Modelling Horizontal Gene Transfer and CRISPR-Cas Activity in Microbial Populations

02-731: Evolution Final Project

Siddharth Reed
Department of Computational Biology
Carnegie Mellon University
slreed@andrew.cmu.edu

May 7, 2021

1 Problem Scenario

1.1 CRISPR-Cas Systems

In the new people often discuss CRISPR associated protein (CRISPR-Cas) technology for the purpose of gene editing. However CRISPR was originally discovered in bacteria as a form of adaptive immunity against bacteriophage infection. Bacteriophages are viruses that target bacteria, injecting their DNA/RNA into the target cell and having it integrate into the genome and eventually kill the target. Bacteria want to protect against such infection so being able to target bacteriophage DNA for degradation can improve fitness, especially in phage-dense environments. If a cell survives an infection CRISPR can take up part of the defeated phage DNA and use it as a kind of identifier to prevent future infections. This "identifier" is integrated into the genome and can act as a guide for the Cas9 protein (nuclease), guiding Cas9 to any matching DNA sequence in the cell and degrading it. So any DNA matching the identifier (i.e. phage DNA) that enters the cell will be matched by the identifier and targeted for degradation by Cas9 if CRISPR-Cas proteins are being expressed.

However not all failed infections will result in identifier uptake. Further CRISPR-Cas can be metabolically expensive so it can incur a fitness cost in environments with low chance for infection. The most interesting part is that *any* DNA in the cytoplasm can be integrated as a CRISPR identifier. Random DNA floating in the environment, DNA resulting from aborted Horizontal Gene Transfer (HGT) events and even the cell's own DNA can all become guides for Cas9. CRISPR has been shown to interfere with HGT events which creates complex fitness dynamics for bacteria.

1.2 Horizontal Gene Transfer

Unlike eukaryotes, bacteria can engage in what is called HGT, transferring genes between organisms who are not related. There are 3 main ways this is done: 1) taking up DNA from the environment 2) transferring DNA between cells 3) transferred via phage infection. When such external DNA makes it's way inside the cytoplasm it can be integrated into the bacterium's genome and can be expressed.

In some cases random DNA can be inserted inside of important genes and disrupt their functions. In other cases things like antibiotic resistance genes can be gained, much faster than they would evolve naturally. Due to the mechanics of transfer, how often it fails and that events can be either functional or disruptive, fitness impacts often depend on the environment. How much HGT a cell allows (via transcription of necessary HGT machinery) can lead to complicated fitness dynamics.

1.3 CRISPR-Cas vs HGT

The mechanics of how these interactions work have been studied and are much to complicated for this report to discuss, but there is much variety in how CRISPR can interfere with HGT. Since CRISPR limits the integration of DNA into the genome then it is intuitive that it can deter HGT events as well, as well there is research demonstrating this. In fact it has also been shown that in some cases CRISPR can actually increase the rates of HGT in a bacterial population. This is further complicated by the fact that CRISPR-Cas expression can be modulated, as well as that CRISPR-Cas genes can themselves be transferred horizontally between bacteria. The variety of these interactions make understanding the fitness trade-off of HGT and CRISPR-Cas expression quite complex and specific and thus worthy of further study.

2 Model Description

We present a version of a modifier model to model the relationship between HGT and CRISPR-Cas activity in response to environmental threats. For simplicity we consider only a single antibiotic and a single bacteriophage and that the resistance allele confers full resistance to the antibiotic.

2.1 Alleles

We consider resistance allele R that represents whether a bacterium posses a antibiotic resistance gene. We also consider two modifier allele C, H that represent the expression of CRISPR-Cas and HGT machinery and respectively

Allele		Description
Major	Minor	
R	r	has/does not have resistance gene
H	h	HGT machinery is expressed/not expressed
C	c	CRISPR-Cas is expressed/not expressed

Table 1: Allele definitions

2.2 Environment

We consider 3 different environments

- E_n: Neutral, no threats
- E_b: Bacteriophage, increased risk of contact with phage
- E_a: Antibiotic, increased risk of contact with antibiotic

We consider various scenarios that involve different functions that govern the transition from one environment to the next. We consider 3 models of changing environments

Singular Threat: A singular event of antibiotic dosage or phage outbreak $e = (s, l)$ is defined by a length $l \in [0, M]$ and a start time $s \in [0, T]$. The maximum length of a event is M generations and the model runs over T generations.

Cyclical Threat: Regular events of antibiotic dosage or phage outbreak every $2l$ generations. More formally, each event is of length l and we pick a set start times for each event $s_1 = l, s_2 = s_1 + 2l, \ldots, s_i = s_{i-1} + 2l$, and continue until $s_i + 2l > T$. Note you could very easily adapt this to randomly sample start times and lengths by alternatively sampling a first event $e_1 = (s_1, l_1) \in ([0, T], [0, M])$, a second event $e_2 = (s_2, l_2) \in ([s_1 + l_1, T], [0, M])$ and so on until $s_i + l_i > T$.

Alternating Threat: It is defined similarly to the Cyclical Threat model but with switching between antibiotic dosage and phage outbreaks at each event.

2.3 Fitness

Now we want to know the fitness of each genotype in each environment, which we will define with the parameters s_p, s_m. Here s_p is the fitness benefit of protecting against the environmental threat (antibiotic or phage) and s_m is the metabolic cost of expressing either HGT or CRISPR-Cas machinery. [1] Both H and C express different proteins so any -HC genotype bacterium incurs a $2s_m$. The H allele acts like a modifier allele, it does not directly impact fitness outside of the metabolic cost but instead increases the success of gaining R via HGT. To ensure that we do see the invasion of protective alleles we also define $s_p >> s_m$ to better reflect the biological reality. Given the above we define the fitness of each genotype in each environments in the following table

[1] Note that while we call s_m the metabolic cost there are other fitness costs expressing CRISPR-Cas or HGT machinery such as taking in toxic gene products, CRISPR-Cas autoimmunity, potential genome instability etc. but we group them all under the s_m penalty.

Genotype	*Environment*		
	E_n	E_b	E_a
RCH	$1-2s_m$	$1+s_p-2s_m$	$1+s_p-2s_m$
RCh	$1-s_m$	$1+s_p-s_m$	$1+s_p-s_m$
RcH	$1-s_m$	$1-s_m$	$1+s_p-s_m$
Rch	1	1	$1+s_p$
rCH	$1-2s_m$	$1+s_p-2s_m$	$1-2s_m$
rCh	$1-s_m$	$1+s_p-s_m$	$1-s_m$
rcH	$1-s_m$	$1-s_m$	$1-s_m$
rch	1	1	1

Table 2: Relative fitness values for each genotype in each environment

2.4 Model Behaviour

We are modelling bacteria, so we have an infinite haploid population that undergoes *asexual* reproduction (no random mating). We have 8 genotypes so we must keep track of 8 frequencies $[x_1(t), \ldots, x_8(t)]$. For every generation we have 1) gene transfer 2) mutation and 3) selection, here gene transfer is like a random mating step. Note that x_g represents the frequency of the genotype g.

2.4.1 Gene Transfer

The benefit of the H allele over h is defined later as it relates to the transfer step. The probabilities of a successful transfer event between between two genotypes G_r =r- - and G_R =R- - (i.e. $r \to R$ transfer) are g_H and g_h for the -H-, -h- genotypes respectively where $g_H > g_h$. Note that the H allele matters for the both G_r, G_R bacteria, so given two -H- bacteria the success probability is $2g_H$, given that only the receiving bacteria is $-H-$ we have just g_H and given neither we have g_h. The population frequency after selection is defined as

$$x_g^t = x_g + \delta(g) \sum_{x_R} x_{\neg g} x_R T(x_{\neg g}, x_R)$$

where $T(g, \neg g)$ is defined as

$$T(g, \neg g) = \begin{cases} g_h & h \in g, \neg g \\ g_H & h \in g, H \in \neg g \\ g_H & H \in g, h \in \neg g \\ 2g_H & H \in g, \neg g \end{cases}$$

and $\delta(g)$ is defined as

$$\delta(g) = \begin{cases} 1 & R \in g \\ -1 & R \notin g \end{cases}$$

Note that if $g = RCH$ then $\neg g = rCH$, defined similarly for cH, Ch, ch genotype and $x_R = \{RCH, RcH, RCh, Rch\}$.

2.4.2 Mutation

Our mutation step is the same as in a regular modifier model. A bacterium can evolve the resistance gene at a rate $P(r \to R) = \mu_R$ and a rate $\sqrt{\mu_R}$ with the selective pressure (in E_a). Similarly it can lose the resistance gene at a rate $P(R \to r) = \mu_r$ and a rate $\sqrt{\mu_r}$ without the selective pressure (not in E_a). The genotype frequencies after mutation are

$$x_g^s = (1 - \mu(g, e))x_g^t + \mu(g, e)x_{\neg g}^t$$

where $\mu(g, e)$ is defined as

$$\mu(g,e) = \begin{cases} \mu_R & r \in g, e \neq E_a \\ \sqrt{\mu_R} & r \in g, e = E_a \\ \mu_r & R \in g, e \neq E_a \\ \sqrt{\mu_r} & R \in g, e = E_a \end{cases}$$

2.4.3 Selection

Our selection step is also the same as in a regular modifier model using our fitness values. The genotype frequencies after selection are

$$x'_g = \frac{x_g^s f(g,e)}{\bar{w}}$$

where $f(g)$ picks the correct fitness coefficient from Table 2 and $\bar{w}$ is the average fitness $\bar{w} = \sum_g x_g^s f(g)$.

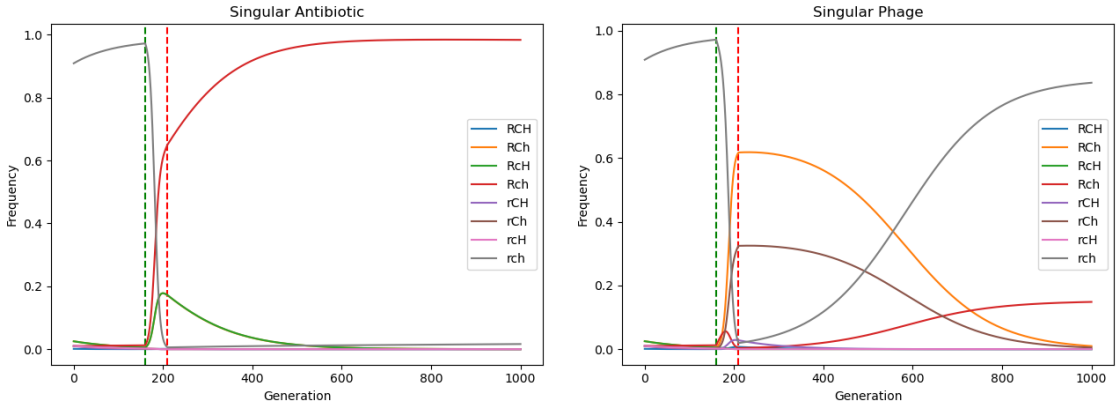
3 Results

We know present the results of the simulation under each environmental model. Note that all simulations were run with the initial population frequencies `[0.001,0.025,0.025,0.01,0.01,0.01,0.01,0.919]` (Table 2 order) and the following parameter values

g_h	g_H	μ_r	μ_R	s_p	s_m	n
0.000001	0.00005	0.00001	0.00001	0.2	0.01	1000

Table 3: Simulation parameter values.

The singular event models confirm that the model works as expected in response to events, although it appears that the antibiotic applies significantly more selective force on the population than the phage.

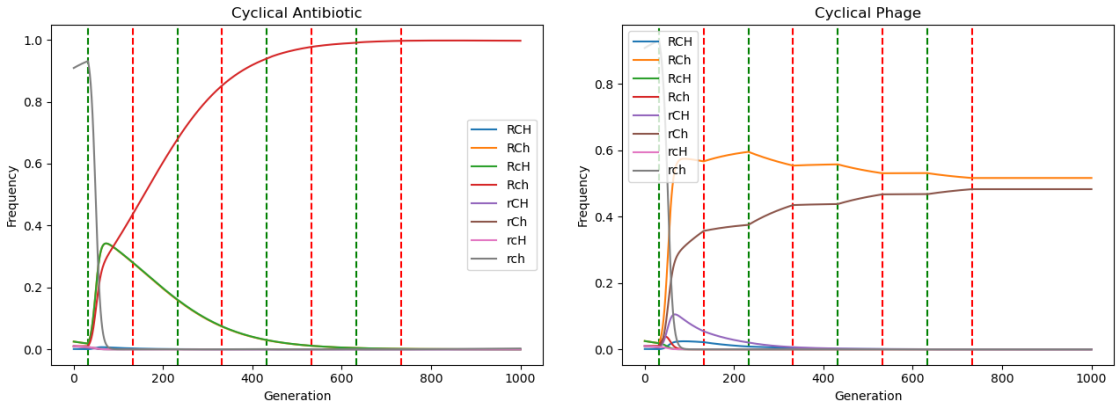


(a) Single Antibiotic Event, event starts at green dotted line and ends at red dotted line.l=50

(b) Single Phage Event, event starts at green dotted line and ends at red dotted line. l=50

Figure 1: Single event simulations.

With cyclical events, even with sustained selective pressure and a small metabolic cost that C, H allele are quickly selected against relatively quickly. We see undulations in the phage threat but unperturbed fixation with the antibiotic threat. Since the rCh, RCh are both the most fit genotypes in the cyclical phage environment it makes sense that they are the fixed alleles.

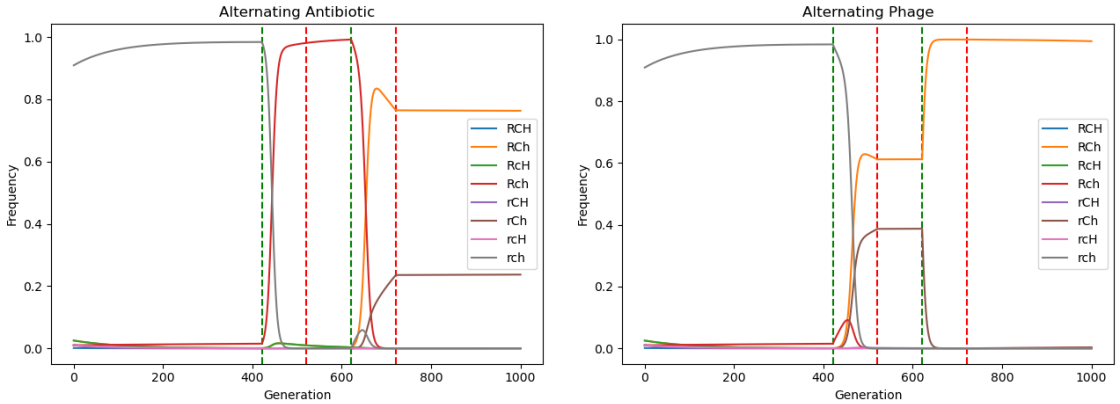


(a) Cyclical Antibiotic Events, events start at green dotted line and ends at red dotted line. l=100

(b) Cyclical Phage Events, events start at green dotted line and ends at red dotted line. l=100

Figure 2: Cyclical event simulations.

It is interesting to see the selective effect of the initial antibiotic breeding out most r genotypes and then seeing further selection of RC upon phage outbreak.



(a) Alternating Events, events start at green doted line and end at red dotted line, first event is Antibiotic. l=100

(b) Alternating Events, events start at green doted line and end at red dotted line, first event is Phage. l=100

Figure 3: Alternating event simulations.

The environmental stability show that no matter the frequency only a single dosage of antibiotic is enough to fix the Rch genotype in the population. However we see that the RCh, rCh genotypes competing even under the alternating model for long turnover periods, but see fixation of RCh with more consistent exposure to the antibiotic

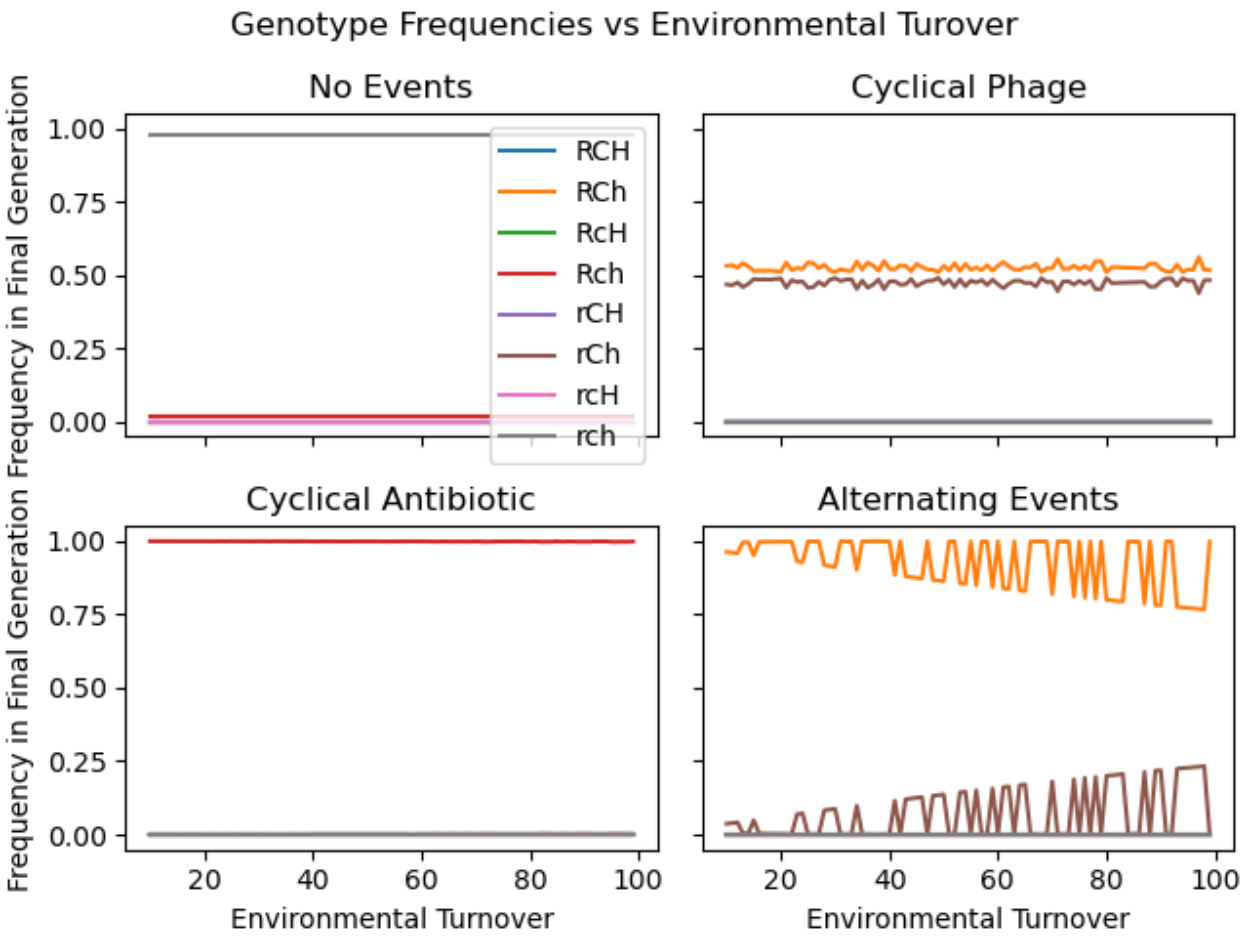


Figure 4: Environmental Stability, plot the genotypes frequencies in the 1000th generation for different event frequencies (l).

4 Discussion

Ultimately I learn that a large part of defining a model is knowing where to create abstractions for simplicity. CRISPR-Cas and HGT are both highly complex and have unintuitive interactions so trying to include all of these interactions would results in too many parameters to examine and difficulty in understanding.

I also found that often parameter spaces for models need to be explored thoroughly. Looking at how parameters vary with each other or what ranges of parameters can lead to chaotic behaviour are often the most interesting results from a paper.

Finally I realized the importance of defining parameters in your model such that they correspond nicely with the phenomena you are trying to study. This helps with determining how studying the behaviour of the model can lead to interpretable conclusions about the phenomenon being studied.