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.

2.3. Personalized network inference and visualization

Network inference and visualization procedure consists of three subsequent steps. In the first step we stack all oral samples together such that each individual will be represented by his oral microbiota as a row in the dataset. The second step is to cluster individuals based on their microbiota to find groups with similar microbial composition. Also, in this step we transpose the dataset per each cluster we found. Next, we again perform clustering but now on OTU level. Clustering and construction of co-occurence matrix is based on the work of Luxburg [14] and Tsivtsivadze et al. [15]. In the third step we visualize resulting co-occurence matrices as a network of interactions. In such networks OTU are represented by network nodes while similarities between them are represented by connecting edges. Short summary of the method workflow is depicted in Fig. 1.

2.4. Methods summary

To summarize the method description we point out following details. As an input for our method we use HMP data on OTU counts. As an output we get a network of OTUs according to the personalized profile of individuals. This network is visualized by a figure where OTUs are nodes and their mutual co-occurence scores are edges. We used HMP database [22,1] to get OTU counts dataset. As a clustering tool we used spectral clustering based on the work of Luxburg [14] and Tsivtsivadze et al. [15]. Visualization of clustering as a network was done via using two different tools, such as filtering complex information systems by Tumminello et al. [18] and Cytoscape by Shannon et al. [19]. 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/.

3. Results

3.1. Clustering of individuals and defining groups with similar microbial composition

To construct co-occurence matrix per individual, the co-regularized algorithm was applied to the preprocessed dataset. We ran clustering with predefined number of clusters ranging from 2

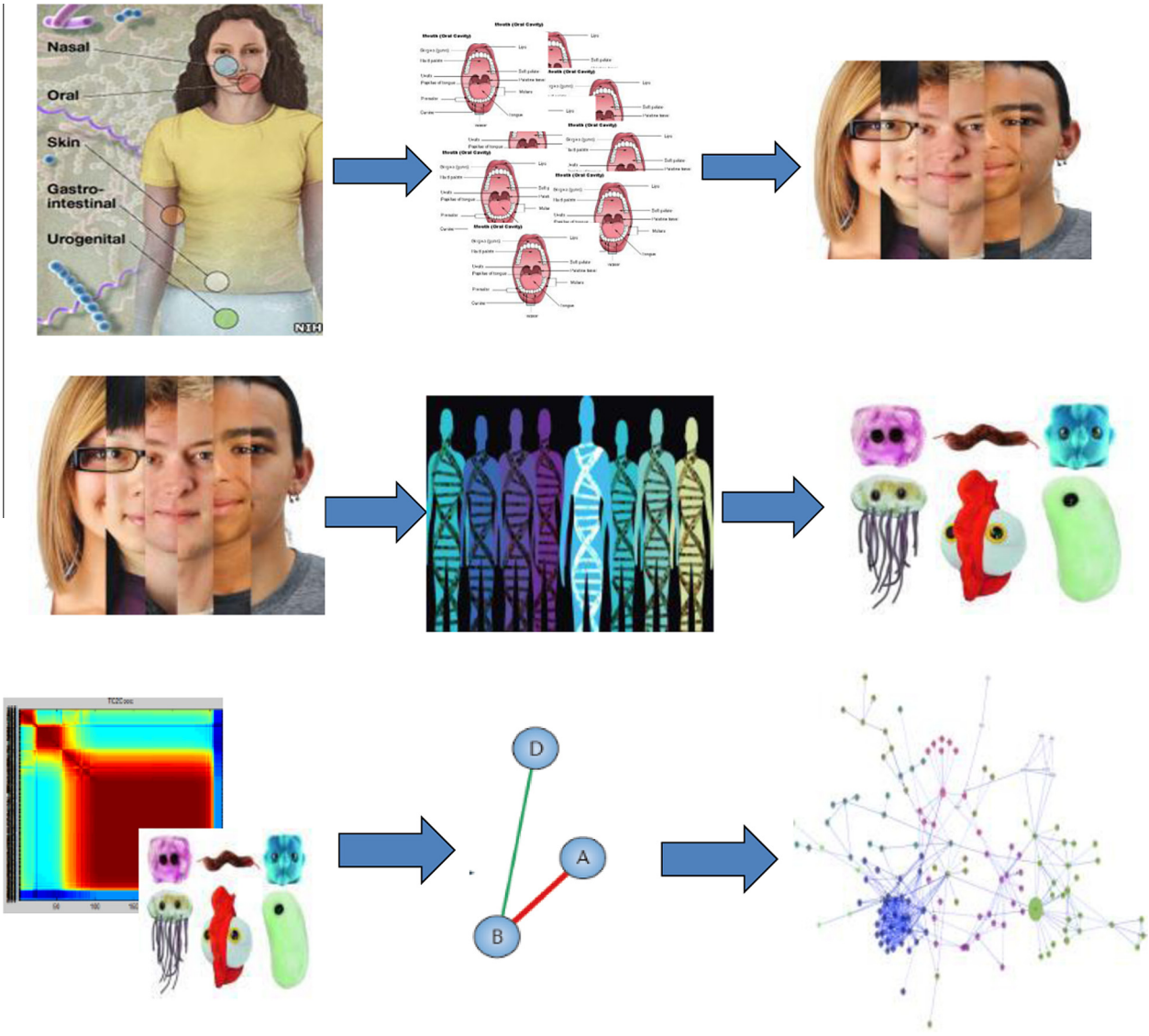

Fig. 1. Workflow of the method. The workflow consists of three steps. In the first step we stack all oral samples together such that each individual will be represented by his oral microbiota. The second step is to cluster individuals based on their microbiota to find groups with similar microbial composition, and later to transpose the dataset per each cluster to perform clustering again, but on OTU level. In the third step we visualize resulting co-occurence matrices as a network of interactions.

to 5. Next, we compute the co-occurrence matrix according to consensus algorithm described by Grotkjaer et. al. [23]. Afterwards, resulting co-occurence matrix was rearranged so that the clusters of co-occurences would be seen immediately. Heatmap plot of the ordered co-occurence matrix is depicted in Fig. 2.

The co-occurence matrix is a representation which summarizes multiple clustering results into a single array. In symmetric case it has as many rows and columns as examples in the clustered dataset. Element $A_{i,j}$ of a co-occurence matrix A represents co-occurence score of examples i and j. Co-occurence score is calculated as a total count of how many times these two examples were labeled with the same clustering label divided by the total count of clustering runs. Obviously, if two examples are labeled by the same arbitrary label in all simulations then co-occurence score has a value of 1. For the same reason all diagonal elements of the matrix are always equal to 1. Normally co-occurence score varies between 1 and 0, and some sorting of both rows and columns is required for the better visualization. The best way to visualize co-occurence matrix is to plot a heatmap of the sorted co-occurence matrix where axes represent indices of examples. This was done in figures below.

Based on Fig. 2 we see two groups of individuals. Group 1 clusters about 61 individuals together and group 2 clusters 116 individuals. It is assumed that within each of these two groups individuals share similar microbial composition while individuals in different groups have distinct microbial composition. Therefore, we can split initial dataset into two groups and apply co-regularized spectral clustering on OTU per each group of individuals treating them as separate datasets. These datasets were used in the next experiment reported in the following subsection.

3.2. Clustering of microbial species per cluster of similar individuals

After clustering individuals, we need to examine a difference in microbial compositions between group 1 and group 2. Therefore, we repeated co-regularized spectral clustering experiment as we did in previous subsection. However, the datasets were transposed so that the clustering was done on OTU level. Figs. 3 and 4 depict heatmaps of sorted co-occurence matrices of these experiments.

Figs. 3 and 4 allow us to zoom into the differences between groups. We see that the plot depicted on Fig. 3 is different from the one depicted on Fig. 4. On the group 1 plot we do not see clearly distinctive clusters of OTU although we see a big group in the center of the Fig. 3. The Fig. 4, however, has two distinct OTU clusters with roughly equal size. Thus, we expect the networks visualized from such personalized datasets to be very different from each other. We expect to have two distinct clusters in network for group 2 and one cluster for the network visualized from the data of group 1. Those assumptions were checked and reported in the next subsection.

3.3. Personalized network visualization of bacterial species within each cluster

Network visualization requires converting pairwise similarity matrix to a 2-dimensional graph. Such graph can be plotted so that similar nodes will be closer to each other and dissimilar nodes will be further away from each other. The connections are visualized by edges. In this work we used algorithm by Tumminello et al. [18]. The algorithm filters complex datasets by extracting a subgraph of representative links, giving filtered graphs that preserve the hierarchical organization of the minimum spanning tree but containing a larger amount of information in their internal structure. Network plots of co-occurence matrices for group 1 and group 2 are depicted in Fig. 5. As we expected, there are indeed two distinct subnetworks in network for group 2 and one main network for the network visualized from the data of group 1.

To give a better visual impression to a reader we also provide figures with zoomed in network representation for each subnetwork in each group. The network nodes are formatted according to niche so that niche-wise interactions are seen in a more convincing way. Fig. 6 depicts main network of group 1, Fig. 7 depicts subnetwork 1 of group 2, and Fig. 8 depicts subnetwork 2 of group 2.

Network of group 1 does not show any obvious patterns in network topology. There are no obvious hubs and bottlenecks. All OTU are connected with each other without any dominating niche or OTU name. Most frequent OTU are *Prevotellaceae, Flavobacteriaceae, Lachnospiraceae, Leptotrichiaceae, Veillonellaceae, Neisseriaceae, Actinomycetaceae, Campylobacteraceae, Porphyromonadaceae*.

OTU-wise and niche-wise networks of subnetwork 1 of group 2 depict a cluster topology with several OTU closely connected to

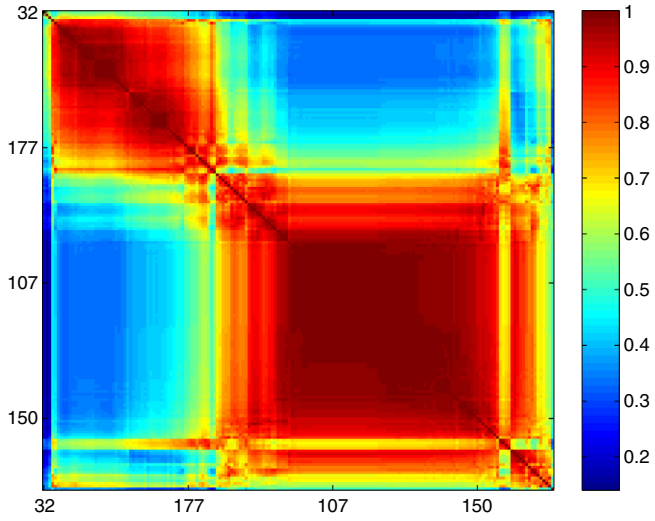

Fig. 2. Sorted co-occurence matrix of clustering of individuals. This plot displays two groups of individuals. Group 1 clusters about 61 individuals together and it is located in the upper left corner of the heatmap. Group 2 clusters about 116 individuals together and it is located in the lower right corner. Numbers on axes are symmetric and they represent sorted indices of individuals.

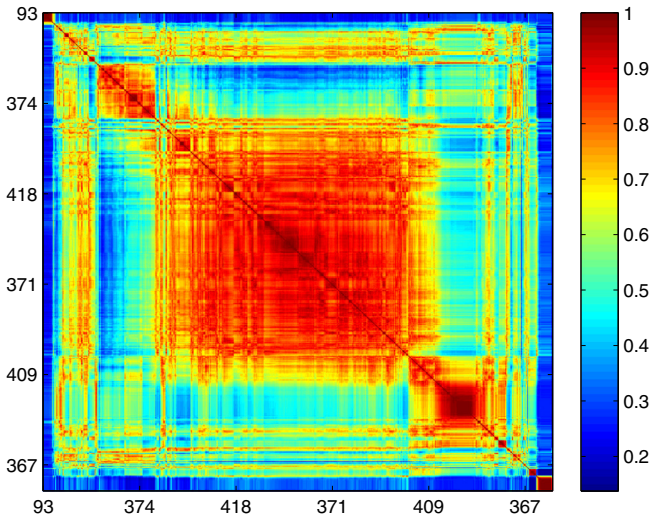

Fig. 3. Sorted co-occurence matrix of clustering of OTU per group 1. This plot do not display clearly distinctive clusters of OTU having only a large cluster of OTU in the center and two smaller clusters in the upper left and lower right corners. Numbers on axes are symmetric and they represent sorted indices of OTU.

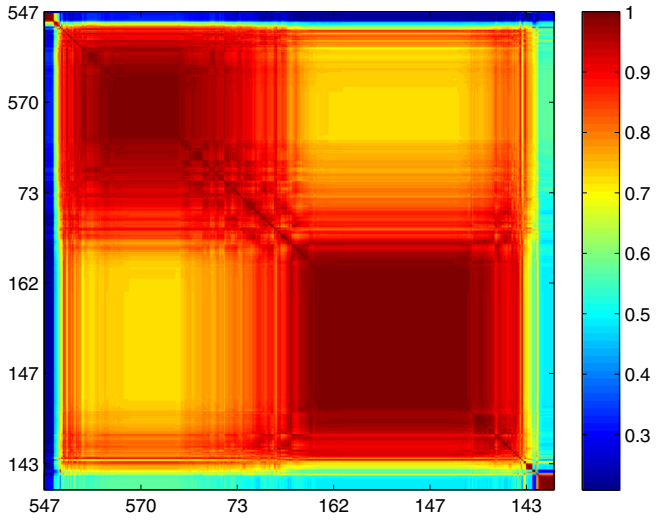

Fig. 4. Sorted co-occurence matrix of clustering of OTU per group 2. This plot displays two distinct OTU clusters with roughly equal size. Similarity scores within clusters are very high suggesting that intracluster similarity between OTU is high. Numbers on axes are symmetric and they represent sorted indices of OTU.

each other. *Prevotellaceae, Flavobacteriaceae, Lachnospiraceae* dominate in this subnetwork. Niche-wise, most of the OTU belong to the following niches: *Muccosa, Throat/Tonsils, Supra/Sub Plaque*. Hub OTU in the center of the dense cluster belongs *Supra/Sub Plaque* OTU which is outside of top 10 OTU list.

OTU-wise and niche-wise networks of subnetwork 2 of group 2 show topology similar to subnetwork 1. There is a cluster of similar OTU and several others which do not belong to this cluster but have high similarity. Like in subnetwork 1, *Prevotellaceae, Flavobacteriaceae, Lachnospiraceae* dominate in this subnetwork. Niche-wise, most of the OTU belong to the following niches: *Throat/Tonsils, Supra/Sub Plaque, Tongue, Hard Palate*. Hub OTU in the center of the dense cluster belongs to *Supra/Sub Plaque* OTU which is outside of top 10 OTU list.

Besides network visualization algorithm described in [18], we also used Cytoscape as a tool to visualize niche-wise interactions [19]. We used the hierarchical layout algorithm for visualization, because it is good for representing main direction or flow within a network. Nodes are placed in hierarchically arranged layers and the ordering of the nodes within each layer is chosen in such a way that minimizes the number of edge crossings. Plots of network visualized from niche-wise lumped co-occurence matrices of clustering OTU per group 1 and group 2 are depicted in Figs. 9 and 10.

To plot these networks we lumped original OTU similarity matrices to the niche-wise levels. For each OTU we have the assignment to particular niche. We calculated sum for all normalized OTU abundances per each pair-wise niche similarity score and divided these sum values by total count per each niche-wise pair so that niche-wise similarities are normalized and can be comparable with each other. As a result, we get 9 by 9 niche-wise co-occurence score matrix where each element represents similarity between all possible pair-wise combinations of all 9 niches. Next, we exported these matrices per each group and visualized network representation of similarities. For better visualization purposes only similarities above certain threshold levels are depicted in the networks. As a similarity measurement we use a co-occurence score, assuming that maximum co-occurence score of 1.0 indicates perfect similarity and, vice versa, minimum co-occurence score of 0.0 indicates perfect dissimilarity.

Fig. 9 depicts a network for group 1 without a clear clustering topology and without a clearly observed single hub, but Saliva, Tongue Dorsum and Gingiva have more connections with Palatine Tonsils and Subgingival Plaque. Almost all nodes have connections with other nodes and with comparable similarity measurement values around 0.63–0.67. Contrary to that, network for group 2 depicted on Fig. 10 has two distinct clusters and number of connections for each node is less that that in group 1 network. Comparing to group 1 both plaque nodes in group 2 have stronger connections with each other via Buccal Mucosa. Also in left-hand cluster of group 2 Palatine Tonsils was not included because its similarity was below the threshold and thus lower than for the other nodes. Again, as in Fig. 9, Saliva and Tongue Dorsum have more intermediate connections than other niches. Similarity measurements are in average higher that that for group 1 and vary in the interval of 0.83–0.90.

4. Discussion

The results of our study of healthy oral samples acquired from various niches during HMP study suggest that microbial interactions on personalized network level differ in several ways. Co-regularized spectral clustering showed two groups of individuals with a different topology of their microbial interaction network; the microbial network suggested that niche-wise interactions differed between them. According to their oral microbiota, healthy individuals have different microbial clusters.

Explaining differences between revealed clusters of individuals would be possible if we had an access to additional metadata of individuals associated with their dental hygiene status and oral

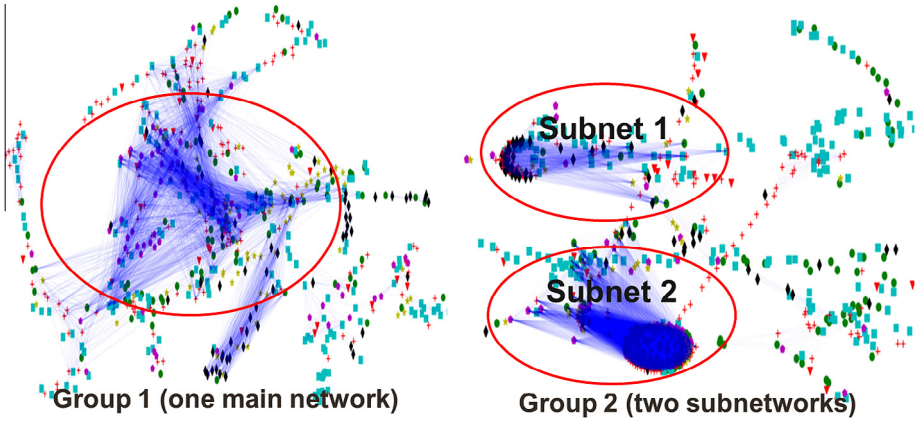

Fig. 5. Networks visualized from co-occurence matrices of clustering OTU per each group. This plot displays two distinct subnetworks in network for group 2 (Right plot) and one main network for the network visualized from the data of group 1 (Left plot).

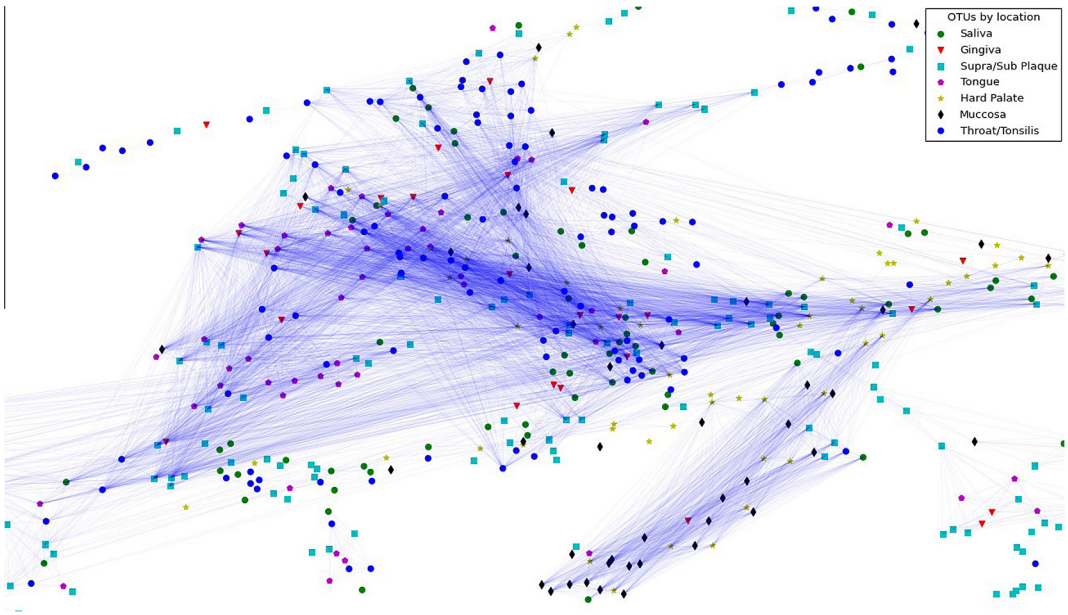

Fig. 6. Zoomed network visualized from co-occurence matrix of clustering OTU in group 1. This plot displays main network of group 1. Nodes are colored according to the corresponding niche.

Fig. 7. Zoomed subnetwork 1 visualized from co-occurence matrix of clustering OTU in group 2. This plot displays subnetwork 1 of group 2. Nodes are colored according to the corresponding niche.

health related lifestyle habits. In our case we can speculate about connections between oral microbiota clusters and dental hygiene habits. For example, plaque-related group differences may arise from the fact how often an individual uses toothpicks to remove plaque between teeth or whether he/she uses an electric toothbrush instead of a mechanical one. Although such type of dental metadata were not collected during HMP study, we know that all participants of HMP study were screened for general health criteria but that does not necessarily mean that they have perfect oral health. Antunes et al. [24] conducted a comprehensive epidemiological survey of oral health of 1799 healthy adolescents. They used a gingival bleeding as an indicator of oral health and they found that over a third of participants had gingival bleeding or dental calculus. Maida et al. [25] conducted a full-mouth periodontal and oral examination in 70 subjects and they found 44.8% bleeding on probing and 22.9% normal pocket depth prevalence at subject level. Moreover, they found a significant linear trend across categories of gingival conditions (healthy, bleeding on probing, calculus presence) for *Prevotella intermedia* and *Aggregatibacter actinomycetemcomitans*. These results may suggest that in our study we found different groups of individuals with different gingival bleeding with corresponding microbial niche interactions.

The idea that microbial network demonstrates organization among closely related individual body sites such as those in oral cavity is not new and it was shown in Faust et al. [17]. However, unlike them, we have demonstrated personalized differences in

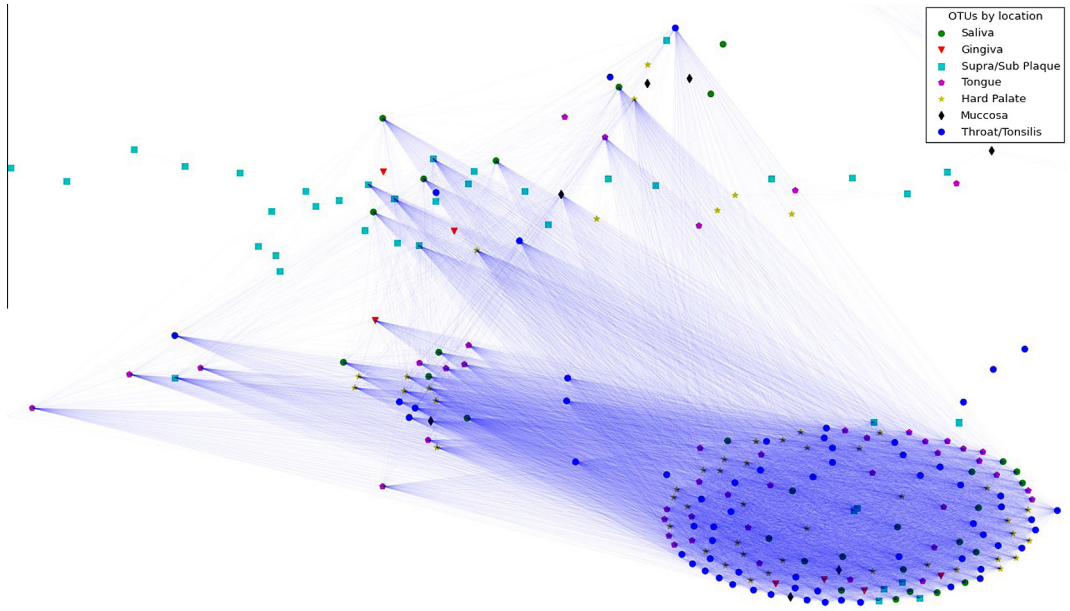

Fig. 8. Zoomed subnetwork 2 visualized from co-occurence matrix of clustering OTU in group 2. This plot displays subnetwork 2 of group 2. Nodes are colored according to the corresponding niche.

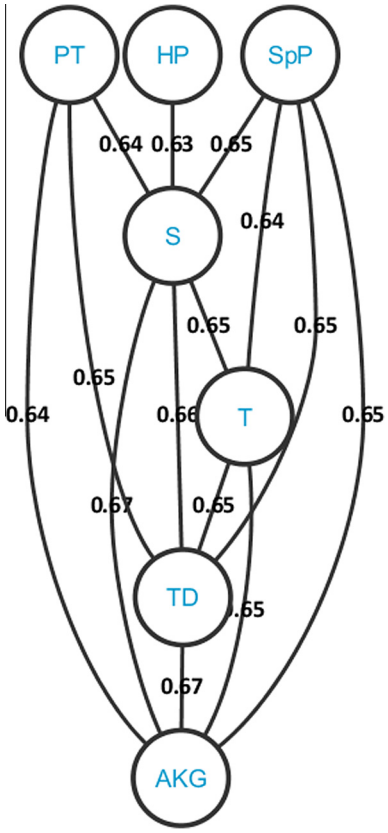

Fig. 9. Network visualized from niche-wise lumped co-occurence matrix of clustering OTU per group 1. Abbreviations refer to the following niches: S – Saliva, BM – Buccal Mucosa, AKG – attached keratinized gingiva, HP – Hard Palate, PT – Palatine Tonsils, TD – Tongue Dorsum, T – Throat, SpP – Supragingival Dental Plaque, SbP – Subgingival dental plaque. Edge labels refer to the mutual similarity score between nodes.

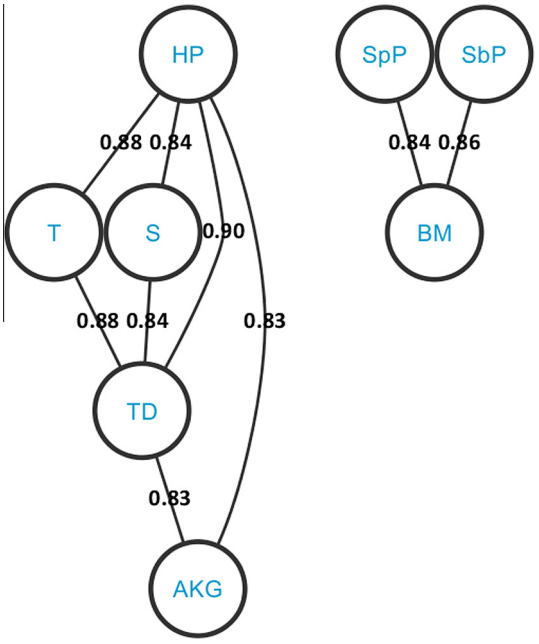

Fig. 10. Network visualized from niche-wise lumped co-occurence matrix of clustering OTU per group 2. Abbreviations refer to the following niches: S – Saliva, BM – Buccal Mucosa, AKG – attached keratinized gingiva, HP – Hard Palate, PT – Palatine Tonsils, TD – Tongue Dorsum, T – Throat, SpP – Supragingival Dental Plaque, SbP – Subgingival Dental Plaque. Edge labels refer to the mutual similarity score between nodes.

such interactions. Faust et al. used all samples in one network without distinguishing them on individuals. We claim that due to this difference we found that Supragingival and Subgingival Plaque show strong interactions with Throat/Tonsislis, Tongue, Hard Palate and Muccosa. In previous studies Plaque was clustered separately and did not show interaction with other oral niches. Another difference in our approach is the use of statistical machine learning techniques that are applicable for analysis of complex metagenomic datasets.

If we lump all similarities on niche level, we found that Saliva, Palatine Tonsils and Tongue Dorsum have the highest number of connections. The fact that Saliva has many connections can be

explained by the fact that it is a medium washing all niches. Higher connections of Palatine Tonsils and Tongue Dorsum are not so straightforward to explain while the fact that plaque is clustered separately could come from the nature of the environment, because plaque is accumulated between tooth and gingiva. This "hard tissue–soft tissue" environment can create unique conditions resulting in unique microbial composition.

Our results are generally consistent with the earlier study of Segata et al. [5] in which they show that Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria and Fusobacteria are largely dominated in oral cavity. We found that there are three OTU dominating clusters in group 2. They are *Prevotellaceae, Flavobacteriaceae, Lachnospiraceae* where the first two belong to Bacteriodetes phylum and the last one belongs to Firmicutes phylum. Segata et al. used clustering algorithm called linear discriminant analysis effect size (LEfSe). This method is a feature selection method and it uses information about taxonomical differences between bacteria. Unlike their method, we used clustering algorithm and it does not incorporate any taxonomical information into account. We consider the latter fact as an advantage because our method is less biased and conclusions we found are based on the patterns hidden behind the data.

We identified presence of network hubs in subnetworks 1 and 2 in group 2. Those hubs belong to Supra- and Subgingival Plaque OTU. Presence of plaque is an important part of etiology of dental diseases such as periodontitis and gingivitis. If we consider biological networks, Yu et al. [10] and Han et al. [26] showed that hubs in protein networks are essential in functions of cell while Barabási et al. [27] showed that the central elements of biological networks in human disease are more important for the development of the disease than ones in the periphery of the network. Although, these findings suggest that hubs of protein interactome networks are important in diseases progression, we can extend this idea to metagenome networks, particularly in drug discovery process. Hopkins [28] suggests that network methods can help to validate target combinations and optimize multiple structure–activity relationships while maintaining drug-like properties in drug discovery industry. Based on our results, we claim that hubs found in our networks can be used as potential targets in treatment and/or prevention of oral diseases such as periodontitis, gingivitis, caries.

In biomedical academic environment oral health has received less attention comparing to other complex diseases such as cancer, diabetes, pneumonia or endocarditis. However, some studies indicate that there is an association between oral infections and systemic diseases e.g. [29]. Also, many academics involved in oral health research realize potential benefits of -omics technologies in personalized dental treatment [12,30]. Such a personalized treatment based on metagenomics data can shift treatment to more fine-tuned tailored approach.

We showed that application of our machine learning methods to complex-omics data can lead to a personalized approach in oral health. Based on our results, we demonstrated that co-regularized spectral clustering on oral metagenomics samples could discover personalized microbial networks. Such networks provide important insights into microbial composition and into understanding of complex microbial interactions, that can be successfully translated into a clinical practice. As their next step authors plan to develop a multi-view spectral feature selection method, because it will allow not only to obtain and visualize clusters, but also to get a list of biomarker features most important for cluster assignments.

Acknowledgment

The study presented in this publication was funded by TI Food and Nutrition, a public–private partnership in pre-competitive research on food and nutrition.

References

[1] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J.A. Schloss, V. Bonazzi, J.E. McEwen, K.A. Wetterstrand, C. Deal, et al., Genome Res. 19 (12) (2009) 2317–2323.
[2] A. Sheiham, Bull. World Health Organ. 83 (2005) 644.
[3] L. Abusleme, A.K. Dupuy, N. Dutzan, N. Silva, J.A. Burleson, L.D. Strausbaugh, J. Gamonal, P.I. Diaz, ISME J. 7 (5) (2013) 1016–1025.
[4] A.L. Griffen, C.J. Beall, J.H. Campbell, N.D. Firestone, P.S. Kumar, Z.K. Yang, M. Podar, E.J. Leys, ISME J. 6 (6) (2012) 1176–1185.
[5] N. Segata, S.K. Haake, P. Mannon, K.P. Lemon, L. Waldron, D. Gevers, C. Huttenhower, J. Izard, Genome Biol. 13 (6) (2012) R42.
[6] S.M. Huse, Y. Ye, Y. Zhou, A.A. Fodor, PLoS One 7 (6) (2012) e34242.
[7] E. Borenstein, Briefings Bioinf. 13 (6) (2012) 769–780.
[8] R. Liu, M. Li, Z.-P. Liu, J. Wu, L. Chen, K. Aihara, Sci. Rep. 2 (2012) 6–9.
[9] J.E. McDermott, M. Costa, D. Janszen, M. Singhal, S.C. Tilton, Dis. Markers 28 (4) (2010) 253–266.
[10] H. Yu, P.M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, PLoS Comput. Biol. 3 (4) (2007) e59.
[11] I. Garcia, R. Kuska, M. Somerman, J. Dent. Res. 92 (7 Suppl.) (2013) S3–S10.
[12] G. Eng, A. Chen, T. Vess, G. Ginsburg, Oral Dis. 18 (3) (2012) 223–235.
[13] C. Huttenhower, A.I. Flamholz, J.N. Landis, S. Sahi, C.L. Myers, K.L. Olszewski, M.A. Hibbs, N.O. Siemers, O.G. Troyanskaya, H.A. Coller, BMC Bioinf. 8 (2007) 250.
[14] U. Luxburg, Stat. Comput. 17 (4) (2007) 395–416.
[15] E. Tsivtsivadze, H. Borgdorff, J. van de Wijgert, F. Schuren, R. Verhelst, T. Heskes, in: Partially Supervised Learning, Springer, 2013, pp. 80–90.
[16] K. Faust, J. Raes, Nat. Rev. Microbiol. 10 (8) (2012) 538–550.
[17] K. Faust, J.F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, C. Huttenhower, PLoS Comput. Biol. 8 (7) (2012) e1002606.
[18] M. Tumminello, T. Aste, T. Di Matteo, R.N. Mantegna, Proc. Natl. Acad. Sci. U.S.A. 102 (30) (2005) 10421–10426.
[19] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Genome Res. 13 (11) (2003) 2498–2504.
[20] A.Y. Ng, M.I. Jordan, Y. Weiss, in: Proceedings of Advances in Neural Information Processing Systems, 14, MIT Press, Cambridge, MA, 2001, pp. 849–856.
[21] A. Kumar, P. Rai, H. Daumé III, in: NIPS, 2011, pp. 1413–1421.
[22] National Institutes of Health Human Microbiome Project, HMMCP – mothur community profiling. <http://www.hmpdacc.org/HMMCP>.
[23] T. Grotkjaer, O. Winther, B. Regenberg, J. Nielsen, L.K. Hansen, Bioinformatics 22 (1) (2006) 58–67.
[24] J.L.F. Antunes, M.A. Peres, A.C. Frias, E.M. Crosato, M.G.H. Biazevic, Rev. Saúde Pública 42 (2) (2008) 191–199.
[25] C. Maida, G. Campus, A. Piana, G. Solinas, E. Milia, P. Castiglia, New Microbiol. 26 (1) (2003) 47–56.
[26] J.-D.J. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, et al., Nature 430 (6995) (2004) 88–93.
[27] A.-L. Barabási, N. Gulbahce, J. Loscalzo, Nat. Rev. Genet. 12 (1) (2011) 56–68.
[28] A.L. Hopkins, Nat. Chem. Biol. 4 (11) (2008) 682–690.
[29] X. Li, K.M. Kolltveit, L. Tronstad, I. Olsen, Clin. Microbiol. Rev. 13 (4) (2000) 547–558.
[30] S. Razzouk, O. Termechi, J. Periodontol. 84 (9) (2013) 1266–1271.