Correlations, interactions, and predictability in virus-microbe networks

Ashley R. Coenen[1] and Joshua S. Weitz[2,1,*]

[1]School of Physics, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA

[*]Corresponding author: jsweitz@gatech.edu

Abstract

Microbes are found in high abundances in the environment and in human-associated microbiomes, often exceeding one million per milliliter. Viruses of microbes are estimated to turn over 10 to 40 percent of microbes daily and, consequently, are important in shaping microbial communities. Given the relative specificity of viral infection and lysis, it is essential to identify the functional linkages between viruses and their microbial hosts. Multiple time-series analysis methods, including correlation-based approaches, have been proposed to infer infection networks *in situ*. In this work, we evaluate the effectiveness of correlation-based inference using an *in silico* approach. In doing so, we compare actual networks to predicted networks as a means to assess the self-consistency of correlation-based inference. Contrary to common use, we find that correlation is a poor indicator of interactions that arise from antagonistic virus-host infections that culminate in lysis. In closing, we discuss alternative inference methods, particularly model-based methods, as a means to predict interactions in complex virus-microbe communities.

Keywords: correlation, inference, interaction network, bacteriophage, viral ecology

1 Introduction

Microbes and viruses are ubiquitous and highly diverse in marine, soil, and human-associated environments. Microbes play important roles in biogeochemical cycling, and viruses of microbes can transfer genes between microbial hosts [1, 2], alter host physiology (*e.g.* via auxillary metabolic genes [3, 4]), and redirect the flow of organic matter in food webs through cell lysis [5, 6]. Viruses therefore are a significant part of microbial communities, and characterizing virus-microbe interactions is necessary for understanding how cellular-level interactions influence community structure and ecosystem function.

Viruses are known to be relatively specific but not exclusive in their microbial host range. Individual viruses may infect multiple strains of an isolated bacteria, or they may infect across genera, *e.g.* cyanophage can infect both *Prochlorococcus* and *Synechococcus* [7]. Analyses of dozens of culture-based studies have revealed structure in virus-microbe interaction networks [8, 9, 10]. Interaction networks are nested at the strain and genus level, which is indicative of rapid coevolution [11], and are hypothesized to be modular at larger phylogenetic scales. Importantly though, these results come from culture-based methods raising the question of how broadly applicable they are *in situ* [1]. Partially culture-independent methods, such as viral tagging [12, 13] and digital PCR [14], overcome some of the hurdles associated with culturability and isolation but do not yet represent a community-based approach to inferring interactions amongst non-targeted viruses and microbes. Community-based inference methods are needed to fully characterize virus-microbe interactions.

Viral metagenomics (viromics) has made it possible to characterize phylogenetic and functional diversity of viruses *in situ*, bypassing culturing altogether [15, 16, 17]. Recent methods for inferring interactions leverage information from time-series obtained via metagenomic sampling. In these methods, population abundances (or marker-based proxies) are estimated directly from viral and cellular metagenomes. Many such inference methods for time-series exist (see reviews [18, 19, 20, 21, 22]) and broadly fall into two categories: model-

based and model-free. Examples of model-free methods include the direct use of correlations, time-lagged correlations (*e.g.* local similarity analysis or LSA [23, 24, 25]), and correlations between log-transformed abundance ratios (*e.g.* sparse correlations for compositional data or SparCC [26]). Other model-free methods which are not correlation-based include cross-convergent mapping (CCM) [27, 28, 29], pairwise asymmetric inference (PAI) [30], and the sparse S-map method (SSM) [31], and many model-based methods exist as well (see Discussion). Out of these, correlation and correlation-based inference methods are widely used with experimental time-series data. Local similarity analysis (LSA) in particular has been widely applied to infer interaction networks in communities of marine bacteria [32, 33]; bacteria and phytoplankton [34, 35]; bacteria and viruses [36]; and bacteria, viruses, and protists [37, 38].

Correlations in time-series are difficult to interpret, despite their widespread use. Positive correlations could indicate "common preferred conditions or perhaps cooperative activities such as crossfeeding" [22], while negative correlations could indicate "opposite seasonality, competition for limited resources or perhaps active negative interactions such as targeted allelopathy or predator-prey relationships" [22]. Correlations are often interpreted as direct interactions, *e.g.* predation, but may instead be indicative of a wide variety of indirect interactions (see [39, 40]). Predicted interactions can vary drastically depending on the particular correlation metric used [19], which further compounds problem of ecological interpretation. Compositional data, which is common in metagenomic time-series, introduces additional complications [26]. Time-series may be strongly and significantly correlated without having any underlying physical or ecological relationship at all, a well-known adage ("correlation does not imply causation") that is often disregarded. Multiple studes have shown that correlations in time-series do not predict interactions in *in silico* microbial communities using a discrete-time Lotka-Volterra model [41] and a parametric statistical model [42].

Despite these challenges, correlation-based inference is rarely verified in advance of its application. Inferred networks are difficult to verify in general, in part because there is no existing "gold standard" interaction network, and as previously mentioned culture-based

methods are not widely applicable. Hence, in this paper, we take an *in silico* approach to assessing the efficacy of correlation-based inference. We use a mechanistic model for a virus-microbe community to generate synthetic time-series in which the interaction network is known *a priori*. Then we apply correlation-based inference to the resulting time-series and compare predicted interactions to the original interaction network. As we show, correlation-based inference fails to recapitulate virus-microbe interaction networks raising substantive concerns over its use in natural systems.

2 Methods

2.1 Modeling the virus-microbe community

We model the dynamics of a virus-microbe community with a system of nonlinear differential equations:

$$\dot{H}_i = \overbrace{r_i H_i \left(1 - \frac{\sum_{i'}^{N_H} a_{ii'} H_{i'}}{K}\right)}^{\text{logistic growth}} - \overbrace{H_i \sum_{j}^{N_V} M_{ij}\phi_{ij} V_j}^{\text{lysis}} \tag{1}$$

$$\dot{V}_j = \underbrace{V_j \sum_{i}^{N_H} M_{ij}\phi_{ij}\beta_{ij} H_i}_{\text{lysis}} - \underbrace{m_j V_j}_{\text{decay}} \tag{2}$$

where H_i and V_j refer to the population density of microbial host i and virus j respectively. There are N_H different host types and N_V different virus types. For our purposes, a "type" is a group of microbes or viruses with identical life history traits, *i.e.* microbes or viruses that occupy the same functional niche.

In the absence of viruses, the hosts undergo logistic growth with growth rates r_i. The hosts have a community-wide carrying capacity K, and they compete with each other for resources both inter- and intra-specifically with competition strength $a_{ii'}$. Each host can be infected and lysed by a subset of viruses determined by the interaction terms M_{ij}. If host i can be infected by virus j, M_{ij} is one; otherwise it is zero. The collection of all the interaction terms is the interaction network represented by matrix $\mathbf{M}$ of size N_H by N_V. The

adsorption rates ϕ_{ij} denote how frequently host i is infected by virus j.

Each virus j's population grows from infecting and lysing hosts. The rate of virus j's growth is determined by its host-specific adsorption rate ϕ_{ij} and host-specific burst size β_{ij}, which is the number of new virions per infected host cell. The quantity $\tilde{M}_{ij} = M_{ij}\phi_{ij}\beta_{ij}$ is the interaction strength between virus j and host i, and the collection of all the interaction strengths is the weighted interaction network $\tilde{\mathbf{M}}$. Finally, the viruses decay at rates m_j.

2.2 Interaction network topology

Virus-microbe interaction networks are represented as bipartite networks or matrices of size N_H by N_V. Here, we generate *in silico* interaction networks given variation in nestedness and modularity. For a given network size N_H by N_V, we first generate the perfectly nested (Fig 1A) and perfectly modular (Fig 1B) networks using the BiMat MATLAB package [43]. For the modular network, we choose a small number of modules relative to the network dimensions N_H and N_V, *e.g.* 2 modules for a 10 by 10 network. In general, the perfectly nested network and the perfectly modular network will have a different number of interactions. In the following rewiring procedure, we treat the two networks separately so that the number of interactions is conserved.

We rewire the perfect network by randomly selecting an interacting host-virus pair ($M_{ij} = 1$) and a non-interacting host-virus pair ($M_{ij} = 0$) and exchanging their interaction values. We do not allow exchanges that would result in an all-zero row or column but do not restrict the exchanges in any other way. We continue the random selection without replacement until all host-virus pairs have been selected no more than once, recording the new interaction network after each exchange. We repeat this procedure several times to generate an ensemble of interaction networks with varying nestedness and modularity.

We measure the nestedness and the modularity of each network in the ensemble using the default algorithms in the BiMat MATLAB package. The nestedness metric used is NODF, and the modularity is normalized [44]. We rearrange the networks in their most nested or

most modular form as determined by the BiMat MATLAB package (see Fig 1 for examples) [43].

2.3 Generating life history traits

The life history traits for a given interaction network are chosen to ensure that all host and virus types can coexist in the long term [45], as summarized here.

First, we randomly sample target steady-state densities H_i^* and V_j^* for each host and virus. We also sample some of the life history traits, in particular the host carrying capacity K, adsorption rates ϕ_{ij}, and burst sizes β_{ij}. All of these values are chosen by independent random sampling from ranges as specified in Table 1.

Next, we sample the host competition terms $a_{ii'}$. To begin, we set all intraspecific competition to one ($a_{ii} = 1$) and all interspecific competition terms to zero ($a_{ii'} = 0$ for $i' \neq i$). To ensure coexistence among the hosts in the absence of viruses, we randomly sample target virus-free steady-state densities H_{0i}^* from the range specified in Table 1. Coexistence is satisfied when

$$K = \sum_{i'}^{N_H} a_{ii'} H_{0i'}^* \tag{3}$$

for each host i. Thus, for each host i, some interspecific competition terms must be made non-zero. We randomly choose an index $k \neq i$ and randomly sample a_{ik} between zero and one. If the new $\sum_{i'}^{N_H} a_{ii'} H_{0i'}^*$ does not exceed the carrying capacity K, we repeat for a new index k. Once the carrying capacity is exceeded, we adjust the most recent a_{ik} so that Eqn 3 is satisfied exactly.

The remaining life history traits, the viral decay rates m_j and the host growth rates r_i, are solved for using the steady-state versions of the virus-host differential equations (Eqns 1 and 2):

$$m_j = \sum_{i}^{N_H} M_{ij} \phi_{ij} \beta_{ij} H_i^* \tag{4}$$

$$r_i = \left(\sum_j^{N_V} M_{ij} \phi_{ij} V_j^* \right) / \left(1 - \frac{\sum_{i'}^{N_H} a_{ii'} H_{i'}^*}{K} \right) \tag{5}$$

2.4 Simulating community dynamics

We use MATLAB's ODE45 to numerically simulate the virus-host dynamical system (Eqns 1 and 2) with *in silico* interaction network and life history traits generated as described in §2.2 and §2.3. We use a relative error tolerance of 10^{-8} and specify regularly spaced time-points on the order of $\left\langle \frac{1}{r} \right\rangle$. Initial conditions are chosen by perturbing the target steady-state densities H_i^* and V_j^* by a multiplicative factor $\delta = \pm 0.3$ where the sign of δ is chosen randomly for each host and each virus. After generating the time-series, we sample from the transient dynamics using a fixed sample frequency.

2.5 Calculating correlation networks

Let $\mathbf{t} = \{t_1, \ldots, t_N\}$ be the collection of sample times. Let $H_i(t_k)$ and $V_j(t_k)$ be the sampled time-series of host i and virus j at the single time-point $t_k \in \mathbf{t}$. We denote the log-transformations of the sample points as $h_i(t_k) = \log_{10} H_i(t_k)$ and $v_j(t_k) = \log_{10} V_j(t_k)$. The Pearson correlation coefficient between host i and virus j is defined as

$$r_{ij} = \frac{\sum_{k=1}^{N} \left(h_i(t_k) - \bar{h}_i \right) \left(v_j(t_k) - \bar{v}_j \right)}{\sqrt{\sum_{k=1}^{N} \left(h_i(t_k) - \bar{h}_i \right)^2} \sqrt{\sum_{k=1}^{N} \left(v_j(t_k) - \bar{v}_j \right)^2}} \tag{6}$$

where N is the number of sampled time points and $\bar{h}_i = \frac{1}{N} \sum_{k=1}^{N} h_i(t_k)$ and $\bar{v}_j = \frac{1}{N} \sum_{k=1}^{N} v_j(t_k)$ are the sample means. The correlation network $\mathbf{R}$ is the collection of Pearson correlation coefficients for all possible host-virus pairs, that is, a matrix of size N_H by N_V.

We also consider correlations given a time-delay τ. The time-delayed Pearson correlation coefficient between host i and virus j is defined as

$$r_{ij}^{\tau} = \frac{\sum_{k=1}^{N} \left(h_i(t_k + \tau) - \bar{h}_i^{\tau} \right) \left(v_j(t_k) - \bar{v}_j \right)}{\sqrt{\sum_{k=1}^{N} \left(h_i(t_k + \tau) - \bar{h}_i^{\tau} \right)^2} \sqrt{\sum_{k=1}^{N} \left(v_j(t_k) - \bar{v}_j \right)^2}} \tag{7}$$

where N is the number of sampled time-points and $\bar{h}_i^\tau = \frac{1}{N}\sum_{k=1}^{N} h_i(t_k + \tau)$ is the mean of the time-delayed host sample. The time-delay is applied to the host time-series so that it is sampled later in time, whereas the virus time-series sample is unchanged. The same number of sampled time-points N is used for both hosts and viruses.

We consider two different approaches for implementing a time-delay τ. If identical time-delays are used for each host-virus pair, τ is a single community-wide time-delay. On the other hand, if time-delays are unique for each host-virus pair, $\tau = [\tau_{ij}]$ is a matrix of size N_H by N_V of unique pairwise time-delays. For both cases, the correlation network $\mathbf{R}^\tau$ is the collection of time-delayed Pearson correlation coefficients.

2.6 Evaluating correlation network efficacy

To evaluate how well the correlation network $\mathbf{R}$ predicts interactions in the weighted interaction network $\tilde{\mathbf{M}}$, we binarize the two networks so that they may be compared directly. For the weighted interaction network $\tilde{\mathbf{M}}$, non-zero values are categorized as interactions while zeros are categorized as non-interactions. The binarized weighted interaction network is $\tilde{\mathbf{M}}_\mathbf{0}$. For the correlation network $\mathbf{R}$, we categorize values according to a threshold c. Correlations greater than or equal to the threshold are categorized as interactions, while those that are less are non-interactions. The binarized correlation network for a threshold c is $\mathbf{R}_\mathbf{c}$.

To compare the two binarized networks $\mathbf{R}_\mathbf{c}$ and $\tilde{\mathbf{M}}_\mathbf{0}$, we count the number of interactions in $\tilde{\mathbf{M}}_\mathbf{0}$ which $\mathbf{R}_\mathbf{c}$ predicts correctly, the true positives, as well as the number of non-interactions which $\mathbf{R}_\mathbf{c}$ predicts incorrectly, the false positives. We normalize the true and false positive counts by the number of actual interactions and non-interactions in $\tilde{\mathbf{M}}_\mathbf{0}$. These are the true positive rates (TPR) and false positive rates (FPR).

We repeat the binarization for many thresholds between -1 and $+1$, the minimum and maximum possible values of Pearson's correlation coefficient. The entire process results in a tradeoff, or receiving operator characteristic (ROC), curve for the correlation network $\mathbf{R}$. We quantify the overall performance of the correlation network $\mathbf{R}$ as the maximum

difference between TPR and FPR, which is known as Youden's J-statistic or simply J [46]. The maximum difference J occurs at a particular threshold, the optimal threshold c^*.

We note that $J = 1$ is a "perfect" recovery, that is, there exists a threshold c for which the thresholded correlation network $\mathbf{R_c}$ perfectly matches the binarized weighted interaction network $\tilde{\mathbf{M}}_\mathbf{0}$. On the other hand, $J = 0$ means that the thresholded correlation networks $\mathbf{R_c}$ have a true positive rate (TPR) less than or equal to their false positive rate (FPR) across all thresholds.

3 Results

3.1 Simple correlation networks

We computed simple Pearson correlation networks (Eqn 6) for two different *in silico* virus-host communities. The two interaction networks each have 10 hosts and 10 viruses. One interaction network is highly nested (Fig 2-A3) and the other is highly modular (Fig 2-B3). The life history traits for the two networks were generated as described in §2.3, with parameter ranges as specified in Table 1. The time-series were generated according to §2.4 with a timestep of $\left\langle \frac{1}{r} \right\rangle \approx 15$ minutes and a relative error tolerance of 10^{-8} (Fig 2-A1 and -B1). The time-series were sampled at a fixed frequency of 8 hours for 100 time-points per host and per virus type, resulting in a sample period of 800 hours $\approx$ 1 month (Fig 2-A2 and -B2).

We evaluated the efficacy of the two correlation networks with the procedure described in §2.6 (Fig 3). For each correlation network, we report the maximum difference between TPR and FPR, that is, Youden's J-statistic or J, as well as the optimal threshold c^*. To compute the p-value for J, we randomly shuffle the identities of the hosts and viruses in the original time-series, and calculate correlation networks for these shuffled time-series. We used $N = 1000$ random permutations and calculated J for each correlation network.

For the two *in silico* communities, simple correlation networks do not predict the interaction network. For the nested community we found $J = 0.09$ ($p = 0.6$; Fig 3-A3), and for

the modular community we found $J = 0.16$ ($p = 0.1$; Fig 3-B3). The low J values mean that the correlation networks have high FPR compared to TPR across all thresholds, whereas a correlation network which successfully recovered the interaction network would have a high TPR and low FPR for at least one threshold, resulting in $J \approx 1$. The large p-values mean that the reported J values are likely to occur by chance (see Supp Fig 1-A1 and -B1 for distributions).

3.2 Community-wide time-delayed correlation networks

We computed time-delayed Pearson correlation networks (Eqn 7) with a community-wide time-delay for the same two *in silico* communities, using the same time-series (Fig 2-A1 and -B1), sample frequency of 8 hours, and sampled time-points of 100. We considered community-wide time-delays that were multiples of the sample frequency up to the sample period, so that there was always some overlap between the sample times. The time-delay was applied to the host time-series, that is, all host time-series were identically sampled later in time, while the virus time-series were sampled as before. We computed correlation networks for all considered community-wide time-delays, that is, we computed 100 correlation networks where $\tau = 8$ hours, 16 hours, ..., 800 hours.

The correlation networks vary with the community-wide time-delay, as can be seen from a few representative networks (Fig 4-A2 and -B2). For the nested community, none of the 100 time-delayed correlation networks successfully predict the interaction network (Fig 4-A3). The best score is $J = 0.22$ ($p = 0.08$) which occurs at a time-delay of $\tau = 768$ hours. The low J value means that the correlation network has a high FPR compared to TPR across all thresholds, and the high p-value means that the reported J value is likely to occur by chance.

For the modular community, the best score is $J = 0.46$ ($p = 0.001$) which occurs at $\tau = 464$ hours. While the measured J value is significant ($p < 0.05$), it is still low, with FPR $= 0.2$ and TPR $= 0.66$ (see Supp Fig 2). Thus, it is not evident that this is a

"successful" correlation network for predicting interactions. Futhermore, without knowing the interaction network *a priori* – as is the case with *in situ* communities – we have no way of identifying the particular community-wide time-delay $\tau = 464$ hours as the best choice.

3.3 Pairwise time-delayed correlation networks

We computed time-delayed Pearson correlation networks (Eqn 7) for the same two *in silico* communities, now with unique time-delays for each host-virus pair. We used the same time-series (Fig 2-A1 and -B1), sample frequency of 8 hours, and sampled time-points of 100. We considered time-delays that were multiples of the sample frequency up to the sample period as before. The time-delays were applied to the host time-series, with each host potentially having a different time-delay, while the virus time-series were sampled as before. For host-virus pair (i, j), we computed correlations for all considered time delays and recorded the maximum correlation coefficient $r_{ij}^{\max}$ and its associated time-delay $\tau_{ij}^{\max}$.

The resulting correlation networks for the two *in silico* communities consist of positive correlations which are almost uniformly high (Fig 5-A3 and -B3). Neither interaction network is successfully recovered. For the nested community, $J = 0.17$ ($p = 0.2$), and for the modular community, $J = 0.04$ ($p = 0.9$).

3.4 Effects of sampling frequency

We repeated the three procedures for calculating correlation networks with and without time-delays (§3.1, 3.2, and 3.3) for varying sample frequencies. We used the same two *in silico* communities and time-series as before (Fig 2-A1 and -B1). We examined sample frequencies between 15 minutes and 48 hours. For each sample frequency considered, we sampled for 100 time-points. Since the number of sampled time-points was fixed, the sample periods varied between 25 hours and 4800 hours ($\approx$ 6 months).

For the simple correlation networks without time-delays, we calculated the networks in the same way as described in §3.1. For the time-delayed correlation networks, we used

time-delays as described in §3.2, that is, we considered time-delays that were multiples of the sample frequency up to the sample period. For the community-wide implementation, we report the maximum J across all considered time-delays for each sample frequency. For the pairwise implementation, we determined time-delays for each host-virus pair such that correlation was maximized as described in §3.3 for each sample frequency.

For both the nested and modular communities, moderate sample frequencies (1−12 hours) performed slightly better than very high sample frequencies (15 minutes) or very low sample frequencies ($24 - 48$ hours) across the three implementations (Fig 6). For the nested community, the highest score was $J = 0.23$ ($p = 0.1$) for a sample frequency of 4 hours, using a community-wide time-delay. The simple and pairwise time-delay implementations had only slightly lower scores. For the modular community, the highest score was $J = 0.6$ ($p = 0.001$) for a sample frequency of 12 hours, using a community-wide time-delay, which outperformed both the simple and pairwise time-delay implementations.

4 Discussion

Using *in silico* virus-host communities, we calculated correlation networks amongst viral and microbial population time-series using both simple correlations and time-delayed correlations. In the case of time-delayed correlations, we considered a two implementations: a single community-wide time-delay and unique pairwise time-delays. The correlation networks for all three implementations failed to effectively recover the original interaction networks, as quantified by the efficacy score J, the maximum difference between true positive and false positive rates. There was a single test scenario involving a modular network where inference was found to be statistically significant (Supp Fig 2). Although significant, the efficacy score was still low. Furthermore, without knowing the interaction network *a priori* – such as when considering *in situ* communities – it is not clear which particular community-wide time-delay and threshold should be used, nor that they would be robust properties of the community (*e.g.* to different initial conditions or measurement noise). Because we observed low ef-

ficacy scores which were non-significant, we conclude that the correlation networks do not meaningfully predict interactions given this mechanistic model of virus-microbe interactions.

In this work, we examined virus-microbe interaction networks which were highly nested as well as networks which were highly modular. These structures are characteristic of complex networks where each host and virus population interacts with many others, as observed in virus-microbe communities using culture-based assays [8, 9, 10]. Complex interaction networks can drive weak correlations amongst interacting taxa and strong correlations amongst non-interacting taxa, due in part to mutual interactions which act as confounding variables (*e.g.* taxa A and B both interact with taxa C but not with each other) [41, 42]. It is also worth noting that counfounding effects from nonlinearities and feedbacks in the underlying dynamical system might cause correlation networks to fail even for simple interaction networks.

Our results came from generating time-series via a particular dynamical model and applying a correlation-based inference method. There is some evidence that correlation performs poorly regardless of the underlying model (see [41, 42]). This has important implications for using correlation-based methods with experimental time-series: if correlation performs poorly for a wide range of *in silico* models, similar or even worse performance should be expected for *in situ* communities. At the very least, the particular correlation-based method of interest should be benchmarked *in silico* before its use with experimental data, including examining the effects of measurement noise and internal system stochasticity.

Despite the difficulties and ambiguities associated with correlation-based inference methods, they are still widely used within *in situ* studies of virus-microbe interactions and microbe-microbe interactions more generally. In light of the poor performance of correlation-based approaches, we advocate for increased studies of model-based inference. In essence, such model-based approaches ask the question: which interaction network is compatible with the observed changes in populations arising from an underlying dynamical model? Thus model-based approaches avoid the assumption that correlations provide direct information

on interactions. Given favorable results of *in silico* benchmarking of model-based methods [41, 42, 47, 48, 49, 50], it will be important to take the next step: to investigate the efficacy of model-based inference of virus-microbe interaction networks *in situ*.

Acknowledgements

We thank Yu-Hui Lin and Ben Bolduc for their helpful comments and review. This work was supported by the Simons Foundation (SCOPE award ID 329108, J.S.W.).

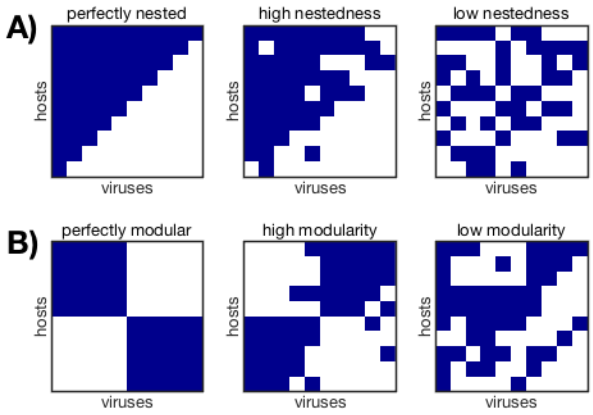

Figure 1: Six example interaction networks characterized by A) nestedness and B) modularity. These are 10 by 10 networks, with rows representing host types and columns representing viral types. Navy squares indicate interaction ($M_{ij} = 1$), that is, the virus can infect and lyse the host. From right to left, the nestedness values are NODF $= 1, 0.8,$ and 0.3, and the modularity values are $Q_{\text{norm}} = 1, 0.8,$ and 0.3. The networks are generated as described in §2.2 and are arranged in their most nested or most modular forms using [43]. Nestedness and modularity values are measured using [43].

	parameter	sampling range	units
H_i^*	host i target steady-state density	$10^3 - 10^4$	cells/mL
V_j^*	virus j target steady-state density	$10^6 - 10^7$	virions/mL
K	community-wide host carrying capacity	10^6	cells/mL
ϕ_{ij}	adsorption rate of virus j into host i	$10^{-7} - 10^{-6}$	mL/ (virion $\cdot$ day)
β_{ij}	burst size of virus j per host i	$10 - 100$	virions/cell
H_{0i}^*	host i target steady-state density in the absence of viruses	$10^3 - 10^6$	cells/mL
$a_{ii'}$	competitive effect of host i' on host i	$0 - 1$	

Table 1: Sampling ranges for the target steady-state densities and the life history traits in the virus-host dynamical system (Eqns 1 and 2). The remaining life history traits, the viral decay rates m_j and host growth rates r_i, are solved for in terms of these sampled parameters and the interaction network M (Eqns 4 and 5). Sampling ranges are taken from [45].

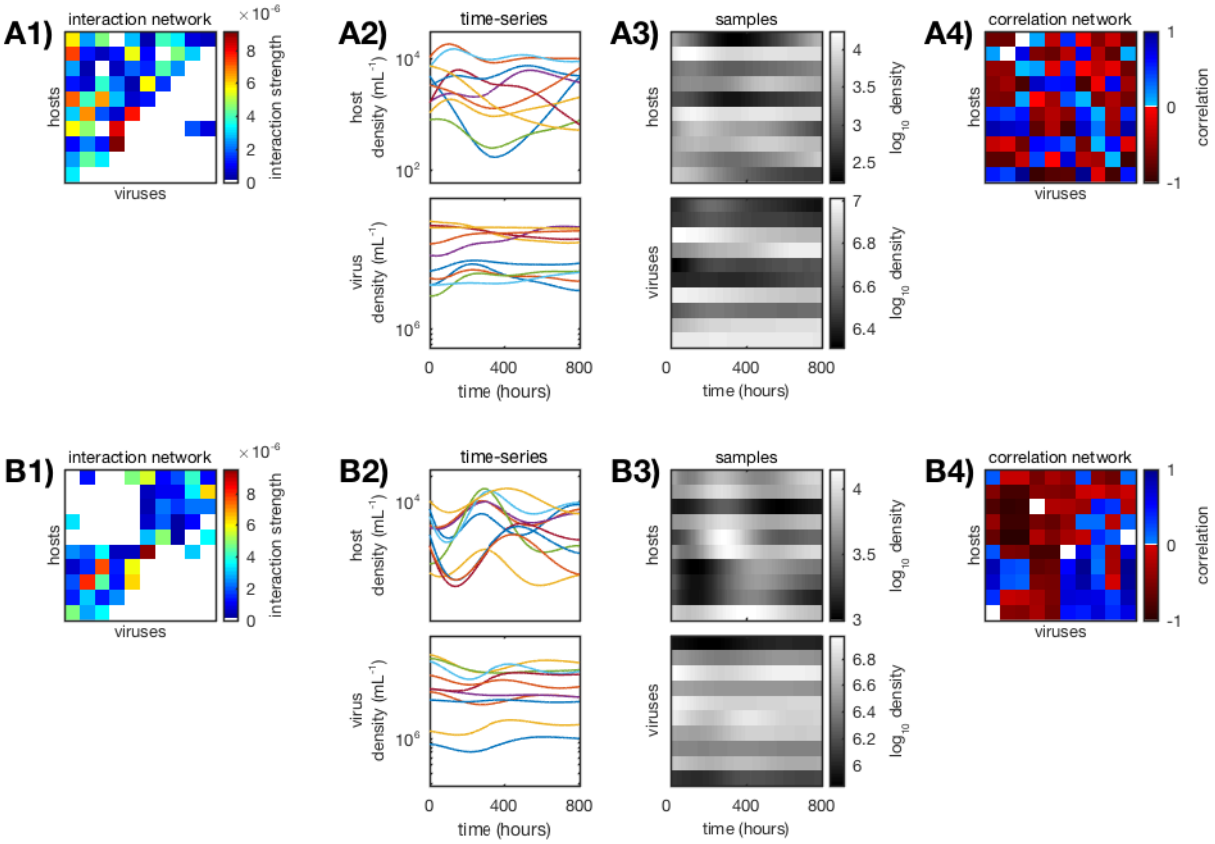

Figure 2: Calculating simple correlation networks for the *in silico* A) nested and B) modular communities. A1-B1) Original weighted interaction networks, generated as described in §2.3 from sample ranges in Table 1. A2-B2) Simulated time-series of the virus-host dynamical system (Eqns 1 and 2), generated as described in §2.4. A3-B3) Log-transformed samples of the time-series, sampled every 8 hours for 800 hours. A4-B4) Simple correlation networks, calculated from the log-transformed time-series samples (Eqn 6).

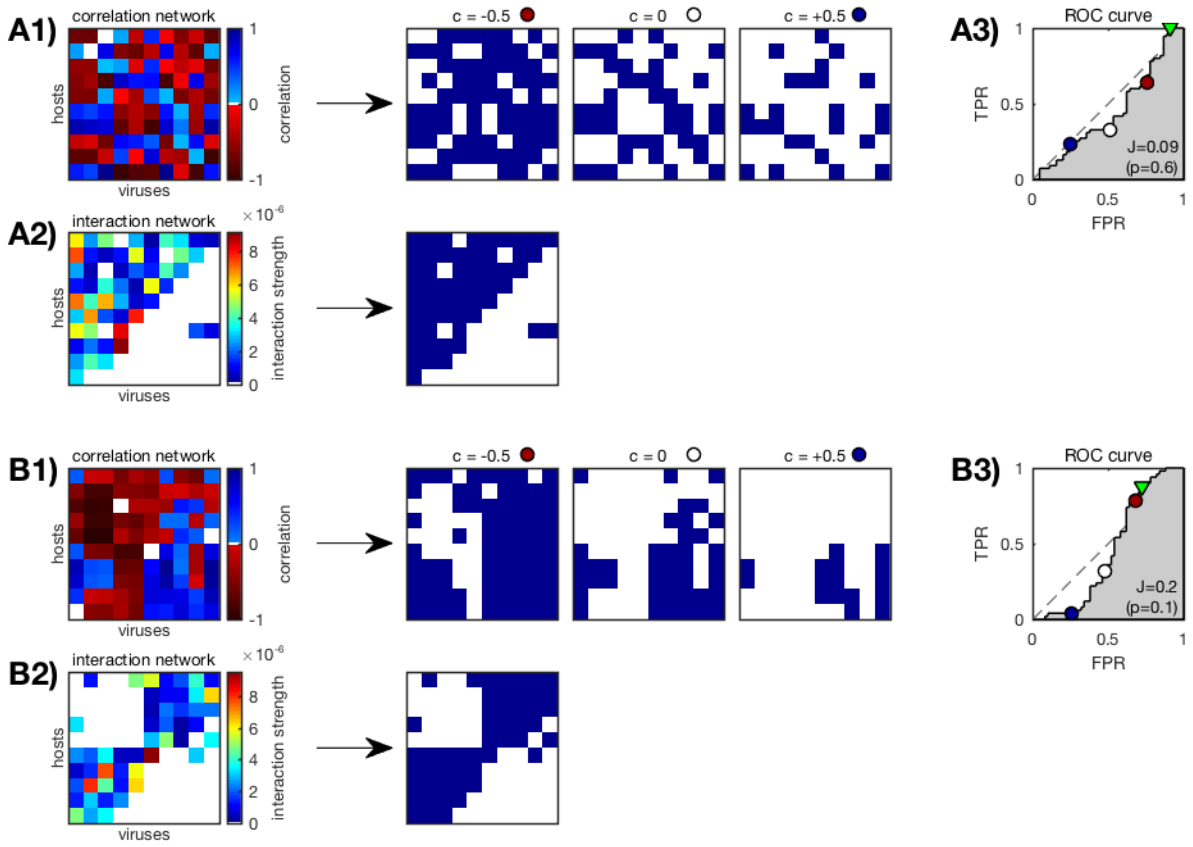

Figure 3: Evaluating correlation network efficacy for the *in silico* A) nested and B) modular communities. A1-B1) The correlation networks are binarized according to thresholds c between -1 and $+1$, three of which are shown here ($c = -0.5, 0,$ and 0.5). A2-B2) The original interaction networks are also binarized. A3-B3) For each threshold, the two binary networks are compared by calculating true positive rates (TPR) and false positive rates (FPR). On the resulting ROC curve, the three example thresholds ($c = -0.5, 0,$ and 0.5) are marked (red, white, and blue circles). The dashed line, or "non-discrimination" line, is where TPR is equal to FPR. The optimal threshold c^* is marked (green triangle) and yields the greatest difference J between TPR and FPR. Distributions for the reported p-values are shown in Supp Fig 1-A1 and -B1.

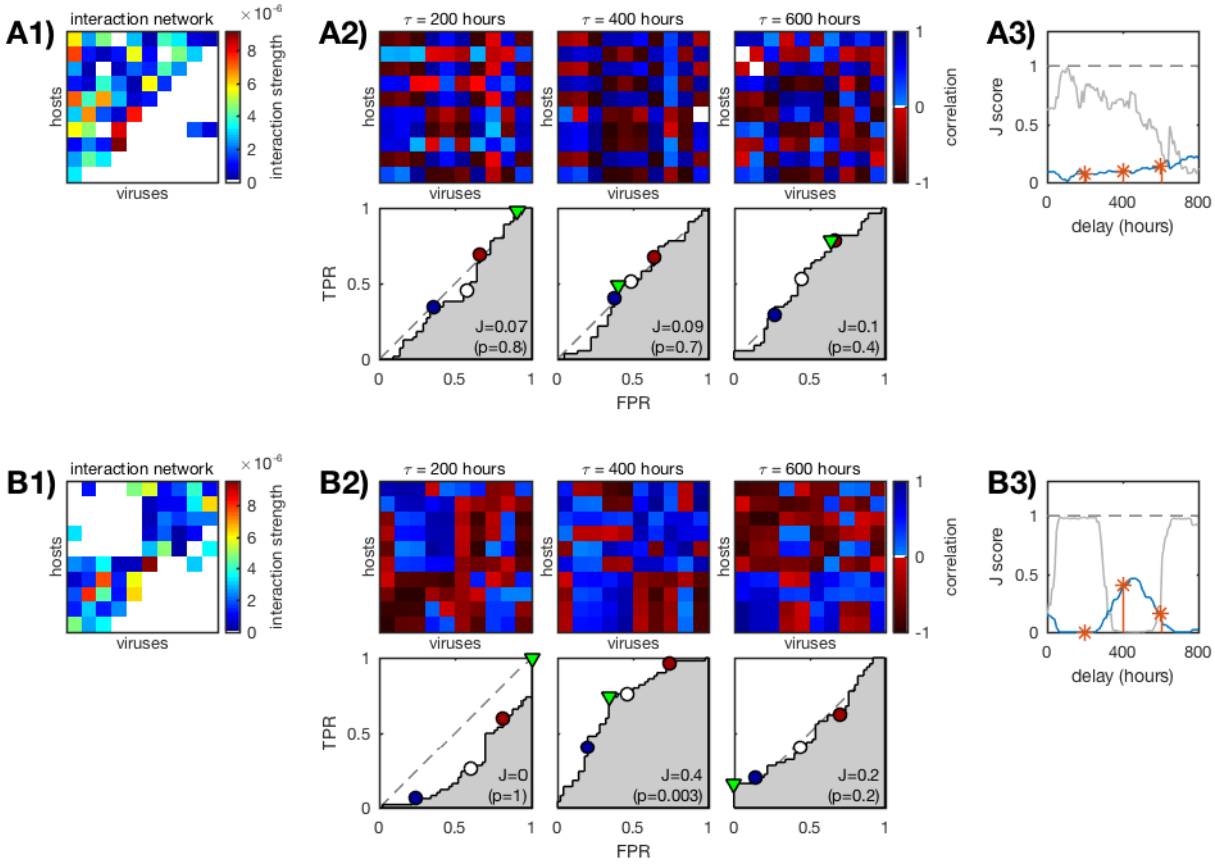

Figure 4: Time-delayed correlation networks with a community-wide time-delay for the *in silico* A) nested and B) modular communities. A1-B1) Original weighted interaction networks. A2-B2) Representative correlation networks for community-wide time-delays $\tau = 200, 400,$ and 600 hours (Eqn 7), with ROC curves below. On the ROC curves, the thresholds $c = -0.5, 0,$ and 0.5 are marked (red, white, and blue circles), as well as the optimal threshold c^* (green triangle). Distributions for the reported p-values are shown in Supp Fig 1-A2 and -B2. A3-B3) Scores J for every considered community-wide time-delay (blue line) and associated p-values (grey line). The representative time-delays $\tau = 200, 400,$ and 600 hours are marked (orange stars).

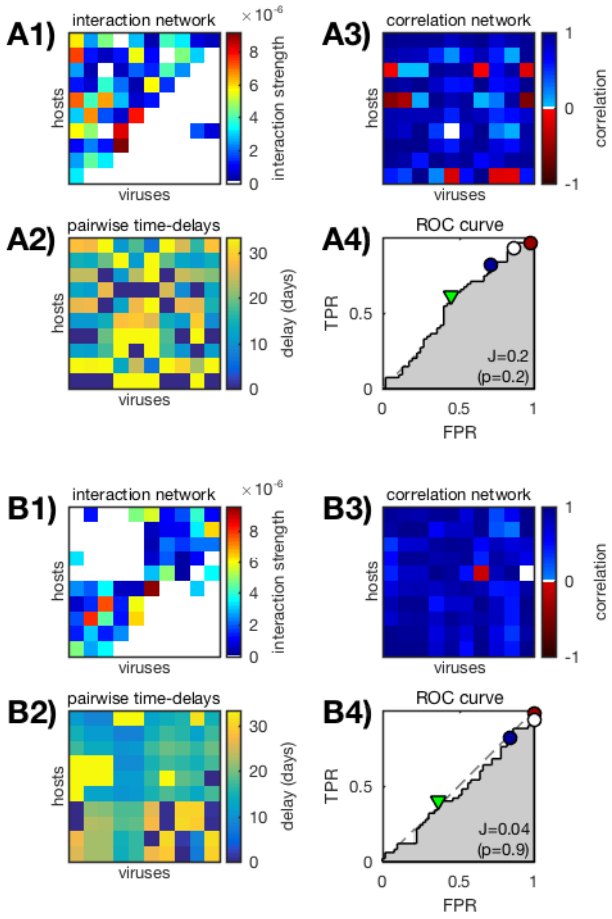

Figure 5: Time-delayed correlation networks with pairwise time-delays for the *in silico* A) nested and B) modular communities. A1-B1) Original weighted interaction networks. A2-B2) Pairwise time-delays applied to the host time-series samples. Time-delays are chosen so that the host-virus pair's correlation is maximized as described in §3.3. A3-B3) The time-delayed correlation networks (Eqn 7). A4-B4) ROC curves for the correlation networks. The thresholds $c = -0.5, 0,$ and 0.5 are marked (red, white, and blue circles), as well as the optimal threshold c^* (green triangle). Distributions for the reported p-values are shown in Supp Fig 1-A3 and -B3.

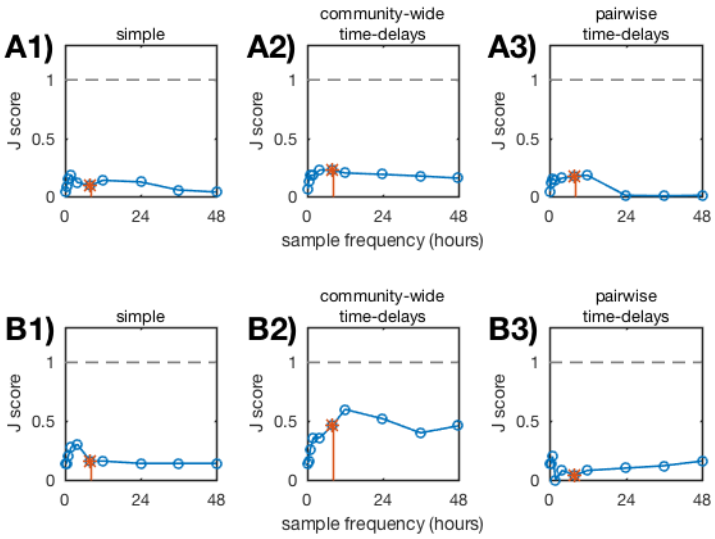

Figure 6: Correlation network efficacy J for varying sample frequencies for the *in silico* A) nested and B) modular communities. The correlation networks were implemented with A1-B1) no time-delays A2-B2) community-wide time-delays, and A3-B3) pairwise time-delays. The 8-hour sampling frequency used in earlier results (§3.1, 3.2, 3.3) is marked (orange star). For the community-wide time-delayed correlation networks, we report the maximum J across all considered time-delays for each sample frequency. Pairwise time-delays were chosen as described in §3.3 for each sample frequency. The number of sampled time-points was held constant at 100 as described in §3.4.

References

[1] F. Rohwer and R. V. Thurber, "Viruses manipulate the marine environment," *Nature*, vol. 459, no. 7244, pp. 207–12, 2009.

[2] L. D. McDaniel, E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul, "High frequency of horizontal gene transfer in the oceans," *Science*, vol. 330, no. 6000, p. 50, 2010.

[3] K. D. Bidle and A. Vardi, "A chemical arms race at sea mediates algal host-virus interactions," *Curr Opin Microbiol*, vol. 14, no. 4, pp. 449–57, 2011.

[4] D. Lindell, M. B. Sullivan, Z. I. Johnson, A. C. Tolonen, F. Rohwer, and S. W. Chisholm, "Transfer of photosynthesis genes to and from prochlorococcus viruses," *Proc Natl Acad Sci U S A*, vol. 101, no. 30, pp. 11013–8, 2004.

[5] J. S. Weitz and S. W. Wilhelm, "Ocean viruses and their effects on microbial communities and biogeochemical cycles," *F1000 Biol Rep*, vol. 4, p. 17, 2012.

[6] C. A. Suttle, "Viruses in the sea," *Nature*, vol. 437, no. 7057, pp. 356–61, 2005.

[7] M. B. Sullivan, J. B. Waterbury, and S. W. Chisholm, "Cyanophages infecting the oceanic cyanobacterium prochlorococcus," *Nature*, vol. 424, no. 6952, pp. 1047–51, 2003.

[8] J. S. Weitz, T. Poisot, J. R. Meyer, C. O. Flores, S. Valverde, M. B. Sullivan, and M. E. Hochberg, "Phage-bacteria infection networks," *Trends Microbiol*, vol. 21, no. 2, pp. 82–91, 2013.

[9] C. O. Flores, S. Valverde, and J. S. Weitz, "Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages," *ISME J*, vol. 7, no. 3, pp. 520–32, 2013.

[10] C. O. Flores, J. R. Meyer, S. Valverde, L. Farr, and J. S. Weitz, "Statistical structure of host-phage interactions," *Proc Natl Acad Sci U S A*, vol. 108, no. 28, pp. E288–97, 2011.

[11] M. B. Sullivan, J. S. Weitz, and S. Wilhelm, "Viral ecology comes of age," *Environ Microbiol Rep*, vol. 9, no. 1, pp. 33–35, 2017.

[12] L. Deng, A. Gregory, S. Yilmaz, B. T. Poulos, P. Hugenholtz, and M. B. Sullivan, "Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging," *MBio*, vol. 3, no. 6, 2012.

[13] L. Deng, J. C. Ignacio-Espinoza, A. C. Gregory, B. T. Poulos, J. S. Weitz, P. Hugenholtz, and M. B. Sullivan, "Viral tagging reveals discrete populations in synechococcus viral genome sequence space," *Nature*, vol. 513, no. 7517, pp. 242–5, 2014.

[14] A. D. Tadmor, E. A. Ottesen, J. R. Leadbetter, and R. Phillips, "Probing individual environmental bacteria for viruses by using microfluidic digital pcr," *Science*, vol. 333, no. 6038, pp. 58–62, 2011.

[15] M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer, "Genomic analysis of uncultured marine viral communities," *Proc Natl Acad Sci U S A*, vol. 99, no. 22, pp. 14250–5, 2002.

[16] R. A. Edwards and F. Rohwer, "Viral metagenomics," *Nat Rev Microbiol*, vol. 3, no. 6, pp. 504–10, 2005.

[17] M. R. Clokie, A. D. Millard, A. V. Letarov, and S. Heaphy, "Phages in nature," *Bacteriophage*, vol. 1, no. 1, pp. 31–45, 2011.

[18] M. Layeghifard, D. M. Hwang, and D. S. Guttman, "Disentangling interactions in the microbiome: A network perspective," *Trends Microbiol*, vol. 25, no. 3, pp. 217–228, 2017.

[19] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight, "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision," *ISME J*, vol. 10, no. 7, pp. 1669–81, 2016.

[20] K. Faust, L. Lahti, D. Gonze, W. M. de Vos, and J. Raes, "Metagenomics meets time series analysis: unraveling microbial community dynamics," *Curr Opin Microbiol*, vol. 25, pp. 56–66, 2015.

[21] K. Faust and J. Raes, "Microbial interactions: from networks to models," *Nat Rev Microbiol*, vol. 10, no. 8, pp. 538–50, 2012.

[22] J. A. Fuhrman, "Microbial community structure and its functional implications," *Nature*, vol. 459, no. 7244, pp. 193–9, 2009.

[23] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun, "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors," *Bioinformatics*, vol. 22, no. 20, pp. 2532–8, 2006.

[24] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun, "Extended local similarity analysis (elsa) of microbial community and other time series data with replicates," *BMC Syst Biol*, vol. 5 Suppl 2, p. S15, 2011.

[25] L. C. Xia, D. Ai, J. Cram, J. A. Fuhrman, and F. Sun, "Efficient statistical significance approximation for local similarity analysis of high-throughput time series data," *Bioinformatics*, vol. 29, no. 2, pp. 230–7, 2013.

[26] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS Comput Biol*, vol. 8, no. 9, p. e1002687, 2012.

[27] G. Sugihara, R. May, H. Ye, C. H. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.

[28] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara, "Distinguishing time-delayed causal interactions using convergent cross mapping," *Sci Rep*, vol. 5, p. 14750, 2015.

[29] T. Clark, H. Ye, F. Isbell, E. R. Deyle, J. Cowles, G. D. Tilman, and G. Sugihara, "Spatial convergent cross mapping to detect causal relationships from short time series," *Ecology*, vol. 96, no. 5, pp. 1174–81, 2015.

[30] J. M. McCracken and R. S. Weigel, "Convergent cross-mapping and pairwise asymmetric inference," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 90, no. 6, p. 062903, 2014.

[31] K. Suzuki, K. Yoshida, Y. Nakanishi, S. Fukuda, and S. McMahon, "An equation-free method reveals the ecological interaction networks within complex microbial ecosystems," *Methods in Ecology and Evolution*, 2017.

[32] C. E. Chow, R. Sachdeva, J. A. Cram, J. A. Steele, D. M. Needham, A. Patel, A. E. Parada, and J. A. Fuhrman, "Temporal variability and coherence of euphotic zone bacterial communities over a decade in the southern california bight," *ISME J*, vol. 7, no. 12, pp. 2259–73, 2013.

[33] J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbruck, J. Reeder, B. Temperton, S. Huse, A. C. McHardy, R. Knight, I. Joint, P. Somerfield, J. A. Fuhrman, and D. Field, "Defining seasonal marine microbial community dynamics," *ISME J*, vol. 6, no. 2, pp. 298–308, 2012.

[34] L. Liu, J. Yang, H. Lv, and Z. Yu, "Synchronous dynamics and correlations between bacteria and phytoplankton in a subtropical drinking water reservoir," *FEMS Microbiol Ecol*, vol. 90, no. 1, pp. 126–38, 2014.

[35] S. F. Paver, K. R. Hayek, K. A. Gano, J. R. Fagen, C. T. Brown, A. G. Davis-Richardson, D. B. Crabb, R. Rosario-Passapera, A. Giongo, E. W. Triplett, and A. D. Kent, "Interactions between specific phytoplankton and bacteria affect lake bacterial community succession," *Environmental Microbiology*, vol. 15, no. 9, pp. 2489–2504, 2013.

[36] D. M. Needham, C. E. Chow, J. A. Cram, R. Sachdeva, A. Parada, and J. A. Fuhrman, "Short-term observations of marine bacterial and viral communities: patterns, connections and resilience," *ISME J*, vol. 7, no. 7, pp. 1274–85, 2013.

[37] C. E. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, and J. A. Fuhrman, "Top-down controls on bacterial community structure: microbial network analysis of bacteria, t4-like viruses and protists," *ISME J*, vol. 8, no. 4, pp. 816–29, 2014.

[38] J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C. E. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, J. M. Rose, I. Hewson, A. Patel, F. Sun, D. A. Caron, and J. A. Fuhrman, "Marine bacterial, archaeal and protistan association networks reveal ecological linkages," *ISME J*, vol. 5, no. 9, pp. 1414–25, 2011.

[39] T. Miki and S. Jacquet, "Complex interactions in the microbial world: underexplored key links between viruses, bacteria and protozoan grazers in aquatic environments," *Aquatic Microbial Ecology*, vol. 51, pp. 195–208, 2008.

[40] T. Miki and S. Jacquet, "Indirect interactions in the microbial world: specificities and similarities to plantinsect systems," *Population Ecology*, vol. 52, no. 4, pp. 475–483, 2010.

[41] C. K. Fisher and P. Mehta, "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression," *PLoS One*, vol. 9, no. 7, p. e102451, 2014.

[42] Z. D. Kurtz, C. L. Muller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, "Sparse and compositionally robust inference of microbial ecological networks," *PLoS Comput Biol*, vol. 11, no. 5, p. e1004226, 2015.

[43] C. O. Flores, T. Poisot, S. Valverde, and J. S. Weitz, "Bimat: a matlab package to facilitate the analysis of bipartite networks," *Methods in Ecology and Evolution*, vol. 7, no. 1, pp. 127–132, 2016.

[44] M. E. J. Newman, *Networks: An Introduction*. New York: Oxford University Press Inc., 2010.

[45] L. F. Jover, M. H. Cortez, and J. S. Weitz, "Mechanisms of multi-strain coexistence in host-phage systems with nested infection networks," *J Theor Biol*, vol. 332, pp. 65–77, 2013.

[46] M. D. Ruopp, N. J. Perkins, B. W. Whitcomb, and E. F. Schisterman, "Youden index and optimal cut-point estimated from observations affected by a lower limit of detection," *Biom J*, vol. 50, no. 3, pp. 419–30, 2008.

[47] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Ratsch, E. G. Pamer, C. Sander, and J. B. Xavier, "Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota," *PLoS Comput Biol*, vol. 9, no. 12, p. e1003388, 2013.

[48] L. F. Jover, J. Romberg, and J. S. Weitz, "Inferring phage-bacteria infection networks from time-series data," *R Soc Open Sci*, vol. 3, no. 11, p. 160654, 2016.

[49] P. Dam, L. L. Fonseca, K. T. Konstantinidis, and E. O. Voit, "Dynamic models of the complex microbial metapopulation of lake mendota," *NPJ Syst Biol Appl*, vol. 2, p. 16007, 2016.

[50] S. Marino, N. T. Baxter, G. B. Huffnagle, J. F. Petrosino, and P. D. Schloss, "Mathematical modeling of primary succession of murine intestinal microbiota," *Proc Natl Acad Sci U S A*, vol. 111, no. 1, pp. 439–44, 2014.