A data-driven search for functional bacterial populations in aquatic systems using flow cytometry, 16S rRNA gene amplicon sequencing and the Randomized Lasso

Extended abstract*

Peter Rubbens
KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University
Ghent, Belgium
Peter.Rubbens@UGent.be

Marian Schmidt
Department of Ecology and Evolutionary Biology, University of Michigan
Ann Arbor, MI, United States of America
marschmi@umich.edu

Ruben Props
Center for Microbial Technology and Ecology (CMET), Ghent University
Ghent, Belgium
Ruben.Props@UGent.be

Nico Boon
Center for Microbial Technology and Ecology (CMET), Ghent University
Ghent, Belgium
Nico.Boon@UGent.be

Vincent Denef
Department of Ecology and Evolutionary Biology, University of Michigan
Ann Arbor, MI, United States of America
vdenef@umich.edu

Willem Waegeman
KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University
Ghent, Belgium
Willem.Waegeman@UGent.be

ABSTRACT

Microbial communities can be characterized by flow cytometry (FCM), a single-cell technology which measures thousands of individual cells in seconds of time. This technique can be used in microbial ecology studies to investigate the dynamics of communities in relation to the environment. When applying FCM in aquatic environments, it gives rise to two distinct microbial groups, known to correspond to a different ecological functioning. By measuring FCM in parallel with 16S rRNA gene amplicon sequencing, one can try to associate individual bacteria with one of these functional groups. We propose to address this problem from a machine learning based variable selection perspective. Results confirm a strong correspondence between 16S rRNA gene sequencing and flow cytometry cell measurements. The Randomized Lasso allows for an effective screening of individual bacteria, but its results are affected by spatio-temporal patterns in the data.

CCS CONCEPTS

•**Computing methodologies** → **Model development and analysis;** *Machine learning;* •**Applied computing** → Life and medical science;

KEYWORDS

16S rRNA gene amplicon sequencing, flow cytometry, microbial community dynamics, microbial ecology, randomized lasso, variable selection

ACM Reference format:
Peter Rubbens, Marian Schmidt, Ruben Props, Nico Boon, Vincent Denef, and Willem Waegeman. 2016. A data-driven search for functional bacterial populations in aquatic systems using flow cytometry, 16S rRNA gene amplicon sequencing and the Randomized Lasso. In *Proceedings of FEED'18, KDD 2018, London, UK, August 19 2018,* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Microbial communities are vital components in the Earth's ecosystem functions. They are primary contributors to most biogeochemical cycling processes [3]. The field of microbial ecology tries to understand the relationship between microbial diversity and ecosystem functioning [20]. The upcoming of 16S rRNA gene amplicon sequencing allows the quantification of microbial community composition and diversity based on the '16S rRNA gene' [13]. Therefore, microbial ecologists have uncovered the identity of microbial communities to a large extent, or in other words, microbiologists know "who are there". However, it is much more challenging to link specific bacterial taxa to ecosystem processes, or to state differently: "who is doing what?" [8]. Therefore, advances are needed to be able to associate certain bacterial populations with ecosystem functions.

Flow cytometry is an alternative technology which is frequently used in the field of aquatic microbiology [25, 26]. It offers high-throughput measurements of single cells, resulting in a multivariate description of their optical properties. In aquatic applications, cells are often treated with a nucleic acid stain (e.g., SYBR GREEN I), which results in two separated groups of cells, the high nucleic

*Results of this report are based on a full manuscript, of which a preprint can be accessed via bioRxiv [22].

FEED'18, KDD 2018, London, UK

DOI: 10.1145/nnnnnnn.nnnnnnn

acid (HNA) and low nucleic acid (LNA) group. This dichotomy has been established for various aquatic environments, ranging from marine and freshwater ecosystems to measurements of drinking water communities [2, 16, 17]. For the latter these can be used as proxies for drinking water stability [1, 16]

In an environmental setting, these groups have been associated with different ecosystem functioning. Traditionally, HNA bacteria have been classified as 'active' bacteria, whereas LNA bacteria have been characterized 'inactive' [4, 11]. This is related to the positive correlation between heterotrophic bacterial production (BP) and HNA abundances, whereas there is no correspondence between BP and LNA abundances [9, 11]. Various scenarios have been proposed to explain these findings in relation to the identity of these functional groups [2], yet a clear explanation is still not established.

Recently, it has been shown that biodiversity estimations based on FCM correspond with those based on 16S rRNA gene amplicon sequencing [18, 19]. By exploiting machine learning-based approaches, one can go one step further and attempt to associate abundances of individual bacteria, identified by 16S rRNA gene amplicon sequencing, with a specific functional group in FCM data. Doing so, one can try and associate individual bacterial populations with ecosystem functioning in a data-driven way.

Concretely, we applied an ensemble variable selection approach, called *stability selection* or the *Randomized Lasso* [10], to 16 rRNA gene amplicon sequencing data in function of abundance variations of HNA and LNA fractions. This was done for a freshwater lake system, Muskegon Lake, an estuary of Lake Michigan. Our approach was motivated by a strong correlation between BP and HNA variations for this specific lake system. Results of the variable selection strategy were evaluated using a recursive variable elimination strategy, in which the optimal amount of variables was determined to predict HNA and LNA abundances. We show that there is a strong correspondence between these two types of data, for which the Randomized Lasso enabled effective screening of individual bactera. This resulted in a considerably higher predictive performance. In addition, we show that variable selection results are subject to spatio-temporal structure in the data, which supports the hypothesis that the identity of HNA and LNA bacterial populations changes according spatio-temporal trends.

2 DATASET DESCRIPTION

Data was collected from samples taken at Muskegon Lake, a freshwater estuary of Lake Michigan. This was done at five different sampling sites across three years in three different seasons. An overview of the lake ecosystem can be found in Figure 1. 16S rRNA gene amplicon sequencing and flow cytometry data collection and analysis is described in [19]. After preprocessing, sequencing data gave rise to 482 different Operational Taxonomic Units (OTUs, i.e., the lowest taxonomic level at which bacterial taxa can be identified). HNA and LNA cell counts (HNAcc/LNAcc) were determined using fixed gates, as determined by guidelines of Prest et al. (2013) [16]. In addition, 20 samples from Muskegon Lake samples were processed for heterotrophic bacterial production (BP), as measured by radiolabeled leucine [23].

Figure 1: Overview of sampling sites in Muskegon Lake, a freshwater lake which interacts with Muskegon River and Lake Michigan.

3 WORKFLOW

Compositional data exhibit a *negative correlation bias* [5]. Recent reviews considering the analysis of 16S rRNA gene amplicon sequencing data suggest to address this issue by performing a centered log-ratio (CLR) transformation before data analysis [5, 14]. This means that the abundance x_i of an individual OTU is transformed by calculating the logarithm of the ratio between its abundance and geometric mean:

$$x_i' = \log\left\{\frac{x_i}{\prod_{j=1}^{p} x_j}\right\}, \tag{1}$$

in which p denotes the number of variables. As the logarithm cannot deal with zero values, each zero was replaced by $\delta = 1/p^2$.

Variables were selected based on an extension of the Lasso estimator, which is called *stability selection* [10]. The Lasso fits a regularized linear regression model, making use of an l_1-penalty:

$$\hat{\beta_\lambda} = \underset{\beta \in \mathbb{R}^p}{\arg\min} ||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2}$$

X denotes the compositional abundance table, Y the target to predict and λ is a penalty term which controls the complexity of the model. By applying stability selection to the Lasso, one retrieves the Randomized Lasso. This is done by performing two different kinds of randomization in order to assign a score to each variable. This resembles the probability a variable will be included in the Lasso model (i.e., its corresponding weight is non-zero). If n denotes the number of samples, B subsamples are created of size $n/2$. A second form of randomization is added by using a weakness parameter $\alpha \in [0, 1]$. In each subset, certain variables will be randomly penalized with λ/α. The Randomized Lasso therefore becomes:

$$\hat{\beta_\lambda} = \underset{\beta \in \mathbb{R}^p}{\arg\min} ||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{W_j}, \tag{3}$$

in which W_j is a random variable which is either α or 1. Next, the score coming out of the Randomized Lasso, denoted by π, is determined by counting the number of times the weight of a variable was

non-zero for each of the models and divided by B. The Randomized Lasso was implemented using the scikit-learn machine learning library, with $B = 500$ and $\alpha = 0.5$ [15]

The scores of the Randomized Lasso were evaluated using a recursive variable elimination strategy [6]. Variables were ranked according to π and were iteratively eliminated until the highest-ranked variables remained. Predictive performance was evaluated using a regular Lasso model, for which λ was tuned using the lassoCV() function. A blocked cross-validation scheme was used, which incorporated to some extent the spatio-temporal structure of the data [21], i.e. samples were grouped according the sampling site and year they were measured, giving rise to 10 spatio-temporal groups. Similarity or *stability* between sets of scores resulting from the Randomized Lasso were quantified using the Pearson correlation coefficient ρ_P [12].

4 RESULTS & DISCUSSION

Although Muskegon Lake was dominated by LNA cell counts (LNAcc, 69.6%), heterotrophic bacterial production (BP) showed significant correlation with HNA cell counts (HNAcc) ($R^2 = 0.65, P < .001$), but not with LNAcc ($R^2 = 0.005, P > 0.1$). OTUs were associated with HNAcc and LNAcc using the Randomized Lasso, which resulted in scores between 0 (unimportant) and 1 (important) for individual OTUs. Results indicated that although OTUs can be ranked from unimportant to important, the score π hardly exceeded the threshold of 0.5. Meinshausen & Bühlmann (2010) proved that in order to control the number of falsely selected variables, π needs to be at least higher than 0.5.

As noted, π can be used to rank variables according to their importance. A recursive variable elimination strategy was next employed to evaluate the sensibility of the Randomized Lasso scores. The performance was expressed in terms of the R^2_{CV}, which denotes the R^2 between predicted and true values, for samples that were left out according to a blocked cross-validation scheme (Figure 2). Results indicate that removal of OTUs based on the Randomized Lasso greatly improved predictive performance of a Lasso model. Only a subset of variables wass needed, as a fraction of 25% gave rise to optimal predictions (R^2_{CV}(HNAcc) = 0.85, R^2_{CV}(LNAcc) = 0.91). In other words, π can be used as tuning parameter from a prediction point of view. Scores in function of HNAcc were correlated with those for LNAcc ($\rho_P = 0.52, P < .001$). This might indicate that OTUs switched between groups. An alternative explanation might be that 'keystone' OTUs were selected that were predictors for HNA/LNA compositions[7], but were not necessarily present in those groups. No relationship could be established between individual abundances of OTUs and π.

To further quantify spatio-temporal effects on variable selection performance, a perturbation experiment was carried out. Samples were grouped according to their spatial or temporal annotation (DEPTH, SEASON, SITE and YEAR). Next, the Randomized Lasso was run after leaving out a specific group, until every group was left out once. In this way, spatio-temporal effects on Randomized Lasso scores could be quantified. This was done by calculating the similarity or *stability* using the Pearson correlation coefficient ρ_P between sets of scores (Figure 3). This analysis shows that especially the DEPTH at which a sample was collected, which was either deep

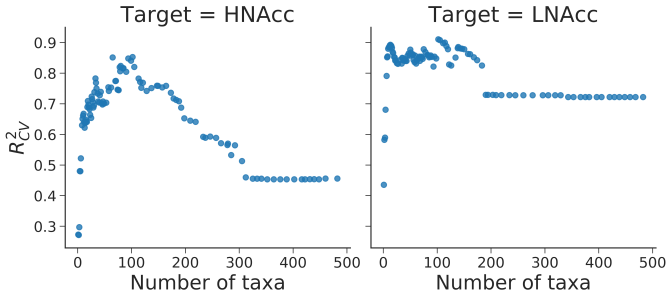

Figure 2: R^2_{CV} in function of the number of the number of selected OTUs, based on the Randomized Lasso. Subsets of variables were created by iteratively removing OTUs based on the ranking according to π and evaluated the predictive performance of the Lasso.

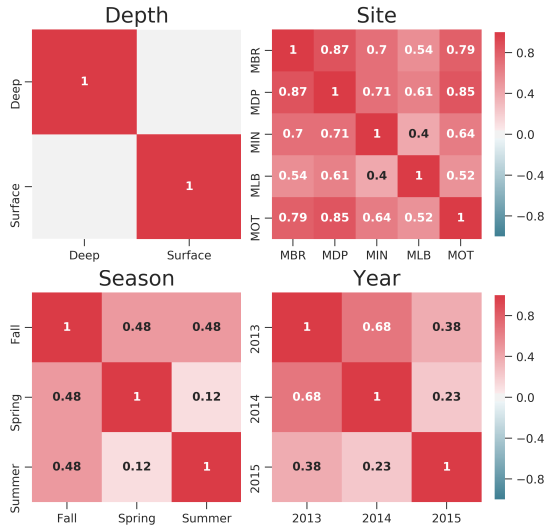

Figure 3: Stability assessment of a perturbation experiment in which samples were left out according the their spatial (first row) or temporal (second row) classification. Note that labeling can be counterintuitive, as a label denotes the group of samples that were left out before analysis. Values denote ρ_P and are visualized when significant ($P < .05$).

or at the surface, gave rise to two sets of scores which did not show similarity. Other sets remained significantly correlated, yet variable selection results were still affected.

Results of the Randomized Lasso were evaluated from an ecological perspective as well. The identity of high-ranked HNA OTUs revealed that most of them were part of the phylum[1] Bacteroidetes, which agrees with previously reported research [24]. LNA OTUs were more scattered across different phyla. Although this group of bacteria showed no significant correlation to BP measurements,

[1]The phylum level is the first level of taxonomic classification at which bacterial taxa can be identified.

when the Randomized Lasso was run at the phylum level in function of productivity measurements, this phylum was ranked second out of 22. The top-ranked phylum, Proteobacteria, was significantly correlated with BP ($\rho_P = 0.72, P < .001$).

5 CONCLUSION

By integrating flow cytometry with 16S rRNA gene amplicon sequencing, machine learning can assist in the association of specific OTUs with ecosystem functioning in aquatic environments. We proposed an approach based on stability selection. A strong correspondence was established between abundances of individual bacteria as measured by 16S rRNA gene amplicon sequencing and functional groups in flow cytometry data, and these can be predicted using machine learning models. Data-driven variable selection methods can be used to associate specific OTUs to functional groups, and may highlight 'keystone' bacterial taxa for specific ecosystem processes. Our results indicate that there are taxonomic differences between HNA and LNA fractions in freshwater lake systems, yet these are not universal and are subject to spatio-temporal changes in the environment. Therefore our results further strengthen the hypothesis proposed by by Vila-Costa et al. (2012) [24], in which a taxonomic division was found between HNA and LNA fractions, which was influenced by seasonal trends and therefore not universal.

REFERENCES

[1] M. D. Besmer, J. A. Sigrist, R. Props, B. Buysschaert, G. Mao, N. Boon, and F. Hammes. 2017. Laboratory-scale simulation and real-time tracking of a microbial contamination event and subsequent shock-chlorination in drinking water. *Frontiers in Microbiology* 8 (2017), 1–11. DOI: http://dx.doi.org/10.3389/fmicb.2017.01900

[2] T. Bouvier, P. A. Del Giorgio, and J. M. Gasol. 2007. A comparative study of the cytometric characteristics of High and Low nucleic-acid bacterioplankton cells from different aquatic ecosystems. *Environmental Microbiology* 9, 8 (2007), 2050–2066. DOI: http://dx.doi.org/10.1111/j.1462-2920.2007.01321.x

[3] P. G. Falkowski, T. Fenchel, and E. F. Delong. 2008. The microbial engines that drive earth's biogeochemical cycles. *Science* 320, 5879 (2008), 1034–1039. DOI: http://dx.doi.org/10.1126/science.1153213

[4] J. M. Gasol and P. A. Del Giorgio. 2000. Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Scientia Marina* 64, 2 (2000), 197–224. DOI: http://dx.doi.org/10.3989/scimar.2000.64n2197

[5] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. 2017. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology* 8 (2017), 1–6. DOI: http://dx.doi.org/10.3389/fmicb.2017.02224

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46 (2002), 389–422. DOI: http://dx.doi.org/10.1155/2012/586246 arXiv:1111.6189v1

[7] C. M. Herren and K. D. McMahon. 2018. Keystone taxa predict compositional change in microbial communities. *Environmental Microbiology* (2018), 1–34. DOI: http://dx.doi.org/10.1111/1462-2920.14257

[8] A. Konopka. 2009. What is microbial community ecology? *ISME Journal* 48, 3 (2009), 561–565. DOI: http://dx.doi.org/10.1038/ismej.2009.88

[9] P. Lebaron, P. Servais, H. Agogué, C. Courties, and F. Joux. 2001. Does the high nucleic acid content of individual bacterial cells allow us to discriminate between active cells and inactive cells in aquatic systems? *Applied and Environmental Microbiology* 67, 4 (2001), 1775–1782. DOI: http://dx.doi.org/10.1128/AEM.67.4.1775-1782.2001

[10] N. Meinshausen and P. Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* (2010). DOI: http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x

[11] X. A. G. Moran, A. Bode, Á. Suárez, L., and E. Nogueira. 2007. Assessing the relevance of nucleic acid content as an indicator of marine bacterial activity. *Aquatic Microbial Ecology* 46 (2007), 141–152. DOI: http://dx.doi.org/10.3354/ame046141

[12] S. Nogueira and G. Brown. 2016. Measuring the stability of feature selection. In *Machine Learning and Knowledge Discovery in Databases*, Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken (Eds.). Springer International Publishing, 442–457.

[13] N. R. Pace. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276, 5313 (1997), 734–740. DOI: http://dx.doi.org/10.1126/science.276.5313.734 arXiv:http://science.sciencemag.org/content/276/5313/734.full.pdf

[14] O. Paliy and V. Shankar. 2016. Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* 25, 5 (2016), 1032–1057. DOI: http://dx.doi.org/10.1111/mec.13536 arXiv:15334406

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M.u Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M.u Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: machine Learning in python. *Journal Of Machine Learning Research* 12 (2011), 2825–2830.

[16] E. I. Prest, F. Hammes, S. Kötzsch, M. C M van Loosdrecht, and J. S. Vrouwenvelder. 2013. Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. *Water Research* 47, 19 (2013), 7131–7142. DOI: http://dx.doi.org/10.1016/j.watres.2013.07.051

[17] C. R. Proctor, M. D. Besmer, T. Langenegger, K. Beck, J.-C. Walser, M.n Ackermann, H. Bürgmann, and F. Hammes. 2018. Phylogenetic clustering of small low nucleic acid-content bacteria across diverse freshwater ecosystems. *The ISME Journal* (2018). DOI: http://dx.doi.org/10.1038/s41396-018-0070-8

[18] R. Props, P. Monsieurs, M. Mysara, L. Clement, and N. Boon. 2016. Measuring the biodiversity of microbial communities by flow cytometry. *Methods in Ecology and Evolution* 7, 11 (2016), 1376–1385. DOI: http://dx.doi.org/10.1111/2041-210X.12607

[19] R. Props, M. L. Schmidt, J. Heyse, H. A. Vanderploeg, N. Boon, and V. J. Denef. 2018. Flow cytometric monitoring of bacterioplankton phenotypic diversity predicts high population-specific feeding rates by invasive dreissenid mussels. *Environmental Microbiology* 20, 2 (2018), 521–534. DOI: http://dx.doi.org/10.1111/1462-2920.13953

[20] J. I. Prosser, B. J. M. Bohannan, T. P. Curtis, R. J. Ellis, M. K. Firestone, R. P. Freckleton, J. L. Green, L. E. Green, K. Killham, J. J. Lennon, A. M. Osborn, M. Solan, C. J. van der Gast, and J. P. W. Young. 2007. The role of ecological theory in microbial ecology. *Nature Reviews Microbiology* 5, 5 (2007), 384–392. DOI: http://dx.doi.org/10.1038/nrmicro1643

[21] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. L.-M., B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 8 (2017), 913–929. DOI: http://dx.doi.org/10.1111/ecog.02881

[22] P. Rubbens, M.L. Schmidt, R. Props, N. Boon, W. Waegeman, and V.J. Denef. 2018. Using machine learning to associate bacterial taxa with functional groups through flow cytometry, 16S rRNA gene sequencing, and productivity data. *bioRXiv* 392852 (2018).

[23] M. L. Schmidt, B. A. Biddanda, A. D. Weinke, E. Chiang, F. Januska, R. Props, and V. J. Denef. 2017. Microhabitats shape diversity-productivity relationships in freshwater bacterial communities. *bioRxiv* (2017). DOI: http://dx.doi.org/10.1101/231688 arXiv:https://www.biorxiv.org/content/early/2017/12/11/231688.full.pdf

[24] M. Vila-Costa, J. M. Gasol, S. Sharma, and M. A. Moran. 2012. Community analysis of high- and low-nucleic acid-containing bacteria in NW Mediterranean coastal waters using 16S rDNA pyrosequencing. *Environmental Microbiology* 14, 6 (2012), 1390–1402. DOI: http://dx.doi.org/10.1111/j.1462-2920.2012.02720.x

[25] J. Vives-Rego, P. Lebaron, and Caron Nebe-von. 2000. Current and future applications of flow cytometry in aquatic microbiology. *FEMS microbiology reviews* 24, 2000 (2000), 429–448. DOI: http://dx.doi.org/10.1016/S0168-6445(00)00033-4

[26] Y. Wang, F. Hammes, K. De Roy, W. Verstraete, and N. Boon. 2010. Past, present and future applications of flow cytometry in aquatic microbiology. *Trends in Biotechnology* 28, 8 (2010), 416–424. DOI: http://dx.doi.org/10.1016/j.tibtech.2010.04.006