MODULARITY IN COEVOLVING HOST-PATHOGEN SYSTEMS

Devin M. Drown[1], Molly Rorick[2], Rob Mills[3], and Alex Moffett[4]

[1] School of Biological Sciences, Washington State University
[2] Genetics Department, Yale University
[3] Natural systems Research Group, University of Southampton
[4] Department of Philosophy, University of Texas

Introduction

Modularity can be observed in biological systems at many different levels (*e.g.*, in gene regulatory networks (Hartwell et al. 1999) and phenotypic development (Wagner and Altenberg 1996)), and why it is there is an open question. Modularity allows one portion of a system to function in partial or full isolation from other portions of the system (Simon 1969).

In addition to functional decomposition, modularity also often refers to the repeated use of a modular component (Lipson, 2007). As Schlosser (2002) notes, in order to have multiple instantiations of a module, the modules must be decomposable -- indicating that functional decomposition is in some sense a more fundamental concept (see Simon (1969) and Schlosser (2002) for a discussion of modularity within dynamical systems, and Watson and Pollack (2005) for static fitness surfaces).

This study focuses primarily on modularity as functional decomposition: when interactions between genes are limited to a subset of genes in the system. When epistatic interactions are clustered within modules (subsets of genes) evolutionary constraints are consolidated: the smaller the set of genes that are affected by a substitution at a given locus, the less constrained this locus is, because its substitution has fewer pleiotropic effects on other loci that may be essential to organismal fitness.

In our model we want to allow some alleles to be compatible with a greater number of genetic backgrounds than other alleles. This broader compatibility is our way of capturing the context-independence and decomposability of a more modular allele. A genetic background can be simply represented by the identity of an allele at a second locus. For the minimal system we therefore only need two loci X and Y. A particular allele at X can then be compatible or not with a particular allele at Y. Hence we define allelic compatibility as follows:

> *Allelic compatibility:* given two loci X and Y, the allelic compatibility of allele i at X, or X_i, is the proportion of Y_j alleles that give rise to viable organisms.

Modeling modularity differences as differences in compatibility highlights the relevance of modularity in determining fitness in the context of recombination.

Recombination, in one form or another, is present in almost all forms of life (Hadany and Comeron 2008, Otto and Lenormand 2002). If selection for successful recombination or some other effect of recombination can foster the evolution of modularity, it might be possible to explain the prevalence of modularity in biological systems. Dynamic coevolutionary

environments are also a part of the greater context in which many forms of life evolve (Salathé et al 2008).

Research by Deem and co-authors indicates that modularity enhances evolvability (Bogorad and Deem 1999, Earl and Deem 2004, Sun and Deem 2007), although these studies use abstract dynamic environments to obtain such results. In this paper, we aim investigate the impact of modularity in a more specific and realistic biological scenario where we can explicitly include recombination. We use a coevolutionary system of host-pathogen interactions to provide dynamic environmental conditions for both the host and pathogen populations (Thompson 1982, Van Valen 1973). We model a pathogen population with two different levels of modularity, and we use this experimental set up to answer the question: does greater modularity in the pathogen lead to a selective advantage?

While interactions between species such as hosts and their pathogens (or plants and pollinators, or herbivores and plants) have long been implicated as a means of generating patterns of diversification (Ehrlich and Raven 1964; Thompson 1994, 2005), the process by which forces generate robustness and evolvability is not well understood, particularly for coevolutionary systems. Previous authors (see variational adaptation in Wagner et al. 2007) have suggested that environmental variation leading to temporal changes in fitness on different traits can select for modularity in genetic architecture. In the simplest form, host-pathogen interactions can lead to temporal variation in fitnesses. However, it remains unclear if this type of variation will generate the same kind of selective forces for modularity. These interactions provide an opportunity to study a different form of temporal variation in fitnesses and allow us to test this potentially strong form of selection for modularity.

We propose to test whether interactions between coevolving hosts and pathogens provide the conditions necessary for selection of pathogen modularity. In this paper, we model an abstract interaction between a host and pathogen. The pathogen alleles that are responsible for interaction with the host can be modular or non-modular. We compare the selective advantage of modular pathogen alleles over non-modular pathogen alleles. The working hypothesis is that antagonistic coevolutionary interactions generate substantial amounts of temporal variation in selection which selects for modularity of the pathogen alleles responsible for the interaction with the host.

Replicator Model

We have modeled an antagonistic interaction with two players (host and pathogen). In this system all individuals are haploid. Hosts have only one locus which is the interaction locus with six allelic variants ($A_{1\text{-}3}$ and $B_{1\text{-}3}$). Pathogens have two loci (X and Y). Each locus has three alleles. We include two levels of modularity in our system by reducing the combination of potential pathogen haplotypes. In our model, the X_1 allele is only compatible with the Y_1, whereas the X_2 and X_3 alleles are compatible with both Y_2 and Y_3 alleles. This reduces the total number of potential pathogen haplotypes to five. For the purposes of this paper we classify the X_1Y_1 pathogen as non-modular and the four other pathogen haplotypes as modular. Strictly speaking, it is the alleles that are expressing an increase or decrease in their allelic compatibility and so the alleles X_1 and Y_1 have a lower modularity than $X_{2\text{-}3}$ and $Y_{2\text{-}3}$. The frequency of each

host (H_i) and pathogen (P_i) haplotype is given in Table 1. As H_i and P_i are frequencies, it thus follows that $(\forall i)(0 \leq H_i \leq 1)$, $(\forall i)(0 \leq P_i \leq 1)$, $\sum_{i=1}^{6} H_i = 1$, and $\sum_{i=1}^{5} P_i = 1$.

Table 1: Haplotype variables

Symbol	Haplotype frequency
H_1	Frequency of hosts with the A_1 genotype
H_2	Frequency of hosts with the A_2 genotype
H_3	Frequency of hosts with the A_3 genotype
H_4	Frequency of hosts with the B_1 genotype
H_5	Frequency of hosts with the B_2 genotype
H_6	Frequency of hosts with the B_3 genotype
P_1	Frequency of pathogens with the X_1Y_1 genotype
P_2	Frequency of pathogens with the X_2Y_2 genotype
P_3	Frequency of pathogens with the X_2Y_3 genotype
P_4	Frequency of pathogens with the X_3Y_2 genotype
P_5	Frequency of pathogens with the X_3Y_3 genotype

The fitness outcome of an interaction between a host and pathogen is determined by allele combinations at the interaction loci. We use a straightforward inverse matching allele model for the interaction (Frank 1991). In our model, the host *A* alleles ($A_{1\text{-}3}$) match the pathogen alleles at the *X* locus ($X_{1\text{-}3}$) on a pair-wise basis (e.g. A_1 matches X_1). The host *B* alleles ($B_{1\text{-}3}$) match the pathogen alleles at the *Y* locus in the same way. If the host allele matches either of the pathogen alleles, the host resists infection by the parasite and has a fitness of 1. If neither of the pathogen alleles matches the host allele, the pathogen has a fitness of 1 because it can subvert the host immune system and infect the host without resistance. Host resistance decreases the pathogen's fitness by α and pathogen infection reduces the host's fitness by β. Table 2 explicitly identifies all of the potential interactions between host and pathogen genotypes in our system.

Table 2: Infection interactions

Host	Pathogen genetics				
	X_1Y_1	X_2Y_2	X_2Y_3	X_3Y_2	X_3Y_3
A_1	R	I	I	I	I
A_2	I	R	R	I	I
A_3	I	I	I	R	R
B_1	R	I	I	I	I
B_2	I	R	I	R	I
B_3	I	I	R	I	R

R = Host resists infection, pathogen fitness is reduced by α and host fitness is unaffected.
I = Host is infected, host fitness is reduced by β and pathogen fitness is unaffected.

We assume that the fitness of any particular host (W_i) or pathogen (V_i) will be determined by this interaction matrix as well as the frequency of encounter. From that we can derive the frequency dependent fitness of each of the hosts

$$W_1 = 1 - \beta(1 - P_1) \quad \text{(A1)}$$
$$W_2 = 1 - \beta\big(1 - (P_2 + P_3)\big) \quad \text{(A2)}$$
$$W_3 = 1 - \beta\big(1 - (P_4 + P_5)\big) \quad \text{(A3)}$$
$$W_4 = 1 - \beta(1 - P_1) \quad \text{(A4)}$$

$$W_5 = 1 - \beta\big(1 - (P_2 + P_4)\big) \quad \text{(A5)}$$
$$W_6 = 1 - \beta\big(1 - (P_3 + P_5)\big) \quad \text{(A6)}$$

and pathogens

$$V_1 = 1 - \alpha(H_1 + H_4) \quad \text{(A7)}$$
$$V_2 = 1 - \alpha(H_2 + H_5) \quad \text{(A8)}$$
$$V_3 = 1 - \alpha(H_2 + H_6) \quad \text{(A9)}$$
$$V_4 = 1 - \alpha(H_3 + H_5) \quad \text{(A10)}$$
$$V_5 = 1 - \alpha(H_3 + H_6) \quad \text{(A11)}$$

The host mean fitness is defined by

$$\bar{W} = \sum H_i W_i \quad \text{(A12)}$$

and the pathogen mean fitness is defined by

$$\bar{V} = \sum P_i V_i \quad \text{(A13)}$$

We take an abstract approach and cast this model in terms of replicator dynamics of each player in the interaction. The replicator equations capture the dynamics of the six host haplotypes are given by

$$\frac{dH_i}{dt} = H_i(W_i - \bar{W}) \quad \text{(A14)}$$

and five pathogen haplotypes by

$$\frac{dP_i}{dt} = P_i(V_i - \bar{V}) \quad \text{(A15)}$$

Recombination can be included in the replicator equation by altering the frequency terms to reflect changes in genotype frequencies that occur independently of their immediate effects on fitness. Assuming that recombination occurs and that it is limited to the pathogen genotypes, we can thus rewrite the pathogen frequencies from (A14) and (A15) in the following way. First, as X_1Y_1 is unable to recombine with any of the other pathogen genotypes, there is no effect of recombination on the frequency of this genotype. However, all other genotypes are affected by recombination. The pre-recombination haplotype frequencies are indicated by the lower case p, and D is the current linkage disequilibrium and given by the standard population genetic formula $D = p_2p_5 - p_3p_4$ and r is the recombination rate between the two loci. After recombination, the new frequencies of the modular pathogens are given by

$$P_2 = p_2 - rD \quad \text{(A16)}$$
$$P_3 = p_3 + rD \quad \text{(A17)}$$
$$P_4 = p_4 + rD \quad \text{(A18)}$$
$$P_5 = p_5 - rD \quad \text{(A19)}$$

These can be used to substitute into (A14) and (A15) to include recombination.

Results for replicator approach

Theoretical analysis

As is typical of most host-pathogen interactions, our system exhibits strong frequency dependent cycling. As a host becomes rare, it allows the pathogen that it detects to increase in frequency. This is due to the model of our interaction, inverse matching alleles (table 2) where a host is resistant to pathogens it can detect. Figure 1 shows sample dynamics of these cyclical interactions. It is also obvious from figure 1 that the magnitude of change in the frequencies become larger and larger through time, but due to the simulations being based on an infinite population size the frequencies of the alleles will never fix or become extinct.

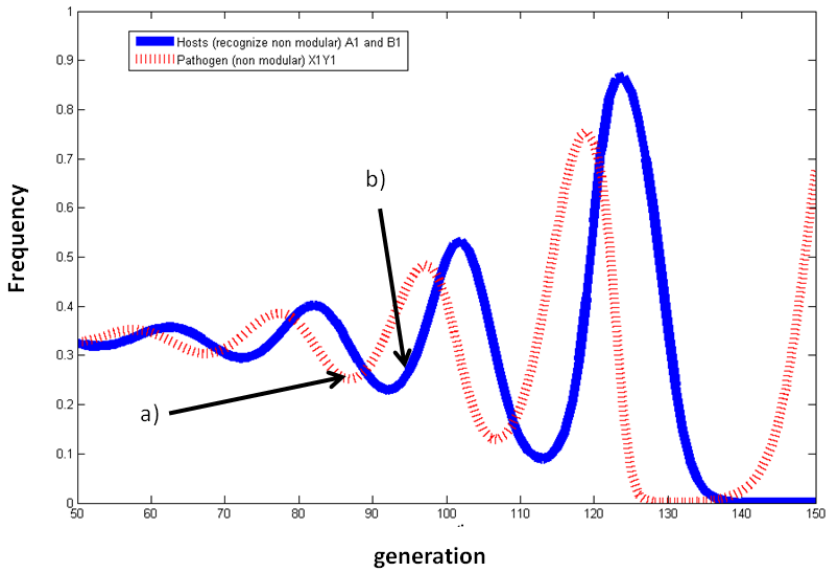

Figure 1: Sample temporal dynamics of replicator equations. Plotted is a time series of the frequency of hosts (A_1 or B_1) that recognize non-modular pathogens (solid blue line) and the frequency of the non-modular pathogen genotype X_1Y_1 (dashed red line). During the course of an interaction, frequency dependent selection is observed which produces cyclical interactions. At some point (**a**) the pathogen escapes selection when the host that detects it becomes rare. The pathogen starts increasing in frequency. When the pathogen becomes common, (**b**) the host that can detect this pathogen type will start increasing in frequency because this host is resistant.

To explore the dynamics of the replicator model, we recast the full model equations (A14-15) in terms of two new variables measuring the "degree of modularity" (ε) in the pathogen and the "degree of pathogen-modularity-detecting alleles" (δ) in the host populations. It holds that $-0.5 \leq \delta \leq 1$. When $\delta = -0.5$, then the A_1 and B_1 alleles are absent ($H_1 + H_4 = 0$) which means that there are only hosts that detect modular pathogen alleles. When $\delta = 0$, the average frequency of hosts of each type is equal ($\frac{H_1+H_4}{2} = \frac{H_2+H_3+H_5+H_6}{4}$). When $\delta = 1$, the $A_{2\text{-}3}$ and $B_{2\text{-}3}$ alleles are absent ($H_2 + H_3 + H_5 + H_6 = 0$) which means that there are no hosts that detect the modular pathogen alleles. It holds that $-0.5 \leq \varepsilon \leq 1$. When $\varepsilon = -0.5$, then the X_1 and Y_1 alleles are absent, which means that there are only pathogens with modular alleles ($P_1 = 0$). When $\varepsilon = 0$, pathogens alleles are equally frequent. When $\varepsilon = 1$, the $X_{2\text{-}3}$ and $Y_{2\text{-}3}$ alleles are absent which means that there are no pathogens with modular alleles($P_2 + P_3 + P_4 + P_5 = 0$).

We used these new variables to explore the system for potential equilibrium points in the system. We used to Matlab (MathWorks 2007) to simulate the direction of change of the new variables (δ and ε) over an arbitrarily small time step. We explored the full range of potential values for δ and ε while starting the host and pathogen frequencies at values based on the following

$$H_1 = H_4 = \frac{1+2\delta}{6} \quad \text{(A20)},$$

$$H_2 = H_3 = H_5 = H_6 = \frac{1-\delta}{6} \quad \text{(A21)},$$

$$P_1 = \frac{1+2\varepsilon}{3} \quad \text{(A22)},$$

$$P_2 = P_3 = P_4 = P_5 = \frac{1-\varepsilon}{6} \quad \text{(A23)}.$$

This ensures that all of the alleles are equally frequent at each locus when δ or ε = 0. Figure 2 shows the results of this analysis as a vector-field plot. It is clear from this plot that the cyclical dynamics we see in figure 1 are still captured by this reduced set of variables.

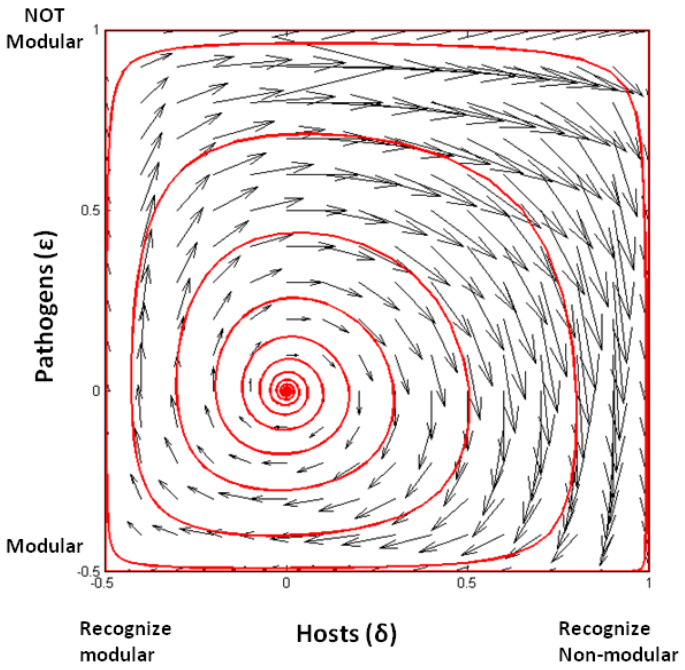

Figure 2: Vector field plot of change in pathogen modularity and hosts that detect modular pathogens. Arrows indicate the vector of change of δ and ε across their ranges in a single time step. An example trajectory of a single run which started just off the center equilibrium point (δ = ε = 0.01) over time (10,000 generations) with parameters (α = β = 0.1, r = 0.5) is shown (solid red line).

As a way to check that the vector-field plot is actually showing the evolution of host and pathogen populations in our system, we also ran several simulations over many more time steps (10,000 generations). A sample run is plotted across the range of δ and ε on figure 2 (red line). Changing the starting point did not change the general outcome which was cyclical changes in the values of δ and ε from their maximum to their minimum which represents a cycling of the modular and non-modular pathogens (and their respective hosts that detect them).

Based on our vector-field plots, we identified five equilibrium points: 1) all alleles in host and pathogen are equal (figure 2: center of spiral), 2-5) four different combinations where either the modular or non-modular alleles are fixed in the pathogen and host becomes fixed as well (figure 2: four corners). We used Mathematica (Wolfram Research 2007) to verify that these are all equilibrium points. Figure 2 reveals that none of these equilibria are stable (to determine if any are saddle or unstable equilibria will require further analysis).

Stochastic simulations

To explore the impact of a finite population size on our system, we created a stochastic version of our model. To accomplish this, at each time step iterated, we drew a sample host (N = 2000) and pathogen population (N = 2000) from the calculated genotype frequencies produced by the replicator equations. We then recalculated the resulting genotype frequencies from these finite populations to plug back into the replicator model. This method allowed us to explore whether any of the equilibria (albeit unstable) are larger attractors than any of the others. We could not address this in our previous analysis because we had an infinite population size which made it impossible to actually reach cases where alleles would fix.

We ran two types of simulations based on the above protocol. They differ in the starting frequencies of the host and pathogen alleles. In the first, we started each simulation replicate with all alleles at the same frequency per locus (host = 1/6, pathogen = 1/3). In the second, we started each simulation replicate with randomly generated initial starting frequencies. For both starting conditions, we ran 10,000 replicate runs and summarized the results. The two different conditions did not differ qualitatively in their results. Figure 3 indicates that the equilibria at the four corners are all attractors, but that some are occupied more frequently than others.

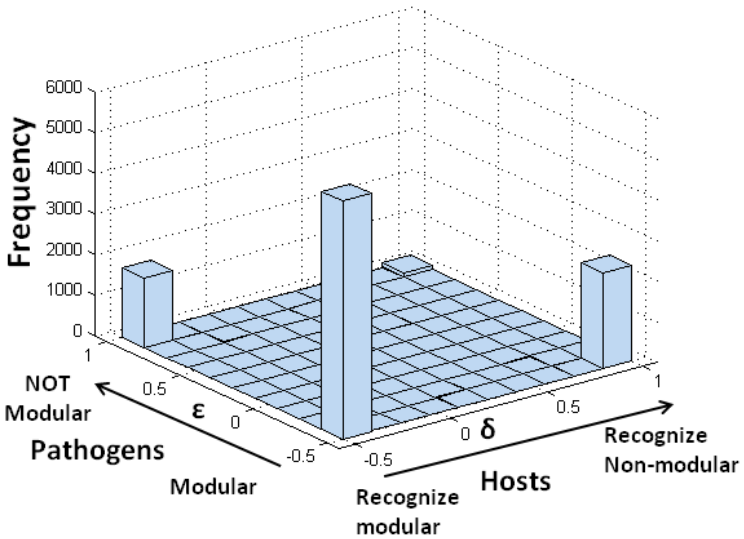

Figure 3: Stochastic simulation results based on 10,000 replicate runs with random initial allele frequencies for both host and pathogens with parameters ($\alpha = \beta = 0.1$, $r = 0.5$). The plot indicates the number of runs with values of δ and ε after 10,000 generations.

The attractor where non-modular pathogens and the hosts that recognize them are all fixed was the least frequently occupied. The attractor where the modular pathogens and the hosts that recognize them are all fixed was the most frequently occupied. It should be noted that because we have only presented the reduced system in terms of δ and ε, we are not stating that all of the modular pathogens are present at this most frequent attractor. These variables state that some unspecified variety of modular pathogens and modularity-matching hosts are present, and that none of the non-modular pathogens and hosts are in the system at this attractor. We present only the random starting conditions in figure 3, but the other results do not differ qualitatively.

Figure 4 presents the full results for both ways in which we initialized the system (random or equal allele frequencies). It is clear from a comparison of the stochastic simulations to the neutral case that the host-pathogen interaction results in the modular alleles becoming more frequent in the system.

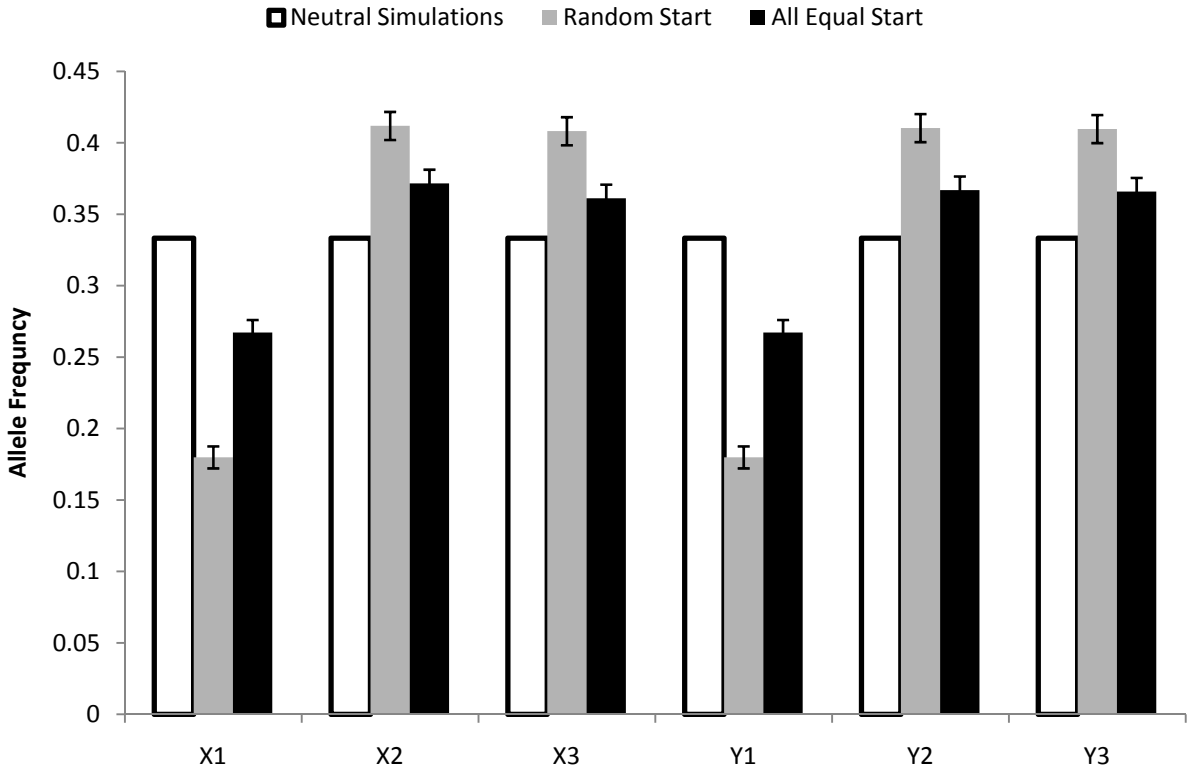

Figure 4: Stochastic simulation results based on 10,000 replicate runs with parameters ($\alpha = \beta = 0.1$, $r = 0.5$). The bars indicate mean frequency (± 1 st dev) of each of the pathogen alleles after 10,000 generations. Two different simulations (random initial allele frequencies: grey; equal allele frequencies: black) were run and compared to the neutral case (where $\alpha = \beta = 0$) where all the allele frequencies are equal.

Lotka-Volterra based model

We imagine a finite population of hosts (h) and pathogens (p). As in the classic Lotka-Volterra predator-prey model, the consumption rate (g) of the host by the pathogen is described by

$$g\big(p(t),h(t)\big) = acp(t)h(t) \quad \text{(M1)}$$

which, according to the mass-action assumption, is a simple function of the host and pathogen population sizes, the constant *c,* and the probability that a pathogen successfully infects a host, *a*. The rate of change of the pathogen (predator) population size *p* at time *t* is

$$\frac{dp}{dt} = \varepsilon g\big(p(t),h(t)\big) - l(p(t)) \quad \text{(M2)}$$

which is a function of the product of the conversion rate ε from host into pathogen, given successful infection, and the consumption rate as well as a constant per capita pathogen death rate *l*. The rate of change of the host (prey) population size *h* is

$$\frac{dh}{dt} = m\big(h(t)\big) - g\big(p(t),h(t)\big) \quad \text{(M3)}$$

which assumes a function a constant per capita growth rate *m* and a death rate, which is the consumption rate *g* of the host by the pathogen (M1).

We imagine that one host locus and two non-linked pathogen loci determine the probability, *a*, that a pathogen successfully infects a host (consumer successfully utilizes prey). We track the allele population sizes for this single host locus, and the allele population sizes for both pathogen loci. This infection probability for a given virus allele, *a*, is determined by 1) the allele frequencies at the other pathogen locus (because these determine the probability of various combinations of pathogen alleles (i.e. genotypes)), 2) the ability to recombine with the alleles present at the other loci, 3) the frequency of host genotypes in the population, and 4) a matrix of a priori determined infection abilities for specific pathogen-host interactions. Given successful infection, we assume a conversion rate, ε, of host into pathogen, equal to 1. We assume that the host and pathogen populations are both well mixed, and that many pathogen genotypes infect every host individual such that the population of pathogen genotypes within each host is representative of the whole population of pathogens. In one approach we allow non-viable pathogen allele combinations (genotypes) to reduce *a* (M4). The logic behind this is that we imagine that all the pathogen alleles freely recombine with each other within the host in accordance with the frequencies of the alleles, and the resulting recombinant genotypes survive according to their host-dependent fitnesses. Non-viable recombinants obviously do not survive, so the alleles that participate in these recombinants are lost from the population. In a second approach we control for the damage caused by non-viable recombinants. In other words, for a given pathogen allele, so long as there is some non-zero number of alleles in the population at the other locus with which it can successfully recombine, there is no harm caused by a prevalence of non-viable recombinants (M5). In this case our logic is that non-viable recombinants simply do not occur, that the viable recombinants occur in proportion to the frequencies of the alleles from which they are composed, and that these recombinants survive according to their host-dependent fitnesses.

$$a_n = \sum_{k=1}^{r} W_{i,j} f_k \quad \text{(M4)}$$

$$a_n = \sum_{k=1}^{r} W_{i,j} \frac{f_k}{\sum_{k=1}^{r} f_k} \quad \text{(M5)}$$

W is the fitness of the pathogen genotype j when it attempts to infect host genotype i, f_k is the frequency of allele k that viably recombines with allele n, and r is the total number of alleles with which allele n can recombine.

The matrix of W_{ij} values is determined by the inverse matching alleles model, which states that a pathogen genotype can successfully infect a host with fitness 1 if it can subvert the host immune system by containing no alleles that match the host's alleles. If, on the other hand, one of the alleles in the pathogen genotype matches the host's alleles, the pathogen can only infect the host with some fitness 1-s, where s is the selection coefficient.

$$W_{ij} = \begin{bmatrix} 0.9 & 1 & 1 & 1 & 1 \\ 1 & 0.9 & 0.9 & 1 & 1 \\ 1 & 1 & 1 & 0.9 & 0.9 \\ 0.9 & 1 & 1 & 1 & 1 \\ 1 & 0.9 & 1 & 0.9 & 1 \\ 1 & 1 & 0.9 & 1 & 0.9 \end{bmatrix}$$

We specify that the pathogen alleles 1, 2 and 3 are alleles at one locus, and pathogen alleles 4, 5 and 6 are alleles at the other locus, and that allele 1 can only recombine with allele 4 to produce a viable pathogen, whereas alleles 2 and 3 can both recombine with either of alleles 5 and 6. In this way we say that pathogen alleles 1 and 4 are "less modular" with respect to recombinability than pathogen alleles 2, 3, 5 and 6.

When W_{ij} is as shown below, and the other parameters are set as follows: c=.01, s=.1, l=.2, m=.4, and the number of each host and pathogen allele type is set to be 10 at t=1, the solution to the equations, numerically solved for t=1 through t=500, are shown in figures 5,6, and 7 (they differ depending on whether we use approach 1 or 2 for a).

Results for Lotka-Volterra Continuous Time Finite Population Approach

When we allow for non-viable recombinants to have an effect on pathogen allele population rates of change (a defined by approach 1, (M4)), modularity has a clear benefit. It allows alleles to recombine with a larger portion of the population, thus allowing these modular alleles to have more viable progeny (figure 5 and 6).

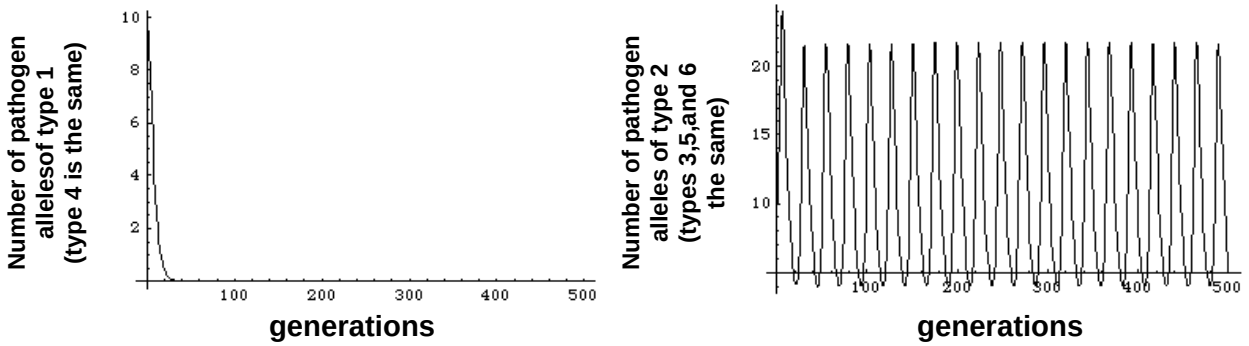

Figure 5: With a defined by (M4), the number of pathogen alleles of a given type within the population as a function of the number of generations.

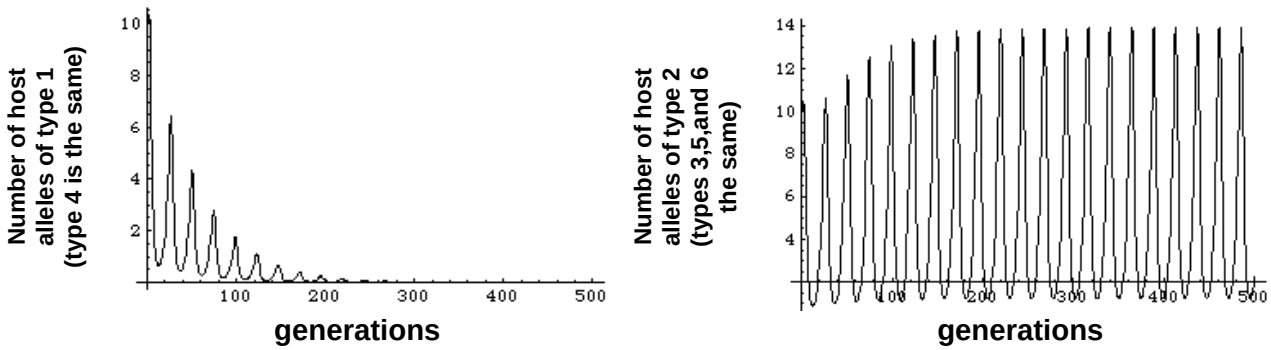

Figure 6: With *a* defined by (M4), the number of host alleles of a given type within the population as a function of the number of generations.

On the other hand, when we control for this, and only allow for viable recombination events to occur in the first place (*a* defined by approach 2, (M5)), there appears to be no benefit to modularity in this host-pathogen system (figure 7 and 8). The primary effect of modularity appears to be that it increases the likelihood of viable recombination, and thus, allows for a higher growth rate.

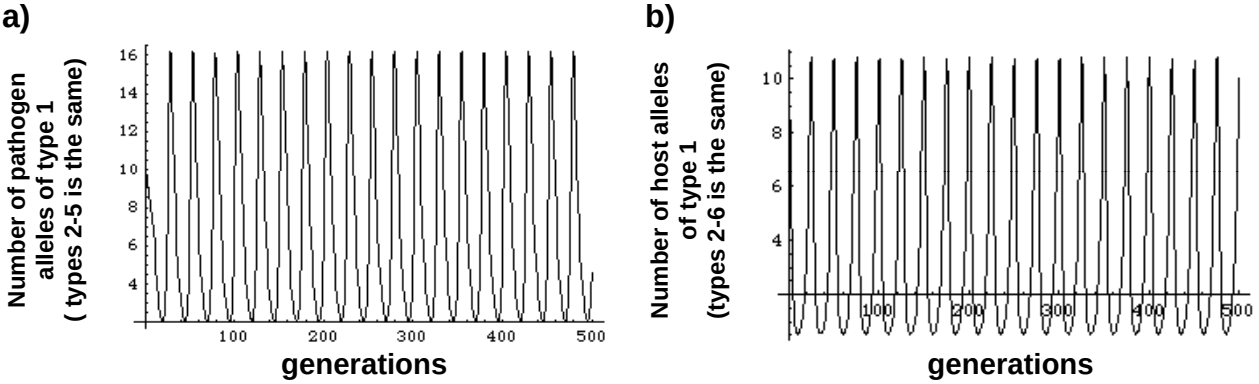

Figure 7: With *a* defined by (M5), the number of pathogen (a) or host (b) alleles of a given type within the population as a function of the number of generations.

Comparison to previous results

A study into evolvability by Earl and Deem (2004) uses a 'DNA swap' mechanism that makes large, but non-random genetic changes in addition to small-scale changes by mutation. The DNA swap is a modular form of variation, involving the substitution of genetic material for a particular genetic subdomain from a pool of low-energy alternatives for that subdomain.

Their investigation uses an abstract representation of protein sequence configurations, within which several substructures contribute to the energy of a given protein structure. The energy of these substructures is defined by a generalized NK landscape that assigns a random value to each possible configuration. In order to provide a varying environment, these random energy values are perturbed periodically, thus changing the configuration of the lowest energy protein sequences. They find that the more severe the perturbation, the more beneficial the large-scale changes are in finding low energy sequences. (They also find that DNA swaps are selected for in increasingly frequently perturbed environments). This work extends that of

Bogarad and Deem (1999) and similar results are obtained using a spin glass system as the model substrate (Sun and Deem 2007).

A related result is provided by Kashtan and Alon (2005) who show that under varying environmental conditions, genotypes with modular organization arise even in the face of an inherent cost of that modularity. Specific environmental conditions occur repeatedly over time in this study, in contrast to Earl and Deem's result where the environmental perturbations are random.

This body of prior work has constructed a case for the benefit of modularity in abstract dynamic environments. The results that we present here complement this argument by demonstrating that modularity in genetic architecture can be selected for in more specific biological circumstances.

Potential future directions

Our analysis of the effects of modularity on the evolution of coevolving host-pathogen systems has used a number of different modeling techniques. The models upon which our analysis is based range from stochastic individual-based models with finite population sizes to deterministic models based on differential equations and applied to infinite populations. In each of these models we have encountered support for the conclusion that modularity offers pathogens a selective advantage that favors the evolution of modularity within populations in which such modularity is initially present. While the pluralism thus employed in our approach to mathematical modeling has provided a greater degree of conformation of our central result than would be possible through the application of a single modeling technique, further support could nonetheless be obtained by studying additional models. Understanding of the full range of compatible models would indicate the range of conditions under which modularity is favored. Further work could also identify the regions of parameter space within each model that are consistent with our result.

In addition to considering whether our result is consistent with the particular host and pathogen populations presented in this analysis, further questions remain about whether modularity is also favored within other types of populations. For instance, while the pathogen population studied in this analysis consisted of pathogens that were either modular or non-modular, it is also possible to represent modularity as a quantitative trait in which pathogens can vary across a whole continuum of modularity values. As the models considered in this analysis can be easily updated to represent a wide range of different populations, our analysis offers a solid foundation upon which to study the selective advantage of modularity within different types of populations.

In considering the range of populations in which modularity is favored it might also be worthwhile to study the evolutionary dynamics between populations that include both modular hosts and modular pathogens. Whereas the present analysis considers only modular pathogens and non-modular hosts, the emergence of modularity within natural populations might more often result from the coevolution of modularity that exists within both host and pathogen populations. The selective advantage associated with the evolution of modularity within either a host or pathogen population may increase the selective pressure upon the competing population

to be able to evolve quickly in response to its quickly evolving competitors, thus increasing the likelihood that modularity will evolve within the population. The present analysis provides a strong foundation with which to study the impact of coevolution upon the evolution of modularity.

Conclusions

Our results suggest that 1) host-pathogen interactions are sufficient to select for modularity in the pathogen, 2) hosts that resist infection by modular pathogens are selected, and 3) significant insight can be gained through exploring simulations with finite population sizes.

Acknowledgements

The authors would like to thank Sarah Cobey for initial discussions and Jon Zelner for helpful comments. This project was undertaken as part of the Complex Systems Summer School 2008. This work was partially supported by the Santa Fe Institute whose research and education programs are supported by core funding from the National Science Foundation and by gifts and grants from individuals, corporations, other foundations, and members of the Institute's Business Network for Complex Systems Research. The IGERT Program in Evolutionary Modelling (IPEM) provided financial support to DMD.

References

Bogorad, L.D. and Deem, M.W. (1999) A hierarchical approach to protein molecular evolution. *PNAS* 96 2591-2595

Earl, D.J. and Deem, M.W., (2004) Evolvability is a selectable trait. *PNAS* 101(32):11531-11536

Ehrlich, P.R. and Raven P.H. (1964) Butterflies and plants: a study in coevolution. *Evolution* 18:586-608

Frank, S.A. (1991) Ecological and genetic models of host-pathogen coevolution. *Heredity* 67, 73-83

Hadany, L., and J. M. Comeron. 2008. Why Are Sex and Recombination So Common? *Ann NY Acad Sci* 1133:26-43.

Hartwell, L.H., Hopfield, J.J., Leibler,S., and Murray, A.W. (1999) From molecular to modular cell biology. *Nature*, 402:C47-C52.

Kashtan, N. and Alon, U. (2005) Spontaneous evolution of modularity and network motifs. *PNAS* 102(39):13773-13778

Lipson, H. (2007) Principles of modularity, regularity, and hierarchy for scalable systems, *Journal of Biological Physics and Chemistry*, 7(4):125-128.

Mathworks. (2007) *Matlab*. The MathWorks, Cambridge MA.

Otto, SP and Lenormand, T (2002) Evolution of sex: Resolving the paradox of sex and recombination. *Nature Reviews Genetics* 3:252-261.

Salathé, M., R. D. Kouyos, and S. Bonhoeffer. 2008. The state of affairs in the kingdom of the Red Queen. *Trends in Ecology & Evolution* 23:439-445.

Schlosser, G. (2002) Modularity and the Units of Evolution. *Theory in Biosciences* 121(1):1-80

Simon, H.A. (1969) *The sciences of the Artificial*. MIT Press, Cambridge, MA.

Sun, J. and Deem, M.W. (2007) Spontaneous Emergence of Modularity in a Model of Evolving Individuals. *Phys. Rev. Lett* 99:228107

Thompson, J. N. (1982) *Interaction and coevolution*. Wiley, New York

Thompson, J.N. (1994) *The Coevolutionary Process*. Chicago, University of Chicago Press

Thompson, J.N. (2005) *The Geographic Mosaic of Coevolution.* Chicago, University of Chicago Press.

Van Valen, L. (1973) A New Evolutionary Lawl. *Evolutionary Theory* 1, p. 1-30.

Wagner, G.P. and Altenberg, L. (1996) Complex adaptations and the evolution of evolvability. *Evolution*, 50(3):967-976.

Wagner GP, Pavlicev M and Cheverud JM. (2007) The road to modularity. *Nature Reviews Genetics*. 8: 921.

Watson, R. A. and Pollack, J.B. (2005) Modular Interdependency in Complex Dynamical Systems. *Artificial Life* 11 (4). pp. 445-457

Wolfram Research, Inc. (2007) *Mathematica*. Wolfram Research, Inc., Champaign, Illinois.