The effects of epidemic dynamics on MHC diversity

Nancy Hernandez-Ceron[1], Jesse E. Taylor[2], Anuj Mubayi[3]

[1] Department of Mathematics, Purdue University, West Lafayette, IN, USA.
[2] School of Mathematical and Statistical Sciences , Arizona State University, Tempe, AZ, USA.
[3] Department of Mathematics, Northeastern Illinois University, Chicago, IL, USA.

Abstract

The major histocompatibility complex (MHC) is a cluster of genes found in most vertebrate genomes which includes several gene families whose proteins play an important role in the recognition of foreign antigens. Pathogen-mediated selection (PMS) is believed to be responsible for the extraordinary levels of MHC diversity observed in humans and many other vertebrate species. Although there have been many theoretical studies of the relationship between PMS and MHC diversity, most have not incorporated the selective impact of epidemiological dynamics. A new discrete time agent-based model of MHC evolution, which includes epidemic processes, is introduced. In our model, we consider a finite population of diploid hosts which can be infected by one or more pathogen species that differ in transmission rate, virulence and mortality rate. Both the MHC genes of each host individual and the antigen gene sequences of each parasite are represented by binary sequences which can change over time as the result of mutations. Matches between the host's MHC sequence and a parasite's antigen sequence accelerate recovery from an infection.

1 Introduction

The molecules of the major histocompatibility complex (MHC) are proteins found on the surfaces of cells that help the immune system recognize foreign substances. MHC proteins can do this by binding to self-peptide and non-self-peptide antigens and presenting them to T-cells. The process of binding triggers a cascade of immune responses.

In humans these molecules are encoded by several genes all clustered in the same region on chromosome 6. Each gene has hundreds of alleles (alternate forms of a gene that produce alternate forms of the protein). As a result, most individuals are heterozygous at most MHC loci. It is rare for two individuals to have the same set of MHC molecules.

High MHC allelic diversity continues to challenge evolutionary biologists for an explanation. Recent models ([4], [8], [9], [10], [12]) suggest that a high number of alleles is implausible via heterozygote advantage alone (an heterozygote individual has two different forms of a particular gene). Studies of mice MHC suggested that with populations, individuals that are heterozygous at the MHC loci will have a higher fitness in pathogen-rich environments. Pathogen-mediated selection (PMS) is thought to be responsible for MHC diversity. There are several ways in which PMS may operate:

- Heterozygote advantage (heterosis) that arises because heterozygotes are able to recognize a larger class of antigens;
- Rare allele advantage which arise from host-parasite coevolution (pathogens evolve along with hosts, so when an allele is common pathogens avoid this "trend" in order to more effectively affect hosts);
- Fluctuating selection arising from spatial and temporal variation in parasite communities.

An agent-based model, discrete in time, is used to describe the long term effect of epidemics on MHC diversity. A population of a thousand hosts, each one with a "genetic sequence" representing their MHC genes, is analyzed. Twenty five species of pathogens with multiple strains are introduced into the host population, each species has specific epidemiological impact reflected in its parameters.

The dynamics of pathogens-host interactions is captured by an underlying stochastic SIR model with disease deaths. Immunity, which depends on the host's and pathogen's genes, is included in the recovery process. Immunity and disease death are the only sources of natural selection. A parameter of migration controls how often a pathogen species is introduced in the host population. The infection process is computed at a macro level but the recovery and death processes are verified at the individual level, depending on the genetic sequences carried by the host and the pathogen.

The aim of this project is to compare the relation between the level of MHC diversity, measured by how many different genetic sequences are present in the host population, with the parameters that controls disease death. We also compare the MHC diversity under neutral selection (no epidemics) and host and pathogen interactions.

2 Description of the model

We consider a population of N diploid hosts (hosts have two chromosomes). In order to keep the computations feasible we let $N = 1000$ and we model each of the two MHC loci carried by each individuals by a binary string of 16 bits. There are several ways to model this genetic sequence, however we chose this representation to be consistent with the work previously done by Borghans and Beltman [4] and Ejsmond el al [7].

$$\begin{array}{lcc} & \text{copy 1} & \text{copy 2} \\ i^{th} \text{ host} & g_{i,1}^{(1)} \cdots g_{i,16}^{(1)} & g_{i,1}^{(2)} \cdots g_{i,16}^{(2)} \end{array} \qquad i = 1, \cdots, 1000.$$

where $g_{i,j}^{(k)} \in \{0, 1\}$. In contrast, we assume that each pathogen species is haploid and carries a single antigen which is also represented by a binary string of 16 bits. There are 25 pathogen species and each species is characterized by three epidemiological parameters: β_s transmission, γ_s recovery and δ_s disease death (see section 2.1 for details).

The host population has a constant size of $N = 1000$. At every time step a host individual can die either from "natural causes" or from infection, and each deceased individual is replaced by a new (susceptible) host. This "offspring" chooses at random, without replacement, its parents. Both parts of the genetic sequence of the newborn (copy 1 and 2) is inherited from one of the parents chosen at random. There is a probability μ_H of mutation per site, so that the number of sites on which mutation (from 0 to 1 or from 1 to 0) occurs follows a binomial distribution with parameters (size=16, probability=μ_H). A Poisson distribution is used to approximate this binomial, as its mean $16\mu_H \ll 1$ (see poisson convergence in [6]).

At any given time step t there may be more than one strains of pathogens, from one or several species, interacting with the host population. The pathogen species defines the epidemiological impact through the parameters, β_s, γ_s and δ_s for $s \in \{1, \cdots, 25\}$, which are set at time zero (see section 2.1). See Table 1 for a description of parameters and symbols commonly used. New strains are originated by pathogen mutation so two strains have different genetic sequences. The infection process is done at a population level: we count the number of individuals infected with a given pathogen as well as the number of individuals that are susceptible to that pathogen. This information is then used to determine the number of new infections and that many susceptible individuals are chosen at random and infected with the pathogen (see section 2.1). We repeat this procedure for all the pathogens that are affecting the population at this time step.

We assume that mutation of the pathogen antigen sequences occurs with probability μ_P during transmission to a new host, in which case one of the sites in the antigen sequences is chosen at random and changed from 0 to 1 or vice versa. Each mutation is considered to generate a new strain which, however, has the same epidemiological parameters as the parental strain and belongs to the same species. This model does not consider cross-immunity, so that an individual previously infected by a particular species can be re-infected by another strain belonging to the same species. Superinfection is permitted. Every infected individual is assumed to be infectious for at least one time step following their own infection.

Table 1: Definition of symbols and parameters frequently used.

Symbol	Definition
β_s	Transmission of s^{th} pathogen species
γ_s	Probability of recovery of s^{th} pathogen species
δ_s	Probability of disease death by s^{th} pathogen species
μ_H	Host mutation ($= 10^{-3}$)
μ_P	Pathogen mutation ($= 10^{-3}$)
d	Probability of natural death in hosts population ($= 0.01$)
m	Pathogen migration ($= 0.001$)
k	Index used for chromosome copy ($k = 1, 2$).
i	Index used for host ($i = 1, \cdots, 1000$).
j	Index used for site in the genetic sequence ($j = 1, \cdots, 16$).
s	Index used for pathogen species ($s = 1, \cdots, 25$).

Recovery and death are simulated at the individual level. If immunity is present (as described below), an infected host is immediately removed from the infected class (for that particular pathogen). If immunity is not present then recovery occurs with probability γ_s. Immunity is assumed to be present if at least 5 *consecutive* bits in one of the host genetic copies and the pathogen genetic sequence are equal, as in the following examples

host (copy 1 or 2) 0 0 **0 1 1 0 1 0 1** 1 0 1 1 1 0 1
pathogen 0 1 $\underbrace{\mathbf{0\ 1\ 1\ 0\ 1\ 0\ 1}}$ 0 1 0 1 0 1 1

host (copy 1 or 2) 0 0 1 0 0 **0 0 1 0 1** 1 1 1 1 0 1
pathogen 0 1 1 0 1 $\underbrace{\mathbf{0\ 0\ 1\ 0\ 1}}$ 0 1 1 0 1 0

This procedure is repeated for every pathogen present at that time. The probability that a host with a randomly generated MHC copy is immune to a pathogen with a randomly generated sequence is about 20%. For more realistic values of immunity present in a human population see [4] and [7].

Natural death occurs with probability $d = 0.01$, per individual per time step, so that on average the life span of an uninfected host is 100 time steps. Disease death is also present in the model, with probability δ_s for all strains of species s. Specifically, an individual infected with species s survives for one additional time step with probability $(1 - d)(1 - \delta_s)$. More generally, we assume that the different sources of mortality act independently of one another, so that if an individual is infected by species $s_1, \cdots, s_a$, then the probability that they survive to the next time step is

$$(1 - d) \prod_{h=1}^{a} (1 - \delta_{s_h}).$$

The parameters δ_s will prove to be important as discussed in section 3. It is relevant to mention that death due to disease is verified after recovery, so no immune host can die in this way.

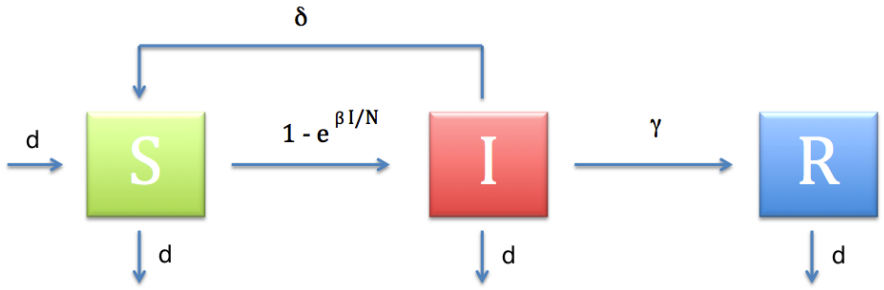

Figure 1: *SIR diagram of the discrete model.*

However, we allow immune host to be infected for one step to account for the new infections that can be produced by them, even though the disease does not represent any danger for them.

Finally, every species has probability m per time step of being reintroduced into the population. The number of pathogens that migrate into our population at time t follows a Binomial distribution, which is approximated by a Poisson random variable with parameter $\lambda = 25m$. If this Poisson random variable takes the value $\nu > 0$, then at the end of the simulation we pick ν hosts at random and infect them with a pathogen from species $s_1, \cdots, s_\nu$ taken at random (discretely uniformly) from $\{1, \cdots, 25\}$, one infection per host.

2.1 Single epidemic and species epidemiological parameters

A discrete deterministic model of a single epidemic is used to approximate the basic reproduction number $\mathcal{R}_0$ of our agent based model, when an epidemic is initiated by a single pathogen. The flow chart is given in Figure 1. The equations are as follows.

$$\begin{aligned} S(t+1) &= \underbrace{dN(t) + \delta I(t)}_{deaths} + e^{-\frac{\beta I(t)}{N(t)}} S(t) - dS(t), \\ I(t+1) &= \underbrace{\left(1 - e^{-\frac{\beta I(t)}{N(t)}}\right) S(t)}_{new\ infections} + I(t) - \underbrace{[dI(t) + \delta I(t)]}_{deaths} - \underbrace{\gamma I(t)}_{recovery}, \\ R(t+1) &= \gamma I(t) + (1-d)R(t), \end{aligned} \tag{1}$$

where $N(t) = S(t) + I(t) + R(t) = N$ is indeed constant. The first term in the right hand side equation of S is equivalent to our assumption that each deceased individual is replaced by a new offspring. The number of new cases at time $t+1$ is given by $\left(1 - e^{-\frac{\beta I(t)}{N(t)}}\right) S(t)$, a term commonly used in discrete models (see [2] and [5]).

Although this model is interesting by itself, we included it with the single purpose of computing $\mathcal{R}_0$ to impose a relationship between the epidemiological parameters β_s, γ_s and δ_s. We use the next generation matrix method described in [2]. The disease free equilibrium is given by $S = N, I = R = 0$. Then

$$I(t+1) = \underbrace{\left(1 - e^{-\frac{\beta I(t)}{N(t)}}\right) S(t)}_{new\ infections} + \underbrace{I(t) - dI(t) - \delta I(t) - \gamma I(t)}_{transitions}$$

$$\frac{\partial}{\partial I}\left[\left(1 - e^{-\frac{\beta I}{N}}\right) S\right] = \frac{\beta S}{N} e^{-\beta I/N} \quad \text{and} \quad \frac{\partial}{\partial I}\left(I - dI - \delta I - \gamma I\right) = 1 - d - \delta - \gamma.$$

At disease free equilibrium we get $F = \beta$ and $T = 1 - d - \delta - \gamma$, so that

$$\mathcal{R}_0 = \frac{\beta}{d + \delta + \gamma}.$$

To see precisely how this relates to our model, assume that at time t we have I individuals infected with a pathogen with parameters β_s, γ_s and δ_s. Assuming that we have S susceptible, then the number of new cases follows a Binomial $\left(S, 1 - e^{-\frac{\beta I}{N}}\right)$, because $e^{-\frac{\beta I}{N}}$ may be interpreted as the probability of staying susceptible for the next step. In a way, this part of the model can be seen as a chain binomial SIR model (see [1], [3] and [11]). As described before, in our model we simulate this binomial random variable and randomly pick susceptible individuals to infect. We repeat this for all pathogens interacting with the host population. Performing the infection process at the population level reduces the time needed to run the simulation. A more expensive, and perhaps more realistic, way to simulate it is by choosing the contacts (using perhaps a network matrix) and test if each contact does produce an epidemic. Figures 2, 3 and 4 show the number of infected, natural and disease deaths of a single epidemic, for different values of $\mathcal{R}_0, \beta, \delta, \gamma$ and β.

For every species s the parameters β_s, δ_s and γ_s are chosen in the following way: (1) choose $\mathcal{R}_0$ from a log-normal distribution around 3. (2) Choose δ_s, the disease death from a truncated exponential (bounded by 1) with mean parameter $\frac{1}{\lambda} = 0.05, 0.10, 0.15$ and 0.20 (see section 3.2). (3) Choose γ_s from a Uniform in (0,0.5). (4) Choose β_s so that $\mathcal{R}_0^{(s)} = \frac{\beta_s}{d+\gamma_s+\delta_s}$. One would expect that the choices of the distributions used do not change the model qualitatively. It would be interesting to verify robustness by changing such distributions and comparing the results in section 3. However, our focus has been on the qualitative nature of the effect of epidemics on MHC diversity.

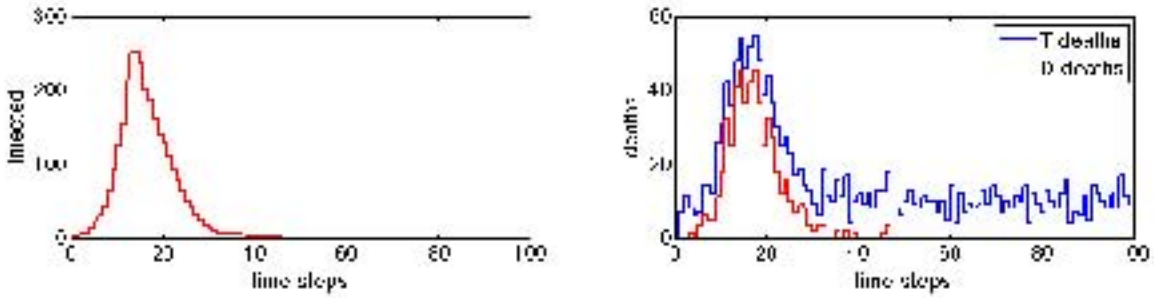

Figure 2: *A single epidemic with parameters* $\mathcal{R}_0 = 2.1849, \delta = 0.2350, \gamma = 0.2492, \beta = 1.0798$. *In the right hand side graphic, the upper line represents the number of deaths due to infection and the lower line total number of deaths.*

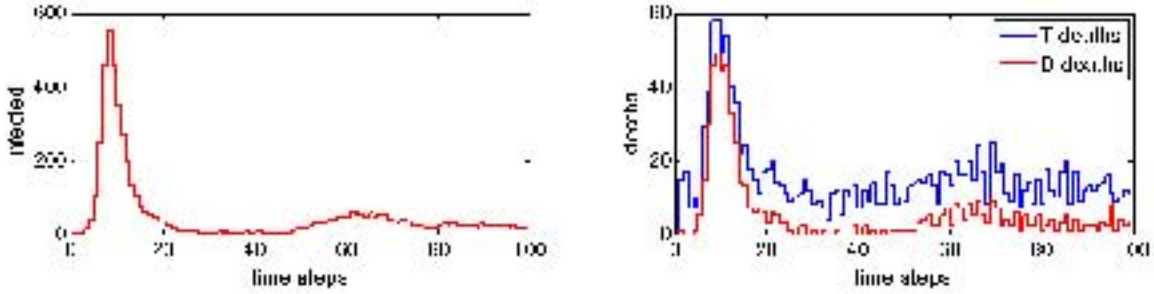

Figure 3: *A single epidemic with parameters* $\mathcal{R}_0 = 5.6634, \delta = 0.1856, \gamma = 0.0980$ *and* $\beta = 1.6625$

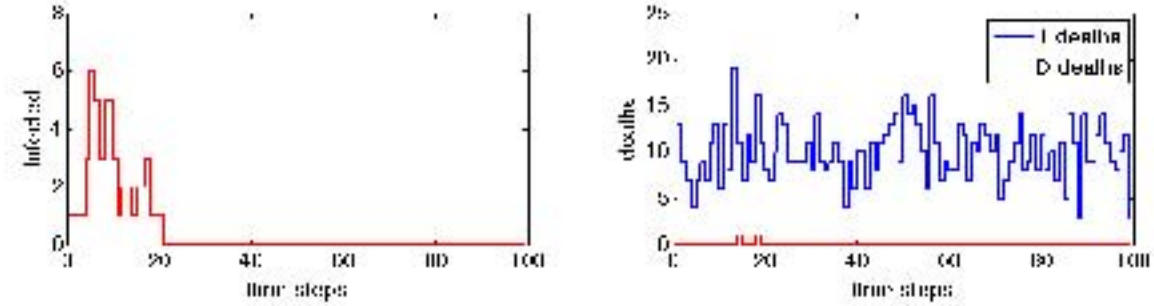

Figure 4: *A single epidemic with parameters* $\mathcal{R}_0 = 1.4590, \delta = 0.0515, \gamma = 0.2794$ *and* $\beta = 0.4974$

2.2 Mutation and Migration of pathogens

Our model allows parasite species to repeatedly immigrate into the population. At random times one or several species of pathogens may be introduced. A Poisson random variable with parameter $25m$ is generated after infection, recovery and death. If the value the generated number is larger than zero, say k, we choose k individuals and k species and infect each person with a particular strain of that species (one species per person). Figure 5 shows the process of introduction of a pathogen in the host population for 100 time steps.

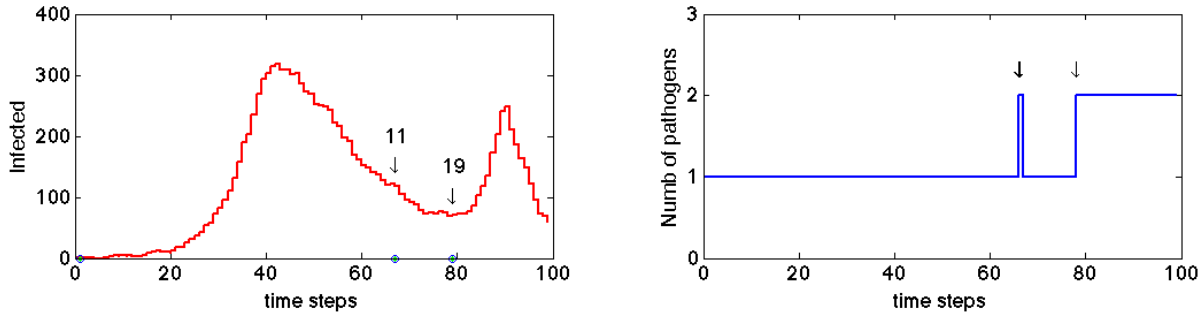

Figure 5: *At $t = 0$ we infect a host with a pathogen. Then, at $t = 66$ species 11 is introduced. At $t = 78$ species 19 is introduced. The right hand side graphic shows the number of pathogens present in the population, quickly after species 11 is introduced one of the two pathogen dies out. On the other hand, at $t = 78$, after species 19 was introduced, both pathogens survive.*

The parameter μ_P controls the probability of pathogen mutation per time step, a mutation generates a new strain of the same species, thus same epidemiological parameters. Mutation is verified after infection as follows: a Poisson random variable with parameter $\lambda = \mu_P I^*(t)$ is generated, where $I^*(t)$ represents the number of newly infected at time t. If the Poisson random variable is larger than zero we randomly choose a newly infected host and infect it with the new strain of the same species. For simplicity, mutation is restricted to only one site:

$$\begin{array}{rc} \text{Species } s & 0100110\underline{\mathbf{1}}11001011 \\ & \Downarrow \\ \text{new strain of species } s & 0100110\underline{\mathbf{0}}11001011 \end{array}$$

Figure 6 and 7 show the process of migration and mutation of a pathogen in the host population for 100 time steps and 500 time steps respectively.

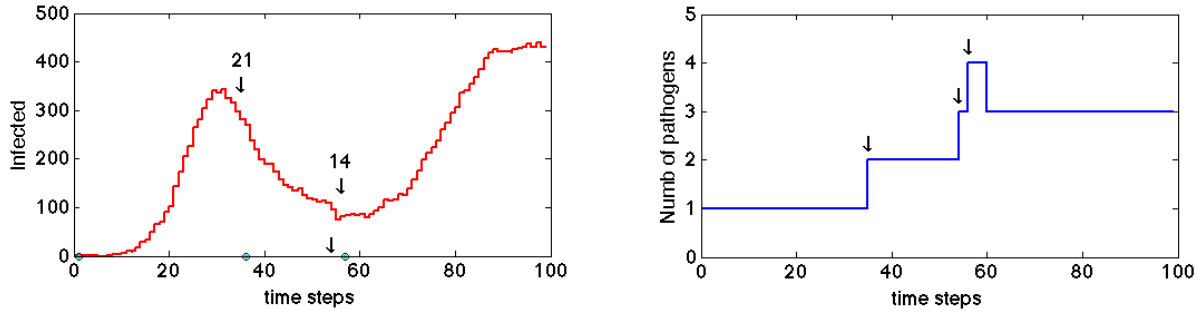

Figure 6: *At $t = 0$ we infect a host with a pathogen. Then, at $t = 34, 55$ species 21 and 14 are introduced respectively. Species 21 mutated at $t = 53$, this can be seen in the second graphic as the number of pathogens at that time $t = 53$ goes up from 2 to 3. However a pathogen vanishes at around $t = 60$.*

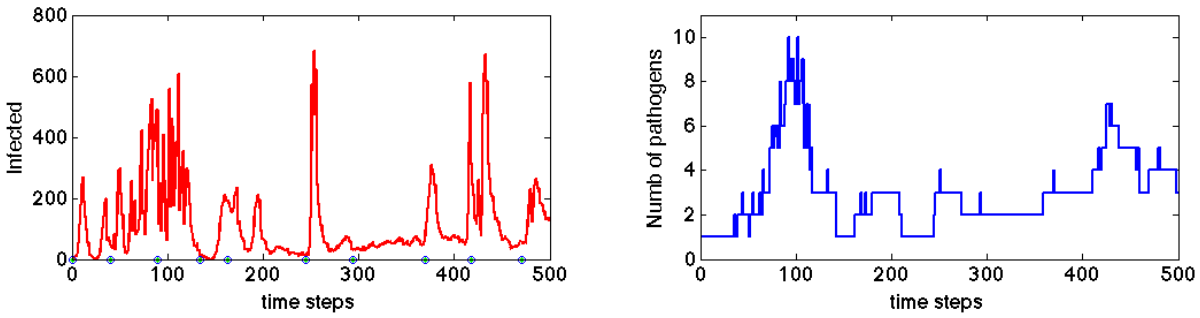

Figure 7: *Simulations with* 500 *time steps. This illustrates how the host-pathogen interaction dynamics become more complicated as time increases.*

Summary. At each time the order of events in the agent based model is:

- Infection to hosts.
- Mutation in pathogens (mutation may occur during transmission to a new host).
- Immunity and recovery in hosts.
- Death in host (disease and natural).
- Birth of new hosts. Mutation in host.

- Migration of a new pathogen.

3 Results

3.1 Comparison: Epidemics vs Neutral Selection

Under neutral selection, i.e. no pathogens interacting with the population, the number of different alleles present in the host population decreases dramatically. At $t = 0$ the genetic sequence assigned to any given individual was generated at random, so when t is relatively small a death results in losing of an allele because the offspring inherits its genetic sequence from its parents, who carry an already existing allele. This explains the exponential decay observed in Figure 8. On the other hand, mutation slowly reintroduces new MHC genotypes into the population, but the expected number of number of alleles present in the population is still small. In Figure 9 we observe that a larger mutation parameter promotes higher levels of MHC diversity. Figure 10 shows that a larger probability of natural dead (d) accelerates the convergence of the number of alleles present in the population. Since the average life span is given by $\frac{1}{d}$, larger values of d imply a much faster replacement of individuals by their offspring.

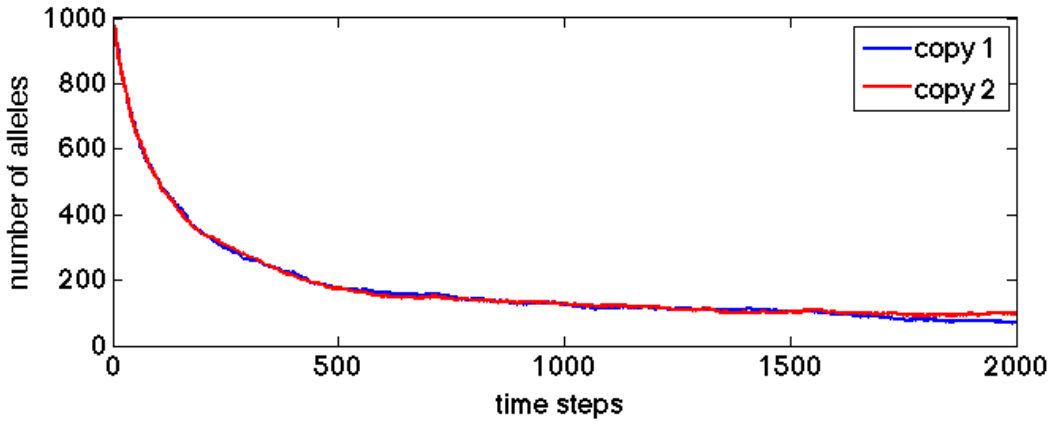

Figure 8: *Allele diversity with no epidemics,* $d = 0.01$ *and* $\mu_H = 10^{-3}$.

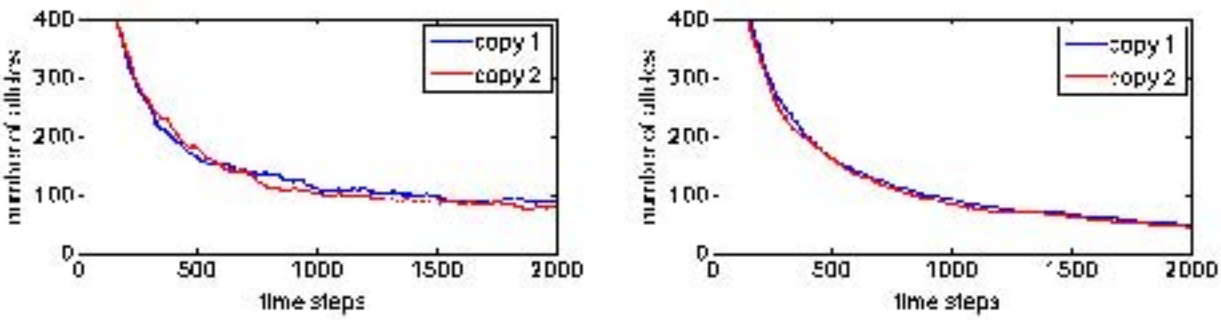

Figure 9: *Allele diversity with no epidemics,* $d = 0.01$. *Left hand side graph* $\mu_H = 10^{-3}$. *Right hand side graph* $\mu_H = 10^{-6}$.

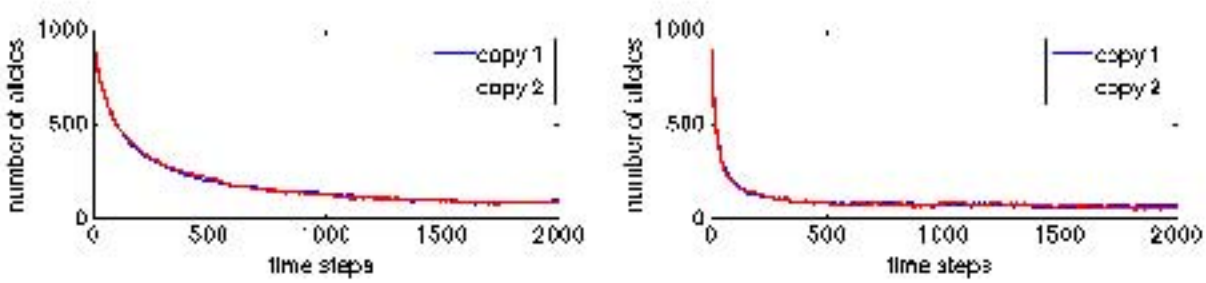

Figure 10: *Allele diversity with no epidemics,* $\mu_H = 10^{-3}$. *Left hand side graph* $d = 0.01$. *Right hand side graph* $d = 0.05$.

When pathogens are present there are many more factors to consider. Since selection is concentrated in the immunity and recovery processes a small disease death may not be enough to promote diversity in the population. On the other hand, MHC diversity will also be reduced if the mortality rates due to infection are very high, since in such cases there will be rapid turnover of the population as new offspring replace deceased individuals. Figures 11 and 12 shows the number of infected (typically by more than one pathogen), the number of deaths both natural and disease related, and the number of pathogens affecting the population at any time t. Figure 13 shows the number of alleles (average between the two hosts chromosomes) under two different severity of epidemics an neutral selection.

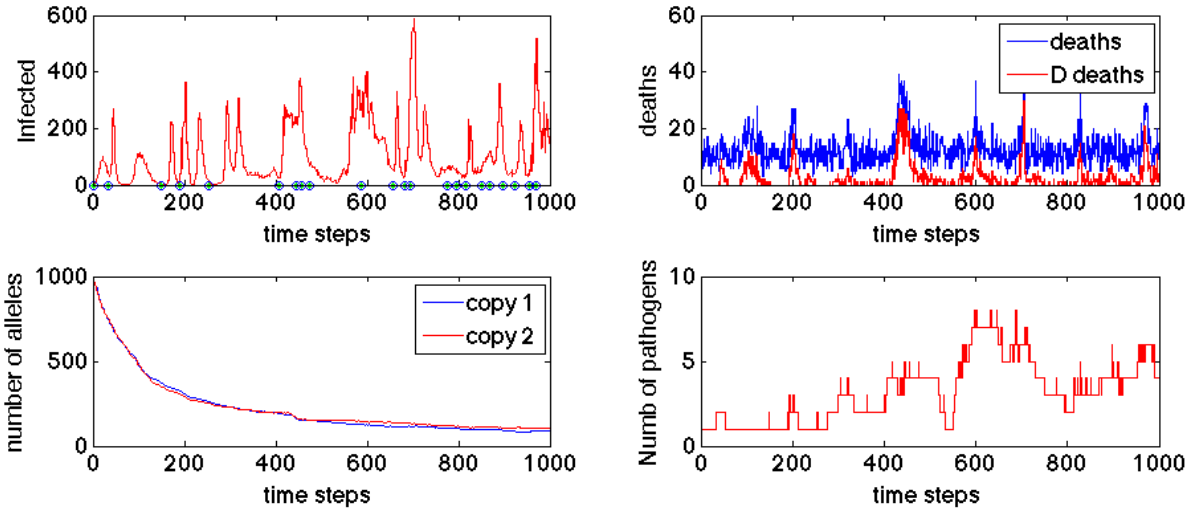

Figure 11: *Parameters:* $\mu_H = \mu_P = 10^{-3}, d = 0.01$ *and* $m = 0.001$. *The green dots in the top left graph (on the x axis) indicate the moment when a new species is introduced.*

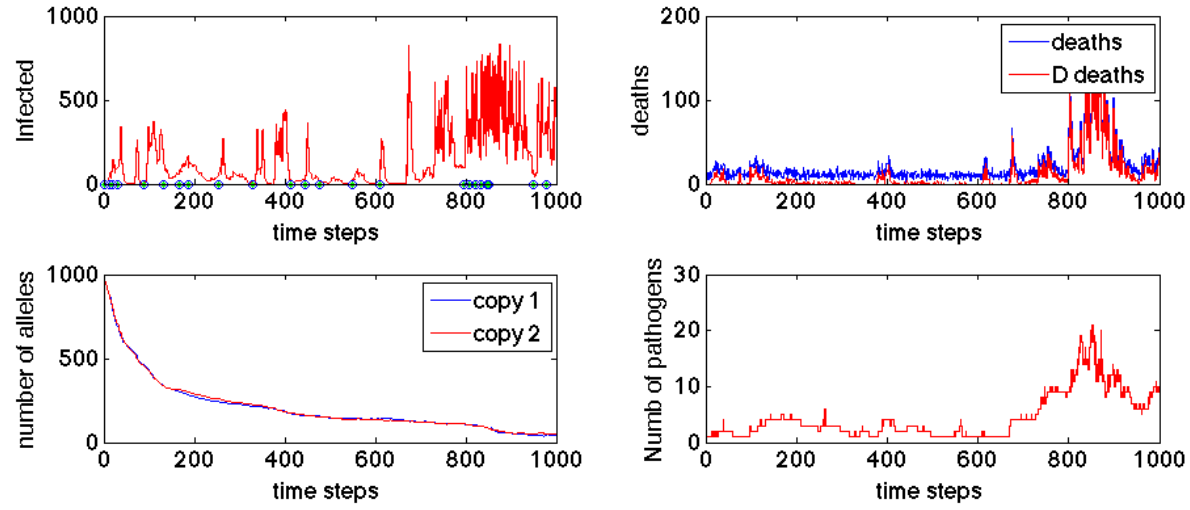

Figure 12: *Parameters:* $\mu_H = \mu_P = 10^{-3}, d = 0.01$ *and* $m = 0.001$

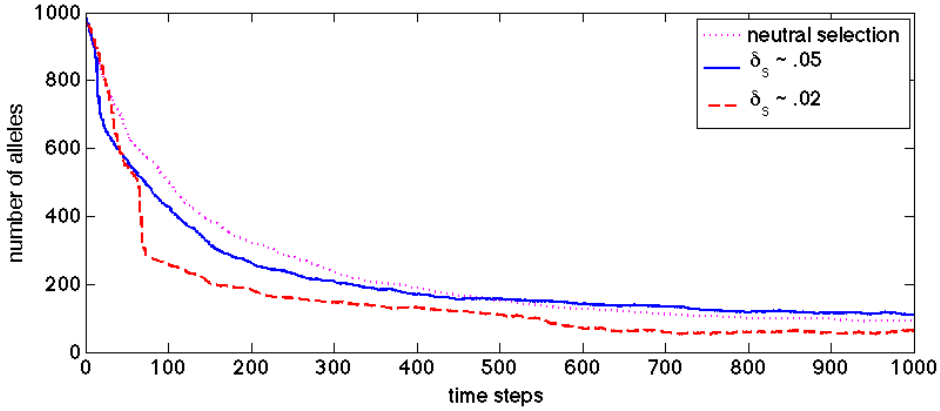

Figure 13: *One sample path of MHC diversity with and without epidemics. Different levels of disease death severity are considered. Parameters:* $\mu_H = \mu_P = 10^{-3}, d = 0.01$

In Figure 13 we compare a sample path of MHC diversity after 1000 time steps under (i) neutral selection (no epidemics) and (ii) epidemics with different levels of disease death severity, see section 3.2.

3.2 Distribution of allele diversity

In this section we present the results of hosts MHC diversity levels after 10,000 time steps under 4 scenarios. Scenarios 1, 2 and 3 include epidemic with different severity levels in terms of disease deaths. In the last scenario there are no epidemics at all (neutral selection). In all cases $d = 0.01, \mu_P = \mu_H = 10^{-1}$ and $m = 0.001$. As discussed in section 2.1 the parameters $\delta_s, i \in \{1, \cdots, 25\}$ are taken from a truncated exponential with mean parameter $\frac{1}{\lambda}$. For scenario a we used

$$\lambda_a = \begin{cases} 0.20 & a = 1, \\ 0.10 & a = 2, \\ 0.05 & a = 3. \end{cases}$$

For each case, 50 observations were simulated (100 data points in each histogram because hosts have chromosomes). Figure 14 suggests larger values of MHC diversity when the severity of the epidemics is low. However, the highest number of alleles present in the population was observed under neutral selection. Our simulations and graphics were performed in Matlab.

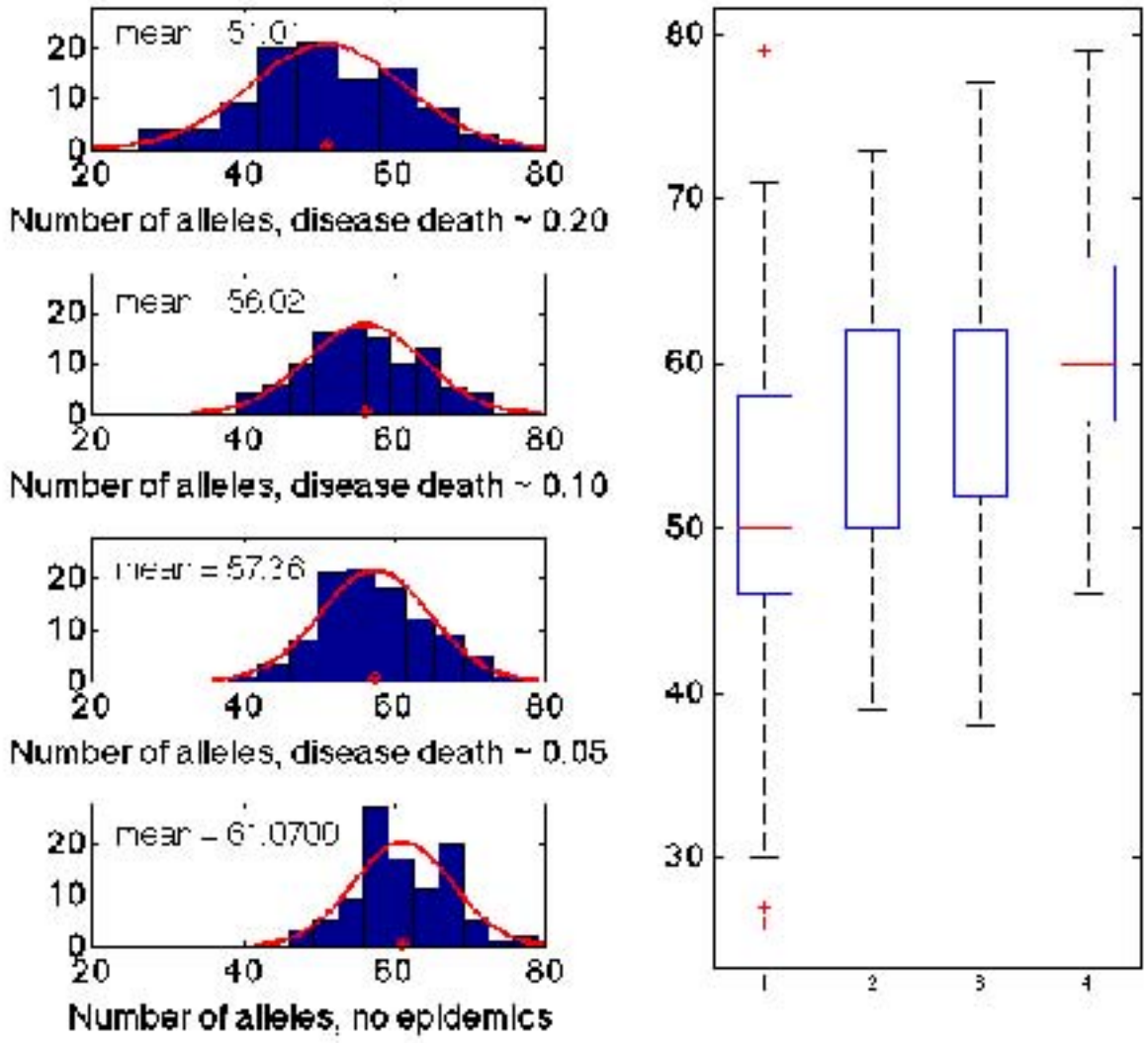

Figure 14: *Distribution of the number of alleles present in the population after 10,000 time steps.*

4 Discussion

Models of host-parasite coevolution, like the models studied in [4] and [7], often assume that

- infection reduces fecundity: a new generation of hosts is crated by a reproduction process

that largely depends on host fitness,

- infection changes the likelihood that an individual dies but not the overall number of deaths per step.

We would expect to see very different results if we incorporated these assumptions into our model.

To the best of our knowledge epidemic dynamics have not been used before, instead a common assumption is to assume that all individuals carry all pathogens. Although we did not prove an increase in MHC diversity in a population under multiple epidemics, we where able to show the importance and impact of severity of epidemics on MHC diversity.

A more realistic approach should include epidemic dynamics and natural selection in both the recovery process, by reducing the chances of disease death, and the fecundity process. Population dynamics may also be considered.

Acknowledgments

We would like to thank Dr. Carlos Castillo-Chavez, Executive Director of the Mathematical and Theoretical Biology Institute (MTBI), for giving us the opportunity to participate in this research program. We would also like to thank Co-Executive Summer Directors Dr. Erika T. Camacho and Dr. Stephen Wirkus for their efforts in planning and executing the day to day activities of MTBI. Special thanks to Dr. Jesse E. Taylor and Dr. Anuj Mubayi for all the help and knowledge provided. Lastly, thanks to Dr. Xiaoxia Wang and Dr. Jose Flores for proofreading and making a lot of suggestions to this technical report. This research was conducted in MTBI at the Mathematical, Computational and Modeling Sciences Center (MCMSC) at Arizona State University (ASU). This project has been partially supported by grants from the National Science Foundation (NSF - Grant DMPS-0838705), the National Security Agency (NSA - Grant H98230-11-1-0211), the Office of the President of ASU, and the Office of the Provost of ASU.

References

[1] L. J. S. Allen. *An Introduction to Stochastic Processes with Applications to Biology.* CRC Press, 2nd edition, 2010.

[2] L. J. S. Allen and P. van den Driessche. The basic reproduction number in some discrete-time epidemic models. *J. Difference Eq. Appl.*, pages 1–9, 2008.

[3] N. Bailey. *Mathematical Theory of Epidemics.* Griffin, London, 1957.

[4] J. Borghans, J. B. Beltman, and R. J. De Boer. MHC polymorphism under host-pathogen coevolution. *Immunogenetics*, 55(11):732739, 2004.

[5] F. Brauer, Z. Feng, and C. Castillo-Chavez. Discrete epidemic models. *Math. Biosc. and Eng.*, 7:1–15, 2010.

[6] R. Durrett. *Probability: Theory and Examples.* Cambridge, 4th edition, 2010.

[7] M. Ejsmond and J. Radwan. MHC diversity in bottlenecked populations: a simulation model. *Conserv Genet.*, 12:129–137, 2011.

[8] P. Hedrick. Pathogen resistance and genetic variation at MHC loci. *Evolution*, 56:19021908, 2002.

[9] A. Sanchez-Mazas, J. Lemaitre, and M. Currat. Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. *Phil. Trans. R. Soc.*, 367:830–839, 2012.

[10] L. Spurgin and D. S. Richardson. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. Lond.*, 277:979988, 2010.

[11] J. E. Taylor. *Lectures for APM 541: Stochastic Modeling in Biology.*

[12] H. Westerdahl, M. Asghar, D. Hasselquist, and S. Bensch. Quantitative disease resistance: to better understand parasite-mediated selection on major histocompatibility complex. *Proc R. Soc. B.*, 279:577–584, 2012.