Machine Learning Approaches in Metagenomic Sequencing Analysis Related to Human Oral Microbiota Composition Associated with Periodontitis

Yaw-Ling Lin, Wen-Pei Chen, Jia-Ying Lin,

*Dept. Computer Science and Information Engineering, Providence University,*yllin@pu.edu.tw

Ming-Li Liou

Dept. Medical Laboratory Science and Biotechnology, Yuanpei University

Abstract

Advances in next-generation sequencing (NGS) has led to the rapid expansion of research in the field and the establishment of "metagenomics", the analysis of DNA from microbial communities in environmental samples without prior need for culturing. In this paper, we propose algorithms for metage-nomic analysis of microbiome composition in human oral environment by various machine learning approaches with selective features reduction. The proposed method can reduce the number of various microbes before constructing machine learning prediction models. By utilizing functions provided by open-source softwares in our platform, these methods analyze the important features by evaluating correlations between microbes and healthy states and other interested phenotypes or environmental variables. These relevant candidate microbes are ranked before constructing the predication models associated with various machine learning approaches. It is shown in these preliminary experimental analyses that several predication models report nearly perfect predication accuracy with just a handful of related observed microbes' features. Furthermore, several tailored virtual machine images are constructed by OpenStack hypervisor that can be download from our web service to provide services for metagenomics analysis by researchers and interested biologists.

Keywords: metagenomics, microbiome community, QIIME, Hadoop, machine learning, support vector machine, logistic regression, feature selection.

Materials and Methods

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. The authors 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 was adopted to produce the prioritized features combination composed by the more informative features. 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.

Experimental Results

In this experiment, the 16S rRNA next generation sequencing run produced 5,026,516 raw 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 OUT 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. 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 finds that, *Filifactor*, *Porphyromonas*, *TG5*, and *Treponema* are related to each other with putatively symbiotic relationship.

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. 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 a few handful of features.

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.