Inferring phage–bacteria infection networks from time-series data

Luis F. Jover¹, Justin Romberg² and Joshua S. Weitz³

¹School of Physics, ²School of Electrical and Computer Engineering, and ³School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA

In communities with bacterial viruses (phage) and bacteria, the phage–bacteria infection network establishes which virus types infect which host types. The structure of the infection network is a key element in understanding community dynamics. Yet, this infection network is often difficult to ascertain. Introduced over 60 years ago, the plaque assay remains the gold standard for establishing who infects whom in a community. This culture-based approach does not scale to environmental samples with increased levels of phage and bacterial diversity, much of which is currently unculturable. Here, we propose an alternative method of inferring phage–bacteria infection networks. This method uses time-series data of fluctuating population densities to estimate the complete interaction network without having to test each phage–bacteria pair individually. We use in silico experiments to analyse the factors affecting the quality of network reconstruction and find robust regimes where accurate reconstructions are possible. In addition, we present a multi-experiment approach where time series from different experiments are combined to improve estimates of the infection network. This approach also mitigates against the possibility of evolutionary changes to relevant phenotypes during the time course of measurement.

1. Introduction

Bacterial viruses are ubiquitous and play an important ecological role at the global scale. In the oceans, viruses are responsible for a significant fraction of bacterial mortality and as a result have an effect on global biogeochemical cycles [1–4]. By killing bacteria, they redirect resources from higher trophic levels. Yet, not all bacteria types are susceptible to each and every virus type. Phage potentially infect a subset of hosts; these relationships constitute complex networks of infection [5]. Quantifying who infects whom remains essential to understanding how individual-based traits affect ecosystem-wide properties in complex environments [6].
For more than 60 years, the host range of phage, i.e. the types of host that a phage type infects, has been measured using plaque assays [7]. A plaque assay is an experimental method in which a growing culture of bacteria on an agar surface is exposed to phage. Clear ‘plaques’ are formed whenever the phage can infect and lyse the target host. Plaque assays are considered the gold standard for determining infection but are hard to scale-up to community levels. The principal reason is that the majority of phage and bacteria in a community sample are not yet available in culture. In response, a number of (partially) culture-independent methods have been proposed, including digital PCR [8], viral tagging [9,10] and PhageFISH [11]. Each of these methods leverages a ‘targeted’ approach, i.e. requiring some degree of culturing or co-visualization of labelled particles. Targeted approaches present challenges for scaling-up to communities. By contrast, an approach that considers community-scale interactions may be feasible, particularly if leveraging information in the temporal dynamics of complex virus–bacteria systems.

The inference of interaction networks from system dynamics is a field of study with widespread applications from inference of gene regulatory networks [12,13] and chemical reactions [14], to neural networks [15]. The key insights, from one class of inference methods is that statistical patterns in dynamics, including cross-correlation and mutual information, can be leveraged to infer interactions [16]. However, such correlation-based approaches can be of limited value when applied to high dimensional systems with nonlinear interactions. As an alternative, Shandilya et al. [17] showed a method for reconstructing interaction networks from discrete measurements of the time series in systems where the underlying functional form of the interactions is known. Similarly, Stein et al. [18] following the work of Monier et al. [19] used discretized Lotka–Volterra equations to estimate interaction networks, model parameters, and time-dependent perturbations in competitive microbial communities.

Here, we adapt the approach of Stein et al. [18] to phage–bacteria systems with antagonistic interactions. The central advance is the use of nonlinear dynamic models and inference methods to estimate quantitative infection and lysis rates given information embedded in community time series. We test the inference capabilities of the method using *in silico* experiments, as a proof of principle. As we show, inferring realistic phage–bacteria infection networks in complex communities may be possible given appropriate deployment of existing culture-independent technologies already available to estimate changing genotype densities over time.

## 2. Method

### 2.1. Model

We model the interaction between $N_h$ host types and $N_v$ virus types using a generalization of the Lotka–Volterra predator–prey equations [20,21]. The densities of multiple host and virus types are described by a system of differential equations that include the effect of competition between host types and the infection of host by multiple virus types [22,23]:

\[
\frac{dh_i}{dt} = r_i h_i \left(1 - \sum_{i'} a_{i'i} h_{i'} / K\right) - h_i \sum_{j} N_v M_{ij} \phi_{ij} v_j
\]

(2.1)

and

\[
\frac{dv_j}{dt} = v_j \sum_{i} N_h \beta_{ij} \phi_{ij} M_{ij} h_i - m_j v_j.
\]

(2.2)

The model consists of $N_h$ equations of the form (2.1) for the density of each host type, $h_i$, and $N_v$ equations of the form (2.2) for the virus densities, $v_j$. In this system: $r_i$ is the growth rate of host $i$ in the absence of viruses and other hosts, $a_{i'i}$ is the competitive effect of host $i'$ on host $i$, $K$ is the system-wide carrying capacity, $\phi_{ij}$ is the adsorption rate of virus $j$ when attaching to host $i$, $\beta_{ij}$ is the burst size of virus $j$ when infecting host $i$, and $m_j$ is the decay rate of virus $j$. Finally, $M_{ij}$ is the infection matrix, i.e. a matrix representation of the infection network, which takes a value of 1 if host $i$ is infected by virus $j$ and zero otherwise. The nonlinearities arise owing to the cumulative effects of pairwise interactions among bacteria ($h_i h_j$) and between viruses and bacteria ($h_i v_j$).
Table 1. Parameter and target steady-state density ranges used to find feasible parameter sets. (Bacteria growth rates, $r_i$, and virus decay rates, $m_j$, were derived using the steady-state equations and the parameters presented here using the feasibility-based framework (see Methods). The range denotes the limits of the uniform distributions used to generate parameters.)

<table>
<thead>
<tr>
<th>Parameter (unit)</th>
<th>range/value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_j$ (ml/(virus · d))</td>
<td>$10^{-8}$ to $10^{-7}$</td>
</tr>
<tr>
<td>$\beta_j$ (viruses cell$^{-1}$)</td>
<td>10–50</td>
</tr>
<tr>
<td>$H_i^*$ (cell ml$^{-1}$)</td>
<td>$10^5$–$10^6$</td>
</tr>
<tr>
<td>$V_j^*$ (virus ml$^{-1}$)</td>
<td>$10^6$–$10^7$</td>
</tr>
<tr>
<td>$K$ (ml)</td>
<td>$\max(H_i^*) \times 100 = 10^6$</td>
</tr>
</tbody>
</table>

2.2. Numerical simulations of the dynamics: infection network ensembles and model parameters

To study the performance of our reconstruction method, we simulated time series of systems where several hosts and virus types interact. We used MATLAB’s ODE45 to numerically integrate systems of equations of the form described in §2.1. In doing so, we use both random infection networks and nested infection networks. Nested interaction networks are commonly observed in culture-based analyses, such that the host range of phage and the phage range of hosts form ordered subsets [24]. Following Jover et al. [25], we generated an ensemble of 100 infection matrices, each one with 10 host types and 10 virus types, spanning a spectrum of nestedness values. The infection matrices were generated by starting with a modular matrix and shifting interactions, through a random process, to regions that increase nestedness [25]. We also found feasible parameter sets (i.e. parameters with positive steady-state densities) for each one of the infection matrices. We followed the procedure described in [25] to find feasible parameter sets. Namely, we select a subset of the model parameters and target densities (table 1) and use the steady-state equations to solve for the rest of the parameters obtaining a feasible parameter set.

2.3. Infection network reconstruction

Our method for reconstructing infection networks requires discrete measurements of the dynamics resulting from the interaction of different host and virus types. This method extends the approach described in [18] to host-phage systems. We use only the equations describing the dynamics of the viruses (equations of the form (2.2)). We start by rewriting equation (2.2) in the form

$$ \frac{d \ln(v_j)}{dt} = \sum_{i=1}^{N_h} \beta_{ij} \phi_i M_{ij} h_i - m_j. \tag{2.3} $$

We assume that we have $N + 1$ measurements of the densities of all virus and host types in the system at times $[t_1, t_2, \ldots, t_{N+1}]$. For time step $\Delta t_n = t_{n+1} - t_n$, we can write a discretized form of equation (2.3):

$$ \frac{\Delta \ln(v_j(t_n))}{\Delta t_n} \approx \sum_{i=1}^{N_h} \tilde{M}_{ij} h_i(t_n) - m_j, \tag{2.4} $$

where we define the quantitative infection network $\tilde{M}_{ij} := M_{ij} \phi_i \beta_{ij}$ and $\Delta \ln v_j(t_n)/\Delta t_n := (\ln (v_j(t_{n+1})) - \ln (v_j(t_n)))/(t_{n+1} - t_n)$. We can write an analogous equation to equation (2.4) for all time steps and all virus types in the system. All of these equations can be written in a compact form using a single matrix.
vector original and reconstructed infection matrices as a metric of the quality of reconstruction, namely infers decay rates for all virus types in the system. We use a normalized Frobenius distance between the

\[ \begin{bmatrix} \Delta \ln(v_1(t_1)) & \Delta \ln(v_1(t_2)) & \ldots & \Delta \ln(v_1(t_N)) \\ \Delta \ln(v_2(t_1)) & \Delta \ln(v_2(t_2)) & \ldots & \Delta \ln(v_2(t_N)) \\ \vdots & \vdots & \ddots & \vdots \\ \Delta \ln(v_N(t_1)) & \Delta \ln(v_N(t_2)) & \ldots & \Delta \ln(v_N(t_N)) \end{bmatrix} \approx \begin{bmatrix} \tilde{M}_{11} & \tilde{M}_{12} & \ldots & \tilde{M}_{1N} & -m_1 \\ \tilde{M}_{12} & \tilde{M}_{22} & \ldots & \tilde{M}_{2N} & -m_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{M}_{1N} & \tilde{M}_{2N} & \ldots & \tilde{M}_{NN} & -m_N \end{bmatrix} \]

or defining the matrices \( W \) and \( H \) with elements \( W_{ij} = \Delta \ln v_i(t_j) / \Delta t_j \) and \( H_{ij} = h_j(t_i) \), and the column vector \( m \) with elements \( m_{ij} \), we can write

\[ W \approx (\tilde{M}^T - m) \begin{bmatrix} H \\ 1 \end{bmatrix}, \]

where \( 1 \) is a vector of ones with dimensions \( 1 \times N \). Given density measurements of the hosts and viruses, we can reconstruct the quantitative infection network using equation (2.6). We solve the following minimization problem to obtain approximations \( \tilde{M}_{\text{rec}} \) and \( m_{\text{rec}} \) of the quantitative infection matrix, \( M \), and the decay rate vector \( m \):

\[ \text{arg min}_{(\tilde{M}^T, m)} \left\| W - (\tilde{M}^T - m) \begin{bmatrix} H \\ 1 \end{bmatrix} \right\|_2 \]

subject to \( M_{ij} \geq 0, \)

\[ m_i > 0. \]

To solve this problem, we used CVX, a package for specifying and solving convex problems [26,27]. In this study, we focus on the reconstruction of the quantitative infection network, but the method also infers decay rates for all virus types in the system. We use a normalized Frobenius distance between the original and reconstructed infection matrices as a metric of the quality of reconstruction, namely

\[ \text{Error}_{\text{rec}} = \frac{\| \tilde{M} - \tilde{M}_{\text{rec}} \|_2}{\| M \|_2}. \]

The inference was implemented in MATLAB, and scripts are available at https://github.com/Weitz Group/infection_network_reconstruction.

3. Results

3.1. Reconstruction quality depends on the variability of the dynamics

We begin with an example in which there are 10 host types, 10 virus types and 20 virus–bacteria interactions. The effective infection rates \( (\phi + \beta) \) vary from \( 10^{-7} \) to \( 5 \times 10^{-6} \). Figure 1 shows an example of a successful infection network reconstruction, using the method described in §2.3. The matrices \( W \) and \( H \) were calculated, using measurements of the dynamics every 6 min for a total of 96 h. This results in a reconstruction error \( \text{Error}_{\text{rec}} = 0.01 \). The method is able to correctly identify all of the interactions. The small error arises from differences in the inferred quantitative values.

In general, there are multiple factors affecting reconstruction quality. One important factor is the variability of the dynamics. For example, if the dynamics start at a fixed point, then there would be no variability in the dynamics, the columns of the matrix \( H \) would all be identical, and it would not be possible to infer the infection network. We test the effect of variability systematically by performing matrix reconstruction for an ensemble of matrices and different levels of variability. To control variability in the dynamics, we change how far the initial densities are from the equilibrium densities. We initialize...
Figure 1. Example of infection network reconstruction. (a) Virus and host dynamics for 96 h. (b) Matrices $W$ and $H$ constructed by taking measurements of virus and host densities every 6 min as described in §2.3. (c) Original and reconstructed infection matrices ($\text{Error}_{\text{rec}} = 0.01$). A feasible parameter set was used in the simulation as described in §2.2.

Figure 2. Effect of deviation from equilibrium on reconstruction. Mean reconstruction error as a function of the fraction away from the equilibrium densities, $\delta$, for an ensemble of 100 matrices. Feasible parameter sets were used in the simulation as described in §2.2.

density of each host and virus type in the system at $x_0 = x_{\text{eq}}(1 \pm \delta)$, where $x_{\text{eq}}$ is the equilibrium density of a given type and $\delta$ is a free parameter that controls the distance from its equilibrium density. We calculated the mean reconstruction error for an ensemble of 100 matrices (figure 2). The reconstruction
Figure 3. Schematic of data aggregation in the multiple-experiment approach. Multiple experiments are performed with the same matrix $M$ and different initial conditions. The host dynamics are concatenated to assemble a single $H$ matrix.

Error has a maximum at $\delta = 0$ (not shown for visualization purposes), which corresponds to starting the system at the equilibrium densities. The quality of the reconstruction increases as the initial conditions move away from the equilibrium densities.

3.2. Reconstruction from multiple experiments: an alternative approach

We propose an improvement to the single experiment approach for reconstruction. In this alternative approach, we combine measurements from different experiments to increase reconstruction quality. One key advantage of this approach is that by increasing the number of experiments used for reconstruction, we can reduce the total time and number of measurements per experiment. This is a crucial advantage in virus–bacteria systems, which are known to evolve rapidly [28–30]. In the multiple-experiment approach, we generate a host matrix $H$ and a virus matrix $W$ by combining matrices from multiple experiments that differ only in their initial conditions (figure 3). This extends equation (2.6) to include information from multiple experiments. Specifically, assuming that we perform $p$ different experiments and calculate matrices $\{H_1, H_2, ..., H_p\}$ and $\{W_1, W_2, ..., W_p\}$ for each experiment, we can write the system

$$ (W_1 W_2 ... W_p) \approx (\hat{M}^T - m) \begin{pmatrix} H_1 & H_2 & \cdots & H_p \end{pmatrix}, $$

(3.1)

where $1$ is a vector of ones with dimensions $1 \times (N_1 + N_2 + \cdots + N_p)$, assuming that we take $N_i$ measurements from experiment $i$. Using the same minimization process presented in §2.3, we can obtain an approximation, $\hat{M}_{rec}$, of $M$.

Figure 4 compares the single and multiple experiments approach for three matrices with different nestedness values. We see how the multiple experiment approach results in lower reconstruction error for the three different cases. Figure 5 extends the comparison with an ensemble of 100 different matrices. We compare the multiple experiment approach with the average result of the single experiment approach. For a given matrix, we performed 20 different experiments. Each experiment has the same infection matrix and the same model parameters but different initial conditions. We compare the performance of the reconstruction, using each experiment individually versus combining the measurements of the 20 experiments as described in equation (3.1). In this comparison, we fix the total number of measurements; we compare the reconstruction error when using 960 measurements from a single experiment (measuring the dynamics every 6 min for 96 h) against the performance when combining the first 48 measurement of all 20 experiments (every 6 min for 4.8 h).
Figure 4. Examples of reconstruction for three different matrices and two different methods. Each row shows the original matrix and the resulting reconstruction for each method. The first column shows the original matrices with values of nestedness (NODF): 0.34, 0.55 and 1, respectively. The middle column shows the reconstructed matrices and corresponding reconstruction errors for the single experiment approach, using 960 measurements. The last column from the right shows the reconstructed matrices and corresponding errors for the multiple experiment approach, using 20 experiments and 48 measurements per experiment. The total number of measurements is the same in the three different methods. The time between measurements is $\Delta t = 6 \text{ min}$.

Figure 5. Reconstruction error versus nestedness for two different methods. The black line denotes the reconstruction error, $\text{Error}_{\text{rec}}$, using the multiple-experiments approach. The blue line describes the mean reconstruction error for the same 20 experiments used in the multiple-experiment approach but using each experiment separately. The total number of measurements is the same in both approaches.

We performed the comparison for 100 different matrices (figure 5). Multiple-experiment reconstruction results in lower error than the average single experiment reconstructions across a wide range of nestedness values. The multiple experiment approach is also more robust; it results in smaller variance in the reconstruction error. Performing more than a few experiments not only decreases the
mean reconstruction error, but also decreases the standard deviation significantly (figure 6). For the specific configuration studied here, reconstruction error minimizes around 18 experiments.

3.3. Identification of optimal sampling intervals

The inference of cross-infection in a complex community also depends on the sampling interval. Here we test the effects of variation in both the sampling interval and the total number of measurements on the quality of reconstruction. First, we vary the sampling interval, $\Delta t$, and total hours, $T$, for an ensemble of 100 matrices. Feasible parameter sets were used in the simulation as described in Methods. Figure 7a shows the variation in reconstruction error as a function of variation in $\Delta t$ and $T$. As is apparent, reconstruction error decreases when increasing the length of sampling given fixed sampling intervals. However, when the total experimental effort is fixed, we find an ‘optimal’ intermediate sampling interval (figure 7b). The inference procedure is not effective given very short sampling intervals. This problem will become particularly acute given noise. Similarly, if the interval is too long, then there is additional error introduced given the linear approximation of a nonlinear model. Refinement of this time should be considered in the design and implementation of experimental protocols, e.g. preliminary estimation of host growth rates and viral latent periods.

Figure 6. Effect of experiment number on reconstruction error. Mean (blue line) and standard deviation (dotted line) of the reconstruction error for 100 infection matrices as a function of the number of experiments used in the multiple-experiment approach. Fixed number of total measurements (960). $\Delta t = 6$ min.

Figure 7. Effect of variation in sampling interval and measurement number on reconstruction. (a) Error of reconstruction given normalized Frobenius distance (colour) given variation in $\Delta t$ and total hours, $T$. The three black lines denote combinations given the same total number of measurements, 100, 200 and 400. (b) Error of reconstruction given normalized Frobenius distance for scenarios in which there are 100, 200 and 400 measurements with alternative sampling intervals, $\Delta t$. 

mean reconstruction error, but also decreases the standard deviation significantly (figure 6). For the specific configuration studied here, reconstruction error minimizes around 18 experiments.

3.3. Identification of optimal sampling intervals

The inference of cross-infection in a complex community also depends on the sampling interval. Here we test the effects of variation in both the sampling interval and the total number of measurements on the quality of reconstruction. First, we vary the sampling interval, $\Delta t$, and total hours, $T$, for an ensemble of 100 matrices. Feasible parameter sets were used in the simulation as described in Methods. Figure 7a shows the variation in reconstruction error as a function of variation in $\Delta t$ and $T$. As is apparent, reconstruction error decreases when increasing the length of sampling given fixed sampling intervals. However, when the total experimental effort is fixed, we find an ‘optimal’ intermediate sampling interval (figure 7b). The inference procedure is not effective given very short sampling intervals. This problem will become particularly acute given noise. Similarly, if the interval is too long, then there is additional error introduced given the linear approximation of a nonlinear model. Refinement of this time should be considered in the design and implementation of experimental protocols, e.g. preliminary estimation of host growth rates and viral latent periods.
3.4. Robustness of inference given noise in measurement

Here we evaluate the effect of measurement of white Gaussian noise on the quality of the inference. We follow the same procedure as in the noiseless case to reconstruct infection networks using multiple experiments. Figure 8 shows mean reconstruction error for an ensemble of 100 matrices as a function of the signal-to-noise ratio (SNR). We see that using 20 experiments and 48 measurements per experiment, network inference is possible for large SNR, but reconstruction error increases significantly when the noise approaches 10% of the signal (SNR = 10 dB).

4. Discussion

We presented a theory-driven method to estimate host–phage infection networks in a community with multiple virus and host types. Current experimental techniques that leverage targeted approaches to measure interactions are difficult to scale to large systems. Our approach addresses this limitation by using time-series measurements involving the whole virus–bacteria community. We presented a series of alternative experimental designs, robustly inferring interactions given a single or multiple time series. The multiple-experiment approach has the additional advantage of requiring shorter measurement times per experiment. As a consequence, there is a lower probability of a host evolving resistance to a virus type or a virus evolving the ability to infect a new host, increasing the chances of reconstructing the infection network of the focal community.

The current method takes as input the measured densities of bacteria and phage in an environmental sample. Next-generation high-throughput sequencing techniques provide a means to characterize bacterial and viral communities in a variety of environmental samples [31–35]. In the past, such characterization has focused on phylogenetics groups, by using RNA and other marker genes. Such markers are insufficiently resolved with respect to differences in relevant phenotypes, e.g. phage–bacteria infectivity. However, new computational approaches are increasingly able to resolve strain-level dynamics from metagenomic datasets [36,37]. The use of quantitative pipelines from sample to strain densities for both bacteria and viruses represents the most promising candidate to enable the inference proposed here [38–40].

Our present approach uses the nonlinear dynamics of virus populations to infer virus–bacteria infection networks. This method can be enhanced by including nonlinear bacterial population dynamics to infer competitive interactions between bacteria types and bacterial growth rates, i.e. by learning from equations (2.1) and (2.2). In addition, it is important to keep in mind that the present approach is adapted to a specific functional form of the virus–bacteria interactions [23]. In the future, it will be important to address the theoretical limits to inference given errors arising from model misspecification, i.e. structural stability and the ecological effects of lysogeny and other persistent infections. In addition, experimental verification [18] is necessary to test whether or not the dynamical model is a sufficiently robust representation of naturally occurring systems, particularly those with high diversity.
In summary, this study presents key steps towards determining quantitative infection and lysis rates in a complex virus–bacteria community. The method has the potential to significantly reduce the experimental burden, by inferring $N_b \times N_v$ quantitative interactions by measuring the dynamics of $N_b + N_v$ populations. Crucially, such inference does not require a culture-based or a targeted approach. Moving forward, incorporating advances in environmental sequencing into a time-series framework may help to realize a long-term goal of inferring community-wide interactions.

Ethics. The authors declare that no licences or approvals were required to complete this work.

Data accessibility. All scripts are available at https://github.com/WeitzGroup/infection_network_reconstruction.

Authors’ contributions. L.F.J. and J.S.W. designed the study. L.F.J., J.R., and J.S.W. designed analytical tools. L.F.J. wrote the code and performed computational analyses, L.F.J., J.R., and J.S.W. interpreted the results, L.F.J. and J.S.W. wrote the manuscript with feedback from J.R. All authors gave final approval for publication.

Competition interests. The authors declare no competing interests.

Funding. Financial support came from NSF grant no. OCE-1233760 (to J.S.W.).

Acknowledgements. The authors thank Sam Brown and Joey Leung for their comments and suggestions and multiple reviewers for their feedback.

References


33. Thurber RV, Hayes M, Brentbart M, Weylger L, Rothner F. 2009 Laboratory procedures to generate


