Supplementary Information: Multi-scale Structure and Geographic Drivers of Cross-Infection within Marine Bacteria and Phages

Cesar O. Flores[1], Sergi Valverde[2], and Joshua S. Weitz*[3,1]

[1]1School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2]Complex Systems Lab and Institute of Evolutionary Biology, University Pompeu Fabra, E-08003 Barcelona, Spain
[3]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA

S1 Dataset

The dataset analyzed here is a subset of the phage-bacteria cross-reaction tests reported by K. Moebus and H. Nattkemper [6]. Among all the datasets reported in this paper, we have focused in the largest collection of tests, i.e., the so-called A-series dataset. This dataset consists of $H = 733$ bacteria and $P = 258$ bacteriophages strains collected at 48 water sample stations in the Atlantic Ocean region (see Figure S1). Only 326 out of the 733 bacteria were found to be susceptible to one or more phages. From the 326 bacteria strains, 250 are unique (the infection pattern is different from each other), 38 are inter-sample doublets (bacteria that have the same infection pattern of another bacteria belonging to a different water sample or station), and 38 intra-sample doublets (doublets from the same water samples). Similarly, there are 224 unique phage strains and 4 inter-sample doublets.

The only source of information about the matrix of cross-reaction tests was the figure shown in the Moebus and Nattkemper paper (see Figure 1 in [6], Figure S2 in this document). We were unable to find other means to access this dataset and thus, we have developed a semi-automatic scanning method to recover this matrix from the printed paper to a digital format suitable for our analysis (see method below). For example, the original paper does not indicate the exact number of bacteria and phages represented in the original figure (see Figure S2). Instead, these numbers have been inferred from the original figure labels and the information given in the whole document (see below). The digitalization process includes the following steps:

1. We scanned the source image from the printed figure in [6] (see Figure 1 in [6] and Figure S2 in this document). The quality of the image made the extraction process difficult. First, the original image is slightly rotated by an angle comprised between 0.4 and 0.6 degrees counterclockwise (depending on what side of the image is chosen as a reference). In addition, there was a tear starting at the bottom (phage station number 484) and running to the left (phage station number 462) of the image that slightly distorts the orientation at the bottom right section. Here, we have estimated the rotation angle to be 0.45 degrees, which is good compromise between the left and bottom orientations. As a consequence of the previous rotation, two bacteria records were lost.

*To whom correspondence should be addressed. E-mail: jsweitz@gatech.edu.

2. We assume that matrix size is approximately equal to the number of columns and rows visible in the source image. We manually cross-checked the row and column counts and find $H = 288$ bacteria and $P = 222$ phages. Further validation comes from a computer program that counts the number of mouse clicks performed by a human over each bacteria/phage label in the "source" (scanned) image. The observed number of bacteria is consistent with the caption of the source figure that reports 288 bacteria strains (250 unique + 38 inter-sample doubles). The case for phages is more ambiguous because the original figure only labels 217 phages out of the 222 (readable) columns. Here, we have only retained labeled and readable phages to yield $H = 286$ bacteria and $P = 215$ phages.

3. We performed a binary thresholding of the source matrix to automatically detect positive interactions of phages with hosts by computing the density of filled pixels at every matrix cell. We delimited the matrix cells by overlaying a grid in the source figure, and the interactions were detected by specifying a threshold of filled pixels inside each cell. This automatic process makes no distinction between matrix cells that denote clear lysis or turbid spots.

4. We manually curated the binary thresholded image to identify and correct any false negatives (undetected interactions) and false positives (empty cells marked as interactions). In addition, empty columns were removed. The output is the curated MN (Moebus and Nattkemper) matrix used for our study (see Figure 1 of main document, and supplementary Figure S2).

S2 Bipartite Modularity

A host-phage interaction matrix can be described as a bipartite network $G = (U, V, E)$ having two disjoint sets of nodes (phages and hosts) and a set of edges ([3]). Here, $H = \|U\|$ is the number of hosts and $P = \|V\|$ is the number of phages and there is an edge $\{u_i, v_j\} \in E$ when phage $v_j \in V$ infects host $u_i \in U$. Notice that interactions between nodes of the same type are excluded. Alternatively, the *adjacency matrix* $A = [A_{ij}]$ indicates whether the $j-$th phage can infect the $i-$th host ($A_{ij} = 1$) or not ($A_{ij} = 0$). Notice that this matrix corresponds to the binary thresholded image obtained in the previous section. A number of useful network measures can be obtained from the adjacency matrix alone. The degree $k_i = \sum_j A_{ij}$ of the $i-$th host is the number of interactions with phages (i.e. how many phages can infect the $i-$th host). The degree $d_j = \sum_i A_{ij}$ of the $j-$th phage is the number of interactions with hosts (i.e. how many hosts can be infected by the $j-$th phage). See Figure S3 for a plot of the cumulative degree frequency of the MN matrix.

An important collection of network measures involves the quantification of interaction patterns in subsets of more than two network nodes. For example, a visual inspection of the infection matrix shown in Figure 3 of main document suggests that there are modules of hosts and phages exchanging many more "ones" between them (a higher density of internal links) than with the rest of types (nodes). Following [2], we assess the quality of a given partition in c (disjoint) modules with the *bipartite modularity*:

$$Q = \frac{1}{m} \sum_{ij} \left(A_{ij} - P_{ij}\right) \delta(g_i, g_j) \tag{1}$$

where A_{ij} is the adjacency matrix, $m = \sum_{ij} A_{ij}$ is the total number of links, $P_{ij} = k_i d_j / m$ is the probability to connect nodes i and j, the node i has been assigned to the module g_i, and $\delta(x, y) = 1$ if $x = y$ and $\delta(x, y) = 0$ when x and y are different. Intuitively, high values of Q will correspond to highly modular partitions of the bipartite network. In this case, node i and j are classified in the same module $g_i = g_j$ (and thus $\delta(g_i, g_j) = 1$) because the probability to

have a link between nodes in the same module is significant (e.g., the difference $A_{ij} - P_{ij}$ is a large, positive value).

For convenience, we use the matrix form of the modularity Equation (1). Here, we replace the function g_i by the $H \times c$ index matrix $\mathbf{R} = [\mathbf{r}_1|\mathbf{r}_2|...|\mathbf{r}_c]$ and the $P \times c$ index matrix $\mathbf{T} = [\mathbf{t}_1|\mathbf{t}_2|...|\mathbf{t}_c]$, for hosts and phages, respectively [2]. Notice that nodes cannot be classified into more than one module. Vectors $\mathbf{r}_i$ and $\mathbf{t}_i$ consist of a single one (corresponding to the chosen module) will all the other entries being zero. For example, $r_{ik} = 1$ if the i-th host belongs to the k-th module and $r_{ij} = 0$ for every other $j \neq k$. Now, we can rewrite the modularity as follows (see Equation (22) in [2]):

$$Q = \frac{1}{m} \operatorname{Tr} \mathbf{R}^T \tilde{\mathbf{B}} \mathbf{T} \tag{2}$$

where $\tilde{\mathbf{B}} = \mathbf{A} - \mathbf{P}$ is the *bipartite modularity matrix*. The goal of the modularity algorithm is to find the optimal assignment of nodes to modules (i.e., the index vectors $\mathbf{R}$ and $\mathbf{T}$) in a way that Equation 2 becomes maximized. However, finding the optimal modularity is a NP-complete problem. In this context, there are a number of practical heuristics that we can use to guide modularity algorithms in the search for good solutions within computational constraints (we always check that the solutions found by the algorithms are meaningful). Next, we discuss the different heuristics explored here.

The original modularity algorithm (called BRIM for Bipartite, Recursively Induced Modules) described in [2] computes the optimal modularity by inducing the division of one set of nodes (say vector $\mathbf{T}$) from the division in the other set of nodes (say vector $\mathbf{R}$). At each step, BRIM assigns nodes of one type to modules in order to maximize the modularity. BRIM iterates this process until a local maximum is reached. However, the choice of a predefined number c of modules limits the efficacy of the algorithm. Barber extended the BRIM algorithm to search for the optimal number of modules along the modularity maximization process [2]. This method, which is called *"adaptive BRIM"*, assumes that there is a smooth relationship between the number of modules c and the modularity $Q(c)$. For continuous and smooth landscapes, a simple bisection method ensures that we will find the optimal value of c corresponding to maximum Q. Starting at $c = 1$ (and modularity $Q(1) = 0$ because all nodes belong to the same module) the adaptive BRIM searches for optimal c by repeatedly doubling the number of modules while modularity increases, $Q(2c) > Q(c)$. At some point, the search crosses a maximum in the modularity landscape, i.e., $Q(2c) < Q(c)$, and we interpolate the number of modules c^* to some intermediate value in the current interval $(c, 2c)$. This heuristic gives very good modularity values for the case of small matrices. For example, we have used the adaptive heuristic in the analysis of the 15 largest modules identified in the MN matrix.

A shortcoming of adaptive BRIM is that its performance degrades for large networks [4]. We propose a recursive algorithm based in [7] to find the optimal number of modules in the full cross-infection matrix. Following [7], we perform repeated divisions of the network until a local maximum of modularity is reached. The algorithm steps are: (i) find all the isolated network components and place them into separated modules, (ii) subdivide each module into $c = 2$ sub-modules using the standard BRIM algorithm and (iii) repeat the subdivision process until there is no improvement in the overall network modularity. The stop condition evaluates if the modularity change ΔQ corresponding to the subdivision event in (ii) is significant or not. That is, $\Delta Q > 0$ means there is still room for further subdivisions. Newman suggests that is not correct to naively remove all edges falling between the subparts and apply the full modularity algorithm to each subpart in isolation [7]. We compute $\Delta Q > 0$ as the difference between the modularity value computed after and before the splitting event:

$$\Delta Q = \frac{1}{m} \left[\operatorname{Tr} {R^{(g)}}^T \hat{B}^{(g)} T^{(g)} - \operatorname{Tr} \hat{B}^{(g)} \right] \tag{3}$$

where $\hat{B}^{(g)}$ is the $h_g \times p_g$ bipartite modularity matrix of the h_g hosts and p_g pages within the module $g \subseteq G$, and $R^{(g)}$ and $T^{(g)}$ are the index vectors describing the splitting of the subgraph

g in two sub-modules. Notice that we can restrict our computation to the subgraph g and thus, the index vectors are subsets of the full index vectors (see Equation 2). This is, to the best of our knowledge, the first time that the Newman's division algorithm has been applied to bipartite networks.

S3 Multi-scale nested analysis

The MN matrix is significantly nested according to initial analysis using both the temperature calculator and NODF. This result is surprising giving the apparent lack of nestedness in visual inspection. However, prior work has noted that standard nestedness measures can signal spurious nested patterns when the network is comprised of nested modules [3]. In this context, Almeida-Neto and co-workers argue that we need specific models for distinct non-nested patterns because there is not an unique, working definition for the opposite of nestedness ("anti-nestedness") [1]. Here, we propose two new approaches (one for each nestedness measurement) to discard any interference of modular organization in the assessment of "true" nestedness.

We start by computing the modular organization of the full network G with our division algorithm (see Section S2). The modules will constrain the space of possible matrix re-arrangements explored by the temperature calculator when searching for the maximum nestedness (minimum temperature). In particular, our proposal for a constrained temperature calculator (i) permutes full modules (or matrix blocks), (ii) permutes rows and columns within a module, (iii) cannot perform any other permutation different from (i) and (ii). Still, the space of possible combinations can be quite large. We developed a heuristic algorithm that obtains good results with simple and deterministic sorting. First, we sort the rows and columns within any module in decreasing degree order (notice that rows and columns are sorted independently). Second, we rank modules according to the (sub-)matrix size and fill. The host (rows) ranking μ_g for the module $g \subset G$ is:

$$\mu_g = \frac{\sum_{i \in g} k_i}{h_g \times P} \tag{4}$$

where h_g is the number of hosts in the module g, k_i is the degree of the $i-$th host and P is the number of phages in the full network. Notice that this score can be seen as the connectance of a network composed of all phages presented in the entire network but only the hosts that belongs to module g. Similarly, there is a phage (columns) ranking ν_g for the module g:

$$\nu_g = \frac{\sum_{j \in g} d_j}{p_g \times H} \tag{5}$$

where p_g is the number of phages in the module g, d_j is the degree of the $j-th$ phage and H is the number of hosts in the full network.

In order to validate this measure of constrained nestedness, we have designed a theoretical experiment with synthetic networks having $2 \leq c \leq 50$ perfectly nested modules without interactions between them. Model networks have the same size as the MN network ($H = 286$, $P = 215$). Notice that $\mu_g = \mu$ and $\nu_g = \nu$ for all modules (blocks) because they have exactly the same size and fill. We place modules along the main diagonal to achieve optimal nestedness (see Figure S5). Every other arrangement (for example with off-diagonal blocks) yields sub-optimal nestedness values.

Our experiment confirms the initial hypothesis, i.e., unconstrained nestedness is higher than constrained nestedness (see Figure S6). This suggests how high unconstrained nestedness of the MN matrix can be a consequence of its nested modular organization. As expected, we achieve maximum nestedness when the matrix is perfectly nested, e.g., there is only $c = 1$ module (see Figure S5 left). At $c = 2$ we have a sudden drop in (both constrained and unconstrained) nestedness because there are interactions below the isocline and absence of interactions above

the isocline (see Figure S5 center). For small values of modularity ($c < 8$), the two null models have significantly lower values of constrained nestedness than the MN matrix. In general, nestedness increases with the number of modules ($c > 20$, see Figure S6) because temperature is directly related to the matrix filling (see Figure S5 right).

S4 Geographical analysis

Both nestedness and modularity are topological, aspatial characteristics of bipartite networks. Here, we investigate the relationship between these network patterns and their spatial context. The MN matrix describes observed infections between host and phages sampled from a set of nearly equally-spaced, numbered stations in the Atlantic ocean. Here, we will review the original hypothesis of the MN study, i.e., to what extent geographical location drives the infection process. In the presence of strong spatial modularity, we should observe significant correlations between stations numbers (a surrogate of geographical location) of nodes within the same module. Otherwise, the geographical biodiversity will be very large.

We will use two different, standard metrics to measure the degree of geographical biodiversity in a topological module. For each module, we will compute the Shannon's entropy index:

$$H_k = -\sum_{i=1}^{R} \frac{n_i}{N} \log \frac{n_i}{N} \tag{6}$$

and the Simpson's diversity index:

$$D_k = 1 - \sum_{i=1}^{R} \frac{n_i(n_i - 1)}{N(N-1)} \tag{7}$$

where N are the number of different strains inside the module, R are the number of stations inside the module, and n_i are the number of strains from the $i-$th station. Low values in both indices indicate low geographical diversity within modules. Using a combination of two diversity indexes will provide additional support for our conclusions.

In order to test the NM hypothesis, we compare the observed diversity indexes (H_1, D_1), (H_2, D_2) ... (H_{15}, D_{15}) for the largest 15 modules found by the BRIM algorithm in the NM matrix (see above) with their expectations coming from an ensemble of 10^6 randomized matrices. We generate each sample by randomly permuting the row and column labels of the NM matrix. Once the random matrix is obtained, we will compare the diversity indexes of each observed module (H_k, D_k) with the pair of indices $(\tilde{H}_k, \tilde{D}_k)$ of random modules having the same size. Figure S7 indicates that, overall, the largest 15 modules display low geographical diversity, i.e., the observed value is lower than expected (considering a one-tailed p-value of 0.05 for statistical significance). This observation appears to be equally valid for hosts and phages (we have analyzed the two types of nodes separately), e.g., compare Figure S7a and Figure S7b.

Table S1: Geographical data of microbial stations

Station	Latitude	Longitude	Station	Latitude	Longitude
454	47.717	-6.633	526	29.600	-57.083
456	44.750	-10.917	531	27.933	-57.733
458	43.200	-14.283	536	30.000	-58.333
460	41.350	-18.067	541	31.500	-59.667
462	39.650	-21.800	547	28.833	-59.633
464	38.000	-24.633	554	26.517	-60.233
465	37.817	-29.050	559	28.500	-61.000
469	37.967	-33.283	564	30.500	-61.000
471	37.333	-37.350	565	32.333	-64.633
472	36.550	-42.383	568	33.050	-59.983
474	35.717	-47.083	570	34.017	-55.317
476	34.867	-51.517	572	36.050	-42.467
478	34.017	-55.317	576	36.433	-39.067
480	33.217	-59.333	581	37.050	-34.350
484	32.567	-62.950	588	37.767	-26.367
489	31.967	-65.183	590	37.333	-22.033
492	30.667	-62.750	593	36.850	-17.417
497	28.783	-60.350	596	36.500	-13.000
501	27.117	-58.550	598	36.117	-8.717
504	26.100	-58.583	600	36.333	-7.467
508	26.417	-58.783	601	41.583	-10.333
513	29.617	-58.883	602	43.617	-9.567
518	31.200	-62.017	603	44.783	-8.833
522	31.067	-57.300	605	47.533	-6.283

Information that were extracted from the original Table 1 [5].

Table S2: Global properties of the extracted modules

Module	H	P	S	I	M	C	L_p	L_h
1	42	23	269	65	966	0.28	6.40	11.70
2	39	12	138	51	468	0.29	3.54	11.50
3	31	31	233	62	961	0.24	7.52	7.52
4	23	13	61	36	299	0.20	2.65	4.69
5	16	20	114	36	320	0.36	7.13	5.70
6	15	5	30	20	75	0.40	2.00	6.00
7	12	7	27	19	84	0.32	2.25	3.86
8	11	8	52	19	88	0.59	4.73	6.50
9	8	6	38	14	48	0.79	4.75	6.33
10	8	11	57	19	88	0.65	7.13	5.18
11	7	5	15	12	35	0.43	2.14	3.00
12	7	7	17	14	49	0.35	2.43	2.43
13	7	9	49	16	63	0.78	7.00	5.44
14	6	7	21	13	42	0.50	3.50	3.00
15	6	6	27	12	36	0.75	4.50	4.50
16	3	4	12	7	12	1.00	4.00	3.00
17	3	3	7	6	9	0.78	2.33	2.33
18	3	1	3	4	3	1.00	1.00	3.00
19	3	1	3	4	3	1.00	1.00	3.00
20	2	1	2	3	2	1.00	1.00	2.00
21	2	3	6	5	6	1.00	3.00	2.00
22	2	1	2	3	2	1.00	1.00	2.00
23	2	1	2	3	2	1.00	1.00	2.00
24	2	2	4	4	4	1.00	2.00	2.00
25	2	2	4	4	4	1.00	2.00	2.00
26	1	1	1	2	1	1.00	1.00	1.00
27	1	2	2	3	2	1.00	2.00	1.00
28	1	1	1	2	1	1.00	1.00	1.00
29	1	1	1	2	1	1.00	1.00	1.00
30	1	1	1	2	1	1.00	1.00	1.00
31	1	1	1	2	1	1.00	1.00	1.00
32	1	1	1	2	1	1.00	1.00	1.00
33	1	1	1	2	1	1.00	1.00	1.00
34	1	1	1	2	1	1.00	1.00	1.00
35	1	1	1	2	1	1.00	1.00	1.00
36	1	1	1	2	1	1.00	1.00	1.00
37	1	1	1	2	1	1.00	1.00	1.00
38	1	1	1	2	1	1.00	1.00	1.00
39	1	1	1	2	1	1.00	1.00	1.00
40	1	1	1	2	1	1.00	1.00	1.00
41	1	1	1	2	1	1.00	1.00	1.00
42	1	1	1	2	1	1.00	1.00	1.00
43	1	1	1	2	1	1.00	1.00	1.00
44	1	1	1	2	1	1.00	1.00	1.00
45	1	1	1	2	1	1.00	1.00	1.00
46	1	1	1	2	1	1.00	1.00	1.00
47	1	1	1	2	1	1.00	1.00	1.00
48	1	1	1	2	1	1.00	1.00	1.00
49	1	2	2	3	2	1.00	2.00	1.00
Average	5.84	4.39	24.88	10.22	75.41	0.83	2.29	2.75
Median	2.00	1.00	2.00	3.00	2.00	1.00	1.00	2.00

H : Number of hosts
P : Number of phages
$S = H + P$: Number of species
I : Number of interactions
$M = HP$: Size
$C = I/M$: Connectance or fill
$L_p = I/P$: Mean phage degree (Average number of susceptible hosts by phage)
$L_h = I/H$: Mean host degree (Average number of virulent viruses by host)

Table S3: Geographical biodiversity indexes

Module	Phages		Hosts	
	Simpson	Shannon	Simpson	Shannon
1	0.953 ($p = 0.086$)	2.487 ($p = 0.040$)	0.970 ($p = 0.272$)	3.048 ($p = 0.221$)
2	0.939 ($p = 0.065$)	2.095 ($p = 0.081$)	0.964 ($p = 0.093$)	2.908 ($p = 0.048$)
3	0.897 ($p = 0.000$)	2.179 ($p = 0.000$)	0.920 ($p = 0.000$)	2.551 ($p = 0.001$)
4	0.808 ($p = 0.000$)	1.479 ($p = 0.000$)	0.909 ($p = 0.000$)	2.198 ($p = 0.000$)
5	0.816 ($p = 0.000$)	1.817 ($p = 0.000$)	0.825 ($p = 0.000$)	1.689 ($p = 0.000$)
6	1.000 ($p = 0.280$)	1.609 ($p = 0.280$)	0.962 ($p = 0.158$)	2.396 ($p = 0.227$)
7	0.714 ($p = 0.000$)	1.004 ($p = 0.000$)	0.833 ($p = 0.000$)	1.517 ($p = 0.000$)
8	0.857 ($p = 0.004$)	1.494 ($p = 0.010$)	0.909 ($p = 0.012$)	1.846 ($p = 0.011$)
9	0.333 ($p = 0.000$)	0.451 ($p = 0.000$)	1.000 ($p = 0.552$)	2.079 ($p = 0.552$)
10	0.909 ($p = 0.020$)	1.768 ($p = 0.005$)	0.893 ($p = 0.013$)	1.667 ($p = 0.027$)
11	0.900 ($p = 0.025$)	1.332 ($p = 0.025$)	0.857 ($p = 0.005$)	1.475 ($p = 0.007$)
12	0.952 ($p = 0.111$)	1.748 ($p = 0.111$)	1.000 ($p = 0.453$)	1.946 ($p = 0.453$)
13	0.889 ($p = 0.010$)	1.677 ($p = 0.013$)	0.857 ($p = 0.006$)	1.475 ($p = 0.008$)
14	0.571 ($p = 0.000$)	0.683 ($p = 0.000$)	0.533 ($p = 0.000$)	0.637 ($p = 0.000$)
15	0.600 ($p = 0.000$)	0.868 ($p = 0.000$)	0.733 ($p = 0.001$)	1.011 ($p = 0.001$)

Small values means low geographical biodiversity. $p < 0.05$ means the module is statistically no geographically diverse. p-values were calculated as the ratio of random permutations index values that are smaller than the real index. See Equation 4 in the main text for a mathematical description of these indexes.

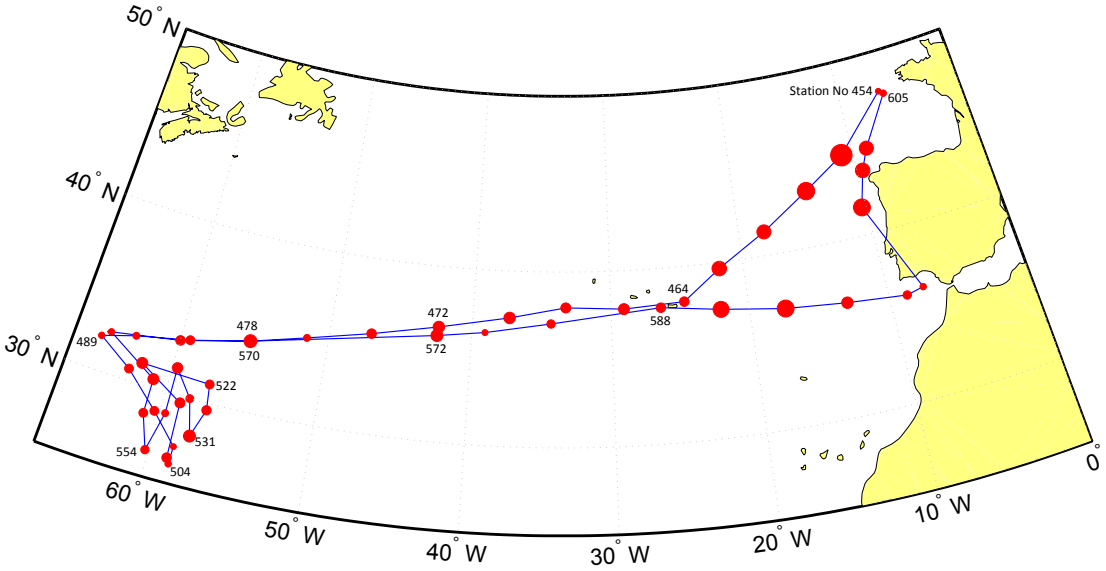


Figure S1: Originally appeared as Figure 1 on [5] with the label *Track of RV "Friedrich Heincke" in the Atlantic Ocean during cruise no. 160 and microbial stations.* Here, each circle represents the geographic location of each station. The radius of the circles corresponds linearly to the number of strains that were extracted in the corresponding station. Some number stations are indicated in order to clarify the direction of the route. Increasing station number indicate the order of visit.

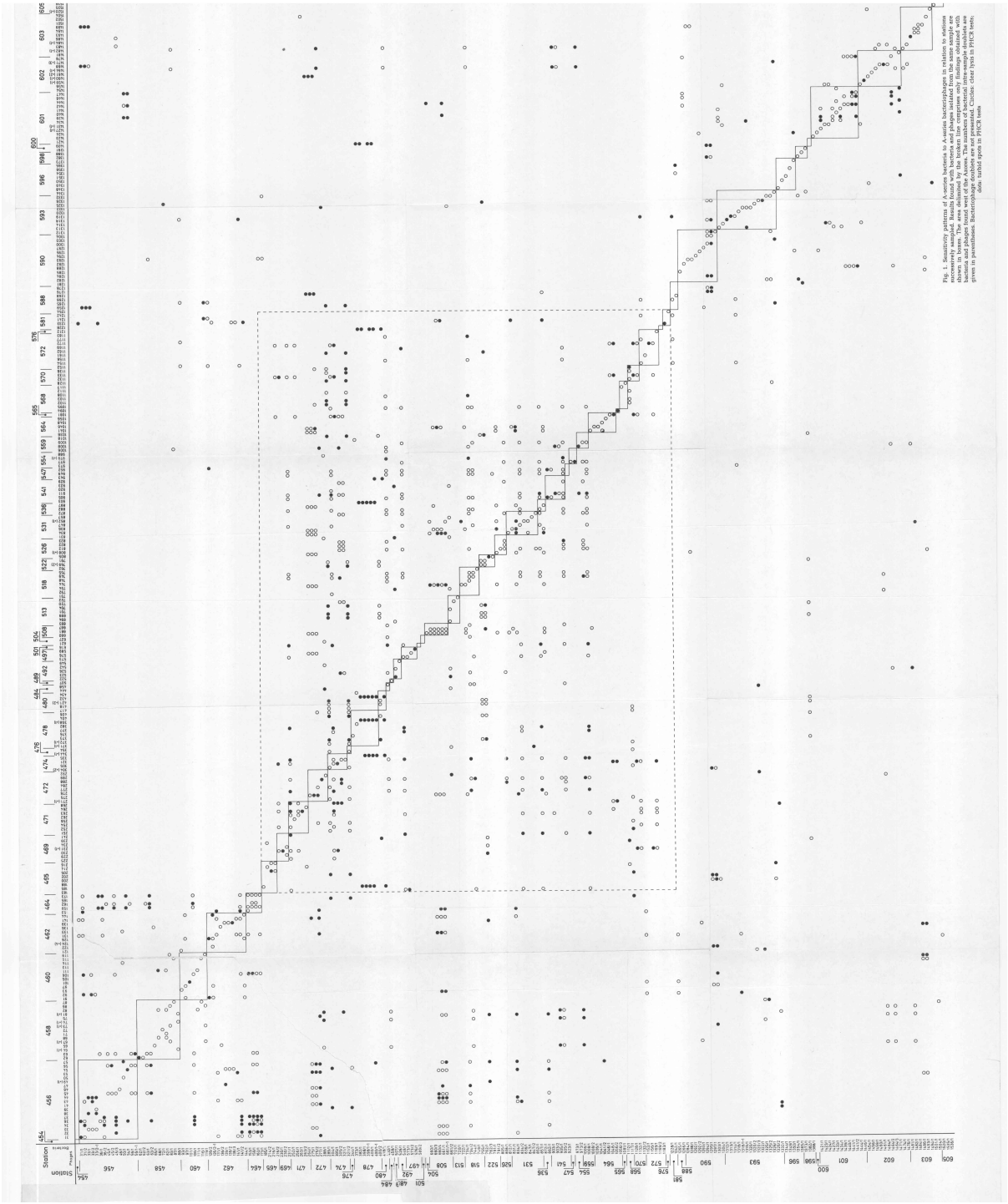
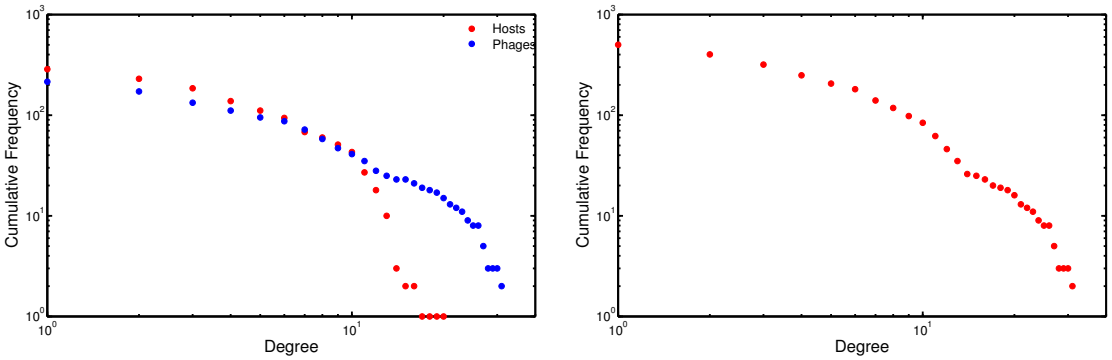


Figure S2: Moebus & Nattkemper [6] cross-reaction test in the Atlantic Ocean region. This matrix is subdivided in different stations, where each square delimits the infections inside strains of the same station. The original label reads: *"Fig 1. Sensitivity patterns of A-seres bacteria to A-series bacteriophages in relation to stations successively sampled. Results found with bacteria and phages isolated from the same sample are shown in boxes. The area delimited by the broken line comprises only findings obtained with bacteria and phages found west of the Azores. The numbers of bacteria intra-sample doublets are given in parentheses. Bacteriophage doublets are not presented. Circles: clear lysis in PHCR tests; dots: turbid spots in PHCR tests."*.



(a) (b)

Figure S3: Cumulative degree frequency of the MN matrix. **a)** Cumulative frequency of the MN matrix with distinction between host and phage nodes. **b)** Cumulative frequency of the MN matrix without distinction between host and phage nodes. Both phages and hosts have a wide range of degree values, in which small degree values are more likely to occur than large degree values.

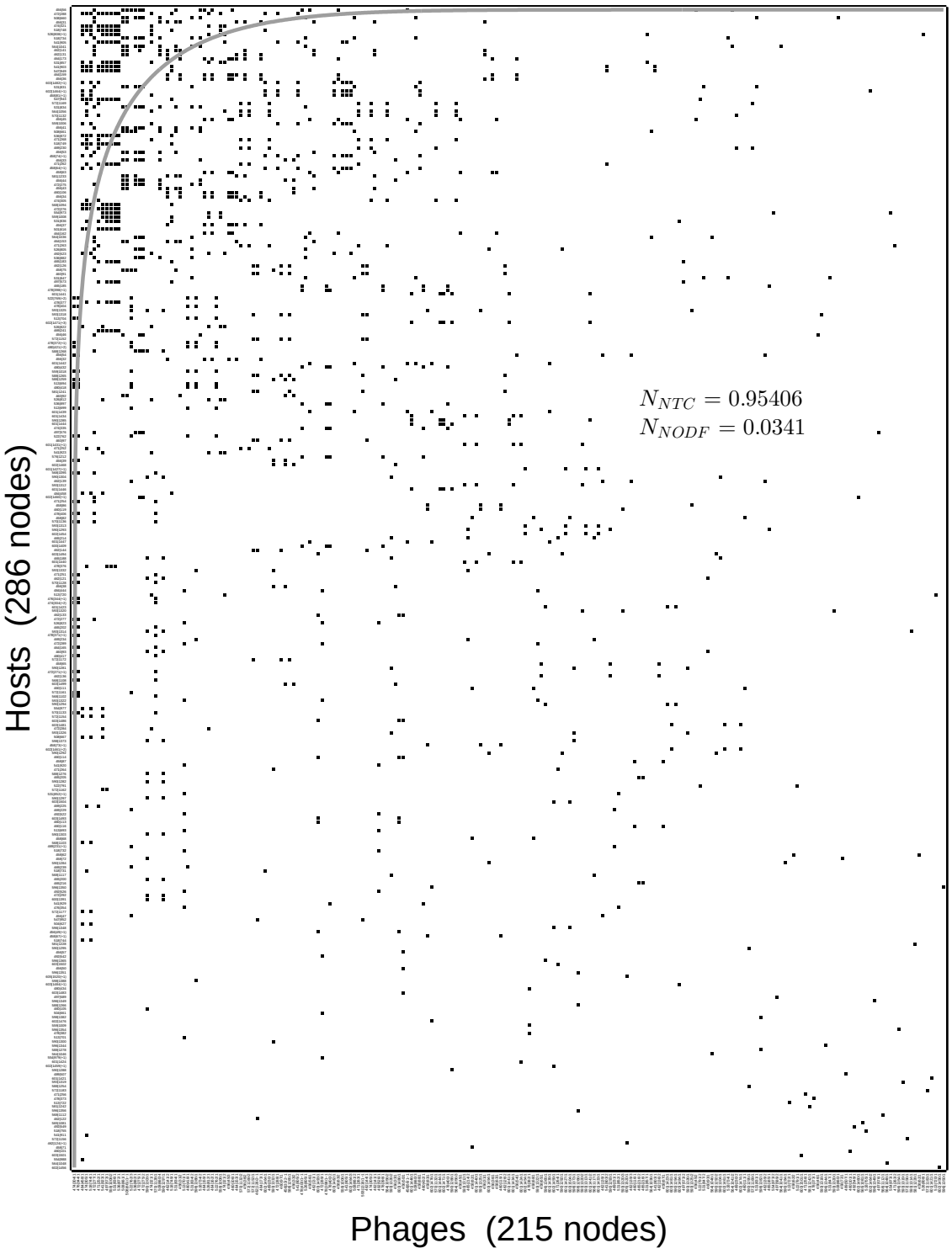


Figure S4: Arrangement of the cross-infection matrix produced with the NTC algorithm. While the nestedness value $N_{NTC} = 0.95$ has a p-value $< 10^{-5}$ in both null models, the nestedness value $N_{NODF} = 0.0341$ has a p-value $< 10^{-5}$ only in the Bernoulli random null model (see text).

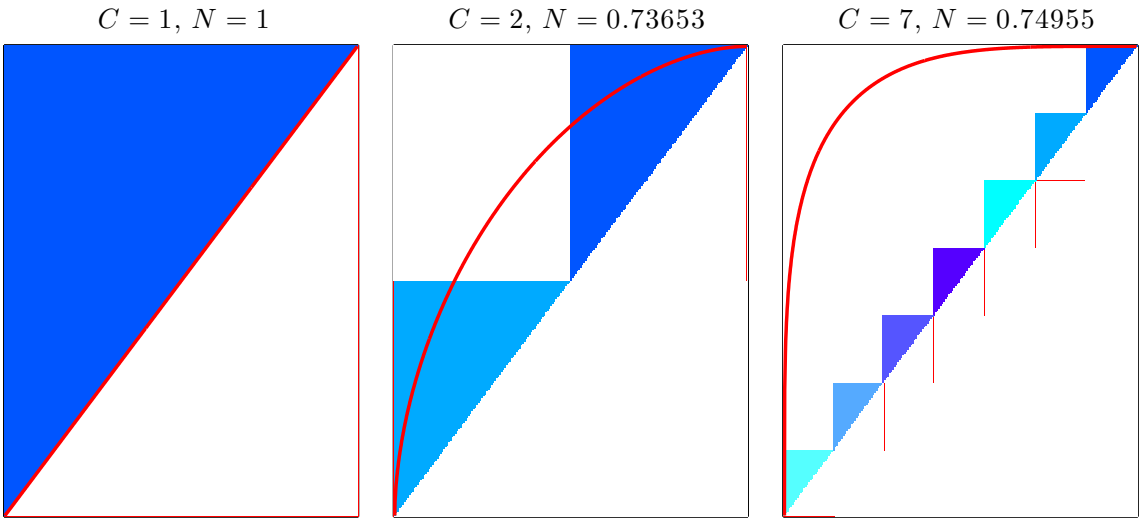


Figure S5: From left to right, correlation between nestedness and modularity in synthetic networks with $c = 1, 2, 7$ perfectly nested modules. Bold red line represents the isocline of perfect nestedness (see material and methods in the main text). Blocks with red outlines indicate modules.

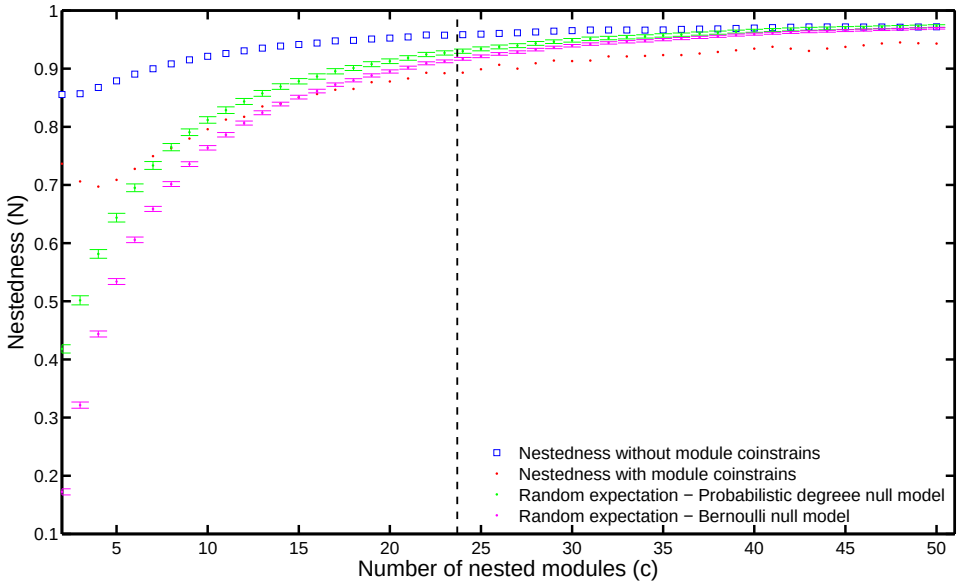


Figure S6: Comparison of constrained vs unconstrained temperature. We analyze synthetic networks with perfect nestedness with varying number of modules $2 \leq c \leq 50$ (see text). The vertical line indicate where the fill of the MN matrix coincides with that of the synthetic networks. Notice that for the corresponding fill, the nestedness of the two random expectations are larger than the value of nestedness with module constraints.

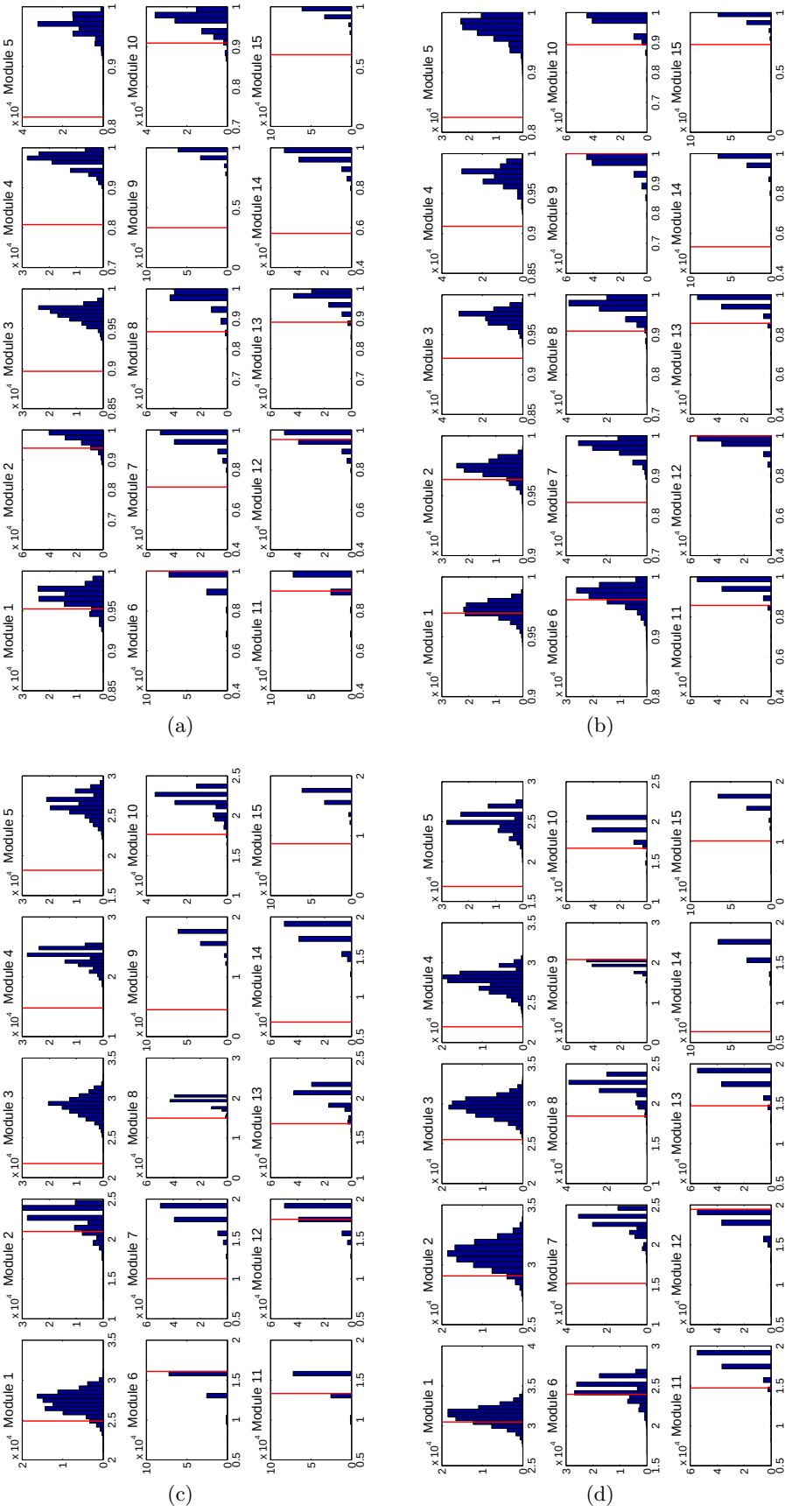


Figure S7: Distribution of geographical diversity for the 15 biggest modules. The index represent the module index. The red lines represent the real geographical diversity value of those modules. **a)** Simpson's index distribution for phages. **b)** Simpson's index distribution for hosts. **c)** Shannon's index distribution for phages. **d)** Shannon's index distribution for hosts.

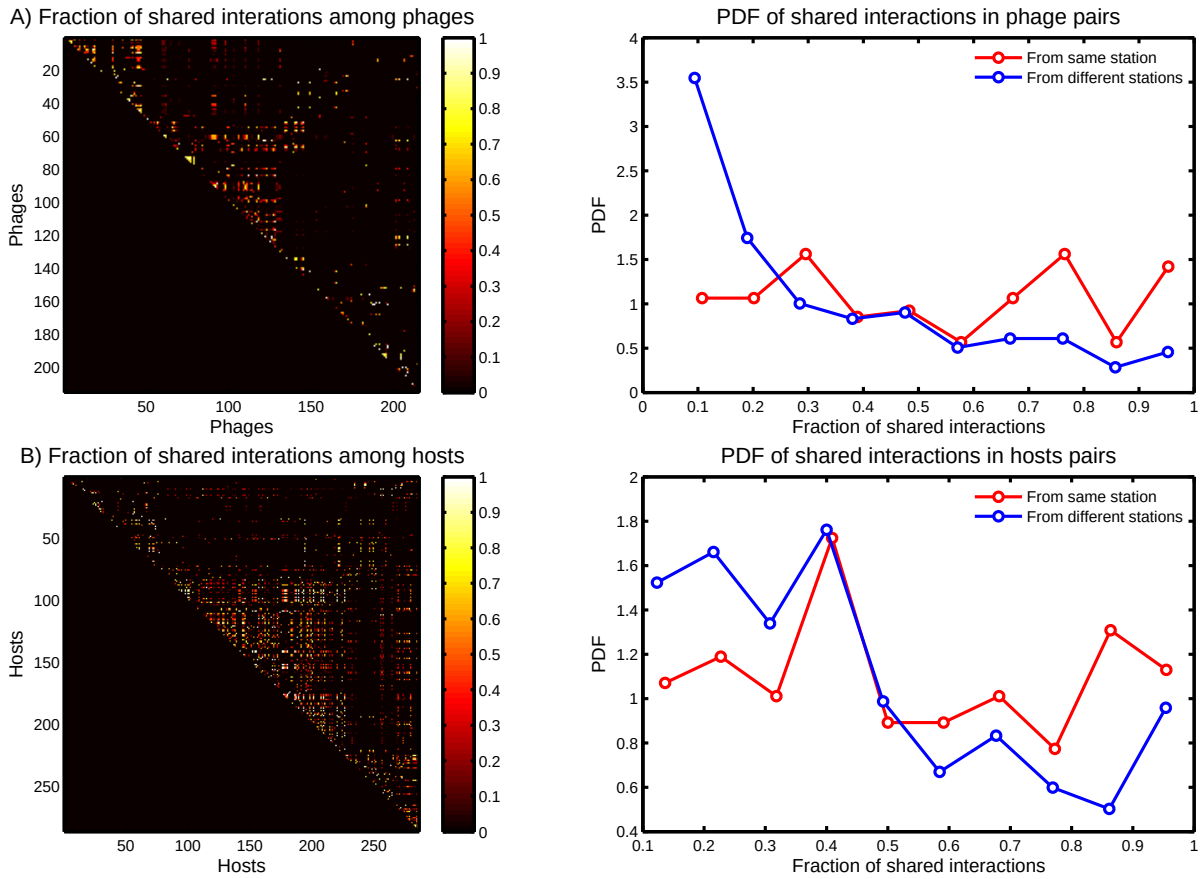


Figure S8: Fraction of shared interactions across pair of nodes. The top shows phage species and the bottom shows host species. The left shows the fraction of shared interactions across every pair of nodes. The right shows the probability density function of shared interaction between pair of nodes given that the pairs shared at least one interaction.

References

[1] M. Almeida-Neto, P. R Guimarães Jr, and T. M. Lewinsohn. On nestedness analyses: rethinking matrix temperature and anti-nestedness. *Oikos*, 116(4):716–722, 2007.

[2] M.J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):66102, 2007.

[3] C.O. Flores, J.R. Meyer, S. Valverde, L. Farr, and J.S. Weitz. Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences*, 108(28):E288–E297, 2011.

[4] X. Liu and T. Murata. Community detection in large-scale bipartite networks. *Information and Media Technologies*, 5(1):184–192, 2010.

[5] K. Moebus. A method for the detection of bacteriophages from ocean water. *Helgolander Meeresuntersuchungen*, 34(1):1–14, 1980.

[6] K. Moebus and H. Nattkemper. Bacteriophage sensitivity patterns among bacteria isolated from marine waters. *Helgolander Meeresuntersuchungen*, 34(3):375–385, 1981.

[7] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.