled in favor of the new emerging species. We propose that the "arms race" between bacteria and viruses may lead to a "positive feedback" mode of evolution, that both enables the emerging viral species to switch hosts, and enables the emerging bacterial strain to evolve at an accelerated pace through selection sweeps to form a new species. Thus, the arms race that bacteria and viruses are locked in is perhaps the engine driving bacterial and viral co-speciation, with selection pressure arising from the environment biasing the direction of evolution. In addition we show that for the simple case of a "butterfly" 2x2 strain interaction network the total concentration of the parental and emerging strains doubles when speciation is complete. We then generalize this result the case of $N_{strains}$ 2x2 interaction networks, with $N_{strains}$ defined as the number of strains per species. Finally we conclude by suggesting an experiment to test our hypothesis regarding "positive feedback evolution".

Summary of findings: Our biophysical model is consistent with an evolutionary model where (1) a bacterial "species" is comprised of bacterial "strains" and where a viral "species" is comprised of viral "strains" (with a "strain" = a quasispecies). (2) New bacterial "species" co-emerge with new viral "species" and vice versa. (2) A bacterial "species" interacts with just one viral "species", however a bacterial "strain" generally interacts with many viral "strains" as it is

part of a network of bacterial-viral interactions. (3) The host range of viruses should be (mostly) species or strain specific. (4) The evolutionary arms-race between phages and hosts (such as the CRISPR warfare) may be a critical part of the stage where bacterial and viral "species" co-emerge out of their parental "strains" by (a) accelerating the evolution of the bacterial strain through selective sweeps and at the same time (b) accelerating the evolution of the virus to switch hosts.

The road map: We will begin by describing the critical features of the biophysical model described in Section 4.4, which our evolutionary model must reproduce when viewed at a coarse-grained level. We will then define the concept of a "strain" and a "species" in such a way that when placed in an evolutionary context produces a "bacterial and phage world" that when coarse-grained is equivalent to the description of our biophysical model. Thus we will use the biophysical model to guide us in selecting a good evolutionary model.

5.2 Definition of a bacterial and viral *strain* and *species*

Critical features of the biophysical phage-host model

The following are the critical features of the biophysical phage-host model described in Section 4.4:

1. **Growth:** All bacteria and viruses are actively replicating in the environment.
2. **Viral control:** Each bacterium is associated with a lytic virus that controls its concentration.
3. **Uniqueness:** Each phage-host system is comprised of a bacterium and a virus that can be distinguished (in some measurable way) from all other bacteria and viruses in the environment. We therefore say that each bacterium belongs to a unique "species" denoted by the index i, and each virus belongs to a unique "species", denoted by the same index.

4. **Symmetry:** There are equal numbers of bacterial and viral "species" (both denoted with the index *i*) .
5. **Independence:** All phage-host systems in a given environment are independent of each other, i.e., there is no cross interaction between one system and another.

An evolutionary model that satisfies these five conditions will be consistent with our biophysical model. We can then use the evolutionary definitions of a bacterial species and a viral species to interpret the meaning of the species in our biophysical model.

Bacteria "take up" concentration: Bacterial "species" in the biophysical model have one additional consequence. A bacterial "species" has the property that it "takes up" concentration in the environment, with the concentration being given by Eq. 7. The reason we say it "takes up" concentration is that any environment has finite resources that can accommodate a finite concentration of cellular organisms (this is how we obtained the number of *species* in the environment, $N_{species}$). Thus, only elements that "take up" concentration contribute to the diversity of the system. A viral "species" also has a concentration, however viruses are not limited by resources and therefore there is no upper bound on the number of viral "species" in a given environment. Therefore viral "species" do not "take up" concentration. Drawing again on an analogy to physics, in this respect, bacteria are like fermions and viruses are like bosons — one can pack an infinite number of bosons into a negligibly small volume, whereas fermions take up volume due to their quantum charges. This is why $N_{species}$ was obtained from c_{bact}^{tot} and not c_{virus}^{tot}, the former has an upper bound whereas the latter does not.

Definition of a bacterial and viral *strains*

We seek to define a bacterial *species*[1] and a viral *species* in such a way that, when placed in an evolutionary context, we are able to reproduce the essential characteristic of the biophysical model described above. To define a *species* we first need to define an auxiliary term, which is a *strain*.

> **Definition of a *strain*:** A genetic element (bacterium or virus) is considered a new *strain* if and only if this genetic element is *distinguishable* from all other *strains* (the first cell is by default a *strain*). To be *distinguishable*, a genetic element needs to have a measurable property that sets that element apart from all other existing *strains*. This measurable property should give consistent results over time despite the mutation load of the genetic element.

This definition of a *strain* is consistent with the biologically intuitive definition of a "strain". Here we have also defined a viral *strain*. A viral *strain* can also be interpreted as a viral quasispecies [1] since each genome in the quasispecies is not *distinguishable* from other elements comprising the quasispecies.

Definition of a bacterial *species*

> **A bacterial *species:*** A bacterial cell constitutes a new *species* if and only if (1) it is actively replicating in the environment; (2) It can be classified as a new *strain* in the environment; (3) It forms a stable association with a virus that can be classified as a new *strain* in the environment.

Criterion 1 is necessary in order to distinguish actively replicating cells that have a finite growth rate from spore cells or inactive (possibly dead) cells [2]. The latter, although possibly alive, cannot be part of a phage-host system since they are not actively growing. Criterion 2 simply

[1] We use italics to distinguish the terms defined in the current model from the colloquial use of these words or from the terms used in the biophysical model.

ensures that the new bacterial *strain* is *distinguishable* from other pre-existing bacterial *strains* in the environment, and thus should receive a new index. Criteria 1+2 define an active strain in the intuitive sense, not a species in the intuitive sense. Why is it that to complete the definition of a bacterial *species* one must talk about its viruses (criterion 3)? The reason is the following: If an environment contains n bacterial *strains* with n infecting viral *strains*, adding a new bacterial *strain* (*strain* # n+1) without a new viral *strain* will lead to an overdetermined system of equations in which one or more bacterial *strains* will become extinct (Section 4.6). Thus, to add a new bacterial *species* one must also introduce a new viral *strain* into the system. This rule can be stated in a more general way (Section 4.6):

> A system of n bacterial *species* must be associated with exactly n viral *species* otherwise the system will be overdetermined, driving excess *species* to extinction.

This definition of a bacterial "species" satisfies the properties of: **bacterial growth** (the bacterial *species* must be growing); **viral control** (each bacterial *species* is associated with a virus); and **bacterial uniqueness** (each bacterial *species* is a new *strain*).

Definition of a viral species

The definition of a viral species is analogous to the definition of a bacterial species:

> **Definition of a viral *species:*** A virus constitutes a new *species* if and only if (1) it is actively replicating in the environment; (2) It can be classified as a new *strain* in the environment; (3) It forms a stable association with a host that can be classified as a new *strain* in the environment.

Criterion 1 is to ensure that we are considering a virus that is active and not a decayed or an inactivated virus. Criterion 2 ensures that the new viral *strain* is *distinguishable* from other pre-

existing viral *strains* in the environment, and thus should receive a new index. Criterion 3 is, as before, required because the system should always have equal number of bacterial and viral *species* otherwise excess *species* will be driven to extinction (Section 4.6).

This definition of a viral "species" satisfies the properties of: **viral growth** (the viral *species* must be replicating); **viral uniqueness** (each viral *species* is a new *strain*); and **symmetry** (if each bacterial *species* is associated with a viral *species* and each viral *species* is associated with a bacterial *species*, there should be equal number of bacterial and viral *species*). The only property that has yet to be satisfied is independence. By constructing an evolutionary model that satisfies this property we will be able to understand the relation between *species* and *strains*.

Note that the definitions of a bacterial and viral *species* suggest that the formation of a new bacterial *species* is linked to the formation of a new viral *species* and vice versa. In the next section we will explain an evolutionary mechanism for this process.

5.3 A model for bacterial-viral co-speciation

5.3.1 Description of the evolutionary model

Stage 1: One bacterial *strain*, one viral *strain* (Fig. 5.1A). Let's assume our environment contains a bacterial *species* (species 1) comprised of a single *strain* (strain 1), and that this bacterial *species* is under the control of a viral *species* (species A), comprised of a single viral *strain* (strain A) (Fig. 5.1A). The concentration of bacterial strain 1 is dictated by Eq. 7, thus viral species A controls bacterial species 1 (the arrow in Fig. 5.1A). Bacterial strain 1 is said to "take up" concentration in the environment.

Stage 2: An incipient bacterial *strain* emerges (Fig. 5.1B). Now let's assume that through some genetic event (e.g., a transposon, a deletion/insertion/inversion event, a recombination event, a new plasmid, etc.), bacterial strain 1 begins to evolve a new bacterial *strain* that is on the verge of becoming *distinguishable* from strain 1 (Fig. 5.1B). The incipient bacterial strain 2 is under the growth control of viral strain A, and will not be "allowed" to take up concentration on its own, independent of bacterial strain 1 — i.e., it will not be allotted a status of a *species* and therefore will not contribute to the diversity of the system (i.e., increase $N_{species}$). Bacterial strain 2 will continue to undergo evolution with time and accumulate more mutations in its process of maturing into a new *strain*. During all this time bacterial strain 1 is under the control of viral s*train* A (Fig. 5.1B).

Stage 3: An incipient viral *strain* emerges (Fig. 5.1C). As the incipient bacterial strain 2 evolves, so does the viral *strain* that infects it (initially viral *strain* A). This viral *strain* (i.e., viral quasispecies) will begin to form a new cluster that will eventually mature into viral strain B. The incipient viral strain B (not yet *distinguishable* from viral strain A) both tracks the evolution of bacterial strain 2 and also drives the evolution of bacterial strain 2. This hypothesis is supported by the following observations. It has been suggested that viruses and bacteria are in a constant state of an "arms race" [3]. Perhaps the best example of this arms race is the CRISPR bacterial defense system. Bacteria continuously acquire CRISPR spacer sequences from viruses to evade these viruses, while viruses rapidly evolve by mutation, homologous recombination, and deletion of the target sequences to evade new acquired spacers [4]. Conversely, CRISPR repeats and their associated proteins undergo evolution to escape shut-down mechanism for the CRISPR system encoded by the phage [3]. There is also evidence that the bacterial population undergoes

sweeping selection events, where potentially only one cell survives (the only cell that had the right spacer) [4]. Such bottlenecks will accelerate the evolution of the emerging bacterial *strain*, driving its evolution forward. This example illustrates how by a process of positive feedback between the new bacterial *strain* (strain 2) and the new viral *strain* (strain B) both elements track each other and push each other to further evolve (Fig. 5.2). **The bacterial-viral "arms race" may therefore be a critical step in forming (or at least accelerating) the formation of new bacterial *species* and new viral *species* from the parental *strains*.** Indeed, CRISPR sequences have been found in nearly half of all sequenced bacterial genomes [3]. While the CRISPR mechanism may contribute to the arms race, it may not be an essential component. Luria and Delbrück have shown that a bacterial strain grown from a single cell will mutates naturally (without interaction with the phage) so that a subpopulation of bacteria will become immune to the virus [5]. Thus, even without a CRISPR system the bacterium can evade the virus. Therefore, this "arms race" may be a fundamental mechanism of evolution to generate new bacterial and viral *species*. Given our interpretation, these events are not a disadvantage in terms of reduction in diversity, as previously proposed [4], since they may provide the mechanism for new *strains* to emerge. Thus ultimately these mechanisms generate diversity.

Stage 4: New bacterial and viral *strains* emerge (Fig. 5.1D). The incipient bacterial strain 2 is now *distinguishable* from strain 1 and can be defined as a new *strain*. The incipient viral strain B emerged as a new viral *strain* (strain B) that initially infects both bacterial strains 1 and 2 (Fig. 5.1C). At this stage, a 2x2 network like interaction emerges. This network can, in principle, persist indefinitely, and as the evolutionary distance between *strain* 1 and *strain* 2 grows, this could lead to the formation of viruses with a wide host range. If the system is stable over time,

the two bacterial *strains* would be under control of two viral *strains* and both would "take up" concentration, thus increasing the diversity of the system ($N_{species}$ in Eq. 16B increases). However, it seems more plausible that as the distance between *strain* 1 and *strain* 2 grows, the cross affinity (B→1 and A→2) will decrease , leading to the emergence of two independent associations (A→1 and B→2). This is because the bacterial strain 2 is driving the evolution of viral strain 2, and it is expected that at some point this virus will lose its ability to infect the parental bacterial strain 1 (Fig. 5.2). Furthermore, it would seem that a 2x2 network of interacting *strains* would not be stable in an open environment for long, since if one of the viral *strains* drifts off, leaving bacterial *strains* 1+2 under the control of the remaining viral *strain*, one or both bacterial *strains* will be driven into extinction over time (Section 4.6). Numerical simulations would be required to see if a network of 2x2 interacting *strains* (*n*x*n* in the more general case) are indeed less stable than two 1x1 associations (or generally *n* 1x1 associations) under loss of a viral *strain*. The fact that in nature, phages typically display a narrow "species" or "strain" level host range [6,7,8] favors the interpretation that indeed independent phage-host system arise. That said, there are a few exceptions and some phages have been found to display a wide host range [8], however this does not seem to be the general case. Therefore we hypothesize that over time, as bacterial strain 2 and viral strain B continue to evolve, the cross infectivity B-1 and A-2 naturally fades, and we will define *strains* 2 and B as new *species* when this cross affinity disappears. Note that this hypothesis ensures that the last property of **independence** is satisfied since we require that emerging phage-host systems lose their dependence on the parental strains to which they were linked initially (discussed further below).

Stage 5: New bacterial species and viral species emerge (Fig. 5.1E). Bacterial strain 2 and viral strain B have evolved sufficiently that cross infectivity has completely faded. At this point bacterial strain 2 is under the exclusive control of viral strain B via Eq. 7 and can "take up" concentration. The association between bacterial strain 2 and viral strain B is stable and lasting. Bacterial strain 2 now answers the definition of a *species* (it is a replicating *strain* stably associated with a viral *strain*) and can be regarded as a new species (species 2). Viral strain B now also answers the definition of a *species* (it is a replicating *strain* stably associated with a bacterial *strain*). At this stage the process can begin again and a new *species* can emerge.

The conclusion from this model is that new bacterial *species* must emerge with a new viral *species* and vice versa. While it has been shown in many experiments that bacteria can evolve in the absence of viruses, this model proposes that in the presence of lytic viruses, the process of evolution may be accelerated.

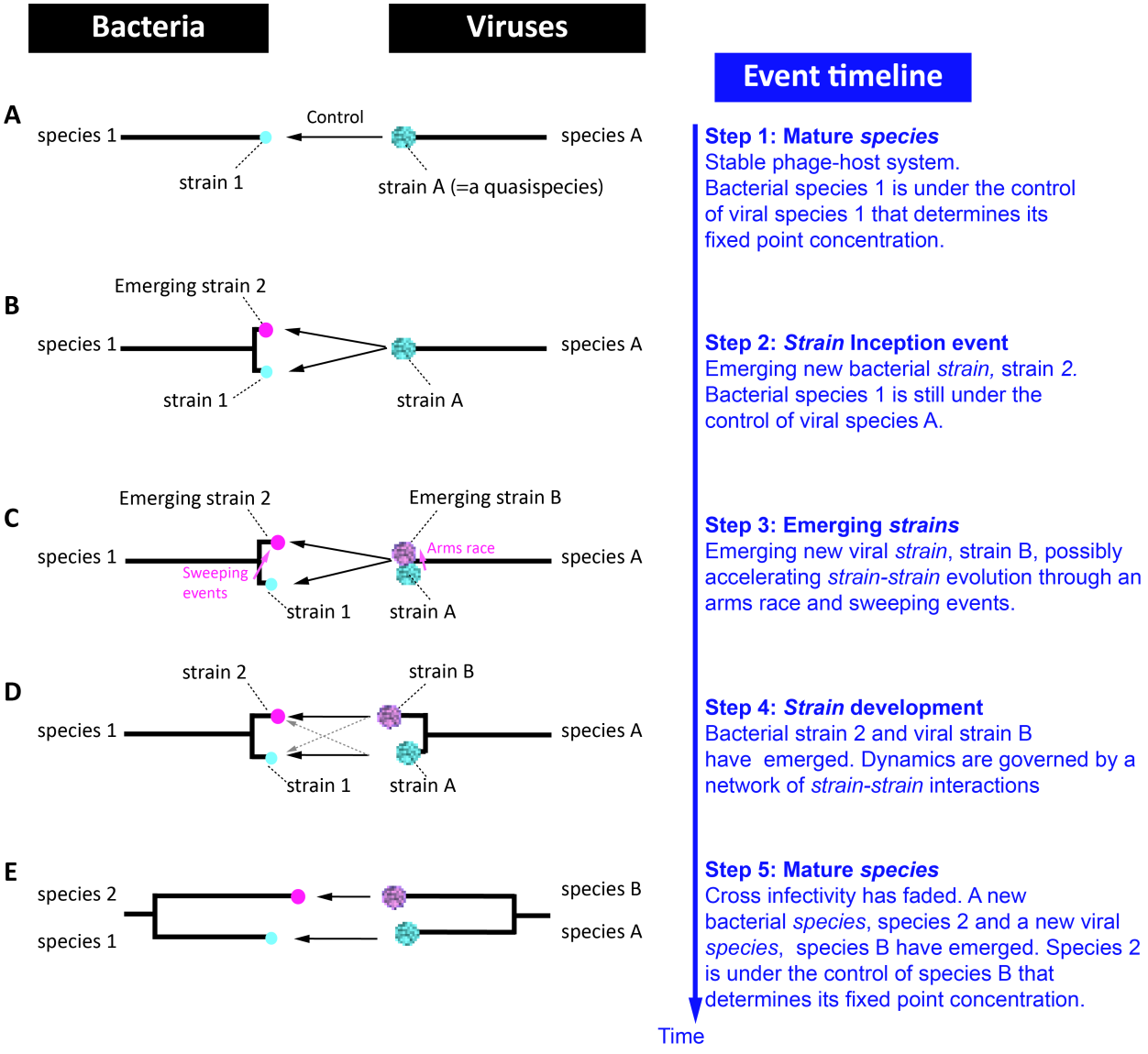

Figure 5.1. A possible evolutionary process of bacterial and viral co-speciation. If species 1 and species 2 have the same size and growth rate, then stage E "takes up" twice the concentration as stage A, with the intermediate states somewhere in between.

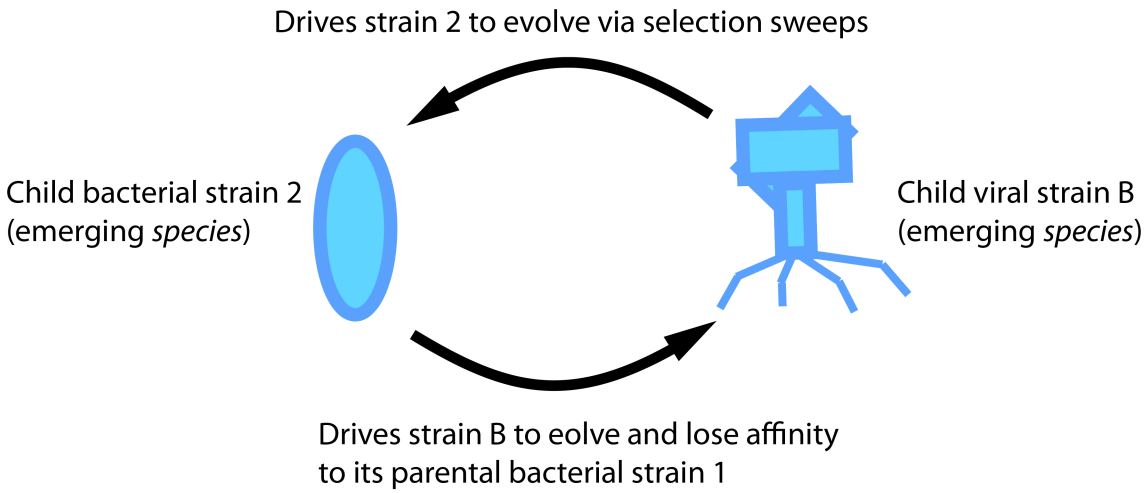

Figure 5.2 Positive feedback evolution model for emerging bacterial and viral *species*. We propose that the arms race between bacteria and viruses may be a critical step in the formation of a new bacterial and viral *species*. This process is critical in order to allow viral strain B to relinquish its 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 viral species have a narrow (*species*) host range. This process may also be critical for the bacterium, where by selective sweeps it drives the bacterium to evolve away from its original parental strain. This positive feedback model may amplify initially "noise". Thus, the process of co-speciation is perhaps equivalent to "amplification of noise" and therefore potentially a chaotic effect. The random trajectory in the genome space may be biased by selection pressure due to environmental factors such as available nutrients, competition and so on. Consequently, phylogenetic trees may have a fractal quality to them, though branches may be biased by selection pressure. 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 at 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).

5.3.2 A coarse-grained view of the evolutionary model satisfies all the properties of the biophysical model

We have seen that all the properties of the biophysical model except for independence were satisfied by the definition of *strains* and *species* that we use. The key point of this model is how the property of independence arises. According to Fig. 5.1, a bacterial *species* is born out of a single parental bacterial *strain*. Initially the new bacterial *strain* is under the control of both a

parental viral *strain* and a new viral *strain* (the latter also born out of a parental viral *strain*). Once the new bacterial *strain* has evolved sufficiently from its parental s*train*, it loses its association with the parental viral *strain*. At the same time, the new viral *strain* loses its control over the parental bacterial *strain* (the transition from Fig. 2D to Fig. 2E). Thus, once both bacterial and viral *species* become an independent pair they are defined to be a new *species*. Therefore, the model described in Fig. 5.1 leads to a "world" where every bacterial *species* is controlled by just one viral *species*, and vice versa. If we view our evolutionary model in a coarse-grained way, and ignore the "structure" within each *species* (shown in Fig. 5.3 and discussed below), we obtain a model where each pair of interacting bacterial and viral *species* are independent of all other pairs (the condition of **independence** is satisfied). Therefore the two models become equivalent in the limit of describing organisms at a low genetic resolution, where the subtle differences between different *strains* within *species* (be it bacterial or viral) are lost. By interpreting the properties of the *species* defined in the current model, we can now answer the question raised in Section 4.4 of what is a "species"?

5.3.3 Revisiting the question of what is a "species"?

5.3.3.1 "Quark-gluon" model of a species

In one of the intermediate stages in the formation of a new bacterial *species* there is a state where a 2x2 network of interactions forms between the new and parental bacterial *strains* and the new and parental viral *strains* (Fig. 5.1D). In the general case, bacterial *strains* are continuously emerging from parental *strains*. Thus in the general case (applying our conservation rule that the number of bacterial *strains* must always equal the number of viral *strains*) we obtain a network of *n* bacterial *strains* infecting *n* viral *strains* (Fig. 5.3). These *n* bacterial *strains* are defined to be a bacterial *species*. Therefore, at any given point in time, a bacterial *species* in nature is

comprised of *strains* that are in the process of maturing into new *species* (Fig. 5.1D). Likewise, the *n* viral *strains* can be technically classified as a viral *species*. Thus, a viral *species* is in essence a collection of viral strains, i.e., a collection of viral quasispecies, infecting the s*trains* of a given bacterial *species* in a network-like fashion (Fig. 5.3).

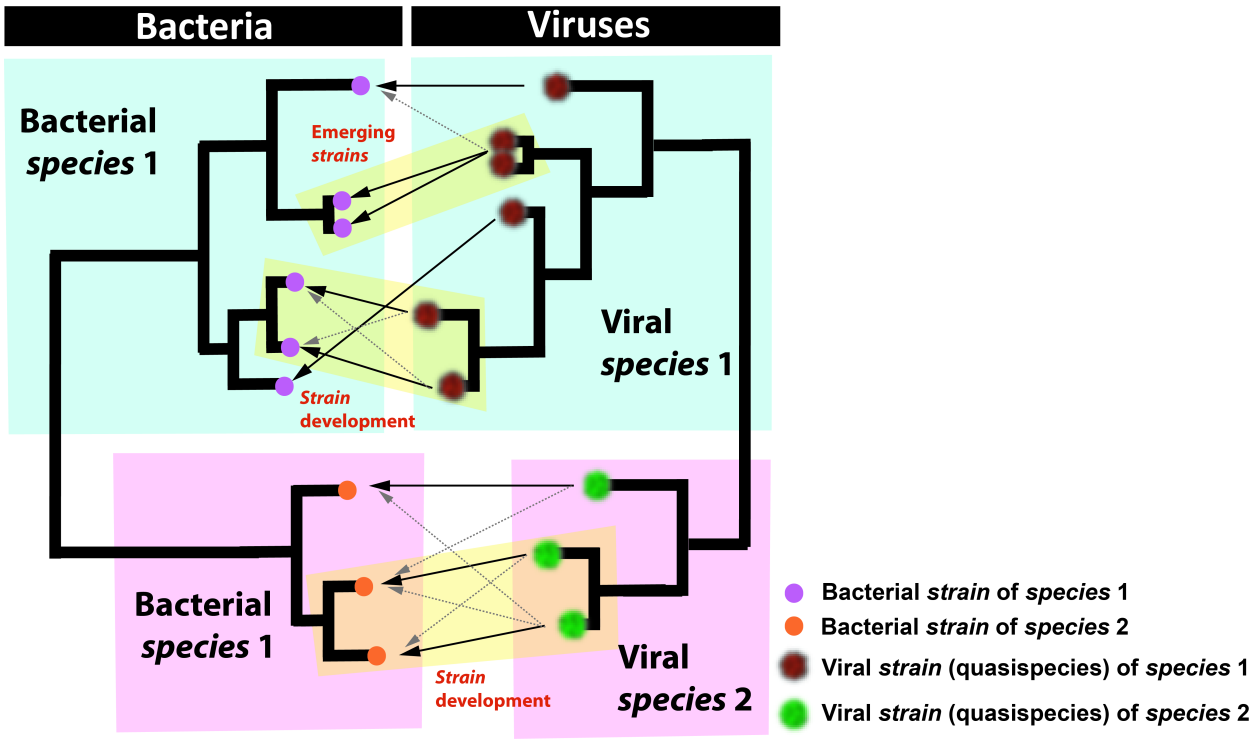

Figure 5.3 The "Quark and gluon" model of a *species*. Hypothetical phylogenetic tree of a conserved bacterial gene (left), revealing two bacterial *species*, paired with a phylogenetic tree of a conserved viral gene (right), revealing two corresponding viral *species*. Each clade of a *species* is comprised of the *strains* of that *species*. Yellow boxes highlight bacterial-viral *strains* in different stages of maturation. The arrows show which bacterial *strain* is infected (i.e., controlled) by which viral *strain*. Solid lines represent primary targets, whereas dashed lines represent secondary, weaker targets. The biophysical model that we propose lumps all *strains* within each *species* clade into one class.

5.3.3.2 The meaning of $N_{species}$

The case of a *species* comprised of one parental *strain*

To understand which entities contribute to $N_{species}$ we need to ask ourselves who "takes up" concentration in the model presented in Fig. 5.1 (or the more realistic view in Fig. 5.2). Let's begin by considering Fig. 5.1 again. Let's assume stage 1 "takes up" concentration x. Stage 2 is still under the control of one virus, so it "takes up" a concentration x as well. We will skip stage 3 for a moment. Stage 4 however is different. In this stage we have two *strains* in a 2x2 "butterfly" network configuration (Fig. 5.4). For simplicity let's assume that the coupling constants (i.e., infection rates) are $k_{11} = k_{22} = \alpha$ and $k_{12} = k_{12} = \beta$. Initially when bacterial strain 2 just emerges (the "child" strain), we have $\beta = \alpha$. This is because the child viral strain B also has just emerged and it is barely *distinguishable* from its parent viral strain A. At this stage we anticipate that both bacterial *strains* (parent 1 + child 2) will contribute together a concentration of x because both are under the control of one viral *strain* (parent A + child B). As the child bacterial strain 2 and child viral strain B evolve, the parent-child coupling constants are hypothesized to fade, and so $\beta \to 0$. When $\beta = 0$ a new *species* of bacteria and viruses has emerged. At this stage, we expect both new bacterial *strains* (*species*) to contribute together 2x to the concentration.

This effect can be readily appreciated by solving the butterfly network: Let B_i be the concentration of bacterial *strain i*, and V_i the concentration of the viral *strain i* , where *i*=1 are the parental strains and *i*=2 are the child strains (Fig. 5.4). The rate equations for the viral *strains* are given in the general case by

$$\frac{dV_1}{dt} = \alpha b V_1 B_1 + \beta b V_1 B_2 - \tilde{\gamma} V_1$$
$$\frac{dV_2}{dt} = \beta b V_2 B_1 + \alpha b V_2 B_2 - \tilde{\gamma} V_2$$

where b is the burst size (assumed to be equal for the two *strains*). Assuming steady-state conditions (to obtain the fixed point concentrations), after some algebra (defining $\gamma \triangleq \tilde{\gamma}/b$), we find that

$$B_{tot} = B_1 + B_2 = \frac{2\gamma}{\alpha+\beta} = \frac{2\gamma}{\alpha}\frac{1}{1+\beta/\alpha} = \frac{2\gamma}{\alpha}\frac{1}{1+\kappa}.$$

where we have defined the normalized parent-child coupling constant $\kappa \triangleq \beta/\alpha$.

Thus, initially, when $\kappa = 1$ we have $B_{tot} = \frac{\gamma}{\alpha}$, and when $\kappa = 0$ we have $B_{tot} = \frac{2\gamma}{\alpha}$. Thus, exactly as we predicted, the total concentration "taken up" by bacterial strains 1+2 increases from $\frac{\gamma}{\alpha}$ to $2\frac{\gamma}{\alpha}$ during the maturation process of the new *species*. We can parameterize this uncertainty with a "maturation factor" μ:

$$B_{tot} = \mu B_{\text{species}}, \quad \text{where} \quad 1 \le \mu \le 2$$

where B_{species} is the concentration one would obtain if one were to coarse grain the system to a *species* level ignoring *strains*. Therefore, in the case of a 2x2 network, if we were to coarse grain bacteria to a *species* level (say an OTU of 3%), we would be underestimating the concentration

taken up by the species by a factor anywhere from μ=1 to μ=2 (Fig. 5.5). Now let's see what happens in a more realistic scenario when a species is comprised of n strain (where in reality n can be very large since it probes the microdiversity of a *species*).

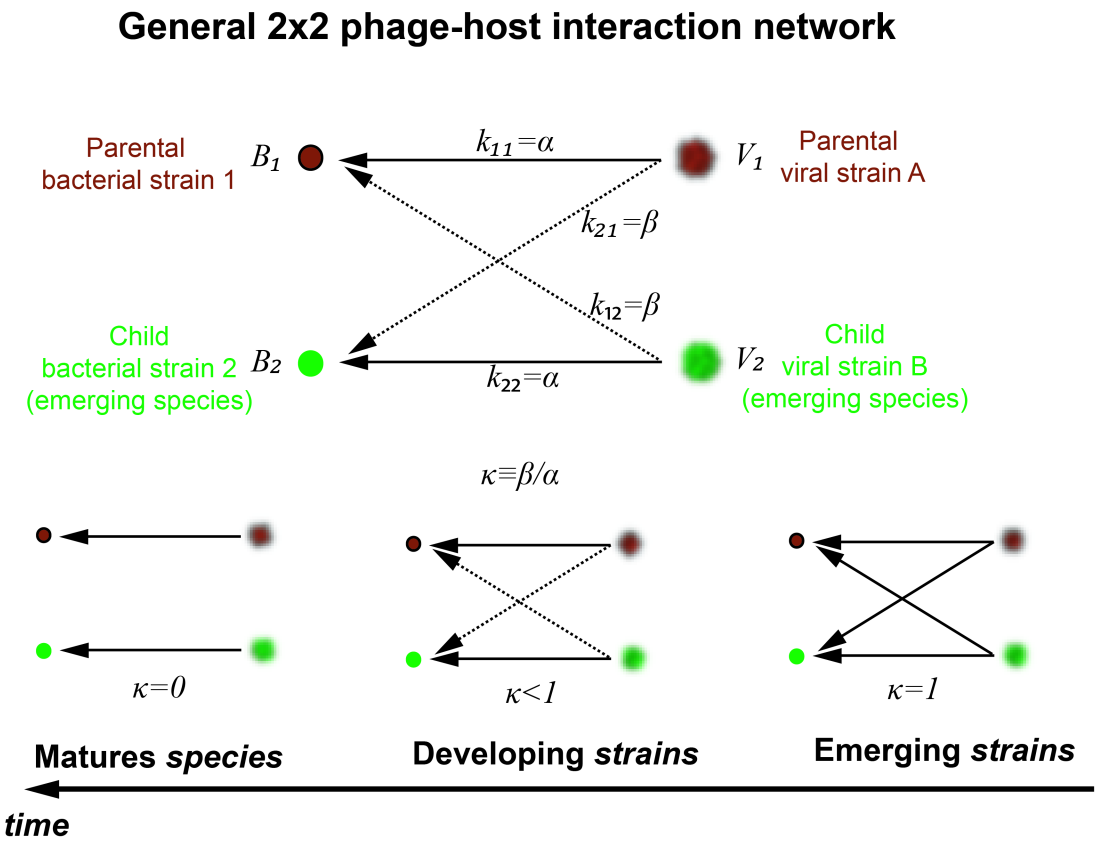

Figure 5.4 A 2x2 phage-host interaction network with event timeline. This diagram is a general 2x2 interaction network between two viral *strains* — a parental viral *strain* (strain A) and the emerging viral *species* (strain B), that are controlling the parental bacterial *strain* (strain 1) and the emerging bacterial *species* (strain 2). The timeline shows the hypothesized evolutionary trajectory of these four *strains*. Initially, as the new (child) *strains* have just emerged, the coupling constants are equal. As the child *strains* evolve, the parent-child coupling constants decrease (dashed lines). Finally the child *strains* have evolved enough so that the parent-child coupling constants are 0 and new *species* of bacteria and viruses have emerged.

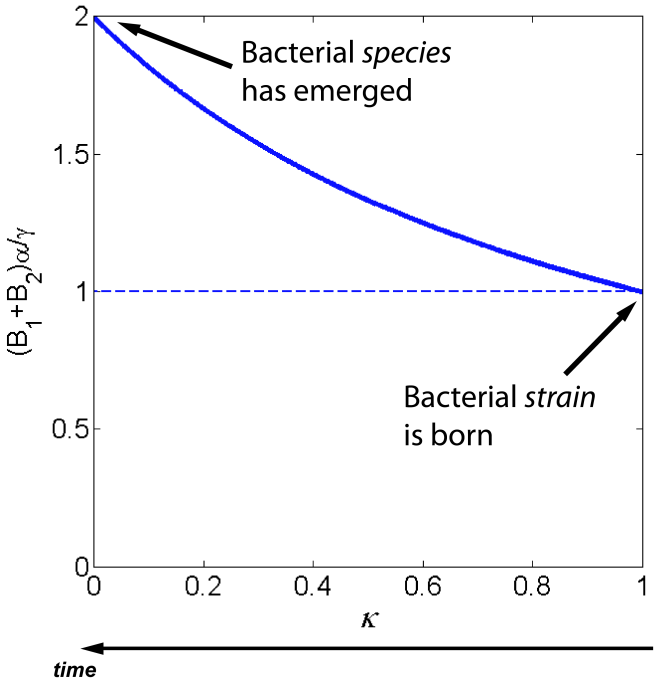

Figure 5.5 Total concentration "taken up" by parent and child bacterial *strains* as child bacterial *strain* evolves towards a new *species*. Here we show how the sum concentration of both parent and child bacterial *strains* changes with time, as the parent-child coupling constant κ goes to 0. Initially, when the bacterial child *strain* is born, it is under the control of the parental viral strain and the parent-child coupling constant is maximal (κ=1). The total concentration at this point is that of a single bacteria *strain* (=1 in normalized units). When the bacterial child *strain* is fully evolved, the parent-child coupling constant equals 0 and a new bacterial *species* under the control of a new viral *species* has emerged. The total concentration at this point has doubled because the new bacterial *species* is allowed by its controlling virus to "take up" a concentration =1 (in normalized units).

The case of a *species* comprised of *n* parental *strains*

In the general case (Fig. 5.3) a bacterial *species* will be comprised of N_{strain} parental *strains*. Each of these parental *strains* is anywhere in the stage between emerging a new *strain* to having a fully emerged *species* (thus the total number of strains will be anywhere between N_{strain} and $2N_{strain}$). We make the approximation that each one of these parental strains is part of a butterfly

network with coupling constant β, which is anywhere between $\beta = \alpha$ to $\beta = 0$. If all *strains* were in a state of $\beta = \alpha$ then the total concentration "taken up" by this *species* would be

$$B_{tot} = \sum_{i=1}^{N_{strain}} B_1^{(i)} + B_2^{(i)} = \frac{\gamma}{\alpha} N_{strain} .$$

If all *strains* were in a state of $\beta = 0$ the total concentration "taken up" by this *species* would be

$$B_{tot} = \sum_{i=1}^{N_{strain}} B_1^{(i)} + B_2^{(i)} = \frac{2\gamma}{\alpha} N_{strain} .$$

Therefore,

$$B_{tot} = \mu N_{strain} B_{\text{species}}, \qquad \text{where } 1 \le \mu \le 2$$

where B_{species}, once again, is the concentration one would obtain if one were to coarse grain the system to a *species* level ignoring *strains*. Therefore the number of "species" in Eq. 16B is given by

$$N_{\text{"species"}} = \mu N_{strain} \approx N_{strain} .$$

Thus, our conclusion from this analysis is very simple and logical. Even though the total number of actual independent phage-host systems is equal to the number of *species* we need to multiply each *species* by a factor which approximately equals the number of strains in that *species*. Thus by probing the "structure" of a *species* (which is the assumed construct in the biophysical model)

we came to the conclusion that one needs to weigh each species approximately by the number of *strains* in that *species*. Since strains are *distinguishable*, indeed each strain should contribute to the total concentration between ×1 and ×2.

5.3.3.3 The dynamics of speciation

The process of *speciation* (i.e., co-formation of new bacterial and viral *species*) is inherently stochastic since a bacterial *strain* can easily become extinct if a viral *strain* is lost, as the system becomes unstable (Section 4.6). We therefore envision the process of *speciation* as one in which new bacterial *strains* continually emerge from extant *strains* (the microclades in Fig. 5.3), with some *strains* evolving to become *species*, and with other *strains* being lost (Fig. 5.6). In principle, one should be able to calculate the rate at which bacterial *species* are formed in the oceans, possibly yielding better bounds on the total diversity in the oceans.

5.3.3.4 Analogy to the conventional concepts of a "species" and "strain"

The intuitive notion of a bacterial "strain" has been familiar to biologists for many years. Indeed genetic microdiversity below the species level has been observed in nature [9,10]. We too have observed such microdiversity in treponeme cells found in the termite hindgut ("Host I" and "Host III" in Fig. 2.2). The concept of a bacterial "species" comprised of "strains" is also well known and widely used by biologists, though the empirical identity thresholds used for classification of new species are somewhat questionable given the lack of a rigorous definition of a species. The concept of a *strain* of viruses is also familiar, this is the well-known quasispecies proposed by Eigen [1]. The definition of viral "species" on the other hand has been quite elusive [11]. If the model we propose proves to be valid, then it would seem that a host-range-based taxonomy [11] should lead to a meaningful organization of viral species, at least for marine

ecosystems. In principle, according to our model, the true classification of marine life-forms (bacteria + viruses) requires both to be classified simultaneously. For example, when two marine bacterial "species" seem very similar (using "species" in the colloquial meaning), then according to our proposed model, if these "species" are infected with different non-overlapping viruses they should be classified as different *species*.

5.3.3.5 The insight for the coarse-grained model

When considering the coarse-grained biophysical model, the most natural definition for a "species" would be "a cell that can be distinguished reproducibly from all other cells", i.e., the definition of a *strain*. The evolutionary model has shown that this is not the case, as one needs a more complex structure, defined here as a *species*, in order to obtain a "world" of non-interacting phage-host systems. Thus, the "species" in the biophysical model are equivalent to the *species* defined in our evolutionary model, however, the concentration of each *species* needs to be multiplied by a weight of $\sim N_{strain}$, which is the number of *strains* in each *species*. This conclusion also leads to a clear distinction between the concept of a bacterial *strain* and a bacterial *species*. While a bacterial *species* interacts with just one viral *species* and vice versa, a bacterial *strain* interacts with several viral *strains* and is not an independent entity.

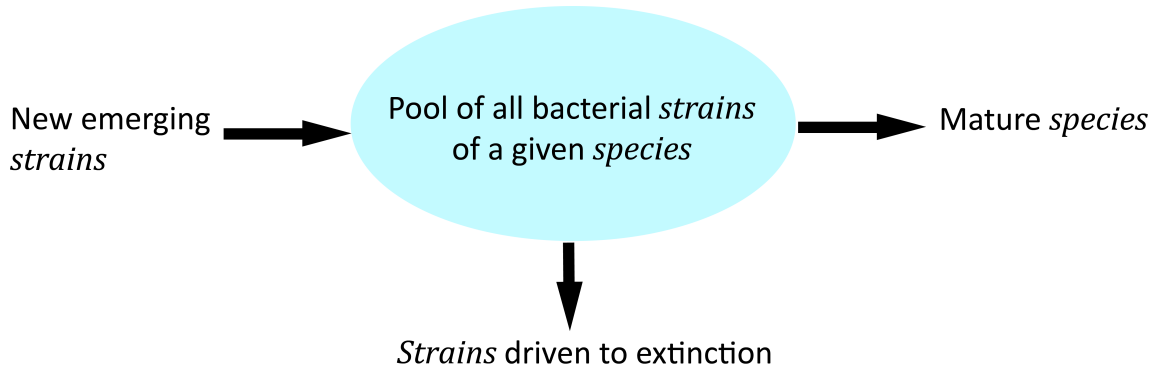

Figure 5.6. Flux of *strains* in the process of bacterial speciation. According to our evolutionary model, bacterial *strains* of a given *species* are cells that are *distinguishable* for all other cells in the population, but do not form a stable (i.e., unique) association with viral *strain* (see Fig. 5.1D). A bacterial *strain* matures into a *species* if it forms a one-to-one association with a viral *strain*. The pool in this figure is the sum of all bacterial *strains* comprising a *species*. The flux into this pool comes from new emerging *strains* (Fig. 5.1B & D). The flux out of this pool is due to either *strains* that have gone extinct (e.g., since the viral network in which they were in became destabilized), or *strains* that have matured into species (Fig. 5.1E).

5.3 Why do phages typically have a narrow host range?

It is a known fact that most phages are species or strain specific (although a few exceptions have been found) [6,7,8]. Naïvely, this observation seems peculiar given that all cellular life forms encode and read information in virtually the same manner (e.g., human genes can be expressed in bacteria). Generally speaking, the genome of phage A could be expressed in many very divergent species, yet phages tend to infect a single species. Why is this the case?

The evolutionary scheme we propose here in fact predicts that phages should have (in the majority of cases at least) a *species*- or *strain*-level host range. According to our model, any given viral *species* is expected to infect a single bacterial *species* (Fig. 5.1 and Fig. 5.3). Thus, the viral *strains* associated with a given viral *species* will infect some (or all) of the bacterial

strains within a given bacterial *species* (Fig. 5.3). Our model therefore predicts that viruses will have either a "species"-specific host range (infecting all *strains* of a given bacterial *species*) or a "strain"-specific host range, infecting a subset of bacterial *strains* (or at the very minimum a single bacterial *strain*).

Mechanisms to generate a wide host range. A viral *species* could in principle evolve to infect another bacterial *species* in addition to its original host (and thus the former bacterial *species* will be susceptible to more than one viral *species*). As long as the viral species is part of an *n*x*n* network of associations, the dynamics are stable (see Section 4.6). However, such a scenario seems to be the exception since in open systems at least, *species* are not spatially constrained. Therefore, if a *species* drifts off, the network will become imbalanced (i.e., *n*x*m* where $n \neq m$) leading to unstable dynamics and, over time, extinction events. This leads to a prediction that in closed systems (for example the gut) there will be more viruses with a wide host range than in open systems. Indeed, phages isolated from sewage appear to display a wide host range [12].

Another possibility for a wide host range is the following: if the cross-species infection in Fig. 2D does not fade away with time as we hypothesized, then in a closed system it is possible to have a lytic viral *species* with a wide host range if it is part of an *n*x*n* network of hosts and viruses. However, in an open environment, where *species* are not spatially constrained, again the system may become unstable as described above. Thus, unless the environment is constrained to a closed volume, it seems that generally a more robust and stable solution (and therefore more likely scenario) would be for phages to have a narrow host range. That said, the scheme we have presented here does not preclude the possibility that a given viral *species* happens to be

successful in infecting many bacterial *species* that are not present in the given environment (e.g., they happen to have the same membrane receptor). Such coincidental events should also be kept in mind.

5.4 Testing the evolutionary model: evolution experiment of a phage-host system

One possible way to test our model is to perform a Lenski-type evolution experiment of a phage-host system (similar to the evolution experiments of Rainey [13]). One choice would be T4 and *E. coli* . To prevent total annihilation of the bacteria, we should add a degradation factor for the phages (or perhaps a chemostat would be sufficient?). *E. coli* is a good choice since its CRISPR system has been investigated [14]. After n generations would expect at least two new bacterial strains to co-emerge with an equal number of viral strains. After enough generations the n emerging strains should be distinguishable (measurable by sequencing). Furthermore, we should observe a decrease in parent-child cross affinity between the new evolving viral strain(s) and the original viral strain. In a different experiment, one can evolve a strain of *E. coli* with a mutation in one of the *cas* proteins inactivating the CRISPR array defense mechanism. We expect that either we will not observe the emergence of new strains, or that it will take a much longer time to obtain the same evolutionary distance between strains.

5.5 References

1. Eigen M (2006) Viral quasispecies. Evolution: a Scientific American reader: 114.
2. Ducklow H (2000) Bacterial production and biomass in the oceans. Microbial ecology of the oceans 1: 85-120.
3. 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.
4. Banfield J, Young M (2009) Variety—the Splice of Life—in Microbial Communities. Science 326: 1198.
5. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28: 491.
6. Suttle C (2000) Ecological, evolutionary, and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae. Viral Ecology: Academic Press. pp. 247–296.
7. Kutter E, Sulakvelidze A (2005) Bacteriophages: biology and applications: CRC Press.
8. Weinbauer M (2004) Ecology of prokaryotic viruses. FEMS Microbiology Reviews 28: 127-181.
9. Moore L, Rocap G, Chisholm S (1998) Physiology and molecular phylogeny of coexisting. Nature 393: 465.
10. Thompson J, Randa M, Marcelino L, Tomita-Mitchell A, Lim E, et al. (2004) Diversity and dynamics of a North Atlantic coastal Vibrio community. Applied and Environmental Microbiology 70: 4103.
11. Lawrence J, Hatfull G, Hendrix R (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. Journal of bacteriology 184: 4891-4905.
12. Jensen EC, Schrader HS, Rieland B, Thompson TL, Lee KW, et al. (1998) Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. Applied and Environmental Microbiology 64: 575.
13. Buckling A, Rainey PB (2002) Antagonistic coevolution between a bacterium and a bacteriophage. Proceedings of the Royal Society of London Series B: Biological Sciences 269: 931.
14. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321: 960.