Dynamic metacommunity model to evaluate HPV strain interactions

Joe Mihaljevic

October 1, 2015

Introduction

The goal of this project is to create a dynamic, statistical, metacommunity model that estimates correlations among species' persistence and colonization probabilities. This will allow for inference into which species may ecologically interact via priority effects, facilitation, or competition. We will then apply this model to a data set of HPV co-infection dynamics across a large number of patients (i.e. patches). The main novelties of this project are: (1) that we have a unique data set with many HPV strains across a large number of patients, and (2) that this is the first stastistical metacommunity model that estimates pair-wise correlations in the probabilities of colonization and persistence.

Model Specification

Here we will outline the dynamic, statistical metacommunity model. This model is based off of Dorazio et al. 2010, Ecology. Importantly, this model does not include the estimation of probability of detection, which corresponds in Dorazio 2010 to the probability of observing rare species. *However, detection probability could later be incorporated in our model to reflect the sensitivity and specificity of HPV genotype testing.*

Our data set consists of K number of patients, who have been scored for the binomial occupancy of $I = 37$ number of HPV strains over T_k number of visits, where T_k is allowed to vary among patients. Thus the data are of type y_{ikt}, which simply consists of a 0 or 1, depending on whether HPV strain i was present in patient k on visit t.

First, we must estimate the initial occupancy probability for each HPV strain, which will then be used in a Markov Process to estimate how strain occupancy changes within and among individuals through time.

$$y_{ik1} \sim Bern(\psi_{ik1})$$

ψ_{ik1} estimates the probability that strain i occupies patient k at the patient's first visit. For now, we will assume that this initial occupancy probability is strain-specific and proportional to each strain's overall occcurance, but is not affected by any specific covariates. Then, the occupancy probabilities for all subsequent visits depend upon the occupancy status of the previous visit, as follows:

$$y_{ik,t+1} \sim Bern(\psi_{ik,t+1})$$
$$\psi_{ik,t+1} = \phi_{ikt}\psi_{ikt} + \gamma_{ikt}(1 - \psi_{ikt})$$

In this formulation, the likelihood of a patient retaining HPV strain i from one visit to the next depends upon the probability of strain i being present at the previous visit, ψ_{ikt}, multiplied by the strain-, patient- and time-specific probability of persistence, ϕ_{ikt}. This persistence probability can also be usefully thought of in terms of an extinction probability, $\xi_{ikt} = 1 - \phi_{ikt}$. Then, the probability that a patient k, who is previously uninfected with HPV strain i, subsequently becomes infected by strain i at time $t + 1$ is determined by the colonization probability, γ_{ikt}.

Covariate effects and correlations

One of our main goals is to determine how strain interactions might affect the ability of a given strain to colonize a new host or to persist in a host. However, we also want to control for the fact that certain strains may respond similarly to certain characteristics of the host environment. For example, if HPV strains A and B both favor young men, we would like to control for this common response when estimating how strain A might affect strain B's colonization probability. In order to account for strain responses to covariates and to simultaneosly estimate covariance among strains, we estimate colonization and persistence probabilities in the following way. For clarity, we will only present the model for persistence probability, although colonization probability is modeled in the same way.

$$logit(\phi_{ikt}) = \alpha_i + \mathbf{B}_i\mathbf{X}_k + \mathbf{B}_i^k\mathbf{X}_{t[k]}$$

This is a multi-level model in which patients have variable numbers of visits and some covariates are measured at the patient level and are static throughout the survey period, while other covariates are dynamic through time. Importantly, the structure of the data makes it easy to incorporate time-varying covariates in this framework, as the full set of covariates are measured at each visit, and time is indexed by visit number. In the above expression, α_i is the strain-specific baseline persistence probability, given average covariate values. $\mathbf{B}_i = (b_{\phi 0i}, b_{1i}, \dots, b_{qi})^{'}$ for q number of covariate effects measured at the patient level that do not change across visits. Thus, $\mathbf{X}_k = (1, x_{1k}, x_{2k}, \dots, x_{qk})$ holds the patient-level values of these covariates.

The next component of the model contains $\mathbf{B}_i^k = (b_{\phi 1i}^k, b_{1i}^k, \dots, b_{ri}^k)^{'}$. These are strain-specific responses to covariates, $(1, ..., r-1)$, that were measured at the patient-level. However, these covariates, $\mathbf{X}_{t[k]} = (1, x_{1t[k]}, x_{2t[k]}, \dots, x_{rt[k]})$, are those covariates which were measured at the patient level and potentially have unique values for each patient visit t.

We account for correlations between strains' colonization and persistence probabilities among and within patientsby drawing random effects b_0 and b_1^k from multivariate normal distributions.

$$b_0 \sim N(0, \Sigma_{b_0}),$$

where Σ_{b_0} is a covariance matrix, which accounts for among-patient correlations between strains' persistence and colonization probabilities. As the simplest example, in a two-strain scenario:

$$\Sigma_{b_0} = \begin{bmatrix} \sigma^2_{\phi_1} & \rho_{\phi_1\phi_2}\sigma_{\phi_1}\sigma_{\phi_2} & \rho_{\phi_1\gamma_1}\sigma_{\phi_1}\sigma_{\gamma_1} & \rho_{\phi_1\gamma_2}\sigma_{\phi_1}\sigma_{\gamma_2} \\ \rho_{\phi_1\phi_2}\sigma_{\phi_1}\sigma_{\phi_2} & \sigma^2_{\phi_2} & \rho_{\phi_2\gamma_1}\sigma_{\phi_2}\sigma_{\gamma_1} & \rho_{\phi_2\gamma_2}\sigma_{\phi_2}\sigma_{\gamma_2} \\ \rho_{\phi_1\gamma_1}\sigma_{\phi_1}\sigma_{\gamma_1} & \rho_{\phi_2\gamma_1}\sigma_{\phi_2}\sigma_{\gamma_1} & \sigma^2_{\gamma_1} & \rho_{\gamma_1\gamma_2}\sigma_{\gamma_1}\sigma_{\gamma_2} \\ \rho_{\phi_1\gamma_2}\sigma_{\phi_1}\sigma_{\gamma_2} & \rho_{\phi_2\gamma_2}\sigma_{\phi_2}\sigma_{\gamma_2} & \rho_{\gamma_1\gamma_2}\sigma_{\gamma_1}\sigma_{\gamma_2} & \sigma^2_{\gamma_2} \end{bmatrix}$$

Here $\sigma^2_{\phi_1}$ is the variance in persistence probability for strain 1 across all patient samples, and so on. Similarly:

$$b^k_1 \sim N(0, \Sigma_{b^k_1}),$$

In this case, the variances on the diagonal of $\Sigma_{b^k_1}$ represent the within-patient variability in persistence and colonization probabilities.

Our model therefore estimates both the among-patient and within-patient strain correlations in persistence and colonization probabilities. It is important to distinguish these among- and within-patient effects because the correlations among strains may be scale-dependent, with different processes leading to different patterns between the two scales (i.e. Simpson's paradox). For example, across all patients and visits, strains A and B may have positively correlated colonization probabilities resulting from the fact that both strain A and B tend to only be able to infect immuno-compromised individuals. However, within a patient, the strains' persistence or colonization probabilities might be negatively correlated if previous infection with strain A hinders the ability of strain B to superinfect.

Simulation as proof of concept

Here we will show that our model works in the sense that we can recover imposed parameters from simulated data.

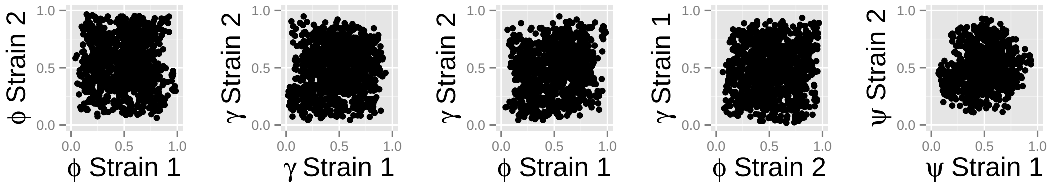
Our $sim_func()$ function simulates the occurrence of two HPV strains across a pre-determined number of patients and visits. The function allows us to manipulate many parameters including the strengths of within- and among-patient correlations in persistence and colonization. It also includes one patient-level and one visit-level covariate, the effects of which can be manipulated for each strain.

This is a simulated 'null' case in which we impose no correlations, and no covariate effects.

```
> source("sim_func.R")
> null <- sim_func(n.pat = 100,
+                        n.vis = 10,
+                        # Within and among host covariate effects for phi and gamma:
+                        bpat1g = 0,
+                        bpat2g = 0,
+                        btime1g = 0,
+                        btime2g = 0,
+                        bpat1p = 0,
+                        bpat2p = 0,
+                        btime1p = 0,
+                        btime2p = 0,
+                        # Correlations:
+                        ## Among-patients:
+                        raG1G2 = 0, #rho among-patient gamma_strain1, gamma_strain2
+                        raP1P2 = 0,
+                        raP1G1 = 0,
+                        raP2G2 = 0,
+                        raP1G2 = 0,
+                        raP2G1 = 0,
+                        ## Within-patients:
+                        rwG1G2 = 0,
+                        rwP1P2 = 0,
+                        rwP1G1 = 0,
+                        rwP2G2 = 0,
+                        rwP1G2 = 0,
+                        rwP2G1 = 0,
+                        # sd across patients for each strain (phi and gamma):
+                        # Assume all sd equal
+                        sa = 1,
+                        # sd within patients for each strain (phi and gamma):
+                        # Assume all sd equal
+                        sw = .4,
+                        #Global probabilities:
+                        globphi = .5,
+                        globgam = .5,
+                        globpsi = .5)
```

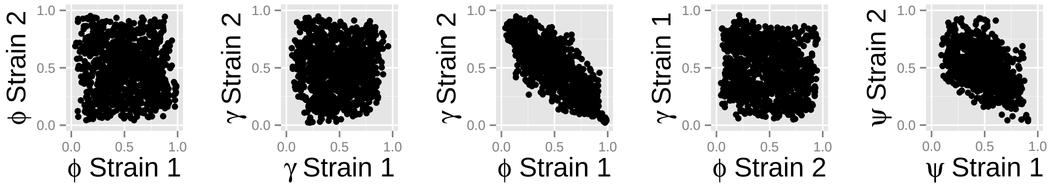
This allows you to see all of the function's options, and you can see that the correlations are all set to zero.

Now we can view the correlations in γ and ϕ, and the induced correlations in ψ. The null example should have no such correlations.



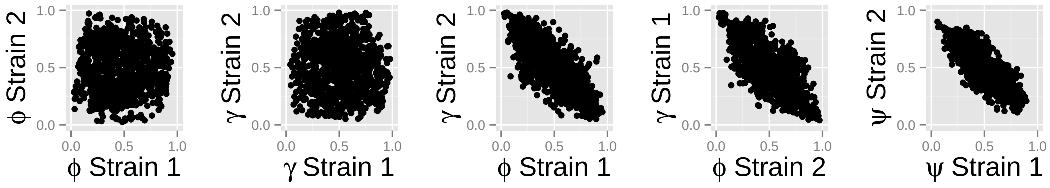
Let's assume a strong negative correlation in among-patient persistence of strain 1 (ϕ_1) and colonization of strain 2 (γ_2). This would represent a situation in which, across all patients, the presence of strain one hinders the colonization of strain 2, a strong priority effect.

```
> corr1 <- sim_func(raP1G2 = -0.9)
```



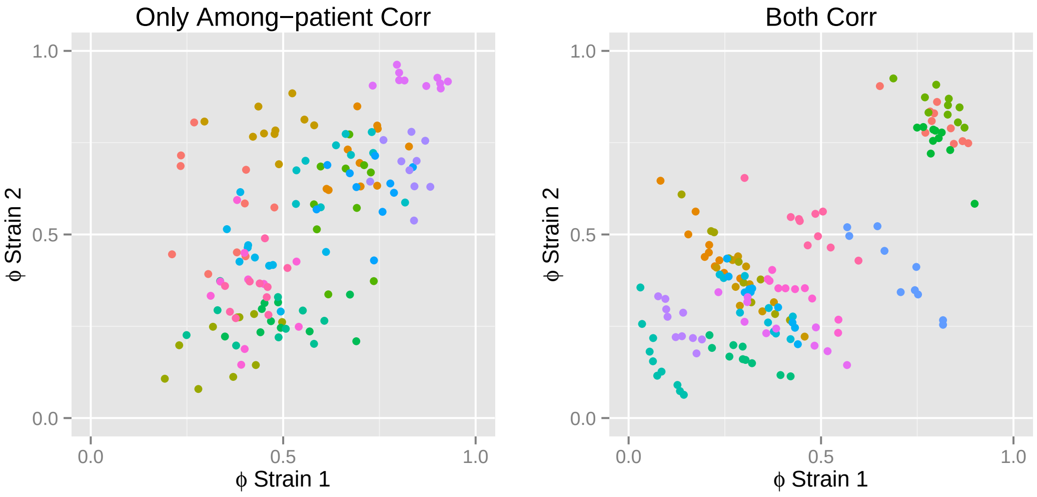
If we have have more than one among-patient correlations that have similar sign, then we see strong induced correlation in ψ.

```
> corr2 <- sim_func(raP2G1 = -0.9, raP1G2 = -0.9)
```



Let's compare if we have within and/or among-patient correlations in ϕ for example. I subsetted only 15 patients for clarity:

```
> Among <- sim_func(raP1P2 = 0.8)
> Both <- sim_func(raP1P2 = 0.8, rwP1P2 = -0.9)
```



In the above figure, you can see that, with among-patient correlations, the ϕ values are grouped by patient (color), but there is random spread within the color grouping. However, when there is both within- and among-patient correlations, ϕ values are grouped by patient, but also the within-patient correlation is visible. You can alter the within and among correlations to be positive or negative and see what the figure looks like. For instance, you can generate Simpson's paradox with correlatinos of different sign.