Identifying synthetic microbial communities by learning in silico communities using flow cytometry

Peter Rubbens PETER.RUBBENS@UGENT.BE
Willem Waegeman WILLEM.WAEGEMAN@UGENT.BE
KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000, Belgium

Ruben Props RUBEN.PROPS@UGENT.BE
Nico Boon NICO.BOON@UGENT.BE
Center for Microbial Ecology and Technology (CMET), Ghent University, Coupure Links 653, B-9000, Belgium

Keywords: flow cytometry, in silico communities, microbiology, multiclass classification, supervised learning, synthetic ecology

Abstract

Single cells can be characterized in terms of their phenotypic properties using flow cytometry. However, up to our knowledge there has not yet been a thorough survey concerning the classification of bacterial species based on flow cytometric data. This paper aims to perform a thorough investigation concerning the identification of bacterial communities of various complexities in species richness. We do this by creating so-called in silico communities, communities created by aggregating the data coming from individual cultures; moreover we show that it is possible to use in silico communities to identify in vitro created communities as well, proving the biological relevance and usability of bacterial in silico communities.

1. Introduction

Flow cytometry (FCM) characterizes the phenotypic properties of single-cells in terms of scatter signals and fluorescence intensity (Müller & Nebe-von Caron, 2010), which results in a multiparametric description of each cell. As FCM is capable of measuring thousands of cells per second, it is an interesting application for the analysis of microbial species (Diaz et al., 2010; Koch et al., 2013).

Accepted for *MLSB 2016*.

Our research focuses on two issues; first, we explore to which extent single-cell predictions can be made when analyzing bacterial FCM data. To do this we create *in silico* communities, communities constructed by aggregating FCM data coming from measuring species separately, after which we perform two supervised machine learning methods in order to classify single cells. Two earlier reports exist using using this approach, however, these reports only considering a binary setting with few bacterial species at their disposal (Davey et al., 1999; Rajwa et al., 2008). We extend this research to a multiclass setting as well, evaluating 150 different in silico communities up to twenty classes.

Second, we show that in silico communities can be used to classify individual cells in their in vitro counterpart communities. This result offers a new approach of setting up controllable synthetic microbial communities, and as these communities consist typically out of a few species only, they can be identified in a supervised way. Therefore they can form the bridge between on the one hand real in vivo ecological systems and on the other hand theoretical models simulating them (De Roy et al., 2014).

2. Learning in silico communities

We have measured 20 individual bacterial cultures through FCM. This dataset offers us the advantage that we can create a vast amount of in silico communities, varying both in species richness S and relative abundances.

First we evaluated to what extent it is possible to make

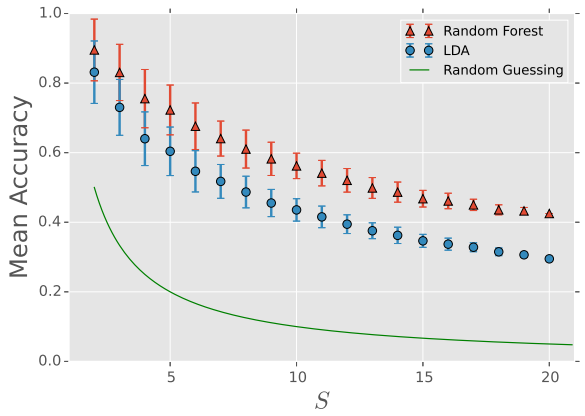

Figure 1. Mean accuracy using LDA and a Random Forest classifier for 150 different in silico communities for increasing S; the dataset consists out of 5000 cells per species, of which 30% were held-out as a test set; figure taken from (Rubbens et al.,).

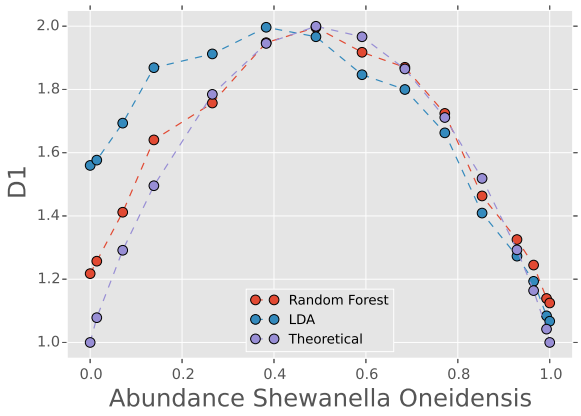

Figure 2. Predicted and target abundance gradients expressed in diversity D_1, for an abundance gradient consisting out of S. Oneidensis and M. Luteus.

single-cell predictions. To do so we made 150 different in silico communities for increasing species richness S (except for $S = 19$ and $S = 20$, for which the maximum number of in silico communities is 20 and 1 respectively). The species have been randomly chosen for every community, but evenly sampled (5.000 cells). By creating training and test sets (70% and 30% of the data respectively), we calculated the accuracy of a classifier on the test set for every in silico community. For now we used Linear Discrimant Analysis (LDA) and a Random Forest classifier alongside all possible features at our disposal (which are eight fluorescence features and four scatter features). Next, we calculated the mean accuracy for every S, for which the results are shown in Fig. 1. We see that for low S single-cell predictions are possible up to high accuracies using 'off-the-shelf' classifiers.

It is yet to prove whether one can use an in silico community to identify single cells coming from an in vitro community, i.e., can we use an in silico community which is in a sense an artificial dataset to analyze a 'real-world' in vitro community? To test this, we created a so-called *abundance gradient* for $S = 2$, meaning that we created various in vitro communities consisting out of two species in varying abundances, going from 1% to 99% (and vice versa for the opposite community). We first measured the bacterial cultures separately, in order to create an in silico community to train a classifier. Using this classifier we predicted the species of all individual cells in the in vitro communities constituting the abundance gradient. From this, we calculated the predicted relative abundances for every in vitro community. An example of this study is shown in Fig. 2, where we created and predicted the abundance gradient for communities consisting out of *Shewanella Oneidensis* and *Micrococcus Luteus*. The abundances are expressed in terms of diversity by means of the first Hill number $D_1 = \exp(-\sum_{i=1}^{S} p_i \ln p_i)$, where p_i denotes the relative abundance of species i (Hill, 1973). We see that we are able to retrieve the abundance gradient with adequate precision. Moreover, this approach offers us the possibility to further quantify the performance of an in silico community by calculating for example the root mean squared error.

3. Conclusion

Single-cell predictions are possible using FCM applied to bacterial species. For low species richness we are able to already achieve acceptable to high performances. More importantly, we have shown that in silico communities can identify corresponding in vitro communities. That means that we can use in silico communities as a relevant representation of synthetic bacterial communities. This result can be used both for experimental studies using low-complexity communities and for studies concerning bacterial flow cytometric data analysis.

Acknowledgments

This research was supported by the Special Research Fund (BOF) from Ghent University.

References

Davey, H., Jones, A., Shaw, A., & Kell, D. (1999). Variable selection and multivariate methods for the identification of microorganisms by flow cytometry. *Cytometry*, *35*, 162–168.

De Roy, K., Marzorati, M., Van den Abbeele, P., Van de Wiele, T., & Boon, N. (2014). Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environmental Microbiology*, *16*, 1472–1481.

Diaz, M., Herrero, M., Garcia, L. A., & Quiros, C. (2010). Application of flow cytometry to industrial microbial bioprocesses. *Biochemical Engineering Journal*, *48*, 385–407.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, *54*, 427–432.

Koch, C., Guenther, S., Desta, A. F., Huebschmann, T., & Mueller, S. (2013). Cytometric fingerprinting for analyzing microbial intracommunity structure variation and identifying subcommunity function. *Nature Protocols*, *8*, 190–202.

Müller, S., & Nebe-von Caron, G. (2010). Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities. *FEMS Microbiology Reviews*, *34*, 554–587.

Rajwa, B., Venkatapathi, M., Ragheb, K., Banada, P. P., Hirleman, E. D., Lary, T., & Robinson, J. P. (2008). Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier. *Cytometry Part A*, *73A*, 369–379. 24th International Congress of the International-Society-for-Analytical-Cytology, Budapest, Hungary, May, 17-21, 2008.

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, *62*, 142–160. Joint Symposium of the Environmental-Microbiology-Group/British-Ecological-Society/Society-for- General-Microbiology, Univ York, York, ENGLAND, SEP 13, 2006.

Rubbens, P., Props, R., Boon, N., & Waegeman, W. Flow cytometric single-cell identification of populations in synthetic bacterial communities. *Under review.*