Metagenomic Analysis and Features Selection in Human Oral Microbiota Associated with Periodontal Disease

Wen-Pei Chen[1], Suh-Jen Jane Tsai[1], Yu-Chen Hu[3], and Yaw-Ling Lin*[1,2]

[1]Department of Applied Chemistry, Providence University, Taichung, Taiwan
[2]Department of Computer Science and Information Engineering, Providence University
[3]Department of Computer Science and Information Management, Providence University

Abstract

Metagenomic information provides deeper understanding of the ecological role, metabolism, and evolutionary history of microbes in a given ecosystem by analyzing environmental DNA directly without prior cultivation. In this paper, we propose methods and implement tools to facilitate the bioinformatics analysis of metagenomic data. The open-source metagenomic sequences data analysis softwares were integrated to construct accessible platforms for metagenomic data analysis. The functionality of the platform is examined by composition analysis of human oral microbiome. Furthermore, a feature selection algorithm was also proposed to choice more informative features among many variables. By using the algorithm, the support vector machine can get absolute accuracy with few features.

Keywords: metagenomics, microbiome community, machine learning, support vector machine, feature selection.

1 Introduction

Metagenomics is the study of genomes of multiple species from environmental samples, such as soil sea water, and the human gut [2, 5, 18, 1, 9]. The link with human body environments generated many studies of microbial community composition designed to assess its role in various metabolic pathway and to determine whether it is involved in inducing and preventing specific pathological conditions. Such investigations could help to clarify the pathogenesis of specific diseases and could also lead to novel disease-markers and to the development of novel therapeutic strategies.

Due to technological improvements in sequencing methods and sample extraction techniques, virtually all the microbes from a given environment can be analyzed in a efficient run, avoiding cultivation steps. In particular, procedures based on 16S rRNA next-generation sequencing, which allow the high throughput microbial identification within a specific metagenome, represent a powerful means to investigate the composition and the biodiversity of microbial communities [12]. The enormous amount of next-generation metagenomic data generated by such procedures necessitates bioinformatic tools and platforms able to analyze them. In fact, an accurate taxonomic assignment of each microbe in a target environment is required to evaluate the structure, the biodiversity, the richness and the role of the community resident in a given environment [13, 6].

The purposes of feature selection include improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [10, 14]. Feature selection methodology can be categorised into three class according to how they combine the feature selection search with the construction of the classification mode: filter methods, wrapper methods and embedded methods. Filter methods estimate the relevance of features by check at the intrinsic properties of the data. They are computationally simple and fast, can scale to very high-dimensional datasets easily, and are independent of the classification algorithm. However, some techniques in filter methods can not be applied to the case of contiguous variables. For instance, the most popular χ^2 suppose variables to be categorical data. Filter methods have

*Corresponding author. Email:yllin@pu.edu.tw.

also been used as a preprocessing step for wrapper methods, allowing a wrapper to be used on larger problems.

In machine learning, support vector machines [16] (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, since in general the larger the margin the lower the generalization error of the classifier. SVMs have been extensively used as a classification tool with a great deal of success in a variety of areas.

In this paper, we provide metagenomic analysis platforms constructed by integrating QIIME [3], PEAR [19], UCHIME [8], UPARSE [7], and other open-source tools. The system is integrated with Hadoop cloud platform and provides efficient and reliable solutions. We also introduce feature selection algorithms for SVMs. The method was based on correlation coefficient between microorganism and healthy state associated with periodontal disease. Bioinformatic analysis of human oral metagenomic data are conducted on the platform to identify the microbiome composition. As a result, the characteristics of human oral environment and analysis of the diversity and richness of the microbial community is reported in the paper.

2 Materials and Methods

2.1 16S rRNA Sequence Dataset

We constructed a dataset containing the 16S rRNA sequence data obtained from the analysis of subgingival plague samples of twenty unrelated persons: ten patients with severe periodontal disease and ten healthy controls. The next generation sequencing evaluation of their oral microbial communities was carried out by using *Illumina MiSeq* after performing amplicon sequencing on 16S rRNA V1-V2 region and PCR reaction of 10 to 18 cycles to enrich the adapter-modified DNA fragments. The minimum length = 35 and error probability < 0.05 was adopted as the criteria for quality trim processing.

2.2 Bioinformatics Analysis

2.2.1 Pre-analysis Step

The pre-analysis step includes paired-end reads assembly, barcodes filtering and trimming, and chimeras removing. The goal of this step is to filtering out noise sequences; and then, once denoising and additional quality control processes are completed, chimeric sequences should be removed from the dataset. The following parameters were set for our experiments: (1) a minimum average quality Phred score of 25 allowed in reads ; (2) 10 bases minimum overlap required in assembly processing; (3) a minimum and maximum sequence length in the range of 50-1000 bases; and (4) a maximum number of ambiguous bases and length of homopolymers equal to 6. In addition, to be as stringent as possible, no any primer mismatches was allowed in our experiments and only a 1.5 maximum number of errors in barcodes was allowed. The "Gold" database which is a FASTA file containing the ChimeraSlayer reference database in the Broad Microbiome Utilities[11] (http://microbiomeutil.sourceforge.net/) was used for chimeras detection and removing.

2.2.2 16S rRNAs Detection, Clustering, and Identification

The freeware UPARSE was used to perform 16S rRNAs detection. The OTU picking procedure consists of dereplication, abundance sort and discard singletons, and OTU clustering. Reads that are singletons after quality filtering and global trimming are discarded after the removal of duplicated sequences. Then, reads with abundances of two or more are sorted by decreasing abundance and are used as input for OTU clustering. In OTU clustering precess, reads are assigned to OTUs by clustering the reads that match the OTU with $\geq$ 97% identity. A sequence is taken in a sequence collection that represents the presence of a taxonomic unit when it shows a similarity level above the required threshold (97% identity). After the OTU picking step, the representative sequence for each OTU, namely, the most abundant sequences in that OTU, is chosen for subsequent analyses in order to reduce the computational power and the analysis time, without losing the frequency information.

GPC($\langle a_1, a_2, \ldots a_n \rangle$) ▷ Generate prioritized features order combination.
Input: $\langle a_1, a_2, \ldots a_n \rangle$ a feature list with n features in prioritized order.
Output: A queue Q used to store $2^n - 1$ features combination.
1 $Q \leftarrow \langle \emptyset \rangle$ ▷ Enqueue empty set $\emptyset$ into queue Q
2 **for** $i \leftarrow 1$ **to** n **do** ▷ Generate attribute combinations according to each feature in the list.
3 $T \leftarrow Q$ ▷ Copy Q into T
4 **for** each s in T **do**
5 Enqueue(Q, $s \cup \{a_i\}$)
6 Dequeue(Q) ▷ Delete first empty set $\emptyset$ from queue Q
7 **return** Q

Figure 1: The prioritized features combination generated algorithm. As an example, when n equals to four, the generated list will be $\langle$ 1000,0100,1100,0010,1010,0110,1110,0001, 1001,0101,1101,0011,1011,0111,1111$\rangle$.

2.2.3 Taxonomic Classification

QIIME can perform the taxonomy assignment using different methods such as RDP[17], BLAST, Mothur[15] and Rtax. In this study, we adopted the BLAST against the Human Oral Microbiome Database[4] (available at http://www.homd.org/), setting the Maximum e-Value Cutoff to 0.001. Reads assigned to the Bacteria root but not attaining the threshold at the chosen taxonomic level fell in the category "Unclassified", while sequences not assigned to the Bacteria root were classified as "No Hits". After taxonomic assignment, QIIME generates a Biological Observation Matrix (BIOM) file useful to transfer the obtained data to other tools for analysis purposes.

2.3 Feature Selection

The correlation coefficient of two variables in a data sample is their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related. If the correlation coefficient is close to 1, it would indicates that the variables are positively linearly related. For -1, it indicates that the variables are negatively linearly related. And for zero, it would indicates a weak linear relationship between the variables.

We calculated the correlation coefficients between the microbes and healthy state associated with periodontal disease. The microbe with higher correlation coefficient was selected to be a more informative feature. Then, the prioritized features combination generated algorithm shown in Figure 1 was adopted to produce the prioritized features combination composed by the more informative features.

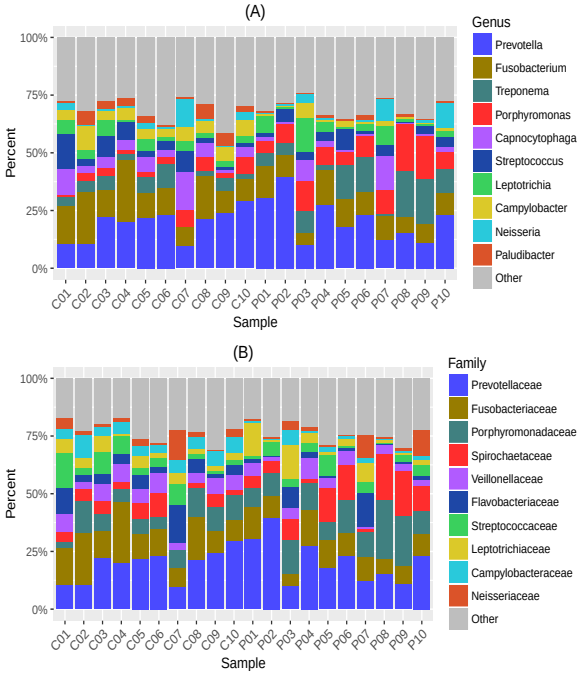

Figure 2: The top 10 taxonomic composition at genus (A) and family (B) level in each sample.

The feature combinations were used to build classifier with SVMs, each sample was selected to be testing sample by turn and others were training samples, and the accuracy of the classifier can be obtained by calculate the average accuracy of all training model. Each combination was assessed until the accuracy exceed the threshold θ.

3 Experimental Results

In this experiment, the 16S rRNA next-generation sequencing run produced 5,026,516 raw

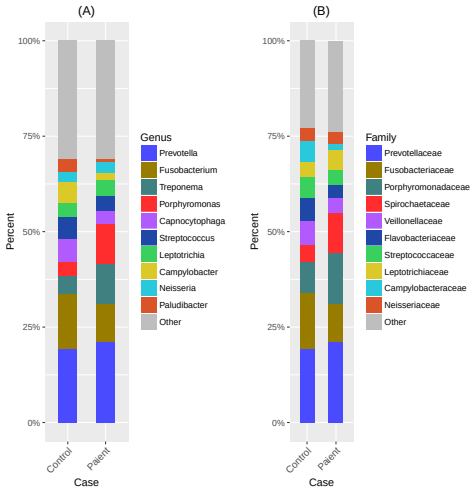

Figure 3: The top 10 genus (A) and family (B) taxonomic composition in healthy and patient case.

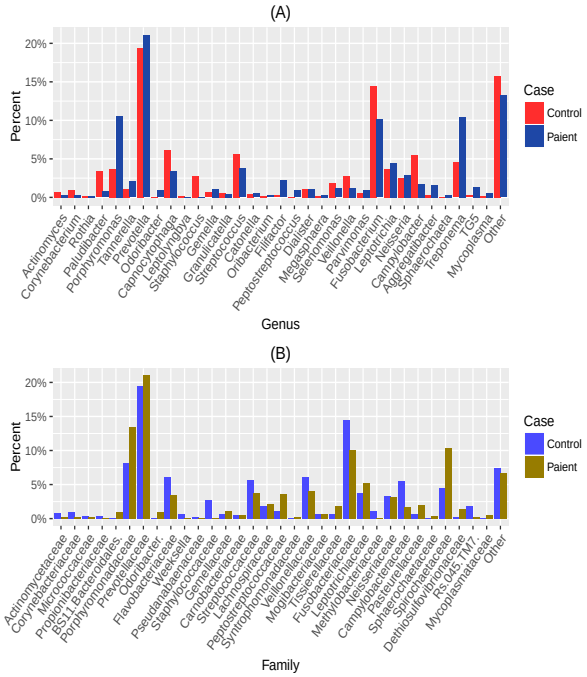

Figure 4: The genus (A) and family (B) level taxonomic composition between healthy and patient.

paired-end sequences belonging to the twenty samples. After merging these raw Illumina paired-end reads by using open-source software, PEAR, it gots 4,536,431(90.25%) assembled sequences and 490,085(9.75%) unassembled reads. In filtering and trimming step, total assembled sequences have been parsed according defined quality thresholds and 2,694,715 sequences have been assigned to appropriate sample ID. The minimum and maximum length of there sequences are 54 and 544, respectively, and the average length is 313. After removing chimeras by using software UCHIME, it obtained 2,560,229 post-filtering reads for OTU clustering process, 134,486(5%) chimeras were found in this step. The freeware, UPARSE was used to perform clustering process which includes dereplication, abundance sort, OTU clustering, and mapping reads back to OTUs steps. Total of 938 OTUs were clustered in this process.

Taxonomy assigning was performed by using BLAST method within QIIME. It identified 7 main phyla within the root Bacteria: *Bacteroidetes*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Spirochaetes*, *TM7*, and *Actinobacteria*. At deeper phylogenetic levels, 118 distinct bacteria families were identified by QIIME, while when considering only families with more than 1% richness in any sample, 32 distinct bacteria families were identified. In our dataset, *Prevotellaceae*, *Fusobacteriaceae*, *Porphyromonadaceae*, *Spirochaetaceae*, and *Veillonellaceae* are obvious families; especially, average 20.23% of sequences belong to *Prevotellaceae* family. At genus level, there are 32 genuses with more than 1% richness in any sample, and *Prevotella*, *Fusobacterium*, *Treponema*, *Porphyromonas*, and *Capnocytophaga* are obvious genuses. Figure 2 shows the top 10 taxonomic composition at family and genus level according to the number of sequences identified by QIIME. The top 10 family and genus taxonomic composition of healthy and patient case is shown in Figure 3.

The family and genus level taxonomic composition between healthy and patient case is reported in Figure 4. Nine families were identified with a widely different score: *Pseudanabaenaceae*, *Syntrophomonadaceae*, *Sphaerochaetaceae*, *BS11[Bacteroidales]*, *Staphylococcaceae*, *Odoribacter*, *Methylobacteriaceae*, *Propionibacteriaceae*, and *Rs-045[TM7]*; especially, *Pseudanabaenaceae*, *Syntrophomonadaceae* , *Sphaerochaetaceae*, and *BS11[Bacteroidales]* were just only found in healthy or patient sample. At genus level, *Leptolyngbya* was only found in healthy sample and *Sphaerochaeta* was only discovered in patient sample. Figure 5 is the heat map of taxonomic composition at genus level, *Prevotella*, *Fusobacterium* and *Fusobacterium* are most rich genuses in both healthy and patient case, moreover, *Treponema* and *Porphyromonas* are more common in patient sample.

In order to understand which microbes are play

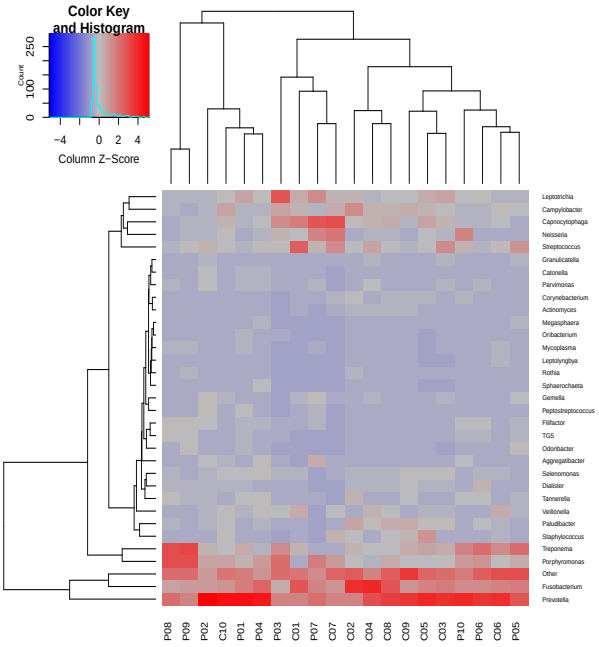

Figure 5: The heat map of taxonomic composition at genus level.

an important role in the study of periodontal disease, we calculated the correlation coefficients between microbes and healthy state of sample. Here, healthy control samples were assigned value 1 to its healthy state and patient samples were denoted by -1. Table 1 shows the top 10 features with higher correlation coefficient.

Furthermore, by using algorithm shown in Figure 1, the features chosen were used to produce feature combinations and build classifier with SVMs. In this study, the predictor can get absolute accuracy just only use *Filifactor* and *Porphyromonas* two features.

The correlation coefficient between this features were analyzed, Figure 6 shows the correlation between the top 10 informative features. It can find that, *Filifactor*, *Porphyromonas*, *TG5*, and *Treponema* have more symbiotic relationship.

4 Discussion and Conclusions

In this paper, a metagenomic analysis method was proposed to solve the problem of microbiome composition. As an example, the methodology is used to analyze the microbiome composition of human oral environment by utilizing functions provided by open-source softwares on our platform.

Feature (Genus)	Correlation coefficient
Filifactor	-0.828
Campylobacter	0.744
Porphyromonas	-0.667
Paludibacter	0.645
Staphylococcus	0.633
Actinomyces	0.591
TG5	-0.573
Corynebacterium	0.565
Treponema	-0.505
Aggregatibacter	-0.465

Table 1: The correlation coefficient between the genuses and healthy state associated with periodontal disease .

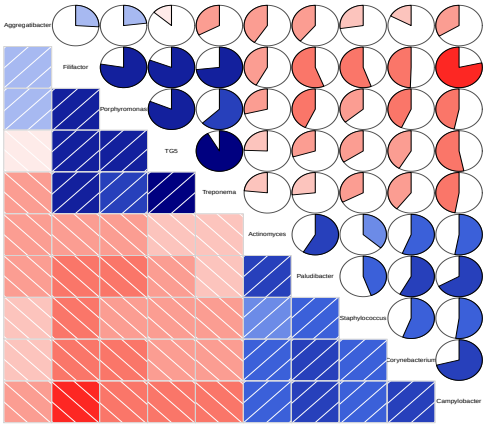

Figure 6: The correlation between the top 10 informative features.

In our dataset, there are 7 main phyla were detected. At deeper phylogenetic levels, it discovered 32 main families and 32 main genuses. *Prevotellaceae*, *Fusobacteriaceae*, *Porphyromonadaceae*, *Spirochaetaceae*, and *Veillonellaceae* are obvious families; especially, average 20.23% of sequences belong to *Prevotellaceae* family. At genus level, *Prevotella*, *Fusobacterium*, *Treponema*, *Porphyromonas*, and *Capnocytophaga* are obvious genuses.

The difference of microbiome composition can be distinguished between patient and healthy samples. Nine families were identified with a widely different score, *Sphaerochaetaceae*, and *BS11[Bacteroidales]* were just only found in healthy or patient sample. At genus level, *Leptolyngbya* genus was only found in healthy sample and *Sphaerochaeta* genus was only discovered in

patient sample.

Furthermore, a feature selection algorithm was also proposed to choice more informative features among many variables. The correlation coefficient of microbes and healthy state were taken as evaluating criterion for feature selection. Using the algorithm, the predictor can get absolute accuracy just only use two features.

5 Acknowledgment

This research was partially supported by the Ministry of Science and Technology under the Grants MOST 103-2632-E-126-001-MY3. The 16S rRNA sequence dataset used in this paper was offered by the project which sponsored by the Ministry of Science and Technology under the Grants MOST 103-2622-E-126-002-CC1.

References

[1] L. D. Alcaraz, P. Belda-Ferre, R. Cabrera-Rubio, H. Romero, . Simn-Soro, M. Pignatelli, and A. Mira. Identifying a healthy oral microbiome through metagenomics. *Clinical Microbiology and Infection*, 18:54–57, 2012.

[2] B. J. Baker, C. S. Sheik, C. A. Taylor, S. Jain, A. Bhasi, J. D. Cavalcoli, and G. J. Dick. Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *The ISME Journal*, 7(10):1962–1973, 2013.

[3] J. G. Caporaso, J. Kuczynski, and J. Stombaugh. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.

[4] T. Chen, W.-H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan, and F. E. Dewhirst. The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*, 2010, 2010.

[5] I. Cho and M. J. Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, 2012.

[6] C. De Filippo, M. Ramazzotti, P. Fontana, and D. Cavalieri. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in Bioinformatics*, 13(6):696–710, 2012.

[7] R. C. Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature Methods*, 10(10):996–998, 2013.

[8] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, 2011.

[9] S. Fang and R. M. Evans. Microbiology: Wealth management in the gut. *Nature*, 500(7464):538–539, 2013.

[10] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, Mar. 2003.

[11] B. J. Haas, D. Gevers, A. M. Earl, and M. Feldgarden. Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome Research*, 21(3):494–504, 2011.

[12] B.-S. Kim, Y.-S. Jeon, and J. Chun. Current status and future promise of the human microbiome. *Pediatric Gastroenterology, Hepatology & Nutrition*, 16(2):71–79, 2013.

[13] P. Ribeca and G. Valiente. Computational challenges of sequence classification in microbiomic data. *Briefings in Bioinformatics*, 2011.

[14] Y. Saeys, I. Inza, and P. Larraaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[15] P. D. Schloss, S. L. Westcott, T. Ryabin, and J. R. Hall. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.

[16] V. N. Vapnik. Statistical learning theory. 1998.

[17] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole. Nave bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007.

[18] P. L. Zeeuwen, M. Kleerebezem, H. M. Timmerman, and J. Schalkwijk. Microbiome and skin diseases. *Current Opinion in Allergy and Clinical Immunology*, 13(5), 2013.

[19] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis. Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics*, 30(5):614–620, 2014.