Methods 83 (2015) 28–35

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

# Personalized microbial network inference via co-regularized spectral clustering

Sultan Imangaliyev [a,b,c,*], Bart Keijser [a,b], Wim Crielaard [a,c], Evgeni Tsivtsivadze [a,b]

[a] Top Institute Food and Nutrition, Wageningen, The Netherlands
[b] Research Group Microbiology and Systems Biology, TNO Earth, Environmental and Life Sciences, Zeist, The Netherlands
[c] Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

*Article history:*
Received 29 January 2015
Accepted 24 March 2015
Available online 2 April 2015

*Keywords:*
Spectral clustering
Personalized network
Metagenomics
Oral health

ABSTRACT

We use Human Microbiome Project (HMP) cohort (Peterson et al., 2009) to infer personalized oral microbial networks of healthy individuals. To determine clustering of individuals with similar microbial profiles, co-regularized spectral clustering algorithm is applied to the dataset. For each cluster we discovered, we compute co-occurrence relationships among the microbial species that determine microbial network per cluster of individuals. The results of our study suggest that there are several differences in microbial interactions on personalized network level in healthy oral samples acquired from various niches. Based on the results of co-regularized spectral clustering we discover two groups of individuals with different topology of their microbial interaction network. The results of microbial network inference suggest that niche-wise interactions are different in these two groups. Our study shows that healthy individuals have different microbial clusters according to their oral microbiota. Such personalized microbial networks open a better understanding of the microbial ecology of healthy oral cavities and new possibilities for future targeted medication. The scripts written in scientific Python and in Matlab, which were used for network visualization, are provided for download on the website http://learning-machines.com/.



## 1. Introduction and background

Although oral health has a great influence on an individual's quality of life [2], dental medicine has focused mainly on disease states, comparing healthy and unhealthy individuals [3,4]. While this approach has proved its practical usefulness, it has not explained why certain healthy individuals are prone to disease than others. Recent changes in computational biology methods now provide oral health scientists with access to large amounts of -omics data, which promise potentially novel insights into the underlying patterns of the healthy oral state. A rapid reduction in costs and the rapid development of computational algorithms now allow oral microbiota to be analyzed at a metagenomics level. However, mining metagenomics datasets is not a trivial task, and there are many questions on how to get a general overview of microbial communities on the OTU (Operational Taxonomical Unit) level.

We aim to answer an important question in the study of metagenomics: whether microbial species are present either in a single homogenous population in their own ecological niche, or in several distinct interacting communities. Understanding such interactions could lead to more effective treatment strategies of dental diseases in which microorganisms play role in their progress. We see three main challenges on the way to achieve goal of having novel insights in microbial network interactions of bacterial communities. The first concerns the best method of finding groups or clusters hidden behind the data. Publicly available datasets such as those collected during the Human Microbiome Project (HMP) [1] provide excellent opportunities to test the applicability of various clustering approaches. For instance, it is possible to incorporate existing knowledge about phylogenetic tree-based distances [5]. However, a single method can have disadvantages, and it is sometimes better to combine several clustering algorithms to get more accurate results [6].

The second challenge is that although clustering by itself provides better understanding of the groupings of microbial communities, a more comprehensible means of visualization is frequently required. A microbial network is one such mean because a network representation of bacterial community gives an intuitively

* Corresponding author at: Research Group Microbiology and Systems Biology, TNO Earth, Environmental and Life Sciences, Zeist, The Netherlands.
*E-mail address:* sultan.imangaliyev@tno.nl (S. Imangaliyev).

http://dx.doi.org/10.1016/j.ymeth.2015.03.017
1046-2023/

better way of understanding interactions between microbial groups. This understanding may lead to better treatment procedures and results. When we talk about networks we mean all biological networks in a wide range. They include protein, gene or microbial interaction networks. Such networks could be constructed at multiple levels, such as cellular, ecological and supra-organismal [7]. Biological networks provide many advantages like, for example, identification of early-warning signals associated with the critical transition in disease progression [8]. By analyzing them it is also possible to discover robust and specific biomarkers of disease [9]. Due to their visual simplicity, biological networks reveal important components of biological systems such as essential genes by identifying important topology nodes such as bottlenecks in regulatory networks [10].

Thirdly, it is possible to build a network based on a single individual's data and apply specific treatment per individual based on his or her -omics profile. Such personalized approach allows applying the right treatment on the right cause at the right moment of time for the particular patient [11]. Thus, clustering algorithms can be combined with network visualization methods to use personalized medicine in any medical field.

Taking into account all these challenges, one might wonder if it is practically possible to combine clustering algorithms, biological networks, complex metagenomics data, and personalized clinical treatment. To support our claim that the answer is "Yes", we wished to establish whether there are any differences in microbial interactions on personalized network level in healthy oral samples acquired from various niches during HMP study. Firstly, we were inspired by Huttenhower et al. [13] who demonstrated the value of Nearest Neighbor Networks (NNN) for generating clusters of genes with similar expression profiles. NNN is a graph-based algorithm which prompted us to use the graph-theory-based clustering method to reveal clusters in metagenomics data. To reveal co-occurence relationships among OTU, we used adapted spectral clustering algorithm [14], particularly suited for complex metagenome data [15] because it outperforms other clustering algorithms such as *K*-means and hierarchical clustering. Furthermore, Faust et al. [16] provided a concise review of co-occurence and correlation networks among bacterial communities derived from 16 s pyrosequencing data. On the basis of HMP data in their other work, Faust et al. [17] also demonstrated how and what kind of co-occurrence relationships can be found in microbial networks in healthy individuals.

A major difference between our study and that of other authors is personalization of a microbial network and the use of state-of-the-art unsupervised machine learning methods. Unlike in previous studies, we first merged niche-based samples to represent human individuals. Only then did we apply statistical machine learning algorithms to stratify individuals according to their oral microbiota. Once such personalized stratification has been completed, we clustered microbial species per group to examine them more closely on different microbial networks that can be visualized in various ways, such as these described by Tumminello et al. [18] or by Shannon et al. [19].

## 2. Materials and methods

### 2.1. Co-regularized spectral clustering algorithm

In this paper we used an adapted version of the multi-view clustering method described in [15]. Consider we are given a dataset containing multiple representations. Let $X^{(v)} = \left\{x_i^{(v)}\right\}_{i=1}^{n}$. Note that here superscript $v$ denotes the representation for a single view. Let $A^{(v)}$ denote an adjacency matrix of the graph constructed using the data representation in a view $v$. We can write the normalized Laplacian matrix as $L^{(v)} = D^{(v)-1/2}A^{(v)}D^{(v)-1/2}$, where $D^{(v)}$ is the corresponding degree matrix. Following [20] the standard special clustering problem (or single view spectral clustering [21]) solves the optimization problem.

$$\min_{Q^{(v)}\in\mathcal{R}^{n\times c}} tr\left(Q^{(v)T}L^{(v)}Q^{(v)}\right), \quad \text{s.t.} \quad Q^{(v)T}Q^{(v)} = I \tag{1}$$

where $Q^{(v)} \in \mathcal{R}^{n\times c}$ denotes the cluster assignment matrix and $c$ is number of predefined clusters. In standard spectral clustering the final cluster membership is obtained by applying the $k$-means algorithm on the rows of the matrix $Q^{(v)}$.

In our work we follow derivation described in [15] to obtain cluster assignment matrix

$$\begin{aligned} J_c &= \sum_{v=1}^{M}\sum_{l=1}^{c}\|H^{(v)}\mathbf{q}_l^{(v)} - \mathbf{q}_l^{(v)}\|^2 \\ &= \sum_{v=1}^{M}\sum_{l=1}^{c}\left[\mathbf{q}_l^{(v)T}((H^{(v)}-I)^T(H^{(v)}-I))\mathbf{q}_l^{(v)}\right] \\ &= tr\left[\mathbf{Q}^T((\mathbf{H}-\mathbf{I})^T(\mathbf{H}-\mathbf{I}))\mathbf{Q}\right], \end{aligned}$$

where $\mathbf{Q}$ is a $(Mn \times c)$ matrix containing the cluster assignments for all views and $\mathbf{H}$ is a $(Mn \times Mn)$ matrix containing predictions of the linear classifiers. Thus, the optimization problem we solve to determine cluster assignment matrices for all views is

$$\min_{\mathbf{Q}\in\mathcal{R}^{Mn\times c}} tr\left[\mathbf{Q}^T((\mathbf{H}-\mathbf{I})^T(\mathbf{H}-\mathbf{I}))\mathbf{Q}\right] \quad \text{s.t.} \quad \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{2}$$

The above problem is closely related to the standard spectral clustering and the solutions are given by top-$c$ eigenvectors of the matrix $\mathbf{L} = (\mathbf{H}-\mathbf{I})^T(\mathbf{H}-\mathbf{I})$.

### 2.2. Dataset description and preprocessing

The selected clustering approach was tested on publicly available dataset. The dataset was downloaded from Human Microbiome Project website [22,1]. Namely, we used V35 Mothur Output File originally containing 27,483 OTU counts for 5372 samples collected from eighteen body locations. We subsampled the dataset so that it includes only nine oral samples referring to following niches: *Saliva*, *Buccal Mucosa* (cheek), *keratinized gingiva* (gums), *Hard Palate*, *Palatine Tonsils*, *Tongue Dorsum*, *Throat*, *Supra-* and *Subgingival Dental Plaque* (tooth biofilm above and below the gum). The choice of those particular sites is determined by clinical relevance in understanding mechanisms of oral diseases such as caries, gingivitis and periodontitis.

Then we created a dataset which represents individual persons by their oral samples. In this dataset each row is constructed by stacking all nine oral niches together so that it would be possible to apply clustering on individual level, not on OTU level. Not all individuals had *all* nine oral niches sampled. Since we were not interested in such individuals, we did not include them in the personalized dataset. Some individuals were sampled in a few visits. For such individuals we took only samples collected during the first visit. As a result we obtained a dataset including 177 individuals. To improve speed of calculations and to consider only most abundant OTU, we reduced the amount of features by removing those in which amount of non zero counts per feature was below 60 (roughly one third amount of individuals). Resulting dataset contained 635 most abundant OTU found in all 9 locations for 177 individuals. Then, we normalized the dataset by forcing row-wise sum to be equal to one. Next, all features were linearly scaled between 0 and 100.