Quantitative Biology 2014, 2(4): 127–141
DOI 10.1007/s40484-014-0037-3

RESEARCH ARTICLE

Constructing a Boolean implication network to study the interactions between environmental factors and OTUs

Congmin Zhu[1], Rui Jiang[1,*] and Ting Chen[1,2,*]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic & Systems Biology, TNLIST/ Department of Automation, Tsinghua University, Beijing 100084, China
[2] Computational and Molecular Biology Program, University of Southern California, Loa Angles, CA 90089, USA
* Correspondence: ruijiang@tsinghua.edu.cn, tingchen@mail.tsinghua.edu.cn

Received November 29, 2014; Revised January 15, 2015; Accepted January 17, 2015

Mining relationships between microbes and the environment they live in are crucial to understand the intrinsic mechanisms that govern cycles of carbon, nitrogen and energy in a microbial community. Building upon next-generation sequencing technology, the selective capture of 16S rRNA genes has enabled the study of co-occurrence patterns of microbial species from the viewpoint of complex networks, yielding successful descriptions of phenomena exhibited in a microbial community. However, since the effects of such environmental factors as temperature or soil conditions on microbes are complex, reliance on the analysis of co-occurrence networks alone cannot elucidate such complicated effects underlying microbial communities. In this study, we apply a statistical method, which is called Boolean implications for metagenomic studies (BIMS) for extracting Boolean implications (IF-THEN relationships) to capture the effects of environmental factors on microbial species based on 16S rRNA sequencing data. We first demonstrate the power and effectiveness of BIMS through comprehensive simulation studies and then apply it to a 16S rRNA sequencing dataset of real marine microbes. Based on a total of 6,514 pairwise relationships identified at a low false discovery rate (FDR) of 0.01, we construct a Boolean implication network between operational taxonomic units (OTUs) and environmental factors. Relationships in this network are supported by literature, and, most importantly, they bring biological insights into the effects of environmental factors on microbes. We next apply BIMS to detect three-way relationships and show the possibility of using this strategy to explain more complex relationships within a microbial community.

Keywords: Boolean implication; metagenome; marine OTUs; environmental factors

INTRODUCTION

The recent advancement of next-generation sequencing technology has enabled the direct capture of all genetic materials in a microbial community [1,2]. It is therefore possible to carry out *in vitro* studies involving a small number of microbial species [3–5], as well as sequence and assemble millions of microbial genes, as exemplified by the recent increase of large-scale metagenomic studies in soil [6–8], air [9,10], marine life [11–15], human [16–18] and many others [19,20]. Direct sequencing of hypervariable regions of the 16S rRNA genes, a simple and low-cost approach, profiles the taxonomic composition of the microbial community in an environmental sample [21–24]. Through clustering analysis, 16S rRNA sequencing data can be transformed into operational taxonomic units (OTUs) of microbial species. For example, Dotur [23] and Mothur [25] employ a hierarchical clustering algorithm, which builds a hierarchical tree from the sequencing data, and then report clusters of sequences (OTUs) according to a user-defined sequence dissimilarity threshold. ESPRIT [26] accelerates hierarchical clustering by adopting *k*-mer distance to avoid unnecessary sequence comparisons and performs complete-linkage clustering. UCLUST [26] and CD-HIT [27,28] use a greedy incremental clustering algorithm for

faster clustering. CROP [22] takes a soft-clustering approach known as the Gaussian mixture model to accommodate sequencing errors and genetic variants in the sequencing data. Combining these methods has also been proposed to achieve a reasonable trade-off between efficiency and quality in the inference of OTUs and their abundance levels, resulting in such online pipelines as the Visualization and Analysis of Microbial Population Structure (VAMPS) project (http://vamps.mbl.edu/index.php). With these methods, compositions of a microbial community and abundance levels of microbial species can be inferred.

Microbes seldom live alone; instead, they live in a community composed of many species, forming complex relationships. Mathematically, such relationships are described as complex networks. For example, co-occurrence patterns of microbial species can be inferred and encoded into a co-occurrence network, where vertices are typically OTUs and edges indicate two OTUs co-occurring at high frequency across multiple samples. Co-occurrence networks have been successfully applied to the study of microbial species in soil [6–8], marine life [11–15], and, more recently, human health [16–18]. However, even though co-occurrence networks reveal direct and indirect functional associations between microbes, they cannot capture and explain the asymmetrical effects of environmental factors on microbial species inside a microbial community.

Therefore, in this study, we extend Sahoo's method [29] and propose a bioinformatics approach called Boolean implications for metagenomic studies (BIMS) to detect Boolean implications between microbial species and environmental factors using 16S rRNA sequencing datasets from a microbial community. A Boolean implication can be viewed as following a simple IF-THEN rule. For example, "IF environmental factor A is high, THEN microbial species B is abundant." Such simple rules describe intrinsic mechanisms that govern microbial communities and are essential to the understanding of relationships not only between microbes but also between microbes and the environment they live in. It should be noted that Boolean rules are general in nature in contrast to the positive and negative correlations widely used in identifying co-occurrence relationships. Under these circumstances, that A and B are positively correlated would indicate that "IF A is high, THEN B is high, and IF A is low, THEN B is low", while the Boolean implication can only state that "IF A is high, THEN B is high", and thus it describes more relationships than the former.

We demonstrated the validity of BIMS through comprehensive simulation studies, showing the reasonable high power it can achieve at a very stringent false discovery rate (FDR). To accomplish this, we applied BIMS to real marine16S rRNA sequencing datasets and detected a total of 6,514 pairwise relationships at the FDR level of 0.01. Based on these high-confidence relationships, we constructed a Boolean implication network between OTUs and environmental factors, demonstrated the consistency between relationships in this network and biological knowledge thus far gained about the effects of microbial interaction with environmental factors. Further, we demonstrated how complicated relationships inside a microbial community can be explained by using three-way Boolean implications.

RESULTS

Data sources

We extracted a dataset that contained abundance levels of 126,999 OTUs and records of 21 environmental factors across 336 environmental samples from the VAMPS project (http://vamps.mbl.edu/diversity/diversity.php). Briefly, from this resource, we selected 21 environmental factors and 27 subprojects that contained marine microbes sampled from different oceans, with each subproject containing a certain number of samples obtained in the same location, but at different times. Removing samples with insufficient measurements of environmental factors, we collected a total of 336 samples. For missing environmental factors, we further adopted a linear interpolation strategy to fill in missing values. We then performed an additional filtration step to remove microbes that appeared in less than 30% of the total samples. It should be noted that such filtration can benefit subsequent analysis by (i) significantly reducing the number of OTUs, thus leading to much lower computational burden and (ii) effectively decreasing the noise introduced by OTUs occurring infrequently, thus improving the robustness of the analysis. After the filtration, only 188 types of OTUs were retained for further analysis.

Simulation studies

We analyzed distributions of abundance levels for the 188 OTUs and 21 environmental factors across the 336 samples. The results show that the fraction of low-abundant (abundance level = 0) samples for OTUs is on average 87.71% (±0.23%), while that for environmental factors is on average 51.36% (±6.06%). This observation suggests that the abundance data of OTUs are very sparse, even though we have removed microbes that appeared in less than 30% of conditions. We further enumerated pairwise combinations of these 209 objects (188 OTUs and 21 environmental factors), and analyzed the number of possible Boolean implications based on the statistical tests described in the Method section. Results show that

the number of possible low → high and Boolean opposite implications are relatively small, while the number of possible high → low implications is relatively large (Figure S1).

We then assessed the effectiveness of our approach through simulation experiments. For each type of Boolean implication, we simulated 100 positive cases and 100 negative controls, mixed them up, and applied our method to detect the positive cases. We calculated the power of a method as the fraction of positive cases detected at the FDR 0.01. Results, as shown in Figure 1, suggest that our BIMS method using the Fisher's exact test as the first stage is effective in the detection of Boolean implications, since, in most combinations of statistical thresholds, the method achieves reasonably higher power at a low false discovery level. Overall, we successfully detected 545 true relationships out of a total of 600 simulated implications at a stringent FDR level of 0.01 when $\mu_2 = 9$, yielding a power of 90.83%. On the other hand, the method using the chi-squared-like statistic as the first stage is slightly less effective in that this strategy successfully detected 539 true relationships out of a total of 600 simulated implications at a stringent FDR level of 0.01, yielding a power of 89.83%. This observation is consistent with the common understanding that, in general, an exact test has higher power than an approximation.

To examine the contribution of the tests in the first phase, we repeated the above simulation experiments with the error checking step removed. As shown in Figure 1, the chi-squared-like statistic detected 366 true relationships out of a total of 600 simulated implications at an FDR level of 0.011 when $\mu_2 = 10$, yielding a power of 61%. To examine the contribution of the tests in the second phase, we repeated the above simulation experiments with the error checking step removed. As shown in Figure 1, just with the second phase we successfully detected only 241 true relationships out of a total of 600 simulated implications, yielding a power of 40.17%. We therefore conclude that the two-stage design is more

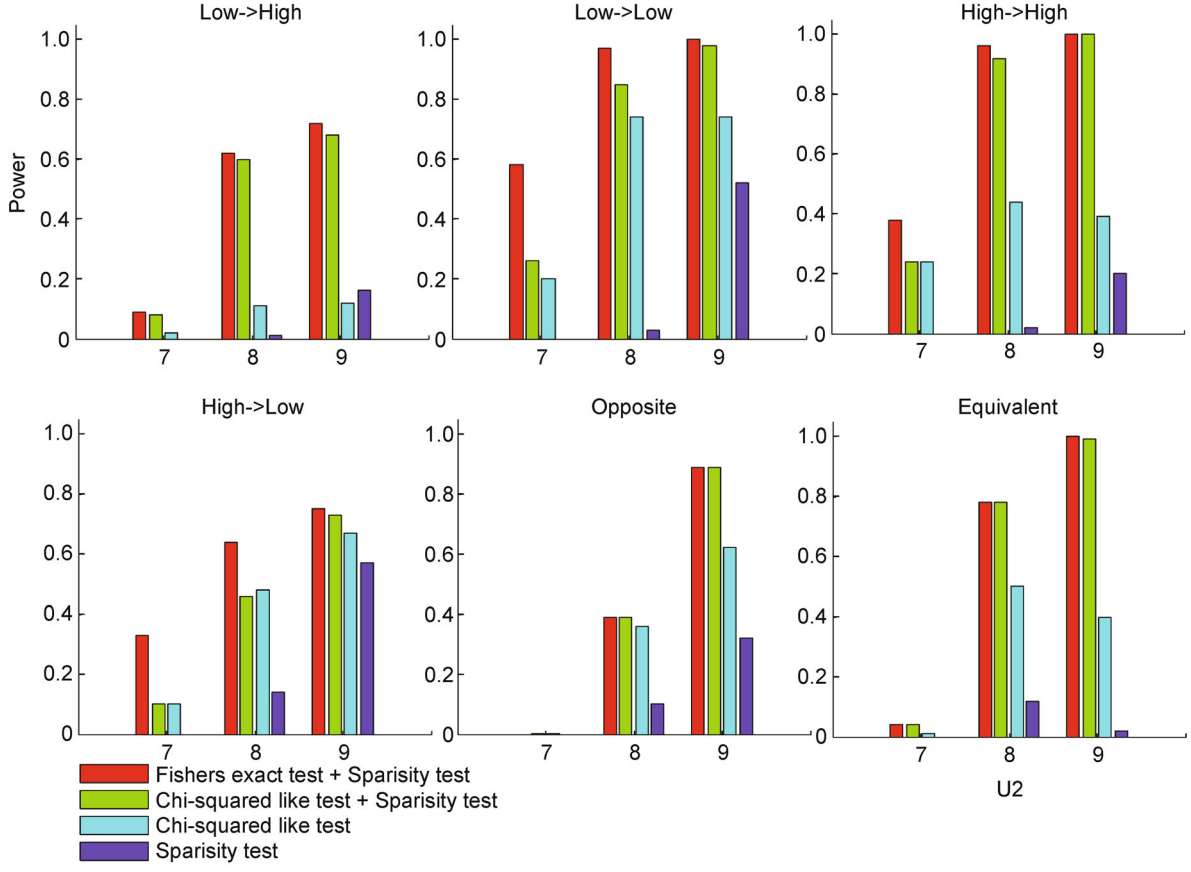

Figure 1. Comparison of the power of four methods potentially used for detecting Boolean implications changes with μ_2 on simulated data. Setting $T_1 = 0.1$ and different values of μ_2, we generate 100 positive relationships for each type of Boolean implication and respectively mix them with 100 negative relationships. We then respectively apply (i) method 1: using the Fisher's exact test as the first stage; (ii) method 2: using the chi-squared-like statistic as the first stage; (iii) method 3: using only the chi-squared-like statistic; (iv) method 4: using only the second phase for each simulated datum. Bars show the power of methods. From the comparison, we can see that the method 1 always has the highest power.

powerful than either the first or the second stage alone.

We also observe that the power of our method is apparently related to the parameters T_1 and μ_2 used to produce simulated data. With the increase of μ_2, the power of our method increases, while the increase of T_1 causes a decrease in power as shown in Figure S2. This phenomenon is consistent with our expectation that a larger μ_2 leads to a higher power by simulating high abundance values that are significantly larger than the low abundance values, while a smaller T_1 also leads to higher power by producing samples that are more consistent with the desired Boolean implication. From the simulation results (Figures 1 and S2), we also notice that the method is relatively less powerful in detecting low → high and high → low relationships. Two reasons may account for this phenomenon. First, the fractions of low-abundant samples for OTUs are in general much higher, leading to the smaller number of possible low → high and Boolean opposite implications. Second, the negative relationships, which are simulated by sampling two objects from the real data and then permuting the labels of one object have a certain degree of similarity with the high → low relationships.

To explore the influence of the sample size to the power of BIMS, we vary the sample size and identify Boolean implications at the FDR of 0.01 in each situation. The results, as summarized in Figure S3, suggest that our method tends to achieve higher power with a sample of larger size. To compare BIMS with the method based on the Pearson's correlation coefficient, we simulated 100 positive cases and 100 negative controls for each type of Boolean implication (with $T_1 = 0.1$ and $\mu_2 = 10$), mixed them up, and detected the positive cases with these two methods. Briefly, BIMS successfully detected 547 true relationships out of a total of 600 simulated implications at a stringent FDR level of 0.01, yielding a power of 91.17%. In contrast, the method based on correlation coefficient detected a total of 217 relationships at the significant level of 0.05, yielding an accuracy rate of only 36.17%. We therefore conclude that our method is capable of capturing relationships missed by the method based on correlation coefficient.

Pairwise Boolean implications detected

Focusing on the 21 environmental factors and a total of 188 OTUs that appear in at least 30% conditions, we identified a total of 6,514 Boolean implications at a low FDR (0.00945) using BIMS with the Fisher's test in the first stage. We also noticed that using the chi-squared-like strategy in the first phase resulted in the detection of only 3,910 implications, all of which were detectable by the Fisher's method, again suggesting the higher power of using the Fisher's exact test. As shown in Table 1, we found that only 7 of these Boolean implications are symmetric (Boolean equivalent or Boolean opposite). In the simulation studies, we showed that the expected number of possible Boolean implications of these two symmetric relationships is not significantly less than the others (Figure S1); therefore, the above finding suggests that microbes are more likely to have complicated asymmetric relationships instead of simple linear ones. For the four types of asymmetric relationships, we only found 16 low → high implications. We can account for this observation in one of two ways, either i) the expected number of possible implications of this type is much smaller than the others (Figure S1) or ii) low-abundant microbes do not, in general, imply high-abundant microbes. Finally, we find that the high → low relationship is dominant. This observation may also be explained in two ways: (i) the expected number of possible implications of this type is itself much larger than the others (Figure S1) and (ii) a high-abundant species of microbe does imply a low-abundant one because most microbes will compete for resources. Based on the converse negative proposition, meaning IF "$A_{low} \rightarrow B_{low}$" THEN "$B_{high} \rightarrow A_{high}$", the number of low → low is equal to the number of high → high (Table 1).

Table 1. Number of different pairwise Boolean implication relationships of marine microbe species and environmental factors.

Sparse quadrant		Relationship	Number
Only one sparse quadrant	low-low	low → high	16
	low-high	low → low	340
	high-low	high → high	340
	high-high	high → low	5804
Two diagonal sparse quadrants	low-low high-high	Boolean opposite	0
	low-high high-low	Boolean equivalent	14
Total			6514

Focusing on the 21 environmental factors and a total of 188 OTUs that appear in at least 30% conditions, we identified a total of 6,514 Boolean implications at a low FDR(0.00945) using BIMS with the Fisher's test in the first stage.

We further varied the threshold in the selection of OTUs from 10% to 50% and identified Boolean implications at the FDR of 0.01 in each situation. As summarized in Figure 2, a looser value of the threshold in general results in the identification of more implications, while a more stringent value usually leads to the detection of fewer implications. However, in all situations, the fractions of symmetric relationships were less than 3%, again suggesting the rarity of the linear implications. Consistent with the previous analysis, Boolean implications of the high → low type were most common, while low → high and Boolean opposite relationships were rare. We conjectured that the sparse nature of OTU abundance levels could account for the rarity of the two types of relationships (low → high and Boolean opposite) that require fewer points in the low-low quadrant. We also noticed that the number of the low → low relationships was almost the same as that of the high → high relationships and that both types of relationships were large, possibly because many OTUs belong to the same colony.

Pairwise Boolean implication network

We then constructed a Boolean implication network among the 209 objects with BIMS relying on the Fisher's exact test as the first stage. The resulted network consists of 181 nodes (162 OTUs and 19 environmental factors) and 6,514 edges (Boolean implications) (Figure S4). Among these edges, 4,934 are between OTUs, 38 between environmental factors and 1,542 connecting OTUs and environmental factors (Figure S5A). Furthermore, on average, each OTU connects with 26.89 OTUs and 4.19 environmental factors, while each environmental factor connects to 39.79 OTUs and 2 environmental factors. This network is fully connected. Without considering environmental factors, the subnetwork consisting of only OTUs can be separated into a giant connected component (175 nodes and 4,934 edges) and 8 singletons.

Observing the network as a whole, it is interesting to note that most Boolean relationships existing among microbes, as shown in Figure S5A, reflect the close and wide connections of OTUs. This is consistent with a previous study [12] that demonstrated the dominance of relationships between microbes rather than those between microbes and environmental factors. One possible reason for this phenomenon is that some environmental factors which may influence microbial abundance are not measured at all. It is also possible that the relatively stable ocean, especially deep sea, environment, compared to changes in environmental factors, nutrients or interactions with other microbes, most likely drives changes of composition and abundance in the marine microbial community [12].

We analyzed network topology characteristics for the subnetwork of OTUs using the Network Analyser plugin in Cytoscape [26]. As shown in Figure S5B, the clustering coefficient of this network is 0.097, much larger than that

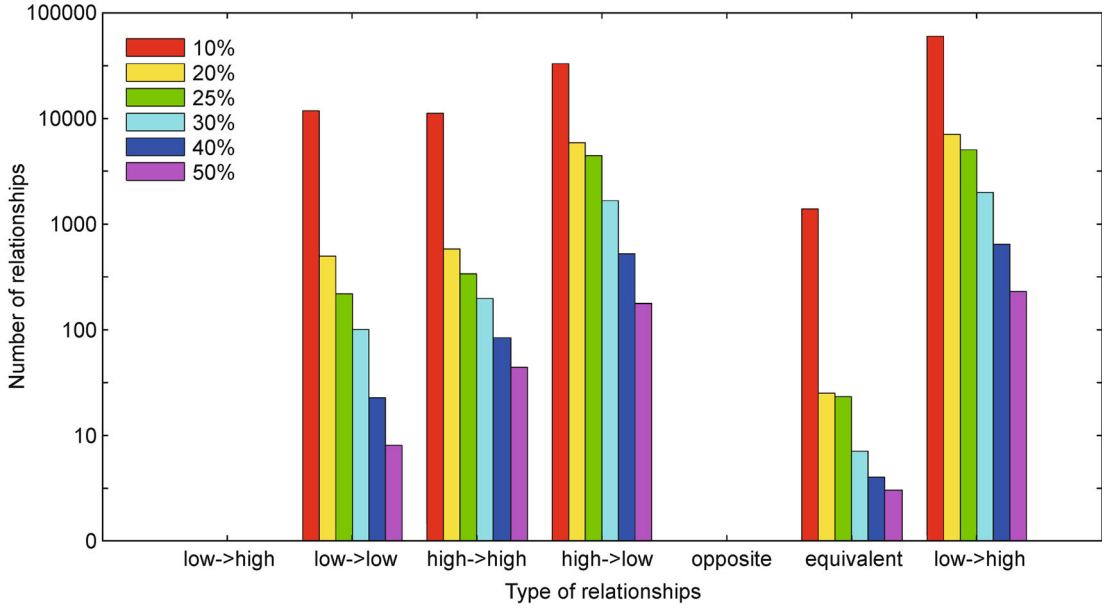

Figure 2. Number of different types of Boolean relationships found in networks. We chose different numbers of OTUs based on occurrences in more than different percent of all samples for constructing Boolean implication network. The number of Boolean relationships found in each network are summarized. In all cases, Boolean implications of the high → low type are most common, while the low → high and Boolean opposite relationships are rare.

of a random network of the same scale (0.036), while the characteristic path length for this network is 2.028, similar to that of a random network (2.242), suggesting that this OTU network may have a "small-world" property [30]. Furthermore, the distribution of the shortest path length (Figure S5C) shows that the shortest path length for this network is between 1 and 3, indicating that most microbes in the resulting network are closely linked by a high level of dependence, a phenomenon consistent with the conclusion from clustering coefficient analysis. The small-world property in ocean microbes was found previously [12], suggesting that most microbes are gathered in a certain community and that very few of them are independent. From Figure S5B, we observe that some microbes with high clustering coefficient (or "hub") might be analogous to microbial 'keystone species' which play central roles in a microbial community. It is also suggested that the small-world pattern makes a network more robust to changes and perturbations, essentially because the network would change dramatically if highly connected nodes were lost [31].

We further visualized a subnetwork and gave several relationships found as examples to show the power of BIMS. As shown in Figure 3A, this network consists of 60 nodes and 228 edges. Among the nodes, 42 are OTUs and 18 are environmental factors. Among the edges, 67 are between OTUs, 16 between environmental factors, and 145 connect OTUs and environmental factors. Particularly, some environmental factors have high degrees (Figure 4) (e.g., depth, degree = 20; chlorophyll, degree = 14; silicate, degree = 13; temperature, degree = 10), indicating their important effect on the abundance of marine microbes. We then extracted these environmental factors and their neighbors to obtain subnetworks that describe Boolean implications between important environmental factors and OTUs (Figure 3B). This network shows that these important environmental factors (e.g., temperature, depth, chlorophyll) play a central role in their small-world pattern, providing awareness of the conditions that either favor or disfavor particular OTUs and may thus be of great significance to our understanding of marine environments.

As shown in Figure 3B, the relationship "depth high → chlorophyll low" is consistent with the conclusion of a previous study [32]. As the sampling depth increases, the intensity of light decreases, leading to weaker photosynthesis. As a consequence, chlorophyll content becomes low. Another high → low relationship that occurs between temperature and an OTU named Alphaproteobacteria_03_29 is illustrated in Figure 5A. Since the sample points in the high-high quadrant are very sparse, the abundance of Alphaproteobacteria_03_29 is low if the temperature is high. The relationship "depth low → salinity low" is also consistent with a previous work [32], which reports that the shallow coastal waters are diluted by the impact of freshwater inflows. The low → low implication also exists between salinity and Gammaproteobacteria_03_58, and the scatter plot in Figure 5B verified this relationship. A high → high relationship between Gammaproteobacteria_03_46 and Gammaproteobacteria_03_75 is also found and verified in Figure5C, indicating that these two OTUs coexist in one community as gammaproteobacteria. An equal relationship is found between the environmental factors depth_start and depth_end, as shown in Figure 5D, which results from the fact that the sampling method restricts the place of sampling within a range of one meter.

Three-way Boolean implications detected

To reduce the computational burden, we detected three-way Boolean implications using the chi-squared-like test as the first stage. To this end, we first assessed the capability of BIMS by performing a simulation study similar to the pairwise case. Briefly, 10 types of three-way Boolean implications exist that corresponded to 10 different sparse situations of eight quadrants among three simulated variables (Figure S6). We then generated 1,000 true Boolean implication relationships (100 for each type) and 1,000 false relationships in a manner similar to that of the pairwise case and applied our method to detect Boolean relationships. The results showed that 864 out of the 1,000 true relationships were successfully found at an FDR of 0.00583, yielding a power of 86.4%, thus supporting the validity of this method in the detection of three-way Boolean implications.

We then assessed various choices of thresholds σ and ρ for the real data in Figure S7, which shows the empirical relationship between these two parameters and the FDR value. The figure suggests that σ between 2.0 and 3.0 and $\rho = 0.1$ can typically control the FDR at a low level (<0.01). In order to examine the robustness of the threshold σ to the final results, we calculated FDRs under different values of σ between 2 and 3. Results, as shown in Figure 6, suggest the robustness of σ to the number of Boolean implications identified. Therefore, we selected $\sigma = 2.2$ and $\rho = 0.1$ as the thresholds for S statistic and R statistic, respectively, in our analysis.

We then applied BIMS to detect three-way Boolean implications in the real data. To reduce the computational burden, we only selected 13 environmental factors and 24 OTUs that appeared in more than 60% conditions. At a stringent threshold value (FDR = 0.0063), we detected a total of 2,186 relationships, as summarized in Table 2. For example, we found a Boolean implication "depth low & temperature low → Betaproteobacteria_03_1 high", suggesting that Betaproteobacteria_03_1 is more abundant in the shallow waters where the temperature is low. The

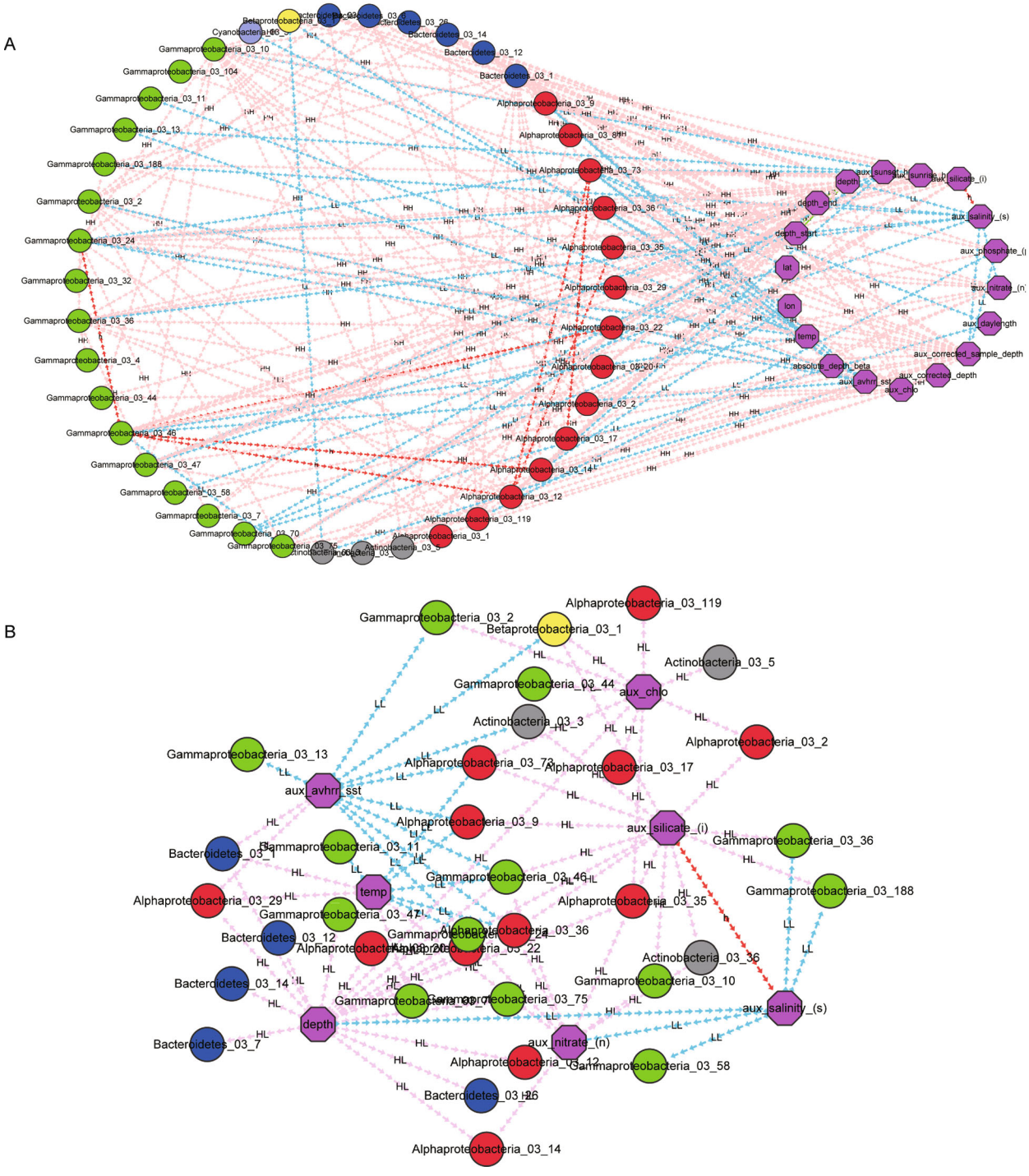

Figure 3. Results of pairwise Boolean implication network. Nodes are either OTUs or EFs, and directed edges with different colors represent different types of Boolean implication relationships. Purple nodes represent environmental factors, and the others in different colors represent different phyla of OTUs. (A) A subnetwork in which the left part describes the relationship network inside marine OTUs and the right part describes the relationship network inside marine EFs. The other edges represent the relationships between OTUs and EFs. (B) Subnetwork including several central environmental factors.

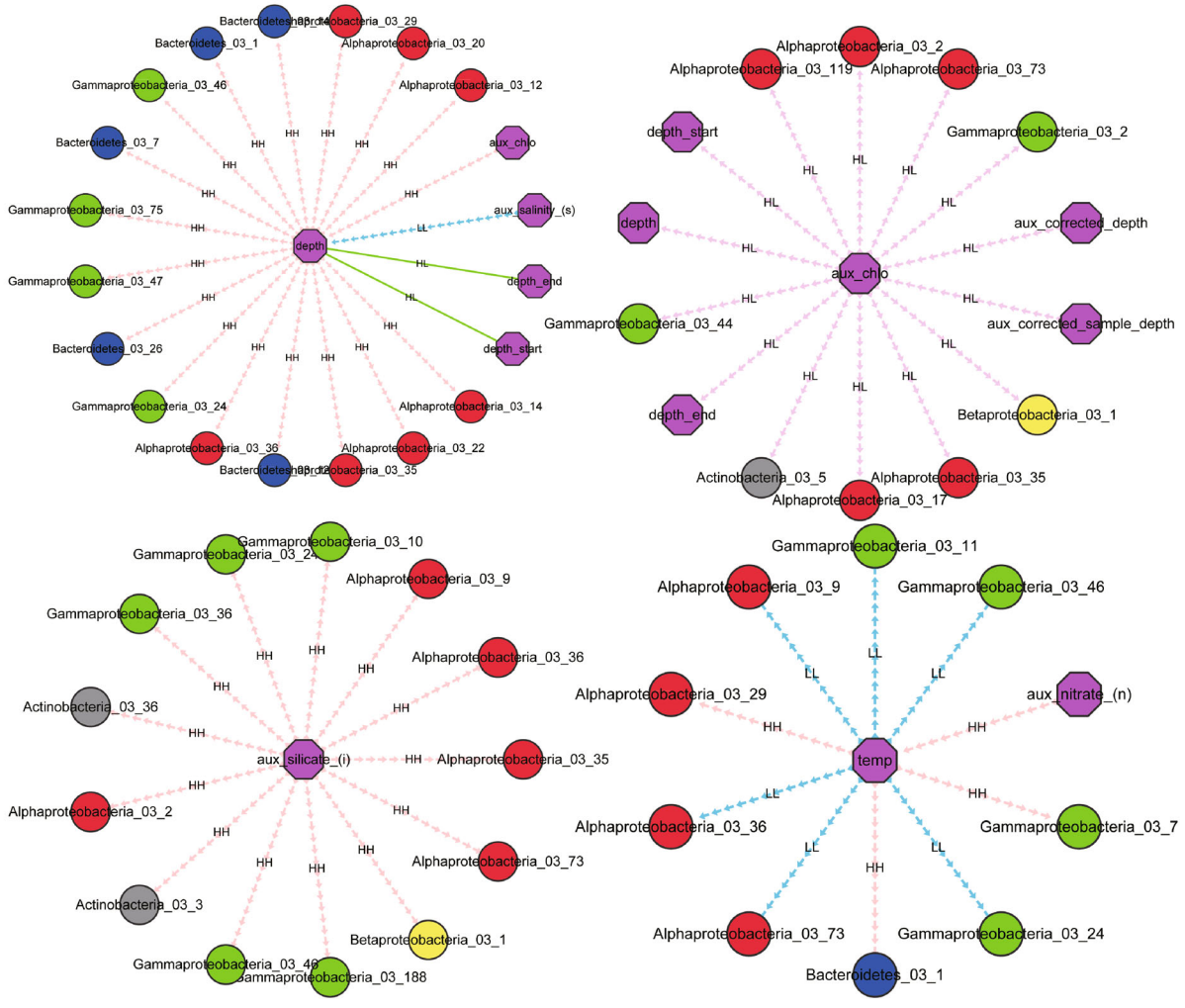

Figure 4. Environmental factors at the centers of four subnetworks. The four environmental factors with high degree, depth (degree = 20), chlorophyll (degree = 14), silicate (degree = 13) and temperature (degree = 10), indicate their important role on the abundance of marine microbes. From these four subnetworks, we can observe OTUs favored or disfavored by the four environmental factors.

implication "temperature high & Bacteroidetes_03_1 high → Bacteroidetes_03_14 high" indicates that Bacteroidetes_03_14 and Bacteroidetes_03_1 may belong to the same species that prefer a warm environment. An implication "chlorophyll high & nitrate high → Gammaproteobacteria_03_32 low" implies that Gammaproteobacteria_03_32 might be a photosynthetic bacterium requiring nitrate to perform photosynthesis. Since little prior knowledge exists about three-way relationships among OTUs or between OTUs and EFs, finding support in the literature for these Boolean implications is impossible. We therefore manually checked several relationships identified by the three-way scatter plot (Figure S8). As shown in Figure S8A, in the plot for the implication of "depth high & latitude low → Alphaproteobacteria_03_1 low", the points in the high-low-high quadrant are fewer than found in other quadrants, leading to the intuitive conclusion that the abundance of Alphaproteobacteria_03_1 is relatively low when the depth is high and the latitude is low. In Figure S8B, we also found that points belonging to the high-high-high quadrant are significantly reduced in comparison to other quadrants in the plot for the implication of "depth high & temperature high → Alphaproteobacteria_03_2 low", Therefore, when the depth is high and the temperature is relatively high, we can conjecture that the abundance of Alphaproteobacteria_03_2 is also relatively high.

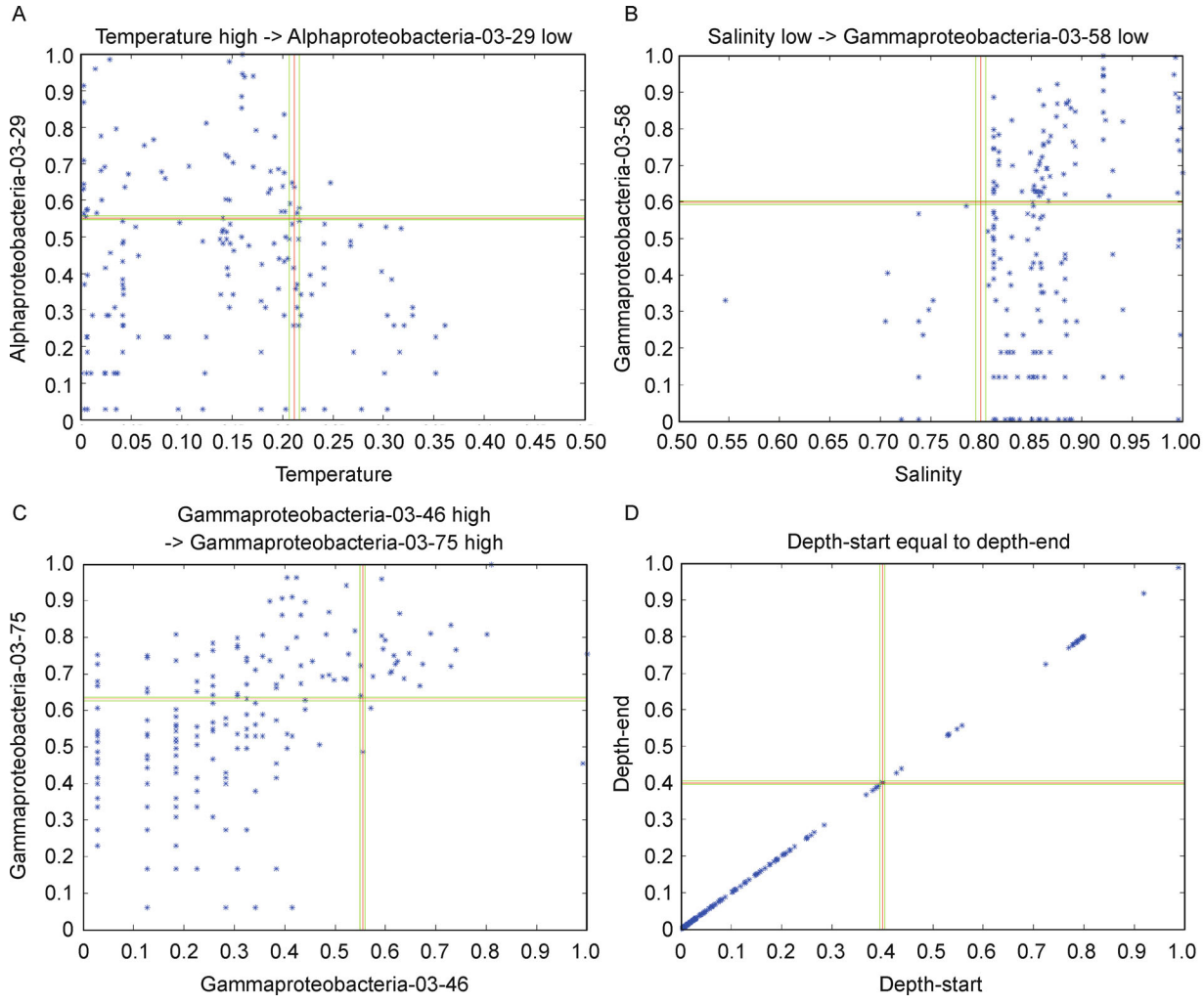

Figure 5. Verification of relationships found by BIMS. A visual examination of the scatter plots is a straightforward way to check the quality of the results. (A) Temperature high → Alphaproteobacteria_03_29 low. The points in the high-high quadrant are sparse, so intuitively we obtain the high → low Boolean implication relationship between *X* variable and *Y* variable. (B) Salinity low → Gammaproteobacteria_03_58 low. The points in the low-high quadrant are sparse, so intuitively we can get the low → low Boolean implication relationship between *X* variable and *Y* variable. (C) Gammaproteobacteria_03_46 high → Gammaproteobacteria_03_75 high. The points in the high-low quadrant are sparse, so intuitively we can get the high → high Boolean implication relationship between *X* and *Y* variable. (D) Depth_start equal to depth_end. The points in the low-low and high-high quadrants are both sparse, so intuitively we can get the Boolean equivalent relationship between *X* and *Y* variables.

Discussion

In this paper, we proposed BIMS to detect Boolean implications between microbial species and environmental factors from multiple 16S rRNA sequencing datasets. We demonstrated the validity of BIMS by simulation studies and then applied it to the real datasets. Based on identified pairwise relationships, we constructed a Boolean implication network between OTUs and environmental factors. Relationships in this network are either supported by literature or provide biological insights into the understanding of interactive effects of microbes and environmental factors. Our BIMS method should shed more light on capturing more complicated relationships than simple liner ones. We further extended our study to three-way relationships and showed the possibility of relying on such combinatorial Boolean implications to explain complicated relationships among multiple factors inside a microbial community.

To explore relationships between microbial species, existing methods typically rely on pairwise correlation coefficients derived from abundance levels. This strategy,

though suitable for explaining such relationships as co-occurrence or co-abundance, cannot explain why a microbial species is abundant or absent in a community and how microbes interact with each other and environmental factors. The opinion of Boolean implication, together with BIMS for detecting such relationships, is most suitable for exploring such complicated logics and hence provides practical reasoning for relationships in a microbial community. Moreover, the network view of all detected Boolean implications further suggest complicated regulation relationships. In this sense, BIMS offers a new platform for mining metagenomics data.

Compared to the raw method of Sahoo's [29] to detect Boolean implication, in BIMS the process of detecting Boolean implication relationships, which relies on exact statistical tests, has slightly higher power, even though it results in slower running time. Therefore, how to improve the power of this process while keeping the merit of low computational burden will be a direction for our future work. In addition, the discretization of continuous abundance levels to binary factors has the benefit of filtering out noise, but may also result in loss of power. It is therefore desirable to develop a model from the viewpoint of regression to handle continuous abundance levels directly. Finally, although we are currently focused on relationships between environmental factors and OTUs, we can extend BIMS to explore relationships between environmental factors and microbial genes assembled from metagenomics data. One caveat to this ambition involves the vast number of microbial genes, which would place an enormous computation burden on the process of detecting Boolean implication with Fisher's exact test.

METHODS

Workflow of BIMS

The study is predicated on the idea that the relationship between an OTU and an environmental factor (EF) can be described using a Boolean implication, and that such Boolean relationship can also be used to characterize the effect of an environmental factor on a microbial species. A total of six types of pairwise Boolean implications between two objects can be identified [29] (Figure S9), and can be described by i) $A_{low} \rightarrow B_{high}$, ii) $A_{low} \rightarrow B_{low}$, iii) $A_{high} \rightarrow B_{high}$, iv) $A_{high} \rightarrow B_{low}$, v) $A_{high} \rightarrow B_{high}$ and $A_{low} \rightarrow B_{low}$, and vi) $A_{low} \rightarrow B_{high}$ and $A_{high} \rightarrow B_{low}$. It should be noted that the last two rules correspond to the positive and negative correlations respectively. Here the relationship v) $A_{high} \rightarrow B_{high}$ and $A_{low} \rightarrow B_{low}$ is said to be 'Boolean equivalent' and vi) $A_{low} \rightarrow B_{high}$ and $A_{high} \rightarrow B_{low}$ is said to be 'Boolean opposite'. We assume that such Boolean relationships, while having exceptions under some situations, should exhibit sufficient stability across a large number of conditions and, hence, be detectable through measured numerical values of these objects.

Therefore, we propose a bioinformatics approach (BIMS), as illustrated in Figure 7, to detect Boolean implications from 16S rRNA sequencing data of a marine microbial community. Two inputs of BIMS include (i)

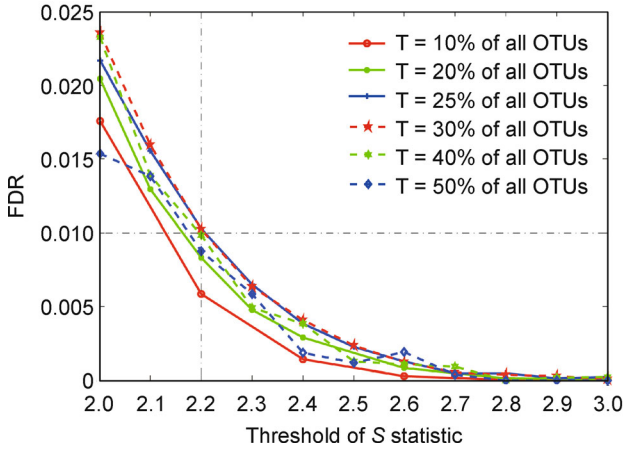

Figure 6. Robustness of the threshold for *S* statistic chosen by FDR. To evaluate the significance of the relationships found by Boolean implication based on our collected data, we computed a FDR for each network constructed with a certain threshold σ. By changing threshold σ between 2 and 3, we obtain different FDR values and then choose σ with an acceptable FDR. The experimentally calculated FDR decreases with the increase of the statistical cutoff threshold. When the threshold of *S* statistic increases by 0.1 from 2 to 3, the calculated FDR decreases. For all networks which consist of different fractions of samples, when the threshold grows to 2.2, the FDR decreases to 0.01. Hence, 2.2 can be chosen as the cut-off threshold of *S* statistic.

Table 2. Number of different Boolean implication relationships in three-way network of marine microbe species and environmental factors.

Sparse quadrant		Relationship	Number
Only one sparse quadrant	low-low-low	x0&y0 → z1	9
	low-low-high	x0&y0 → z0	178
	low-high-low	x0&y1 → z1	23
	low-high-high	x0&y1 → z0	631
	high-low-low	x1&y0 → z1	4
	high-low-high	x1&y0 → z0	256
	high-high-low	x1&y1 → z1	65
	high-high-high	x1&y1 → z0	1020
Two diagonal sparse quadrants	low-low-low high-high-high	x & y opposite to z	0
	low-low-high high-high-low	x & y equal to z	0
Total			2186

To reduce the computational burden, we only selected 13 environmental factors and 24 OTUs that appeared in more than 60% conditions. At a stringent threshold value (FDR = 0.0063), we detected a total of 2,186 relationships.

multiple 16S rRNA sequencing datasets, and (ii) measurements of environmental factors. The output of BIMS is a Boolean implication network for the effects of environmental factors on the OTUs. In this network, nodes are either OTUs or environmental factors, and directed edges represent Boolean implications between the nodes. Furthermore, each edge is labelled with one of the six types of Boolean implications.

We first pool multiple metagenomic datasets (samples) together and cluster sequencing reads using such clustering tools such as Dotur [23], Mothur [25], SLP (single linkage preclustering) [24], Uclust (heuristic clustering) [33], or CROP (Bayesian Clustering) [22] of multiple metagenomic datasets to infer OTUs. Next, we estimate raw abundance levels of the OTUs in a sample by counting the number of reads for each OTU in the sample. In the case that the abundance levels of OTUs have already been inferred, one can simply skip this step. After that, we perform a filtration step by removing OTUs that are absent (0 read count) from a predefined fraction (default set to 30% in our study) of all samples and normalize the abundance levels of OTUs for each sample by dividing their read counts over the total number of reads of all the remaining OTUs in the same sample. Then, we infer pairwise Boolean implications for the remaining OTUs and environmental factors. To accomplish this, both abundance levels of OTUs and numeric measurements of environmental factors are discretized into categorical values of 1 and 0, followed by adopting a two-phase hypothesis testing procedure based on the contingency table of two objects. Finally, we combine all the inferred pairwise Boolean implications into a Boolean implication network.

Data discretization and inference of pairwise Boolean implications

We use StepMiner, which is used to extract binary signals from microarray time-course data [34], to determine threshold values for discretizing abundance levels of OTUs, as well as the values of environmental factors, into binary ("low" versus "high") values. Briefly, we sort abundance levels of an OTU across multiple datasets in non-decreasing order and then fit an increasing step function to the ordered data in order to minimize the difference between the fitted and original values. StepMiner evaluates every possible step position using linear regression in order to identify the optimal position. At each position, it calculates an F statistic, which equals to dividing the regression mean square with the error mean square fitted values. Since this F statistic follows an F-distribution, a corresponding p-value can then be calculated as the tail probability of the realized value in this distribution. The step position having the minimum p-value is chosen as the threshold for discretization. A value above or below the threshold is discretized into 1 (high) or 0 (low), respectively.

We detect the Boolean implication between two objects using a two-phase hypothesis testing procedure [29] (Figure S10). This is achieved by constructing a contingency table to represent occurrence frequencies of combinations of the discretized abundance levels for the

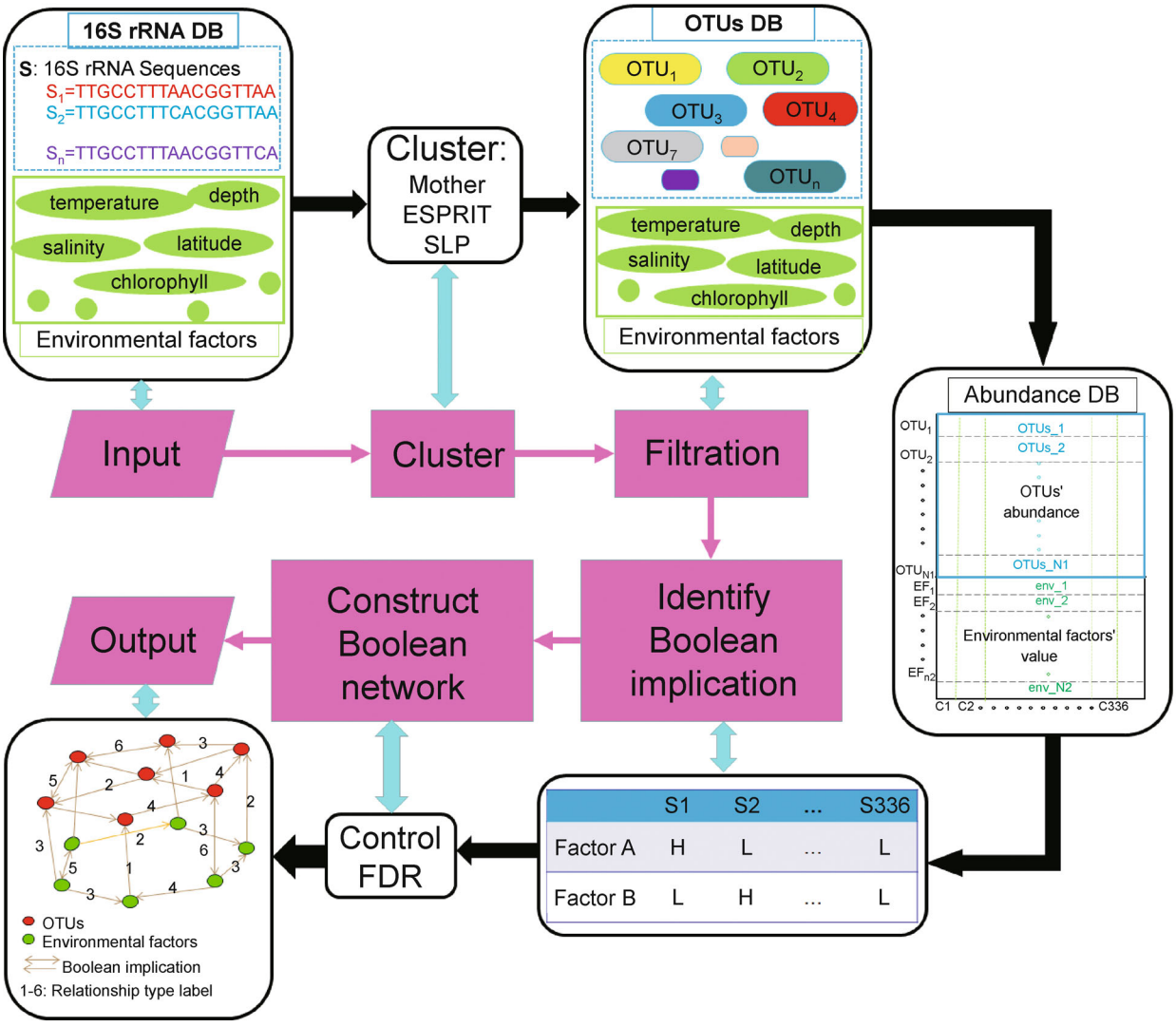

Figure 7. Workflow of BIMS for constructing Boolean implication networks from 16S rRNA sequence and metadata. Inputs of BIMS include two parts: raw sequence reads of 16S rRNA genes across different samples (blue box) and numeric measurements of EFs (green box). Output is a Boolean implication network. Each node represents OTUs or EFs, and each edge is labeled by an integer between 1 and 6, indicating one of the six types of Boolean implications. Since we collected OTUs abundance data from the VAMPs project, the step in the blue-dotted box does not pertain to our work.

two objects. Then, in the first step of the procedure, we apply a series of Fisher's exact tests to detect whether a potential Boolean implication exists between the two objects. If the sample size is large, we use a chi-squared-like statistic in this step, according to the literature [29,35]. In the second step, we employ a sparcity test to determine whether an implication does, indeed, exist.

For example, when testing $A_{low} \rightarrow B_{high}$, we first obtain a contingency table by counting frequencies for each of the four combinations of abundance values, say low-low, low-high, high-low and high-high, denoted by a_{00}, a_{01}, a_{10}, a_{11}, respectively. Then, we test which of these combinations occurs more rarely than the others using the Fisher's exact test. Specifically, we first compute the proportions as $p_{00} = a_{00}/(a_{00} + a_{10})$ and $p_{01} = a_{01}/(a_{01} + a_{11})$, and then define the null hypothesis in $H_0 : p_{00} = p_{01}$, and the alternative hypothesis as $H_1 : p_{00} < p_{01}$, to test whether a_{00} is truly sparse. Repeating this procedure for each of the four situations, we are able to know which combinations are rare and further claim a potential low $\rightarrow$ high relationship if the low-low combination is the only rare case. However, if the sample size is large, the Fisher's exact test may not be practical based on the increased computational burden needed to compute hypergeometric distributions. We therefore follow the literature to perform a test based on a chi-squared-like statistic if every cell in

the contingency table is not very rare (e.g., $\geqslant 5$). In detail, we calculate a test statistic as

$$S_{00} = \frac{(E_{00} - O_{00})}{\sqrt{E_{00}}},$$

where O_{00} and E_{00} denote the observed and expected numbers of occurrence for the low-low combination, respectively, and are calculated as $O_{00} = a_{00}$ and $E_{00} = a_{0\bullet} \times a_{\bullet 0}/a_{\bullet\bullet}$ with $a_{0\bullet} = a_{00} + a_{01}$, $a_{\bullet 0} = a_{00} + a_{10}$ and $a_{\bullet\bullet} = a_{00} + a_{01} + a_{10} + a_{11}$. Since this test statistic indicates the rarity of samples in the low-low combination and can therefore be used with an appropriately selected threshold σ (e.g., between 2.0 and 3.0) to determine the existence of such combination.

In the second phase, we consider erroneous points observed in the sparse quadrant and calculate the maximum likelihood estimate (MLE) of the error rate as:

$$R_{00} = \frac{1}{2}\left(\frac{a_{00}}{a_{0\bullet}} + \frac{a_{00}}{a_{\bullet 0}}\right),$$

where $a_{00}/a_{0\bullet}$ and $a_{00}/a_{\bullet 0}$ are the MLE for the binomial distribution in two combinations, respectively, and R_{00} is the average of the two MLEs. Since R_{00} should be low if the low-low quadrant is sparse, making it possible to reuse this statistic with a predefined threshold ρ (e.g., 0.1) to further filter out false positive cases that pass the test of the first phase.

Construction of a Boolean implication network

We evaluate the significance of each Boolean implication using the following three threshold values, the p-value in Fisher's exact test, σ in the chi-squared-like test, and ρ in the sparsity test. We combine all discovered Boolean implications into a network. To this end, we compute a FDR to indicate the proportion of Boolean implications that may be discovered by chance. In detail, we first permute abundance values of OTUs and numeric measurements of environmental factors separately and then apply the statistical tests to identify Boolean implications in the permuted data. Repeating these two steps a number of n (default $n = 50$) times, we compute the FDR as the average number of Boolean relationships discovered in the randomized datasets divided by the number of Boolean implications identified in the original data. With properly selected threshold values for a small FDR (default < 0.01), we detect pairwise Boolean implications.

Extension to three-way Boolean implications

It is suggested that very rarely do microbes live in a single species community [1]. Existing studies have shown that combinations of environmental variables are more predictive for microbial changes over time when compared with individual environmental variables [36,37]. Therefore, we further study the relationships among multiple objects (OTUs and EFs) by extending the inference of pairwise Boolean implications to three-way relationships such as "A_{low} and $B_{high} \rightarrow C_{high}$" (IF environmental factor A is low and environmental factor B is low, THEN the abundance of microbial species Z is high).

Given the discretized binary abundance levels of each object, we obtain eight combinations of abundance values for every three objects, e.g., low-low-low, low-low-high, low-high-low, low-high-high, high-low-low, high-low-high, high-high-low, high-high-high. We denote the numbers of samples that belong to the eight combinations as a_{000}, a_{001}, a_{010}, a_{011}, a_{100}, a_{101}, a_{110}, a_{111}, respectively (Figure S11). Following the same procedure for detecting the pairwise Boolean relationships, we again adopt the two-phase hypothesis testing procedure to detect the Boolean implication among three objects. Specifically, when one or two or three quadrants are sparse after a statistical test and enough high and low abundance values are available for each object for a threshold, a Boolean implication potentially exists. To this end, we first determine the corresponding relationships, as shown in Table 2, between the sparse situations of eight quadrants and the types of Boolean implication and then examine which quadrant is sparse (Figure S6).

For example, to test if the number of observed data in the low-low-low combination is sparse or not, we will use either the Fisher's exact test or the chi-squared-like method in the first stage. Particularly, using the latter, we calculate a test statistic as

$$S_{000} = \frac{(E_{000} - O_{000})}{\sqrt{E_{000}}},$$

where O_{000} and E_{000} denote the observed and expected numbers of occurrence for the low-low-low combination, respectively, and can be calculated as $O_{000} = a_{000}$ and $E_{000} = a_{0\bullet\bullet} \times a_{\bullet 0\bullet} \times a_{\bullet\bullet 0}/a_{\bullet\bullet\bullet}^2$ with $a_{0\bullet\bullet} = a_{000} + a_{001} + a_{010} + a_{011}$, $a_{\bullet 0\bullet} = a_{000} + a_{001} + a_{100} + a_{101}$, $a_{\bullet\bullet 0} = a_{000} + a_{010} + a_{100} + a_{110}$ and $a_{\bullet\bullet\bullet} = a_{000} + a_{001} + a_{010} + a_{011} + a_{100} + a_{110} + a_{111}$. This test statistic indicates the rarity of samples in the low-low-low combination and can be used with threshold σ to determine the significance of such combination. Repeating this procedure for each of the eight combinations, we are able to determine the potential Boolean relationship. In the second phase, the observed values in the sparse quadrant are considered erroneous points, and the MLE of the error rate is computed as:

$$R_{000} = \frac{1}{3}\left(\frac{a_{000}}{a_{0\bullet\bullet}} + \frac{a_{000}}{a_{\bullet 0\bullet}} + \frac{a_{000}}{a_{\bullet\bullet 0}}\right),$$

which is the average of three MLEs for three dimensions to measure the error rate of the observation for the low-low-low combination. The value should be low if the low-low-low quadrant is sparse. Finally, a three-way Boolean implication is identified if both tests in both steps are passed.

Simulation models

We assess the effectiveness of BIMS through a series of simulation studies. Briefly, for each of the six pairwise Boolean implication models, we mix a number of simulated positive cases with the same number of negative controls and see how many positive cases can be identified at a given level of FDR. In detail, a negative control is simulated by sampling two objects from the real data and then permuting the label of one object. A positive case is simulated using the following probabilistic model. Let X and Y be two objects sampled from the real data and α and β fractions of low abundant samples of X and Y, respectively. Let p be the probability that both X and Y are low abundant. We have the joint probabilities

$$P(X=0,Y=0)=p,\ P(X=1,Y=0)=\beta-p,$$
$$P(X=0,Y=1)=\alpha-p,\ P(X=1,Y=1)=1-\alpha-\beta+p.$$

The range of parameter p should be restricted since, in a Boolean implication, either one or two of the above joint probabilities should be significantly smaller than the others. For example, for $\mathrm{X_{low}} \rightarrow \mathrm{Y_{low}}$, $P(X=0,Y=1)$ should be significantly smaller than the other three, while in an Boolean equivalent relationship, both $P(X=0,Y=1)$ and $P(X=1,Y=0)$ should be significantly smaller than the other two. To simulate this constraint, we set $\mathrm{T_1}$ as the upper limit of the smaller probabilities and $\mathrm{T_2}$ as the lower limit of the larger probabilities, and thus the range of p can be determined. For example, in the $\mathrm{X_{low}} \rightarrow \mathrm{Y_{low}}$ relationship, the constraint is

$$\begin{cases} P(X=0,\ Y=0) > \mathrm{T_2} \\ P(X=0,\ Y=1) < \mathrm{T_1} \\ P(X=1,\ Y=0) > \mathrm{T_2} \\ P(X=1,\ Y=1) > \mathrm{T_2} \end{cases} \Rightarrow \begin{cases} p > \mathrm{T_2} \\ p > \alpha - \mathrm{T_1} \\ p < \beta - \mathrm{T_2} \\ p > \beta + \alpha + \mathrm{T_2} - 1 \end{cases}$$

.

We can therefore sample a p uniformly from its range to simulate the joint probabilities of a positive case and further generate a number of n (the number of samples in the real data) points according to these probabilities. Furthermore, to obtain continuous abundance data, we sample from two normal distributions: $N(\mu_1,1)$ and $N(\mu_2,1)$ $(\mu_1<\mu_2)$ for the low and high abundance points, respectively. Using the same method, we can generate the Boolean implication relationship for other types ($\mathrm{X_{low}} \rightarrow \mathrm{Y_{high}}$, $\mathrm{X_{high}} \rightarrow \mathrm{Y_{low}}$, $\mathrm{X_{high}} \rightarrow \mathrm{Y_{high}}$) with different constraints of p (see Supplementary Materials for details).

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-014-0037-3.

ACKNOWLEDGEMENTS

This research was partially supported by the National Basic Research Program of China (2012CB316504), the National High Technology Research and Development Program of China (2012AA020401), the National Natural Science Foundation of China (61175002), the Recruitment Program of Global Experts of China, and Tsinghua National Laboratory for Information Science and Technology (TNLIST).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Congmin Zhu, Rui Jiang and Ting Chen declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Wooley, J. C., Godzik, A. and Friedberg, I. (2010) A primer on metagenomics. PLoS Comput. Biol., 6, e1000667
2. Chaffron, S., Rehrauer, H., Pernthaler, J. and von Mering, C. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res., 20, 947–959
3. Amann, R. I., Ludwig, W. and Schleifer, K. H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev., 59, 143–169
4. Pace, N. R. (1997) A molecular view of microbial diversity and the biosphere. Science, 276, 734–740
5. Rappé, M. S. and Giovannoni, S. J. (2003) The uncultured microbial majority. Annu. Rev. Microbiol., 57, 369–394
6. King, A. J., Farrer, E. C., Suding, K. N. and Schmidt, S. K. (2012) Co-occurrence patterns of plants and soil bacteria in the high-alpine subnival zone track environmental harshness. Front. Microbiol, 3, 347
7. Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., Loiacono, K. A., Lynch, B. A., MacNeil, I. A., Minor, C., et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol., 66, 2541–2547
8. Voget, S., Leggewie, C., Uesbeck, A., Raasch, C., Jaeger, K. E. and Streit, W. R. (2003) Prospecting for novel biocatalysts in a soil metagenome. Appl. Environ. Microbiol., 69, 6235–6242
9. Nautiyal, C. S., Chauhan, P. S. and Nene, Y. L. (2007) Medicinal smoke reduces airborne bacteria. J. Ethnopharmacol., 114, 446–451
10. Ortiz, G., Yagüe, G., Segovia, M. and Catalán, V. (2009) A study of air microbe levels in different areas of a hospital. Curr. Microbiol., 59, 53–58
11. Martinez, A., Tyson, G. W. and Delong, E. F. (2010) Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. Environ. Microbiol., 12, 222–238

12. Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C. E., Sachdeva, R., Jones, A. C., Schwalbach, M. S., et al. (2011) Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J., 5, 1414–1425
13. Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. and DeLong, E. F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol., 178, 591–599
14. Tseng, C. H. and Tang, S. L. (2014) Marine microbial metagenomics: from individual to the environment. Int. J. Mol. Sci., 15, 8878–8892
15. Woyke, T., Xie, G., Copeland, A., González, J. M., Han, C., Kiss, H., Saw, J. H., Senin, P., Yang, C., Chatterji, S., et al. (2009) Assembling the marine metagenome, one cell at a time. PLoS One, 4, e5299
16. Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P., et al. (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut, 55, 205–211
17. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen,D., et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490, 55–60
18. Cho, I. and Blaser, M. J. (2012) The human microbiome: at the interface of health and disease. Nat. Rev. Genet., 13, 260–270
19. Rasheed, Z., Rangwala, H. and Barbará, D. (2013) 16S rRNA metagenome clustering and diversity estimation using locality sensitive hashing. BMC Syst. Biol., 7, S11
20. Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G. and Grim, S. L. (2013) Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol. Evol., 4
21. Chen, W., Zhang, C. K., Cheng, Y., Zhang, S. and Zhao, H. (2013) A comparison of methods for clustering 16S rRNA sequences into OTUs. PLoS One, 8, e70837
22. Hao, X., Jiang, R. and Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics, 27, 611–618
23. Schloss, P. D. and Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol., 71, 1501–1506
24. Huse, S. M., Welch, D. M., Morrison, H. G. and Sogin, M. L. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol., 12, 1889–1898
25. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol., 75, 7537–7541
26. Assenov, Y., Ramírez, F., Schelhorn, S. E., Lengauer, T. and Albrecht, M. (2008) Computing topological parameters of biological networks. Bioinformatics, 24, 282–284
27. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22, 1658–1659
28. Li, W., Jaroszewski, L. and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 17, 282–283
29. Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R. and Plevritis, S. K. (2008) Boolean implication networks derived from large scale, whole genome microarray datasets. Genome Biol., 9, R157
30. Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. Nature, 393, 440–442
31. Montoya, J. M., Pimm, S. L. and Solé, R. V. (2006) Ecological networks and their fragility. Nature, 442, 259–264
32. Timothy Pennington, J. and Chavez, F. P. (2000) Seasonal fluctuations of temperature, salinity, nitrate, chlorophyll and primary production at station H3/M1 over 1989–1996 in Monterey Bay, California. Deep Sea Res. Part II Top. Stud. Oceanogr., 47, 947–973
33. Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26, 2460–2461
34. Sahoo, D., Dill, D. L., Tibshirani, R. and Plevritis, S. K. (2007) Extracting binary signals from microarray time-course data. Nucleic Acids Res., 35, 3705–3712
35. Sinha, S., Tsang, E. K., Zeng, H., Meister, M. and Dill, D. L. (2014) Mining TCGA data using Boolean implications. PLoS One, 9, e102119
36. Fuhrman, J. A., Hewson, I., Schwalbach, M. S., Steele, J. A., Brown, M. V. and Naeem, S. (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. Proc. Natl. Acad. Sci. USA, 103, 13104–13109
37. Vigil, P., Countway, P. D., Rose, J., Lonsdale, D. J. and Gobler, C. J. (2009) Rapid shifts in dominant taxa among microbial eukaryotes in estuarine ecosystems. Aquat. Microb. Ecol., 54, 83–100

