153. FizzyQIIME: Feature Selection for Metagenomics

Gregory Ditzler[1], J. Calvin Morrison[1], Christopher Blackwood[2], and **Gail L. Rosen**[1]

[1]Department of Electrical & Computer Engineering, Drexel University, Philadelphia, PA
[2]Department of Biological Sciences, Kent State University, Kent, OH

http://www.ece.drexel.edu/gailr/EESI/index.php
http://github.com/EESI/FizzyQIIME

Project Goals: Develop open-source software tools for variable selection with -omics data formats

Sequences from bacterial communities are collected from whole genome shotgun (WGS), or amplicon sequencing runs, which allows researchers to study the taxonomic composition and func- tion of a sample. Ecologists represent the data in the form of an abundance matrix, which usually holds counts of operational taxonomic units (OTUs), but can also hold counts of genes/metabolic pathway occurrences. It is quite common to have different factors in a metagenomic study, such as environmental pH and salinity values, or a health related status [4, 5]. A natural question to ask about these studies with multiple factors is: "which OTUs are important for differentiating the multiple factors?" Knowing the answer to this question can be useful for understanding which conditions are driving/being affected by differences in composition and function across samples. Answering this question can be addressed using feature selection – sometimes referred to as vari- able selection [1, 2].

Motivation Some of the current software tools for comparative metagenomics provide researchers the ability to investigate and explore bacterial communities using α– & β–diversity. Feature se- lection – a sub-field of machine learning – provides an intuitive solution to performing these com- parisons. In particular, these methods pick which OTUs (or functional features) have the most influence on the condition being studied. For example, our previous work has used information theoretic feature selection to understand the differences between protein family abundances that best discriminate between different age groups in humans [3].

Results We have developed a new Python module for the QIIME software package for microbial ecologists that implements information-theoretic feature selection methods. We demonstrate the software tools capabilities on publicly available data sets.

Availability We have made the software implementation freely available under the GNU GPL. The software can be found at http://github.com/EESI/FizzyQIIME.

References

[1] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[2] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature Extraction: Foundations and Applications*. Springer, 2006.

[3] Y. Lan, A. Kriete, and G. Rosen. Selecting age-related functional characteristics in the human gut microbiome. *BMC Microbiome*, 2013.

[4] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Lin- neberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, and S. D. Ehrlich. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65, 2010.

[5] P. Turnbaugh, M. Hamady, T. Yatsunenko, B. Cantarel, A. Duncan, R. Ley, M. Sogin, W. Jones, B. Roe, J. Affourtit, M. Egholm, B. Henrissat, A. Heath, R. Knight, and J. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 475:480–485, 2009.