Bio-Medical Materials and Engineering 24 (2014) 3737–3744
DOI 10.3233/BME-141202
IOS Press

Characterization of microbial associations in human oral microbiome

Min Su Lee[a,b], Sangyoon Oh[a,c,*] and Haixu Tang[a]
[a]*School of Information and Computing, Indiana University, Bloomington, IN, USA*
[b]*Department of Computer Science and Engineering, Ewha Womans University, Seoul, Korea*
[c]*Department of Software Convergence Technology, Ajou University, Suwon, Korea*

Abstract. Microorganisms interact with each other within a community. Within the same community, some microorganisms tend to co-exist, whereas some others tend to avoid each other. The association among microorganisms can be revealed by computing the correlation between their abundance patterns that are measured through metagenomic sequencing across multiple communities. In this paper, we built an *association network* among microorganisms from the human oral microbiome. To improve its accuracy, we adopted a network deconvolution algorithm to filter out indirect associations, and we used an ensemble of three correlation measures to filter out the false-positive associations. When applying on the metagenomic data from human oral samples, experimental results showed that phylogenetically close microorganisms formed highly correlated network clusters. Additionally, most of the identified mutually exclusive associations were related to the order Lactobacillales.

Keywords: Association network, correlation, microbiome, network deconvolution

1. Introduction

Microbes do not live in isolated environments; instead, they co-exist in microbial communities and exhibit various kinds of positive and negative interactions. The microbial communities associated with the human body (i.e., the human microbiome) consist of approximately 100 trillion microbial cells, which is more than ten times the number of human cells. Furthermore, the composition of human microbiome varies greatly across various individuals [1]. The human body is closely associated with microbial organisms, and the human microbiome plays a critical role in human health, and in various diseases such as obesity, diabetes, and neurochemical imbalances. Thus, it is important to characterize the human microbiome in order to better understand the functional aspects of their impact on human hosts.

Positive interactions between microorganisms within the same community can be interpreted as *mutualism.* Examples include cross-feeding, co-colonization, niche overlap, and co-aggregation in biofilms. Negative interactions can take the form of predator-prey or host-parasite relationships, respectively [2]. Although it is difficult to understand the precise ecological relevance of the positive and

*Corresponding author: Sangyoon Oh, Department of Software Convergence Technology, Ajou Univeristy, Suwon, Korea. Tel.: +82-31-219-2633; Fax: +82-31-219-1621; E-mail: syoh@ajou.ac.kr.

negative interactions, the identification of these microbial interactions does provide additional insights into the complex ecological relationships shared by these microorganisms.

Until now, most researchers working in the area of metagenomics have used conventional correlation, information-based theory, and distance measures to identify the relationships among microorganisms. However, various different correlation analysis methods often generate dissimilar outputs from the same exact input data. Also, a high correlation between two microorganisms may be caused by indirect transitive effects (discussed in the following sections). Therefore, a novel approach based on the metagenomic data is necessary to reliably identify the direct associations between various microorganisms.

In this study, we constructed an accurate microbial co-occurrence and mutual-exclusion network for the human oral microbiome. To filter out unreliable pairs of correlated microorganisms, we adopted an ensemble of three correlation measures. We also employed a network deconvolution algorithm to filter out the indirect associations.

2. Related works

2.1. Inferring microbial co-occurrence patterns

Owing to the recent advances in DNA sequencing, the characterization of microbial communities can now be accomplished using high throughput and low cost techniques, including the 16sRNA amplicon sequencing and the shotgun metagenomic sequencing. The microorganismal abundances can be estimated by using the techniques mentioned above. These estimates are based on the frequencies of reads obtained from multiple sequencing datasets. Several groups have developed computational methods for understanding association networks. This is usually done by assessing the co-occurrence and mutual exclusion patterns of two microorganisms in multiple samples. A variety of correlation measures and their corresponding statistical significances are required for the above determinations. Warren *et al*. identified microbe-microbe and host-microbe associations specific to colorectal carcinomas, using the co-occurrence networks based on Pearson's correlation [3]. Faust *et al*. constructed a global co-occurrence network of the human microbiome within and across body sites, using multiple similarity measures in combination with a generalized boosted linear model and various statistical tests [4].

2.2. Identified problems in conventional approaches

Two major problems need to be addressed before using the conventional co-occurrence analyses derived from correlation measures. Firstly, the different correlation measures often produce considerably dissimilar outputs. For example, Pearson's correlation coefficient variables are only suitable for the continuous and normally distributed data. However, if the data are not linear, the results may be leading. Moreover, as Pearson's r is sensitive to the values of the variables, an improper pre-processing step may significantly distort the outputs of Pearson's correlation. Spearman's correlation coefficient (i.e. Spearman's ρ) assesses monotonic relationships, and is simply defined as Pearson's correlation coefficient with ranked values. As it uses rank values instead of real values, it may tend to ignore the causal relationships between two variables. Furthermore, the Spearman's correlation coefficient is also sensitive to rarely occurring variables. As evident from the above examples, every correlation measure generates unique outputs that are not identical to those generated using other methods. In other words,

the choice of an optimal correlation measure has a large impact on the accuracy of the inferred co-occurrence network.

Secondly, the results of conventional correlation analysis may include an enormous amount of indirect association because of the transitive effects of directly co-occurring pairs. For example, if microorganism A interacts with B, and B interacts with C, it is often observed that microorganisms A and C are correlated. In a general case, the degree of observed correlation between two variables (microorganisms) is the sum of associations along all the connecting paths in the network reflecting the actual interactions among microorganisms [5]. Even though the problem of inaccurately inferred network was reported by Sewall Wright in 1921, the problem has not been clearly resolved till now.

3. Proposed methods for identifying microbial direct relationships in human oral microbiome

3.1. Data set and data preprocessing

16S rRNA (1.5 kbp in length) is commonly used as a marker gene for discerning microorganisms in communities because its highly conserved sequences have shown considerable variation with the evolutionary time. We downloaded the 16S rRNA amplicon sequences from the microbiome of healthy human subjects. These sequences to be used in our experiment were generated by the Human Microbiome Project at http://hmpdacc.org [1]. The dataset (release: 1 May 2010) includes the number of sequences that are affiliated with each phylotype from 18 body site samples of 239 healthy people. Nine out of the 18 body sites are from the oral cavity. We selected the following oral sites: buccal mucosa, hard palate, keratinized gingiva, palatine tonsils, saliva, subgingival plaques, supragingival plaques, tongue dorsum, and throat. The total number of oral sample records in the dataset was 2,835. To reduce the sparseness of the occurrence matrix, we excluded the 329 rarely occurring phylotypes (out of a total of 665) that are supported by less than five sequences across the entire dataset. As the collected samples do not possess similar volumes, the absolute counts of each phylotype were normalized by the total counts in each sample. As a result, we arrived at a normalized abundance matrix for 336 phylotypes across 2,835 oral cavity samples.

We then converted the normalized abundance matrix into a 3,024 × 323 matrix in which each row represents the combination of phylotypes and oral sites and each column represents an individual during one sampling visit. Note that each human subject was sampled during multiple visits. However, the number of individual visits varied from one to three for different subjects, and all nine oral sites were not sampled for each individual visit. This caused a lot of missing information in the converted matrix. As the correlation analyses are highly sensitive to missing data, some individual-visit columns that had missing oral site samples were eliminated from our analysis, and only 282 individual-visit columns were retained in the end. After removing null rows (i.e. rows with zero counts), a total of 2,100 rows remained. As a result of the pre-processing, we obtained a 2,100 × 282 microbial abundance matrix for the phylotypes in nine oral sites. The data were used to construct the co-occurrence and mutual exclusion network.

3.2. Co-occurrence / mutual-exclusion analysis

To identify the microbial co-occurrence and co-exclusion patterns, we performed pairwise correlation analyses. As discussed in Section 2.2, there are many correlation measures, and different measures produce significantly different outputs. Thus, to avoid misleading or biased results, we used the

following three correlation measures and aggregated them: mutual information, Pearson's correlation coefficient, and Spearman's correlation coefficient. By using mutual information for identifying the microbial co-occurrence and co-exclusion, we can easily capture the generic dependency among phylotypes. However, it does not state whether the dependency is positive or negative. By examining the sign of Pearson's and Spearman's correlation coefficients, we can determine whether a particular interaction is positive or negative.

3.3. Network deconvolution

A biological correlation network is typically constructed through correlation or by using a similarity analysis between each pair of objects from the observed data across multiple conditions. However, the results include many erroneous links or over-estimated edge weights in many cases due to the indirect information flow (e.g., the transitive relationship in the case of direct interactions).

Network deconvolution is a new method that eliminates the indirect weight from an inferred dependency network [5]. There are several assumptions in the network deconvolution method. Firstly, the measured edge weights from the observed data are assumed to be the sum of direct weights and indirect information flow. Secondly, the indirect flow weights can be approximated as the product of direct edge weight. Under these assumptions, the inferred observed network can be expressed as an infinite sum of true direct networks and all indirect information flows of increasing lengths. Finally, the direct edge weights can be inferred by reversing the effect of transitive information flow across all possible indirect paths. Let $\boldsymbol{G}_{obs}$ be an observed dependency network, $\boldsymbol{G}_{tru}$ be a true dependency network (which includes the direct dependencies only), and $\boldsymbol{G}_{ind}$ be a network, which includes only the indirect dependencies. Then, we can express the observed network ($\boldsymbol{G}_{obs}$) in terms of the true network ($\boldsymbol{G}_{tru}$) and the indirect network ($\boldsymbol{G}_{ind}$), and the indirect network can be expressed in terms of all indirect effects along paths of increasing length as follows:

$$\boldsymbol{G}_{obs} = \boldsymbol{G}_{tru} + \boldsymbol{G}_{ind} = \boldsymbol{G}_{tru} + (\boldsymbol{G}_{tru}^2 + \boldsymbol{G}_{tru}^3 + \boldsymbol{G}_{tru}^4 + \cdots) = \boldsymbol{G}_{tru}(\boldsymbol{I} - \boldsymbol{G}_{tru})^{-1} \tag{1}$$

Thus, the true direct network can be computed using the observed network:

$$\boldsymbol{G}_{tru} = \boldsymbol{G}_{obs}(\boldsymbol{I} + \boldsymbol{G}_{obs})^{-1} \tag{2}$$

If the observed network and the true dependency network are represented with a n × n decomposable matrix (e.g., a symmetric matrix), then each eigenvalue of the true network λ_i^{tru} can be expressed as a function of a single corresponding eigenvalue of the observed network λ_i^{obs} as follows:

$$\lambda_i^{tru} = \frac{\lambda_i^{obs}}{1+\lambda_i^{obs}} \quad \forall 1 \le i \le n. \tag{3}$$

To guarantee the convergence of the Taylor series in Eq. (1), the maximum absolute eigenvalue of the true network should be strictly less than one. Hence, the linear eigenvalue scaling of the observed network is required as described in reference [5] and its supplementary note. In this study, we set the linear eigenvalue scaling factor to 0.9 and the diagonal elements to zero to filter out self-relationships.

3.4. Construction and analysis of direct microbial co-occurrence and mutual exclusion network

After network deconvolution, we selected the correlation patterns that were commonly supported by three direct correlation measures, in order to identify reliable direct co-occurrence and mutual exclusion patterns of microbes in the human oral microbiome. After assuming that the relationship is likely to be a true positive relationship if different measures concur, and likely to be a false positive relationship if the measures differ, we combined the three correlation measures. A pair was regarded as a reliable direct co-occurrence pattern or a mutual exclusion pattern if each of its three direct correlation coefficients were greater than 0.3 or less than -0.3, respectively. The direct correlation coefficients were derived from three correlation analysis methods and the network deconvolution algorithm.

To analyze the direct co-occurrence/mutual exclusion patterns in the context of the network, we clustered the microbial correlation network using the MCODE (Molecular Complex Detection) algorithm [6]. MCODE finds the densely connected regions of a network using a graph-based clustering algorithm. The algorithm makes each cluster containing the most relevant data and it makes visualization of large networks more manageable by dividing whole data into closely connected subsets.

4. Results and discussion

4.1. A network of direct microbial co-occurrence and mutual exclusion relationships within and between oral sites

Using the analysis procedures described in Section 3, we selected 1,326 out of 2,203,950 microbe occurrence patterns within and between human oral sites. Their q-values (i.e., adjusted p-values) in the three correlation analyses were less than 0.05. The averaged strength of co-occurrence (i.e., direct correlation coefficients) of 1,228 co-occurrence patterns ranged from 0.3 to 0.59, and those of the 98 mutual exclusion patterns ranged from -0.47 to -0.33, respectively.

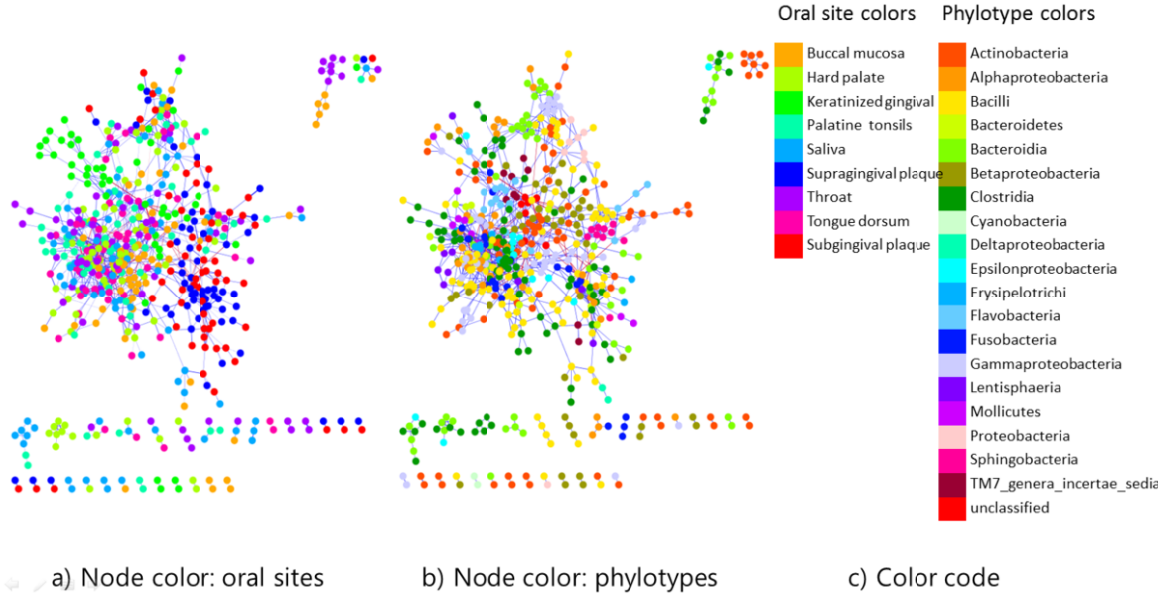

Fig. 1. A global network of microbial direct co-occurrence and mutual exclusion relationships in human oral microbiome.

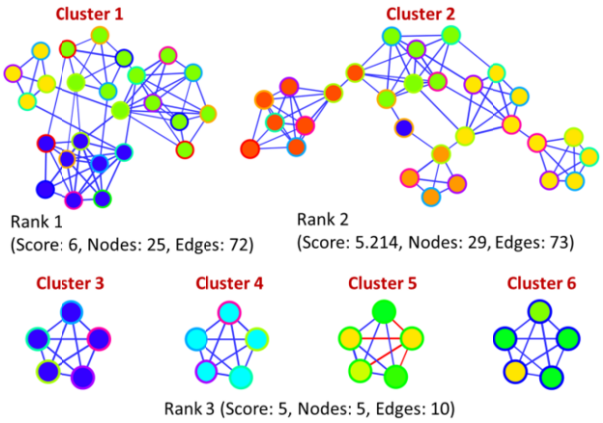

Fig. 2. Top six correlation network clusters.

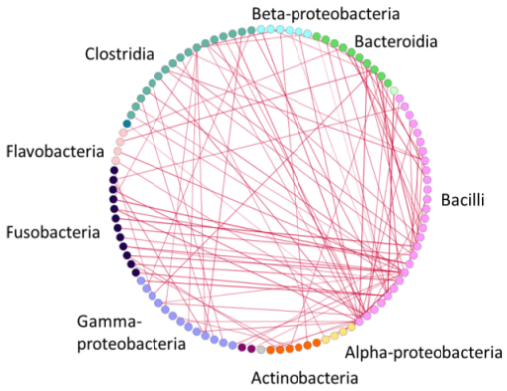

Fig. 3. Mutual exclusive patterns.

Figure 1 shows a global network of microbial direct co-occurrence and mutual-exclusion patterns within and between the oral sites. The network consists of 563 nodes and 1,326 edges. Each node represents a phylotype in an oral site. To visualize the characteristics of the network, it was colored as per the corresponding oral sites (Figure 1(a)) or phylotypes (Figure 1(b)). Each phylotype was annotated using a hierarchical taxonomy, whose depth was six [1]. To capture the overview of the global network, we used taxonomic level three when colouring nodes by phylotypes. By identifying the nodes as phylotypes and oral sites respectively (Figure 1(c)), we could see that there were global co-occurrence tendencies among the phylogenetically close clades and among the same oral sites.

4.2. Network cluster analysis

Figure 2 shows the six highest densely connected complexes obtained through MCODE graph clustering. Different node colours represent different phylotypes in taxonomical level three, and different border colours correspond to different oral sites (see Figure 1(c)). The thickness of the edge represents the correlational strength, and the red edge represents the mutually exclusive patterns.

Cluster 1 and cluster 2 show tightly connected patterns among closely related phylotypes, regardless of the oral sites. On the other hand, the loosely connected patterns within the same oral sites are shown between groups of tightly connected phylotypes. Cluster 3 and cluster 4 show co-occurrence patterns of Gamma-proteobacteria and Epsilon-proteobacteria in five oral sites (viz. saliva, hard palate, palatine tonsils, throat, and tongue dorsum). In contrast, cluster 5 and cluster 6 show the co-occurrence and mutual exclusion patterns between different phylotypes in an oral site. In cluster 5, Lactobacillales, an order of Bacilli, was mutually exclusive with other phylotypes, viz. Clostridia, Bacillales (another order of Bacilli), Bacteroidetes, and Beta-proteobacteria in the keratinized gingival. Cluster 6 shows the co-occurrence patterns for Lactobacillaes, Bacteroidia and Clostridia in the supragingival plaque.

4.3. Mutual exclusion patterns

Mutual exclusion patterns convey the essential information necessary to understand competitive or predator-prey relationships in the microorganism community. We collected 98 direct mutual exclusive patterns and depicted the results graphically, by using a circular layout for the negative interactions among clades in Figure 3. The figure shows that Bacilli (most of them were Lactobacillaes and the rest

were Bacillales) are highly involved in the mutually exclusive interactions. Furthermore, Fusobacteria, Bacteroidia, and Clostridia show higher negative co-occurrence patterns with Bacilli.

4.4. The effectiveness of the network deconvolution and the ensemble of three correlation measures

To analyse the effectiveness of the network deconvolution method, we compared the correlation coefficients before and after network deconvolution for each of the three similarity measures. After the network deconvolution process was applied, we observed a decrease in the higher observed correlation coefficients, and a slight increase in the lower observed correlation coefficients. These results indicate that the overestimated observed weights decreased and the relatively underestimated observed estimated weights increased by removing the indirect information flow.

We also checked the characteristics of the phylotype pairs whose correlation weights were significantly diminished. We checked their direct co-occurrence weight as derived by other similarity measures with network deconvolution, and found that most of them showed lower direct co-occurrence weights. Because these results were not supported by other correlation measures, they could be putative false positives in the co-occurrence inference.

Finally, we computed the correlation of outputs among the three similarity measures before and after network deconvolution. Before the process, the correlation between Pearson's r and Spearman's ρ was high. However, the correlations between the mutual information and Pearson's r, and between the mutual information and Spearman's ρ were very low (-0.08 and 0.07, respectively). After network deconvolution, the correlations between the mutual information and Pearson's r as well as between the mutual information and Spearman's ρ slightly increased (0.10 and 0.26, respectively). However, the correlation patterns among the outputs from different measures still varied considerably. Therefore, our approach, wherein we selected the commonly occurring interaction pairs based on three similarity measures, may be useful for constructing a reliable co-occurrence network.

5. Conclusion

In this study, we identified the direct microbial associations in the human oral microbiome, using the 16S rRNA abundances derived from the metagenomic sequencing data and constructed a direct co-occurrence and mutual-exclusion network. We adopted an ensemble of three correlation measures and the network deconvolution method to remove the unreliable and indirect associations. Our results revealed some characteristics in the direct microbial co-occurrence and mutual exclusion networks: 1) phylogenetically close microorganisms are highly likely to co-occur and form network clusters; 2) phylogenetically distant microorganism groups are only loosely connected by co-occurred pairs in an oral site; and 3) Lactobacillales are highly involved in the mutually exclusive relationships.

Our proposed method can be used to accurately predict unknown functions of microbial species using comparative genomics and can help us understand the interactions between microorganisms within a microbial community. Moreover, this method can be utilized for synthetic ecology, which attempts to manipulate microbial communities in order to enhance the abundance of beneficial species, while suppressing the harmful ones.

Acknowledgement

This work was supported by NRF of Korea and WISET Grant funded by Korean Government (MSIP) (No.KW-2014-PPD-0053)

References

[1] B.A. Methe et al., The human microbiome consortium: A framework for human microbiome research, Nature **486** (2012), 215–221.
[2] K. Faust and J. Raes, Microbial interactions: from networks to models, Nature Reviews **10** (2012), 538–550.
[3] R. Warren, D. Freeman, S. Pleasance, P. Watson, R.A. Moore, K. Cochrane, E. Allen-Vercoe and R.A. Holt, Co-occurrence of anaerobic bacteria in colorectal carcinomas, Microbiome **1** (2013), 1–12.
[4] K. Faust, J.F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes and C. Huttenhower, Microbial co-occurrence relationships in the human microbiome, PLoS Computational Biology **8** (2012), e1002606.
[5] S. Feizi, D. Marbach, M. Medard and M. Kellis, Network deconvolution as a general method to distinguish direct dependencies in networks, Nature Biotechnology **31** (2013), 726–733.
[6] G.D. Bader and C.W. Hogue, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinformatics **4** (2003), 1–27.