Learning in silico communities to perform flow cytometric identification of synthetic bacterial communities

Peter Rubbens[1], Ruben Props[1] Nico Boon[1], and Willem Waegeman[1]

[1]Ghent University, Ghent, Belgium,
Peter.Rubbens@UGent.be

Abstract. Flow cytometry is able measure up to 50.000 cells in various dimensions in seconds of time. This large amount of data gives rise to the possibility of making predictions at the single-cell level, however, applied to bacterial populations a systemic investigation lacks. In order to combat this deficiency, we cultivated twenty individual bacterial populations and measured them through flow cytometry. By creating *in silico communities* we are able to use supervised machine learning techniques in order to examine to what extent single-cell predictions can be made; this can be used to identify the community composition. We show that for more than half of the communities consisting out of two bacterial populations we can identify single cells with an accuracy >90%. Furthermore we prove that in silico communities can be used to identify their in vitro counterpart communities. This result leads to the conclusion that in silico communities form a viable representation for synthetic bacterial communities, opening up new opportunities for the analysis of bacterial flow cytometric data and for the experimental study of low-complexity communities.

Keywords: flow cytometry, microbiology, in silico communities, synthetic bacterial communities, linear discriminant analysis, random forests

1 Introduction

Flow cytometry (FCM) is an experimental technique which characterizes individual cells in terms of fluorescence and scatter signals; this results in a multidimensional description of every cell. As the analysis of cells is increasing (up to 50.000 of cells per second), alongside with the dimensionality of the data (up to 50 dimensions will be available soon), the field of FCM *bioinformatics* is growing accordingly [11]. A promising approach of analyzing FCM data is the use of supervised machine learning techniques in order to identify single cells, an approach which has been used in for example the recognition of leukemia [17] or to identify various populations of phytoplankton [3], [12].

However, applied to bacterial populations this approach seems to be lacking a thorough investigation. Two reports exist, of which the first analyzed the effect of various cocktails of fluorescent staining [6], and the second the extent to which

individual cells can be classified using multiple but only scatter signals [13]. However the number of populations used in latter studies is small, remaining only to the binary setting.

To investigate this issue more thoroughly, we cultivated twenty different bacterial populations and measured them individually through FCM. We propose the use of *in silico communities*, communities we created by aggregating the data coming from these individual cultures. This approach leads to two advantages; first, we are able to use supervised machine learning methods as we know the individual label of every cell. Second, we have the ability to create a wide spectrum of bacterial communities ranging from low complexity to high complexity, and ranging from low evenness (i.e., unevenly distributed populations) to high evenness communities. For example, for a population richness of $S = 2$ we already have the possibility of analyzing 190 different bacterial communities, only considering the population richness.

In the first section we perform a thorough analysis regarding the possibility of making single-cell predictions. We will show that for a binary setting we are able to achieve high accuracies for a majority of possible bacterial communities. Next, we consider a multiclass setting as well, showing that FCM data should be feasible for increasing population richness. We chose methods which extend to a multiclass setting in a natural way. For now we opted to use Linear Discriminant analysis (LDA), which is an established method in microbial ecology to perform multivariate analyses [14], and Random Forests, known for its high performance in various applications with only one hyperparameter to tune [8].

In the second section of the paper we show that we can use the statistical properties of in silico communities in order to classify individual cells contained in resembling (i.e., containing the same bacterial populations) in vitro communities. This is not self-evident for two reasons; first, flow cytometric measurements suffer from technical variations and second, it has been proven that bacterial populations exhibit heterogeneous behavior which is reflected in FCM data [18]. In order to test this hypothesis, we created so-called *abundance gradients*; we define an abundance gradient as a set of in vitro communities which contain the same two bacterial populations, but in varying abundances. We will show that we are able to retrieve these relative abundances using classifiers which are trained on an evenly distributed in silico community. This result forms a strong argument that flow cytometric in silico communities form a proper representation for synthetic bacterial communities, and thus can be used for further study; furthermore, it enables researchers the use of supervised methods combined with FCM in order to analyze low-complexity communities.

2 Exploratory analysis of in silico communities

In order to systematically investigate the possibility of making single-cell predictions, we have cultivated twenty bacterial populations and measured them through FCM; a full list can be found in Tab. 1.

Table 1. List of cultivated bacterial populations measured through FCM (dataset 1).

Bacterial population	Culture collection reference
Agrobacter rhizogenes	UFZ requested [16]
Bacillus subtilis	LMG 7135
Burkholderia ambifaria	LMG 19182
Citrobacter freundii	DSMZ 15979
Cupriavidus necator	LMG 1201
Cupriavidus pinatubonensis	LMG 1197
Edwardsialla ictaluri	LMG 7860
Enterobacter aerogenes	DSMZ 30053
Escherichia coli	DSMZ 2840
Janthinobacterium sp. B3	UFZ requested [16]
Klebsiella oxytoca	LMG 3055
Lactobacillus plantarum	LMG 9211
Micrococcus luteus	UFZ requested [16]
Pseudomonas fluorescens	R 23898
Pseudomonas putida	R 17801
Rhizobium radiobacter	LMG 287
Shewanella oneidensis	LMG 19005
Sphingomonas aromaticivorans	LMG 18303
Streptococcus salivarius	LMG 11489
Zymomonas mobilis subsp. mobilis	LMG 460

For $S = 2$ we have analyzed all possible pairwise combinations (this number equals 190). We created in silico communities sampling an equal amount of 5.000 cells for every population; this means that an in silico community consists out of 10.000 cells. We trained a classifier using LDA and Random Forests on 70% of the in silico community; hereafter we predicted the population to which cells belong to contained in the 30% held out test set. We note that there was no need to tune the Random Forest classifier; using the preset $\sqrt{K}$, with K being the total number of available features to choose from at every split ($K = 12$, of which eight are fluorescence signals and four are scatter signals), gives rise to (near-)optimal results, in accordance with [2]. We note that after having performed ten-fold cross-validation for K on twenty randomly picked in silico communities, the increase in accuracy was 0.007 at the utmost. We expressed our performance for every in silico community in terms of the *area under the receiver operating characteristic curve* (AUC) and the *accuracy* (Fig. 1).

We note that the ensemble of pairwise combinations of populations give rise to performance accuracies ranging from 0.99 to near random guessing predictions; in other words, we were not biased towards highly discriminative populations. We have further summarized our results in Tab. 2, reporting the mean AUC and accuracy, along with their standard deviations, and the percentage of communities giving rise to performances higher than 0.90. Based on these numbers, we conclude that we are able to achieve high accuracies for a significant amount of possible communities. We note that a combination of *E. ictaluri* -

Table 2. Summary of analysis using LDA and Random Forests (RF) for $S = 2$. We denote the mean AUC (μ_{AUC}) and accuracy (μ_{acc}), along with the standard deviation ($\sigma_{\text{AUC/acc}}$) and the percentage of communities reporting a performance of 0.90 or higher.

	μ_{AUC}	μ_{acc}	σ_{AUC}	σ_{acc}	AUC > 0.90	acc > 0.90
LDA	0.90	0.83	0.089	0.088	62%	27%
RF	0.95	0.90	0.071	0.085	82%	65%

S. aromaticivorans results in the highest AUC of 0.999, a combination of *K. oxytoca* - *Z. m. subsp. mobilis* results in the highest accuracy, being 0.996.

Generally using Random Forests results in better performances than LDA, however, this is not always the case. 45% of the possible in silico communities report an increase in AUC of less than 0.03, 17% report an increase in accuracy less than 0.03.

In order to assess the fruitfulness of analyzing bacterial communities in a multiclass setting, we created 150 randomly chosen in silico communities for every increment of S, for which populations are evenly sampled (again 5.000 cells per population); this means that the total amount of cells contained in an in silico community N_{tot} equals $N_{\text{tot}} = S \times 5.000$. We used the same approach as described previously to perform LDA and Random Forests (i.e., creating a training set using 70% of the data and a test set using the other 30% of the data). We calculated the accuracy for every test set, after which we calculated the mean accuracy accompanied with its confidence-interval (CI) for every S (Fig. 2.)

For all values of S one is able to make single-cell predictions significantly better than random guessing. As S increases, both the mean accuracy and the size of the CI decline. This is due to the fact that for growing population richness, the degree in overlap between populations in the multidimensional 'FCM-space' starts growing. Therefore it is harder for classifiers to make a distinction between populations, which results in performances that are lower and more centered.

The difference in performance between the two classifiers increases as S increases. This means that for communities with a low richness ($S = 2, 3$) LDA might every so often be a sufficient method to perform single-cell predictions, however this is not always the case. This also means that although for low S a linear combination of variables already discriminates populations quite well, predictions can be improved by choosing classifiers which are able to combine variables in a non-linear way, especially for higher complexity communities.

3 Identifying bacterial populations in synthetic communities using in silico communities

We created three abundance gradients in order to verify to what extent an in silico community is able to identify its an in vitro community containing the same bacterial populations. We chose combinations which initially (according

Table 3. Three different combinations (Comb.) of bacterial populations used to create abundance gradients (dataset 2).

Comb.	Population 1	Population 2
1	*P. fluorescens*	*P. Putida*
2	*A. rhizogenes*	*Janthinobacterium sp. B3*
3	*M. luteus*	*S. oneidensis*

to the analysis described above) reported a low (Comb. 1), medium (Comb. 2) and high performance (Comb. 3), respectively (Tab. 3). As we do not know the individual labels of the cells contained in these communities, we predicted the relative abundance of populations present in a community, which can be derived by summing the predicted labels of individual cells. We express the performance in terms of the *root mean squared error* (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}{n}}, \tag{1}$$

with p being the target relative abundance to predict, $\hat{p}$ the predicted relative abundance and n the the total number of bacterial communities constituting the abundance gradient; $n = 13$ in this case[1]. Because we measured the populations individually beforehand, we are able to construct the same abundance gradient in silico. This enables us to not only carefully examine to what extent these abundances can be retrieved, but also to compare the in silico results with the in vitro results. The resulting RMSE is summarized in Tab. 4, along with the mean AUC calculated for the ensemble of communities constituting in silico abundance gradients; the predicted abundance gradients are visualized in Fig. 3.

Comb. 2 (Figs. 3CD) and 3 (Figs. 3EF) are well-predicted as opposed to Comb. 1 (Figs. 3AB), which is reflected in the RMSE; the mean AUC however reports quite a high AUC for Comb. 1 when using Random Forests, albeit still lower than for Comb. 2 and 3. The results for the in vitro analysis of Comb. 2 and 3 give rise to a similar RMSE, although we expected from initial performances that these values would be different. To investigate this issue, we added additional results in Tab. 5. We report the performance of a classifier in terms of the accuracy and the AUC trained on 70% and evaluated on 30% of the new in silico communities; in other words, classifiers were trained in the same way as in the previous section, so that a succinct comparison is possible with the originally reported values (*).

We note that although the performances are similar for Comb. 3, this is not the case for Comb. 1 and 2. Whereas the performances for Comb. 1 initially reported higher, the performances for Comb. 2 initially reported lower. This explains why the results for the in vitro analysis for Comb. 2 and 3. report similar

[1] We have constructed communities with the following relative abundances (population 1/population 2): 1%/99%, 5%/95%, 10%/90%, 20%/80%, 30%/70%, 40%/60%, 50%/50%, 60%/40%, 70%/30%, 80%/20%, 90%/10%, 95%/5% and 99%/1%.

Table 4. RMSE and mean AUC (μ_{AUC}) for predicted abundance gradients. RMSE has been calculated between the predicted gradients and the target gradients, both in silico and in vitro, having used LDA and a Random Forest classifier. μ_{AUC} has been calculated by calculating the AUC for every in silico community, and averaging over all in silico communities constituting the respective abundance gradient.

	Comb. 1	Comb. 2	Comb. 3
RMSE LDA in silico	0.29	0.0060	0.10
RMSE RF in silico	0.21	0.0036	0.022
RMSE LDA in vitro	0.51	0.036	0.096
RMSE RF in vitro	0.48	0.036	0.032
μ_{AUC} LDA in silico	0.64	1.0	0.93
μ_{AUC} RF in silico	0.88	1.0	1.0
σ_{AUC} LDA in silico	0.022	0.00070	0.0054
σ_{AUC} RF in silico	0.055	0.00039	0.00088

Table 5. Comparison of performances using LDA and Random Forests using datasets 1 as opposed to dataset 2. Performance using LDA and a Random Forest classifier for in-silico communities created with the same populations as in Comb. 1, 2 and 3, using the data reported in the previous section (dataset 1, denoted with *) and the abundance gradient data (dataset 2). These in silico communities are constructed and analyzed in exactly the same way, that is, they are evenly distributed communities consisting out of the same number of cells. Classifiers are trained on 70% of the data and evaluated on the opposite 30% test data.

	Comb. 1	Comb. 2	Comb. 3
AUC LDA*	0.64	0.82	0.96
AUC LDA	0.62	1.0	0.93
acc LDA*	0.62	0.77	0.92
acc LDA	0.59	0.99	0.91
AUC RF*	0.82	0.94	1.0
AUC RF	0.70	1.0	0.99
acc RF*	0.75	0.87	0.99
acc RF	0.64	1.0	0.97

results. However, this implies that although our approach is fruitful to analyze synthetic communities, performances are not yet exactly reproducible when individual bacterial populations are measured at different time points through FCM.

Furthermore, we emphasize the similar behavior between the in silico analysis (Fig. 3, left panel) and in vitro analysis (Fig. 3, right panel). We see that results are almost identical using either LDA or a Random Forest classifier analyzing Comb. 2. Moreover, inspecting Comb. 3, we note that using Random Forests increases the performance significantly, both for the in silico communities and in vitro communities; LDA suffers from a systematic bias, which is almost entirely (but not in full) reduced when one uses Random Forests.

4 Conclusion

After a thorough survey we can state that it is possible to predict the population to which bacterial single cells belong based on FCM data for low-complexity communities. Furthermore we have shown in a rigorous manner that in silico communities can be used to identify their in vitro counterpart communities. This leads to the conclusion that in silico communities form a viable representation for synthetic bacterial communities, and thus, we are allowed to use these in silico communities for further study. Supervised machine learning methods become therefore available to study issues as FCM data transformation [9] or feature selection, a topic studied in the field of immunology [10], but which seems te be lacking thus far in the field of microbiology. For low-complexity communities 'off-the-shelf' classifiers will most of the time already suffice to identify bacterial single cells. The outcome of this research therefore complies with the motivation to integrate supervised machine learning methods into standard FCM software [4].

A natural extension of this research would be to find the optimal multiclass method to analyze FCM data; a number of possibilities exist, ranging from binary classifiers which are naturally extendable to a multiclass setting or a combination of binary classifiers using a *one-versus-one (OVO)* or *one-versus-all (OVA)* approach [1]. However, it has to be noted that the performance of classifiers is not yet reproducible. The reason behind this is that bacterial populations exhibit heterogeneous behavior, which is reflected in FCM data [18]. However, FCM has been suggested to further quantify bacterial heterogeneity [5],[7], research in which in silico communities in combination with supervised machine learning methods might prove its value.

References

1. Aly, M.: Survey on multi-class classification methods. Tech. rep. (2005)
2. Bernard, S., Heutte, L., Adam, S.: Influence of hyperparameters on random forest accuracy. In: Benediktsson, JA and Kittler, J and Roli, F (ed.) Multiple Classifier Systems, Proceedings. Lecture Notes in Computer Science, vol. 5519, pp. 171–180. Int Assioc Patern Recognit & Tech Comm 1; IEEE Geosci & Remote Sensing Soc, IEEE Iceland Sect; Univ Cagliari; Univ Surrey (2009), 8th International Workshop on Multiple Classifier Systems, Univ Iceland, Reykjavik, ICELAND, JUN 10-12, 2009
3. Boddy, L., Morris, C., Wilkins, M., Al-Haddad, L., Tarran, G., Jonker, R., Burkill, P.: Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. Mar Ecol Prog Ser 195, 47–59 (2000)
4. Davey, H.M.: Prospects for the automation of analysis and interpretation of flow cytometric data. Cytometry A 77A(1), 3–5 (JAN 2010)
5. Davey, H.M., Winson, M.K., et al.: Using flow cytometry to quantify microbial heterogeneity. Curr Issues Mol Biol 5(1), 9–15 (2003)
6. Davey, H., Jones, A., Shaw, A., Kell, D.: Variable selection and multivariate methods for the identification of microorganisms by flow cytometry. Cytometry 35(2), 162–168 (FEB 1 1999)

7. Fernandes, R.L., Nierychlo, M., Lundin, L., Pedersen, A.E., Tellez, P.E.P., Dutta, A., Carlquist, M., Bolic, A., Schapper, D., Brunetti, A.C., Helmark, S., Heins, A.L., Jensen, A.D., Nopens, I., Rottwitt, K., Szita, N., van Elsas, J.D., Nielsen, P.H., Martinussen, J., Sorensen, S.J., Lantz, A.E., Gernaey, K.V.: Experimental methods and modeling techniques for description of cell population heterogeneity. Biotechnol Adv 29(6), 575–599 (NOV-DEC 2011)
8. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15, 3133–3181 (2014)
9. Finak, G., Perez, J.M., Weng, A., Gottardo, R.: Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics 11 (NOV 4 2010)
10. Hassan, S.S., Ruusuvuori, P., Latonen, L., Huttunen, H.: Flow cytometry-based classification in cancer cesearch: a view on feature selection. Cancer Informatics pp. 75–85 (04 2016)
11. O'Neill, K., Aghaeepour, N., Spidlen, J., Brinkman, R.: Flow cytometry bioinformatics. PLoS Comput Biol 9(12) (DEC 2013)
12. Pereira, G.C., Ebecken, N.F.F.: Combining in situ flow cytometry and artificial neural networks for aquatic systems monitoring. Expert Syst Appl 38(8), 9626–9632 (AUG 2011)
13. Rajwa, B., Venkatapathi, M., Ragheb, K., Banada, P.P., Hirleman, E.D., Lary, T., Robinson, J.P.: Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier. Cytometry A 73A(4), 369–379 (APR 2008), 24th International Congress of the International-Society-for-Analytical-Cytology, Budapest, HUNGARY, MAY 17-21, 2008
14. Ramette, A.: Multivariate analyses in microbial ecology. FEMS Microbiol Ecol 62(2), 142–160 (NOV 2007), Joint Symposium of the Environmental-Microbiology-Group/British-Ecological-Society/Society-for- General-Microbiology, Univ York, York, ENGLAND, SEP 13, 2006
15. Rubbens, P., Props, R., Boon, N., Waegeman, W.: Flow cytometric single-cell identification of populations in synthetic bacterial communities. Under review
16. Saleem, M., Fetzer, I., Dormann, C.F., Harms, H., Chatzinotas, A.: Predator richness increases the effect of prey diversity on prey yield. Nat Commun 3 (DEC 2012)
17. Toedling, J., Rhein, P., Ratei, R., Karawajew, L., Spang, R.: Automated in-silico detection of cell populations in flow cytometry readouts and its application to leukemia disease monitoring. BMC Bioinformatics 7 (JUN 5 2006)
18. Vives-Rego, J., Resina, O., Comas, J., Loren, G., Julia, O.: Statistical analysis and biological interpretation of the flow cytometric heterogeneity observed in bacterial axenic cultures. J Microbiol Methods 53(1), 43–50 (APR 2003)

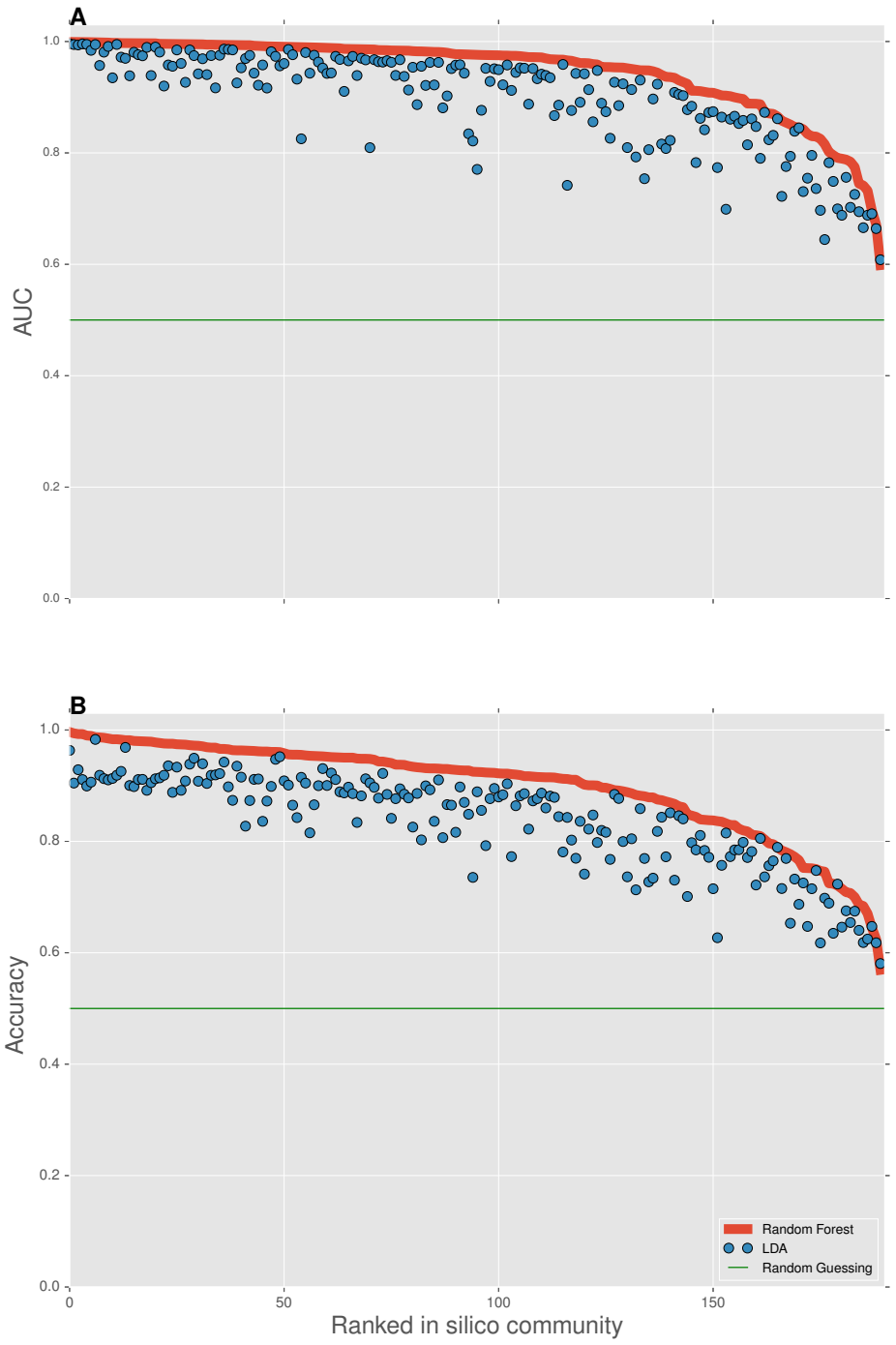

Fig. 1. Ranked performances using LDA and Random Forests for $S = 2$. A Ranked AUC. **B** Ranked accuracy. Performances are visualized for allevenly distributed 190 in silico communities and have been ranked in descending order according to the performances resulting from using Random Forests, accompanied with performances resulting from using LDA on the same in silico community. The performances have been calculated on a 30% held-out test set; figure taken from [15].

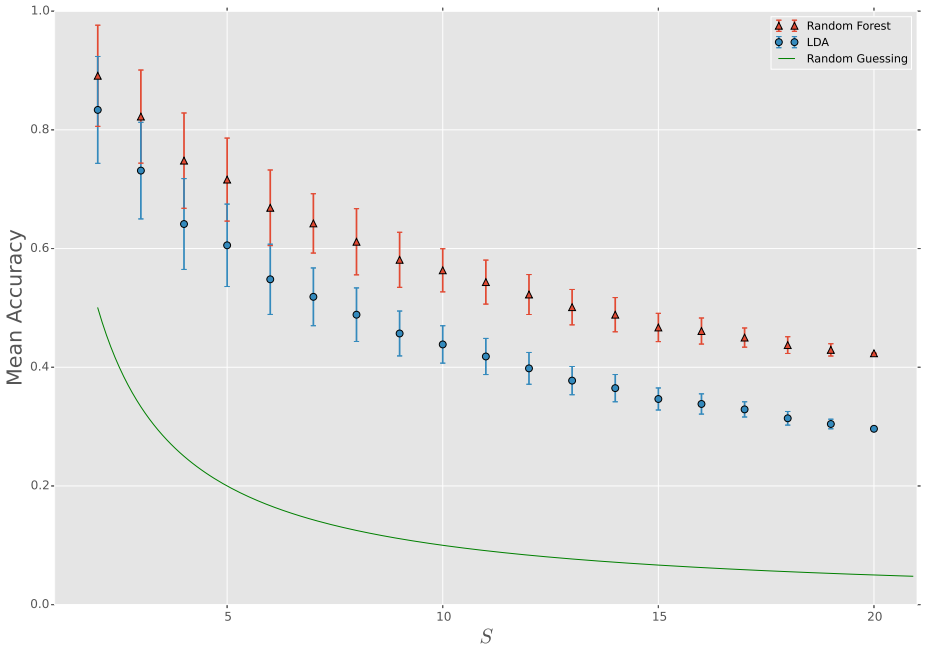

Fig. 2. Mean accuracy for increasing population richness using LDA and Random Forests. Mean accuracy along with a 68%-CI is displayed, resulting from an analysis using LDA and Random Forests for 150 randomly chosen in silico communities for $S = 2, ..., 18$ (for $S = 19$ and $S = 20$ this number becomes 20 and 1 respectively); every in silico community is evenly distributed. The accuracy has been calculated on a 30% held-out test set, after which the mean accuracy is calculated for the ensemble of silico communities for every increment of S; figure taken from [15].

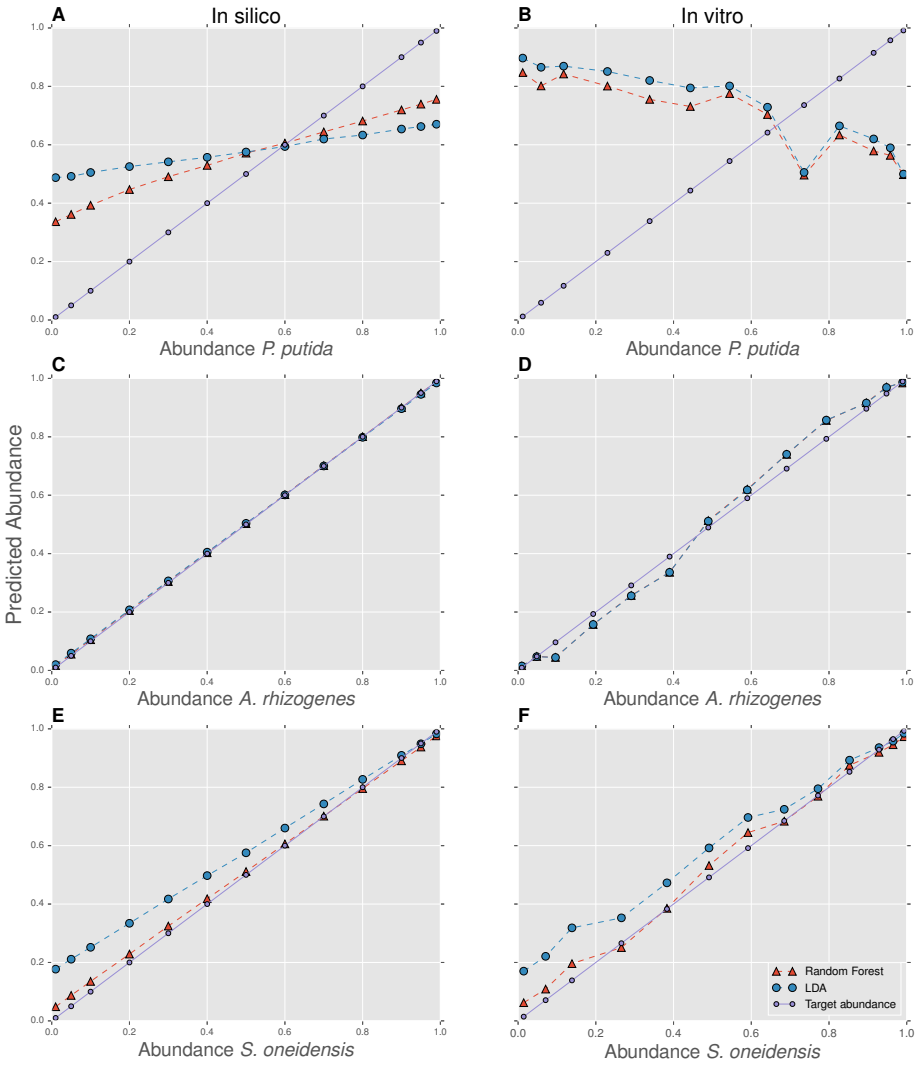

Fig. 3. Predicted abundance gradients. AB Comb. 1: *P. putida* – *P. fluorescens*; **CD** Comb. 2: *S. oneidensis* – *M. luteus*; **EF** Comb. 3: *A. rhizogenes* – *Janthinobacterium sp. B3*. Both the in silico (left panel) and in vitro (right panel) abundance gradients are visualized. The predicted relative abundance gradient is plotted against its target (i.e., designed in silico and in vitro) relative abundance for the first population of the three combinations. It follows that the relative abundance of the opposite population equals one minus the relative abundance of the first population (as $S = 2$); figure taken from [15].