Title

Network construction and structure detection with metagenomic count data

Permalink

https://escholarship.org/uc/item/6qp8z71s

Journal

BioData Mining, 8(1)

ISSN

1756-0381

Authors

Liu, Zhenqiu
Lin, Shili
Piantadosi, Steven

Publication Date

2015-06-01

DOI

10.1186/s13040-015-0072-2

Peer reviewed

Liu *et al. BioData Mining* (2015) 8:40
DOI 10.1186/s13040-015-0072-2

BioData Mining

METHODOLOGY Open Access

Network construction and structure detection with metagenomic count data

Zhenqiu Liu[1*], Shili Lin[2] and Steven Piantadosi[1]

*Correspondence:
zhenqiu.liu@cshs.org
[1] Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
Full list of author information is available at the end of the article

Abstract

Background: The human microbiome plays a critical role in human health. Massive amounts of metagenomic data have been generated with advances in next-generation sequencing technologies that characterize microbial communities via direct isolation and sequencing. How to extract, analyze, and transform these vast amounts of data into useful knowledge is a great challenge to bioinformaticians. Microbial biodiversity research has focused primarily on taxa composition and abundance and less on the co-occurrences among different taxa. However, taxa co-occurrences and their relationships to environmental and clinical conditions are important because network structure may help to understand how microbial taxa function together.

Results: We propose a systematic robust approach for bacteria network construction and structure detection using metagenomic count data. Pairwise similarity/distance measures between taxa are proposed by adapting distance measures for samples in ecology. We also extend the sparse inverse covariance approach to a sparse inverse of a similarity matrix from count data for network construction. Our approach is efficient for large metagenomic count data with thousands of bacterial taxa. We evaluate our method with real and simulated data. Our method identifies true and biologically significant network structures efficiently.

Conclusions: Network analysis is crucial for detecting subnetwork structures with metagenomic count data. We developed a software tool in MATLAB for network construction and biologically significant module detection. Software MetaNet can be downloaded from http://biostatistics.csmc.edu/MetaNet/.

Keywords: Metagenomics data, Networks analysis, Modules

Background

Our human body is a host to various of microbes. Over 90 % of the cells in human body are bacterial or other non-human cells. These microbes have great influence on human physiology and nutrition, and are crucial for our health [1]. Metagenomics, which is the study of genetic material recovered directly from uncultured microorganisms, has accelerated the analysis of functional biodiversity relevant to its ecology. The objectives of human microbiome research are to explore the host-microbiota interactions, associate differences in microbial communities with differences in metabolic functions and diseases, and understand how microbiota changes may affect human health [1]. Massive amounts of metagenomic sequencing data have been generated with advances in next-generation sequencing (NGS) technologies. There are two NGS methods for metagenomics: whole metagenomic shotgun sequencing (WMGSS) and 16S rRNA gene sequencing. 16S rRNA sequencing is an amplicon sequencing method for identifying and comparing bacteria present within a given sample, while WMGSS comprehensively sample all genes in all organisms present in a given complex sample. The two techniques are quite different and intend for answer different biological questions. It has been shown that 16S rRNA sequencing contains hundreds of thousands of 16S RNAs fragments and is an efficient tool to infer bacterial communities, while WMGSS is mainly used for functional delineation and it is generally not deep enough to detect rare species in complex communities [2, 3]. In this paper, we infer network structures and taxa co-occurrence with 16S rRNA sequencing. By examining the relationship of genome structure and function across many different taxa with NGS data, the scope of microbiology and of microbial evolution studies has been greatly broadened, and the field of systems biology has emerged [4, 5].

There have been great strides in determining the taxonomical and functional contents of a sample in the last several years. Many software packages including MOTHUR [6], UniFrac [7], QIIME [8], and SILVAngs [9] have been designed primarily for the analysis of 16S rRNA sequencing data, while the other software packages including MEGAN [5, 10], Phymm [11], NBC [12] were developed mainly for shotgun metagenomic sequencing data. Those tools provide different approaches for the comparison of microbial communities with metagenomic sequence data. One output from some of the software is the abundance counts (sequence reads) for each taxa. These taxa abundance counts can be further analyzed to identify taxa and microbial communities that are associated with human diseases by comparing taxa counts from two or more groups with different disease status. Study of the link between characteristics of a microbiome and human disease is a active area of research. Current approaches such as MetaStats [13] and MetaDistance [14] mainly focus on variations in abundance across different clinical conditions, ignoring the interactions and structural variations among taxa. However, bacteria taxa do not act alone, rather they form part of large interacting (co-occurrence) networks and may function together. Variations in network structures and taxon interactions may be associated with disease status and clinical phenotypes [15–18]. Therefore network methods specifically designed for metagenomic count need to be developed.

Networks methods and graph theory have been widely applied to gene regulatory network construction with expression data [19–21]. Network analysis has also proved powerful for studying the characteristics of metabolic networks and their impact on various functional and evolutionary properties [22–24]. RNA-Seq is a NGS approach to transcriptome profiling. It provides a far more precise measurement of levels of transcripts and their isoforms than other methods [25]. Local Poisson graphical (log-linear) model and Bayesian generalized graphical model for network construction have been developed with RNA-seq data recently [26, 27]. However, the log-linear model is not valid when there are zero counts or measures in the data, which is common in metagenomics. Also, the Bayesian Poisson graphical model is slow when the network size is large. It usually takes hours to construct a network with hundreds of nodes. Those parametric methods can not be applied to metagenomic count data without modification. Moreover, even though there are a few methods available for network construction with microbiome data [28–32], most methods for network analysis are based on pairwise correlations (or distance) and ignore high-order correlations. However, high-order (partial) correlation has the advantage over pairwise correlation, because it measures the conditional dependency between two taxa given the effect of other taxa being removed or fixed, and reflects direct correlation between taxa and excludes the between-taxon dependency due to other taxa. In addition, variance heterogeneity and non-normality of metagenomic count data make standard correlations invalid (e.g. Pearson correlation). One way to deal with the problem is to use proportion and log-ratio transformations [33]. However the log-ratio is not defined when there are zeros in the data and approximation methods have to be used.

In this paper, we propose a nonparametric approach for co-occurrence network construction and subnetwork structure detection. We propose similarity (or distance) measures between taxa derived by adapting distance measures between samples with abundance counts defined in ecology [34]. We also expand the sparse inverse covariance method to sparse inverse of general similarity matrices for high-order correlation. The performance of our methods are evaluated through simulation and publicly available metagenomic data sets. The proposed methods are efficient for detecting true network structures. Even though the co-occurrence network is just a description analysis from temporal snapshots, it may be informative regarding how microbial taxa function together.

Methods

Given samples with or without associated phenotypes, our goal is to study the connectivity and subnetwork structures of bacteria taxa with human microbiome. The final output from 16S rRNA sequencing of the host's microflora is an integral, non-negative number of sequencing reads for each taxon. Such reads are the metagenomic counts represented as

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

where X is the count matrix with n samples and m taxa, and x_{ij} denotes the total number of reads of taxon j in sample i. In case there have been known disease status or clinical conditions ($\mathbf{y}$) available, we will also discuss methods for detecting structural variations across clinical conditions. Constructing a human metagenomic network requires several sequential steps: (i) estimating pairwise similarity (or distance) measures between different taxa, (ii) adjacency matrix construction, (iii) network structure (module) detection and differentiated networks. We will discuss each of these steps.

Pairwise similarity measures

Correlation coefficients are one type of similarity measures that describe the magnitude and direction of association between two variables. Because metagenomic count data typically have variances that are a function of the mean and are not normally distributed, the usual pairwise correlation (e.g. Pearson correlation coefficient)is not appropriate for network analysis. We use two distribution-free nonparametric correlations for count data. Given two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ and their corresponding ranks $\mathbf{R}_x$ and $\mathbf{R}_y$, we have

- Spearman rank-order correlation:

$$R(\mathbf{x},\mathbf{y}) = 1 - \frac{6\sum_{i=1}^{n}(\mathbf{R}_{xi} - \mathbf{R}_{yi})^2}{n(n^2-1)}$$

  We will take the average of the scores when multiple elements have the tied ranks.
- Kendall's τ rank correlation:

$$\tau(\mathbf{x},\mathbf{y}) = \frac{\sum_{i<j}\operatorname{sgn}(x_i < x_j)\operatorname{sgn}(y_i < y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}},$$

  where $T_0 = n(n-1)/2$, $T_1 = \sum_k t_k(t_k - 1)/2$, and $T_2 = \sum_l u_l(u_l - 1)/2$. The t_k is the number of tied $\mathbf{x}$ values in the kth group of the tied $\mathbf{x}$ values, u_l is the number of tied $\mathbf{y}$ values in the lth group of tied $\mathbf{y}$ values, and sign(z) is defined as:

$$\operatorname{sgn}(z) = \begin{cases} 1 & \text{if} x_i < x_j, \\ 0 & \text{if} x_i = x_j, \\ -1 & \text{if} x_i > x_j. \end{cases}$$

Our similarity matrix S can be defined with either $S = \left[\sin\left(\frac{\pi}{2}R\left(\mathbf{x}_i,\mathbf{x}_j\right)\right)\right]_{m\times m}$ or $S = \left[\sin(\frac{\pi}{2}\tau(\mathbf{x}_i,\mathbf{x}_j))\right]_{m\times m}$ [35]. Those distribution-free correlations only utilize rank information, and are more robust than the parametric approach. Even though they are slight less efficiency than Pearson correlation under normal distribution, both Spearman and Kendall correlation coefficients provide a good compromise between robustness and efficiency [36].

Distance measures are commonly used for quantifying the dissimilarities between samples and visualizing the samples in 2D and 3D [37]. They have been modified to measure pairwise similarity between taxa and construct phylogenetic tree recently [38, 39]. Given two n-dimensional column vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ of two taxa, distance measures between taxa can be defined as

- Hellinger distance:

$$D(\mathbf{x}_1,\mathbf{x}_2) = \sqrt{\sum_{i=1}^{n}\left(\sqrt{\frac{x_{i1}}{\mathbf{x}_{+1}}} - \sqrt{\frac{x_{i2}}{\mathbf{x}_{+2}}}\right)^2},$$

  where $\mathbf{x}_{+1} = \sum_{i=1}^{n} x_{i1}$, and $\mathbf{x}_{+2} = \sum_{i=1}^{n} x_{i2}$.
- The χ^2 distance:

$$D\left(\mathbf{x}_1,\mathbf{x}_2\right) = \sqrt{\sum_{i=1}^{n}\frac{(\mathbf{x}_{+1} + \mathbf{x}_{+2})}{(x_{i1} + x_{i2})}\left(\sqrt{\frac{x_{i1}}{\mathbf{x}_{+1}}} - \sqrt{\frac{x_{i2}}{\mathbf{x}_{+2}}}\right)^2}.$$

- Bray-Curtis dissimilarity:

$$D(\mathbf{x}_1,\mathbf{x}_2) = 1 - 2\frac{\sum_{i=1}^{n}\min(x_{i1},x_{i2})}{\sum_{i=1}^{n}(x_{i1} + x_{i2})} = 1 - 2\sum_{i=1}^{n}\frac{\min(x_{i1},x_{i2})}{(\mathbf{x}_{+1} + \mathbf{x}_{+2})}.$$

These distances can be calculated with either raw or relative abundance reads. Even though there is no great difference, we suggest to use relative abundance for sequencing depth adjustment. The relative abundance matrix P is computed from the count matrix X with $P = [p_{ij}]_{n\times m}$, where $p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}}$. Based on the distance measures, we define a similarity measure with the popular Gaussion kernel as

$$S = [S_{ij}]_{m\times m}, \qquad \text{where} \qquad S_{ij} = S(\mathbf{x}_i,\mathbf{x}_j) = e^{-\frac{D^2(\mathbf{x}_i,\mathbf{x}_j)}{\sigma}},$$

where the free parameter σ can be estimated by resampling. We set $\sigma = 1$ for all computations in this paper. This distance based similarity matrix S is a positive (semi)-definite kernel matrix well studied in machine learning and bioinformatics. The kernel function $e^{-\frac{D^2(\mathbf{x}_i,\mathbf{x}_j)}{\sigma}}$ can be treated as an inner product in the high-dimensional feature space, so S can be regarded as the covariance matrix in the feature space.

Similarity to adjacency matrix

To compensate for noise and measure error, we propose two efficient approaches to determine statistically significant nonzero similarities. Unlike most methods in the literature determining the network structure with an arbitrary threshold of pairwise correlation, we are more interested in studying high order correlations. i.e., how $\mathbf{x}_i$ and $\mathbf{x}_j$ associate with each other when information about other variables is taken into consideration. Sparse inverse covariance for graph construction was originally proposed for continuous data with the assumption that the observations are from a multivariate Gaussian distribution [40]. This approach can handle large network efficiently. We extend this method to study the sparse inverse of a general similarity matrix S^{-1}, and evaluate its efficiency using simulation. Unlike S, a value zero in any cell of S^{-1} implies conditional independence among those variables. Mathematically $S^{-1}_{ij} = 0 \Rightarrow P(\mathbf{x}_i, \mathbf{x}_j|\mathbf{x}_{-i,-j}) = 0$, where $\mathbf{x}_{-i,-j}$ denotes all the variables other than $\mathbf{x}_i$ and $\mathbf{x}_j$. The likelihood estimate of $A = S^{-1}$ is

$$\max_{A \succ 0} L = \log\det A - tr(SA),$$

where $tr(SA)$ is the trace of SA. Assuming S is nonsingular, and taking the first order derivative, we have $A^{-1} = S$. However, it is common that $n < m$ in metagenomic data, so S can be singular. In such case, the following l_1 penalized error function can be minimized to obtain maximal likelihood estimates:

$$\min_{A \succ 0} E = -\log\det A + tr(SA) + \lambda||A||_1,$$

where $||A||_1 = \sum_{ij} |a_{ij}|$ is the elementwise l_1 norm for matrix A. The sparse structure of A can be estimated directly. This approach follows the framework of block coordinate descent [40, 41]. Mathematically, we partition the matrices S and A into the following block form:

$$S = \begin{bmatrix} S_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{bmatrix}; \qquad A = \begin{bmatrix} A_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{12} & a_{22} \end{bmatrix},$$

where $S_{11}, A_{11} \in \mathbb{R}^{(m-1)\times(m-1)}$, $\mathbf{s}_{12}, \mathbf{a}_{12} \in \mathbb{R}^{m-1}$, and $s_{22}.a_{22} \in \mathbb{R}$. Then we have

$$\log\det A = \log\det\left[A_{11}\left(a_{22} - \mathbf{a}_{12}^T A_{11}^{-1}\mathbf{a}_{12}\right)\right] = \log\det A_{11} + \log\left(a_{22} - \mathbf{a}_{12}^T A_{11}^{-1}\mathbf{a}_{12}\right).$$

So

$$\min_{A \succ 0} E = -\log\det A_{11} - \log(a_{22} - \mathbf{a}_{12}^T A_{11}^{-1}\mathbf{a}_{12}) + tr(SA) + \lambda||A||_1.$$

Assuming A_{11} is fixed and taking the first order derivative for $\mathbf{a}_{12}$ and a_{22}, we have the sub-differential of E with respect to $\mathbf{a}_{12}$:

$$\frac{\partial E}{\partial \mathbf{a}_{12}} = \frac{2}{a_{22} - \mathbf{a}_{12}^T A_{11}^{-1}\mathbf{a}_{12}} A_{11}^{-1}\mathbf{a}_{12} + 2\mathbf{s}_{12} + 2\lambda \mathrm{sgn}(\mathbf{a}_{12}),$$

where $\mathrm{sgn}(x) = \frac{\partial |x|}{\partial x}$ for $x \in \mathbb{R}$ is defined as $\mathrm{sgn}(x) = \begin{cases} 1 & x > 0, \\ [-1, 1] & x = 0, \\ -1 & x < 0. \end{cases}$ Similarly, since $a_{22} > 0$, the partial derivative of E with respect to a_{22}:

$$\frac{\partial E}{\partial a_{22}} = -\frac{1}{a_{22} - \mathbf{a}_{12}^T A_{11}^{-1} \mathbf{a}_{12}} + s_{22} + \lambda.$$

After finding the derivative, we initialize $A^0 = (S + \lambda I)^{-1}$, and then use the standard decent gradient algorithm to update each row/column repeatedly until the algorithm converges. After obtaining the sparse A and taking the absolute value $A = |A|$, we set the diagonal value of A to zero with $A = A - \mathrm{diag}(A)$ to get the final adjacency matrix A. The adjacency matrix A is a representation of a graph, where the value of a_{ij} represents the connectivity between taxa i and j.

λ Determination

The regularization parameter λ controls the number of nonzero estimated links between nodes and the sparsity of the network. The larger the λ, the sparser the network. A common approach for determining λ is stability selection [42, 43]. This approach seeks the λ leading to the most stable sets of edges. Given data X, stability selection first draws p sub-samples $X_1, X_2, \ldots, X_p$ of size q $(1 < q < n)$, where $q = \frac{2}{3}n$ in this paper, and then estimates one separate network $A^i(\lambda)$ for each sub-sample X_i and a fixed regularization parameter λ. Stability selection then defines the average fraction of disagreements over all edges of the sub-sampled graphs as

$$D(\lambda) = \frac{\sum_{j<k} \bar{a}_{jk}(\lambda)(1 - \bar{a}_{jk}(\lambda))}{\binom{p}{2}}, \quad \text{where} \quad \bar{a}_{jk} = \frac{1}{p}\sum_{i=1}^{p} a^i_{jk}(\lambda).$$

The optimal $\hat{\lambda}$ is then chosen as:

$$\hat{\lambda} = \min\left\{\lambda : \max_{0<t<\lambda} D(t) \le \alpha\right\}, \quad \text{where} \quad \alpha = 0.05.$$

Final network is constructed using whole data X and $\hat{\lambda}$.

Network structure detection and differentiated networks

The problem of subnetwork structure detection requires the partition of a network into communities (subnetworks/modules) of densely connected nodes, while nodes belonging to different communities are sparsely (weakly) connected. Bacterial subnetwork structure detection is very important because we want to know which taxa coexist and function together. One simple approach to accomplish this is modularity function maximization [44, 45]. The modularity function also measures the quality of a partition and can be used to compare the performance of different partition methods. Given a weighted network with adjacency matrix A, to attribute each node to a module c_i, a modularity function can be defined as follows:

$$Q = \frac{1}{2w}\sum_{i,j}\left[a_{ij} - \frac{k_i k_j}{2w}\right]\delta(c_i, c_j),$$

where a_{ij} is the weight of an edge between taxa (node) i and j, $k_i = \sum_j a_{ij}$ is the sum of the weights of edges attached to taxon i, and $w = \frac{1}{2}\sum_{i,j} a_{ij}$ is the total weight. In addition, c_i is the subnetwork (module) to which taxon i is assigned, and the δ function $\delta(u, v) = 1$ if $u = v$ and 0 otherwise. Obviously, $-1 \leq Q \leq 1$ and the larger Q indicates better separation in subnetworks. The subnetwork partition algorithms are designed for maximizing Q. We adopt the two-step iterative local greedy approach [44] in this paper. Unlike K-means or hierarchical clustering, this two-step algorithm automatically determines the number of network modules (clusters) without predefining.

Given metagenomic data from different clinical conditions or different times, we construct a network with similarity S^i and adjacency matrix A^i for clinical condition or time i. We are interested in knowing wether the network structures of a subset of taxa have changed from one clinical condition (time) to another. We may define network statistics to measure the network structure changes either with the similarity matrix S or the adjacency matrix A. Given two networks A^1 (S^1) and A^2 (S^2), our first network statistic is defined by the following mean absolute distance (MAD):

$$\Delta(A) = \frac{1}{m(m-1)} \sum_{i<j<m} |a^1_{ij} - a^2_{ij}|, \text{ or } \Delta(S) = \frac{1}{m(m-1)} \sum_{i<j<m} |s^1_{ij} - s^2_{ij}|,$$

where a^1_{ij} and a^2_{ij} are the interaction score between taxa i and j in network 1 and 2. The networks are considered to be significantly different if the value of $\Delta(A)$ or $\Delta(S)$ is large. Permutation tests can be used to estimate the P-value. Exact permutation test will be used when the sample size is small (< 100), otherwise, the number of permutations used will be $L = 100{,}000$. We first permutate the original data L times and compute the $\Delta(A, \pi)$ or $\Delta(S, \pi)$ for each permutation π, the P-value corresponding to $\Delta(A)$ or $\Delta(S)$ can be computed as:

$$P(\Delta(A)) = \frac{1}{L} \sum_{\pi} I(\Delta(A, \pi) \geq \Delta(A)), \text{or } P(\Delta(S)) = \frac{1}{L} \sum_{\pi} I(\Delta(S, \pi) \geq \Delta(S)),$$

where $I(x) = 1$ if x is true and 0 otherwise, and

$$\Delta(A, \pi) = \frac{1}{m(m-1)} \sum_{i<j<m} \left| a^{\pi,1}_{ij} - a^{\pi,2}_{ij} \right|, \text{or } \Delta(S, \pi) = \frac{1}{m(m-1)} \sum_{i<j<m} \left| s^{\pi,1}_{ij} - s^{\pi,2}_{ij} \right|.$$

MetaNet package

MetaNet toolbox in MATLAB was implemented to construct sparse network from metagenomic count data. The toolbox was tested under MATLAB 2013a, but should also work on the later versions of MATLAB. Implemented functions in this toolbox include several similarity (distance) measures, simulated distributions and network models, sparse inverse covariance estimation, network structure detection and differential networks, and network visualization. The goal of this distribution is to provide an easy-to-use tool for network construction and analysis. Although the package is till under development, the users can construct, analyze, and visualize a network from their own data without much difficulty. MetaNet is provided as is without warranty of any kind. More information and the toolbox can be downloaded from http://biostatistics.csmc.edu/MetaNet/.

Data sets

Simulated data

Simulated count data with different numbers of nodes and sample sizes are generated from a negative binomial (NB) distribution. More specifically, the data sets are generated from a NB distribution $X_{ij} \sim NB(\lambda_{ij}, \gamma)$ with mean λ_{ij} and dispersion parameter γ, and $\log(\lambda_i)$ is from a multivariate distribution $\log(\lambda_i) \sim N(\mu, \Sigma)$ with mean μ and covariance matrix Σ. The graphical structures are constructed through $A = \Sigma^{-1}$, where A is an adjacency matrix with additional diagonal elements $a_{ii} = \sum_{j,i \neq j} a_{ij} + 1, i = 1, \ldots, n$. The adjacency matrices are generated using three different models include small world, scale free, and range dependent networks. The small world network we use only allows a node to connect with its neighborhood node, while the scale-free network has the number of nodes of degree 2 following a power law, and the range dependent network has an edge between nodes i and j with probability $0.9.0.3^{|j-i|-1}$. These three networks are known to mimic the behavior of real biological networks. All count data sets in this paper are generated by setting $\mu = 3$ and the overdispersion parameter $\gamma = 2$.

Real metagenomic data from body habitats

The real data was collected from six body habitats including external auditory canal (EAC), gut, hair, nostril, oral cavity, and skin [46]. The objective of the original study was to estimate the microbial community composition and detect the differentiation in abundance among body habitats. A total of 815 samples were collected for 6 categories of habitat. Networks were constructed from gut, oral cavity (OC), and skin samples with the sample sizes of 45, 54, and 612 respectively. There were total 1713 taxa at the genus level.

Results

Results with simulation data

The proposed approaches were first evaluated using simulated count data. Given the number of nodes $n = 20$ and sample sizes of $m = 500$, we simulated the count data with known network structures. The predicted adjacency matrix was then estimated with the inverse of similarity matrix. The regularized parameter λ was chosen via stability selection. Computational results with different similarity measures and graphic models are given in Fig. 1.

The proposed similarity or dissimilarity measures including Spearman correlation coefficients, Hellinger, and Bray-Curtis performed well to detect the true structures as shown in Fig. 1. The area under ROC curves (AUC) was used to evaluate the performance of detecting proposed network structures, where the specificity for a network measures the proportion of no edges that are correctly detected, while the sensitivity for a network is the proportion of edges that are correctly identified. With the optimal $\lambda^* = 0.2, 0.55,$ and 0.65 for Spearman, Hellinger, and Bray-Curtis, we have the predicted AUCs of 0.91, 0.96, and 0.92 respectively with the small world network. The Bay-Curtis distance performed the best, while the other two measures also performed well (≥ 0.91). Similarly, our proposed approach also performed reasonable well with both scale-free and range dependent networks. We achieved the best predicted AUC of 0.851 and 0.902 with Hellinger distance and $\lambda^* = 0.6$ and 0.45 respectively for scale-free and range dependent networks. Overall Spearman has the lowest predicted AUCs for all models with the negative binomial simulated data as shown on the bottom of Fig. 1, but the differences among all measures are not very significant.

To further evaluate the performance of the method for large networks with small sample sizes and different similarity measures, Small world, scale free, and range dependent networks with 500 nodes and the sample size of 50, 100, and 200 respectively are used for

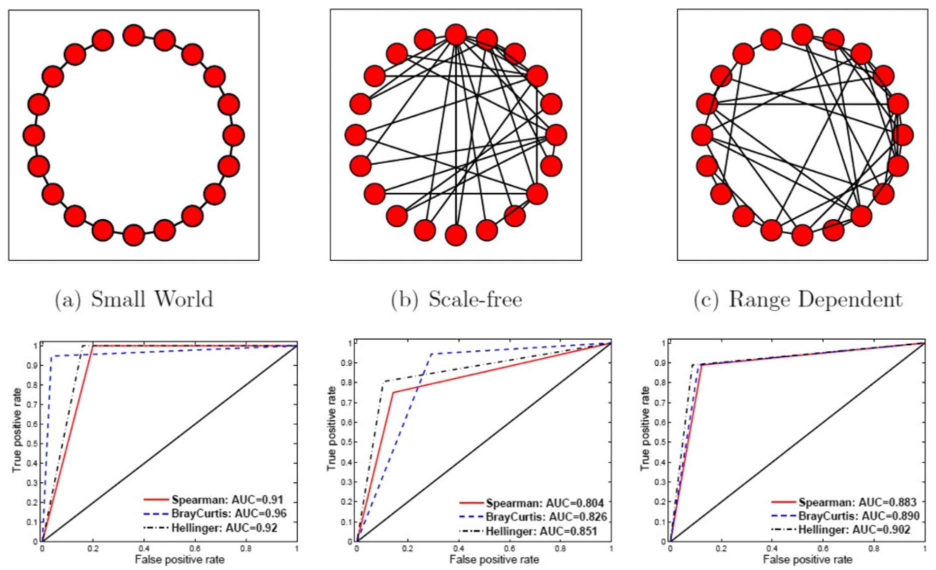
(a) Small World (b) Scale-free (c) Range Dependent

Fig. 1 Simulated networks. Results from simulation data with three network structures: *Top* panel: three different network structures; *Bottom* panel: predicted AUCs for each simulated network

data simulation. The count data are generated from Poisson distribution with $\mu = 3$. We repeated the computational experiments 50 times, the average AUC and their standard deviations with different similarity measures are reported in Table 1.

Table 1 indicates that the predicted AUCs increase and the performance gets better as the sample size increases. The proposed method performs reasonable well for different network structures. While Hellinger distance achieves the best result in small-world network, Spearman's rank correlation has the best performance with scale-free and range dependent networks. Therefore, Spearman's rank correlation is more robust with large networks generated from Poisson distribution, even though differences among different similarity measures are not always statistically significant.

Results with real body habitats data

There were 59 genera left for network construction after discarding genera with average abundance of less than two reads. We first calculated the similarity matrices from Spearman correlation coefficient, Hellinger, and Bray-Curtis distances respectively, and

Table 1 Predicted AUCs with different network structures and sample sizes for large networks with 500 nodes

n	Similarity	Small world	Scale-free	Range-dependent
50	Spearman	0.909 (±.015)	0.930 (±.035)	0.730 (±.021)
	Bray-Curtis	0.894 (±.014)	0.786 (±.017)	0.693 (±.016)
	Hellinger	0.936 (±.018)	0.844 (±.019)	0.709 (±.015)
100	Spearman	0.982 (±.009)	0.972 (±.020)	0.849 (±.014)
	Bray-Curtis	0.975 (±.011)	0.820 (±.016)	0.794 (±.019)
	Hellinger	0.986 (±.016)	0.883 (±.013)	0.812 (±.011)
200	Spearman	0.999 (±.012)	0.994 (±.015)	0.872 (±.015)
	Bray-Curtis	0.996 (±.009)	0.864 (±0.014)	0.828 (±.018)
	Hellinger	0.998 (±.006)	0.944 (±.012)	0.846 (±.015)

then determined the adjacency matrices with sparse inverse of the similarity matrices. The optimal λs for Spearman, Hellinger, and Bray-Curtis were then determined with stability selection. To further reduce the false-positive rate, the final adjacency matrix A was determined by the common edges of three different adjacency matrices. For comparison purpose, we also constructed a common network with all the samples from different habitats. Exact permutation test was used to compare networks constructed from gut, oral cavity, and skin with the common network. The differentiated network structures for skin, gut, and oral cavity were detected with a permutation P value of 0.05. Subnetwork structures were then identified with modularity maximization. Bacteria networks unique for skin, gut, and oral cavity are shown in Fig. 2.

Different colors of the nodes indicate different network modules, and the node size represents relative abundance: The larger the nodes, the higher the relative abundance of a genera. The edges indicate the direct coexistence (co-occurrence) between two genera. The skin subnetwork on the top panel of Fig. 2 has 4 modules colored in green, blue, orange, and red with 6, 6, 2, and 2 genera respectively. Several genera on the network are known to cause skin infections. For instance, *Streptococcus* on the green module is a well-known bacteria directly related to several skin infections including Impetigo, Cellulitis, and Erysipelas. *Actinomcyes* genus causes a chronic (slowly progressive) infection named Actinomycosis, and Fusobacterium genus has been known to cause tropical

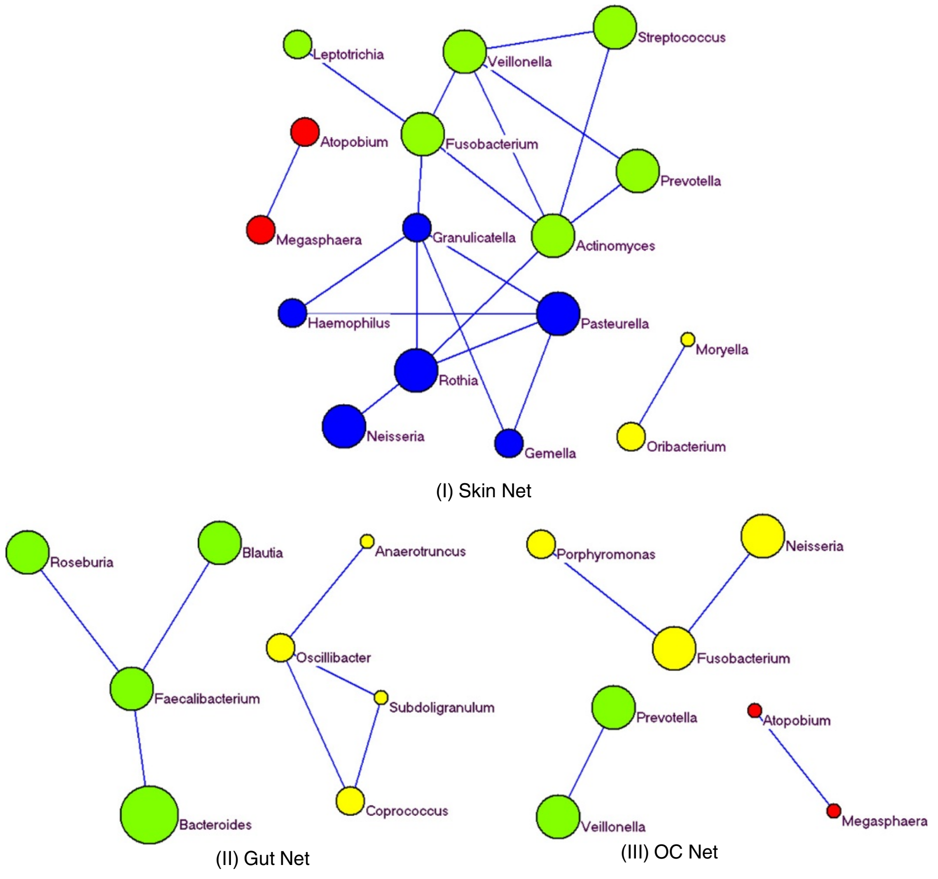

Fig. 2 Differentiated networks for Skin (I), Gut (II), and OC (III)

phagedenic ulcer (http://dermnetnz.org/bacterial/). More interestingly, the direct association of *Actinomcyes* and *Fusobacterium* on the green module was verified by a recent study experimentally [47]. The mixture and co-infection of two bacteria genera cause mastoiditis. In addition, *Pasteurella* genus has also been shown to cause skin disease [48]. Other genera directly connected to *Pasteurella* including *Rothia, Granulicatella, Gemella*, and *Haemophilus* may function together biologically and clinically through 'guilt by association'. Even though the co-occurrences among genera are only verified statistically, their biological and medical implications need to be further validated in a wet lab, our methods provide a guidance for investigators in their research.

The gut network constructed with 45 samples is shown in the bottom left panel (II) of Fig. 2. Two modules colored with green and yellow respectively are identified, each of them with 4 genera. All 4 high abundance genera and their interactions on the green module are associated with gut related diseases. *Faecalibacterium* and *Bacteroidetes* genera are both associated with type 2 diabetics and obeisity [49]. Another recently study also indicates that *Blautia* and *Faecalibacterium* vary together during antibiotic therapy [50], and it has been shown that both *Roseburia* and *Faecalibacterium* have lower abundance in patients with Crohn's Disease compared with their healthy siblings [51]. All these studies support our results that *Faecalibacterium, Blautia, Roseburia*, and *Bacteroidetes* interact with each other. However, interactions among the 4 genera (*Subdoligranulum, Coprococcus, Anaerotruncus*, and *Oscillibacter*) with lower abundance in the yellow module have not been well studied, even though individual genus has been reported in recent literature.

Bacteria genera in oral cavity (OC) play a key role in mouth infections and periodontal diseases. Oral cavity network built with genera from 54 oral cavity samples is shown in the bottom right panel (III) of Fig. 2. Three modules colored with green, red, and yellow were identified with 2, 2, and 3 genera on each subnetwork. Interactions and co-occurrences are identified among both high abundance genera (*Fusobacterium, Veillonella*, Neisseria, Prevotella) and low abundance genera (Atopobium, and Megasphaera) in OC. Genera such as Fusobacterium, Neisseria, Porphyromonas, and *Prevotella* have been known to be significantly different in abundance with different clinical conditions and disease status [52, 53]. The co-occurrences between *Porphyromonas* and *Fusobacterium* on the yellow module has been verified through a mouse model experimentally [54]. However, co-occurrences among *Fusobacterium, Porphyromonas*, and *Neisseria* together have not been explored. One interesting finding is the common interactions between *Prevotella* and *Veillonella* and between *Megasphaera* and *Atopobium* in the oral cavity and skin, indicating that some interactions and co-occurrences may be shared at different body habitats. Therefore, co-occurrence network analysis with proposed approach is useful for determining novel biological interactions that may help to decipher the structure of complex microbial communities. It is also useful for systematically exploring co-existence patterns in big metegenomics data that standard tools may fail to detect.

Conclusions

We have developed a systematic approach for constructing networks and detecting subnetwork structures with metagenomic count data. Our contributions are in two areas: (1) we adapt distance measures between samples from ecology to compute similarity between taxa, and (2) we extend sparse inverse covariance methods for Gaussian models to determine high-order interactions with general similarity matrices. Based on both simulated and real data, our method can identify true and biologically important interactions and associations with limited computational experiments. One advantage of our approach is that it detects the partial (high-order) correlations among the taxa. Unlike pairwise correlation, partial correlation measures the conditional dependency between two taxa given the effect of rest taxa being removed or fixed. Therefore, networks constructed from partial correlation usually have lower false positive connections than those from pairwise correlation. In addition, modularity function maximization for structure detection in MetaNet automatically determines the number of network clusters, while other popular approaches such as K-means and hierarchical clustering require the number of clusters or a cutoff point to be predefined. Even though MetaNet is slightly computational intensive when comparing to the popular pairwise approach, it only takes minutes to construct a network with thousands of nodes. While our method has been developed for 16S rRNA sequencing data from human body, it can be applied to 16S sequencing data from other organisms. It may also be used to analyze whole metagenomic shotgun sequencing or RNA-seq, as long as the sequences are properly aligned. Note that our method constructs networks solely based on their statistical similarity or dissimilarity among taxa, the biological significance of the results has to be further validated in a web lab. Future works wledge and pathway information into our network constructions.

Appendix: network statistics

Given the adjacency matrix A and the their subnetworks, different statistics has been defined to describe the network and taxa. The most important network statistics are [55]:

- Degree: number of links connected to a taxon (node) i, $K_i = \sum_j a_{ij}$.
- Number of triangles around a taxon i: $T_i = \frac{1}{2} \sum_{j,k} a_{ij} a_{ik} a_{jk}$.
- Clustering coefficient: Measuring the segregation of a network, $C = \frac{1}{m} \sum_i C_i = \frac{1}{m} \sum_i \frac{2T_i}{K_i(K_i-1)}$.
- Participation coefficient of taxon i, $Y_i = 1 - \sum_{m \in M} \left(\frac{K_i(m)}{K_i} \right)^2$, where $k_i(m)$ is the within-module degree of taxon i in module m.

We will apply network statistics to rank the bacteria taxa. The larger the network statistics, the stronger connectivity the taxa, and the more important the taxa statistically. The biological importance of those taxa with larger network statistics can be further validated in the lab.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZL conceptualized and designed method, developed the software, and wrote the manuscript. SL and SP helped in method design and manuscript writing and revised the manuscript critically. All authors read and approved the final manuscript.

Acknowledgements

We thank the Associate Editor and anonymous referees for their constructive comments. This work was partially supported by the National Science Foundation (NSF) grants DMS-222381 (ZL) and DMS-1220772 (SL).

Author details

[1]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA.
[2]Department of Statistics, The Ohio State University, Columbus, OH 43210, USA.

Received: 5 June 2015 Accepted: 18 November 2015
Published online: 12 December 2015

References

1. Turnbaugh P, Ley R, Hamady M, Fraser-Liggett C, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449:804–10. doi:10.1038/nature06244.
2. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. Brief Bioinform. 2012;13(6):669–81. doi:10.1093/bib/bbs054.
3. Keller A, Horn H, Förster F, Schultz J. Computational integration of genomic traits into 16S rDNA microbiota sequencing studies. Gene. 2014;549(1):186–91.
4. Wooley J, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010;6(2):e1000667. doi:10.1371/journal.pcbi.1000667.
5. Huson D, Auch A, Qi J, Schuster S. Megan analysis of metagenomic data. Genome Res. 2007;17:377–86.
6. Schloss P, Westcott S, Ryabin T, Hall J, Hartmann M, Hollister EB, et al. Introducing mothur: opensource, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75:7537–41.
7. Lozupone C, Lladser M, Knights D, Stombaugh J, Knight R. Unifrac: an effective distance metric for microbial community comparison. ISME J. 2010;5:169172. doi:10.1128/aem.01541–09.
8. Caporaso J, Kuczynski J, Stombaugh J, Bittinger K, Bushman F, Costello EK, et al. Qiime allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–36. doi:10.1038/nmeth.f.303.
9. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene data base project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(D1):D590–96.
10. Huson D, Mitra S, Weber N, Ruscheweyh H, Schuster S. Integrative analysis of environmental sequences using megan4. Genome Res. 2011;21:1552–60.
11. Brady A, Salzberg S. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. Nat Methods. 2009;6:673–76.
12. Rosen G, Reichenberger E, Rosenfeld A. Nbc: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 2010;27:127–29.
13. White J, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009;5:1000352. doi:10.1038/nmeth.f.303.
14. Liu Z, Hsiao W, Cantarel B, Drbek E, Fraser-Liggett C. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. Bioinformatics. 2011;27(23):3242–49. doi:10.1093/bioinformatics/btr547.
15. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15(3):382–92. doi:10.1016/j.chom.2014.02.005.
16. Boutin S, Bernatchez L, Audet C, Derôme N. Network analysis highlights complex interactions between pathogen, host and commensal microbiota. PLoS One. 2013;8(12):e84772. doi:10.1371/journal.pone.0084772.
17. Hurwitz BL, Westveld AH, Brum JR, Sullivan MB. Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. Proc Natl Acad Sci U S A. 2014;111(29):10714–9. doi:10.1073/pnas.1319778111.
18. Liu Z, Sun F, Braun J, McGovern D, Piantadosi S. Multilevel Regularized Regression for Simultaneous Taxa Selection and Network Construction with Metagenomic Count Data. Bioinformatics. 2015;31(7):1067–74.
19. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol. 2008;4(8):e1000117. doi:10.1371/journal.pcbi.1000117.
20. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13.
21. Krämer N, Schäfer J, Boulesteix AL. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics. 2009;10:384.
22. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. Nature. 2002;420(6912):190–3.
23. Guimera R, Nunes Amaral LA. Functional cartography of complex metabolic networks. Nature. 2005;433(7028):895–900.
24. Kreimer A, Borenstein E, Gophna U, Ruppin E. The evolution of modularity in bacterial metabolic networks. Proc Natl Acad Sci U S A. 2008;105(19):6976–81.
25. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
26. Allen GI, Liu Z. A Local Poisson Graphical Model for inferring networks from sequencing data. IEEE Trans Nanobioscience. 2013;12(3):189–98.
27. Zhang L, Mallick BK. Inferring gene networks from discrete expression data. Biostatistics. 2013;14(4):708–22.
28. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50.
29. Barberán A, Bates ST, Casamayor EO, Fierer N. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J. 2012;6(2):343–51.
30. McMurdie PJ. Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8(4):e61217.
31. Oberauner L, Zachow C, Lackner S, Högenauer C, Smolle KH, Berg G. The ignored diversity: complex bacterial communities in intensive care units revealed by 16S pyrosequencing. Sci Rep. 2013;3:1413.
32. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2013;489(7415):220–30.
33. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687.
34. Clarke KR, Somerfield PJ, Chapman MG. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. J Exp Mar Biol Ecol. 2006;330:55–80.

35. Liu H, Han F, Yuan M, Lafferty J, Wasserman L. High dimensional semiparametric gaussian copula graphical models. 2012. Technical report, Johns Hopkins University.
36. Croux C, Dehon C. Influence functions of the Spearman and Kendall Correlation measures. Stat Methods Appl. 2010;19(4):497–515.
37. Legendre P, Gallagher ED. Ecologically meaningful transformations for ordination of species data. Oecologia. 2001;129:271–80.
38. Mitra S, Gilbert JA, Field D, Huson DH. Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. ISME J. 2010;4(10):1236–42.
39. Parks DH, Beiko RG. Measuring community similarity with phylogenetic networks. Mol Biol Evol. 2012;29(12):3947–58.
40. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–41.
41. Banerjee O, El Ghaoui L. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. J. Mach. Learn. Res. 2008;9:485–516.
42. Liu H, Roeder K, Wasserman L. Stability approach to regularization selection for high dimensional graphical models. Advances Neural Inf Process Syst. 2010;1432–1440.
43. Meinshausen N, Bühlmann P. Stability selection. JR Stat Soc Series B Stat Methodol. 2010;72:417–73.
44. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;10:P10008.
45. Newman ME. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E Stat Nonlin Soft Matter Phys. 2006;74(3 Pt 2):036104.
46. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science. 2009;326:1694–97.
47. Salipante SJ1, Hoogestraat DR, Abbott AN, Sengupta DJ, Cummings LA, Butler-Wu SM, et al. Co-infection of Fusobacterium nucleatum and Actinomyces israelii in mastoiditis diagnosed by next-generation DNA sequencing. J Clin Microbiol. 2014;52(5):1789-92.
48. Hazelton BJ, Axt MW, Jones CA. Pasteurella canis osteoarticular infections in childhood: review of bone and joint infections due to pasteurella species over 10 years at a tertiary pediatric hospital and in the literature. J Pediatr Orthop. 2013;33(3):e34–8. doi:10.1097/BPO.0b013e318287ffe6. Review. PMID: 23482278.
49. Remely M, Aumueller E, Jahn D, Hippe B, Brath H, Haslberger AG. Microbiota and epigenetic regulation of inflammatory mediators in type 2 diabetes and obesity. Benef Microbes. 2014;5(1):33–43.
50. Ferrer M, Martins Dos Santos VA, Ott SJ, Moya A. Gut microbiota disturbance during antibiotic therapy: A multi-omic approach. Gut Microbes. 2013;5(1):64–70.
51. Hedin CR, McCarthy NE, Louis P, Farquharson FM, McCartney S, Taylor K, et al. Altered intestinal microbiota and blood T cell phenotype are shared by patients with Crohn's disease and their unaffected siblings. Gut. 2014. doi:10.1136/gutjnl-2013-306226.
52. Sassone LM, Fidel RA, Faveri M, Figueiredo L, Fidel SR, Feres M. A microbiological profile of unexposed and exposed pulp space of primary endodontic infections by checkerboard DNA-DNA hybridization. J Endod. 2012;38(7):889–93.
53. Nascimento Cd, Pita MS, Fernandes FH, Pedrazzi V. de Albuquerque Junior RF, Ribeiro RF. Bacterial adhesion on the titanium and zirconia abutment surfaces. Clin Oral Implants Res. 2014;25(3):337–43. doi:10.1111/clr.12093.
54. Metzger Z, Lin YY, Dimeo F, Ambrose WW, Trope M, Arnold RR. Synergistic pathogenicity of Porphyromonas gingivalis and Fusobacterium nucleatum in the mouse subcutaneous chamber model. J Endod. 2009;35(1):86–94.
55. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. Neuroimage. 2010;52(3):1059–69.