Learning Single-Cell Distances from Cytometry Data

Bac Nguyen[☆a], Peter Rubbens[☆a,*], Frederiek-Maarten Kerckhof[b], Nico Boon[b], Bernard De Baets[a], Willem Waegeman[a]

[a] *KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure links 653, 9000 Ghent, Belgium*
[b] *Center for Microbial Ecology and Technology (CMET), Department of Biotechnology, Ghent University, Coupure links 653, 9000 Ghent, Belgium*

Abstract

Recent years have seen an increased interest in employing data analysis techniques for the automated identification of cell populations in the field of cytometry. These techniques highly depend on the use of a distance metric, a function that quantifies the distances between single-cell measurements. In most cases, researchers simply use the Euclidean distance metric. In this paper, we exploit the availability of single-cell labels to find an optimal Mahalanobis distance metric derived from the data. We show that such a Mahalanobis distance metric results in an improved identification of cell populations compared to the Euclidean distance metric. Once determined, it can be used for the analysis of multiple samples that were measured under the same experimental setup. We illustrate this approach for cytometry data from two different origins, i.e. flow cytometry applied to microbial cells and mass cytometry for the analysis of human blood cells. We also illustrate that such a distance metric results in an improved identification of cell populations when clustering methods are employed. Generally, these results imply that the performance of data analysis techniques can be improved by using a more advanced distance metric.

Keywords: Metric learning, flow cytometry, mass cytometry, microbiology, synthetic microbial communities, transfer learning

[☆] These authors contributed equally
[*] Corresponding author
Email address: peter.rubbens@ugent.be (Peter Rubbens[☆])

1. Introduction

Automated data analysis techniques are becoming increasingly popular in the field of cytometry. This can be attributed to the increasing dimensionality of cytometric assays and the increasing amount of acquired data per assay [1, 2, 3, 4]. Automated data analysis techniques include a number of preprocessing steps, such as specific transformations and quality controls of the data [5, 6], often followed by dimensionality reduction and clustering techniques to visualize the data and to determine cell populations in an automated way [7, 8, 9, 10]. The latter techniques usually depend on a predefined distance metric to measure the distances between single-cell measurements. In most techniques, the distance metric of choice is simply the Euclidean distance metric. Other distance metrics, such as the Mahalanobis distance metric, have been considered [11, 12, 13], but are rarely employed.

The availability of single-cell annotation has paved the way for supervised or semi-supervised machine learning techniques in the field of cytometry [14, 15]. Among those, distance metric learning exploits such available knowledge by learning a distance metric that appropriately measures the similarity of the data examples. The goal is to learn a distance metric that results in small distances between instances of the same class (in this case cell population) and large distances between instances of different classes. Recent developments in distance metric learning have shown that an advanced distance metric can lead to an increased performance for many distance-based techniques [16, 17, 18].

In this paper, the usefulness of a Mahalanobis distance metric for the analysis of cytometry data is explored. More specifically, a Mahalanobis distance metric is derived from the data using the Distance Metric Learning through Maximization of the Jeffrey divergence (DMLMJ) method [18]. DMLMJ leads to significant improvements in k-nearest-neighbour (k-NN) classification, and, more importantly, DMLMJ is scalable to large datasets. We compare the performance of k-NN classification according to cell population by DMLMJ with

the use of Euclidean distances. This is done for two cytometry applications: the first being flow cytometry in the field of synthetic microbial ecology (two datasets) and the second being mass cytometry or Cytometry by Time-Of-Flight (CyTOF) for the analysis of human cells (three datasets). We assessed the robustness of DMLMJ in the so-called transfer DMLMJ setting (t-DMLMJ). In this case, the Mahalanobis distance metric, which is determined using a specific sample, is used to quantify distances between cells that are part of a different sample that has been measured by the same experimental setup. Ultimately, we evaluated the impact of a Mahalanobis distance metric on the clustering analysis of CyTOF data of a disease versus control group [19], using the PhenoGraph algorithm [10]. Samples from multiple patients were included in the analysis, in order to evaluate an automated clustering analysis with incorporation of a Mahalanobis distance metric.

2. Materials & Methods

2.1. Distance Metric Learning

Many machine learning tasks, including dimensionality reduction, clustering, and visualization, rely on a distance metric that measures the (dis)similarity between data instances [17]. Unfortunately, common distance metrics, such as the Euclidean and Manhattan distance metrics, do not always achieve satisfactory results because they assume that all variables are independent and ignore correlations between them [16]. Therefore, learning an optimal distance metric from data is more desirable than using a default one. The underlying idea is to adjust the parameters of a distance metric that measures the distances between examples in order to improve the performance of a learning method [20, 21]. For example, in supervised classification settings, one tries to learn a distance metric so that the distances between examples of the same class are small, while distances between those of different classes are large [22]. We refer the reader to [17] for a comprehensive survey of distance metric learning.

2.1.1. k-Nearest-Neighbor Classification

We use k-nearest-neighbor (k-NN) classification in order to evaluate the robustness of a distance metric. The k-NN classifier is among the simplest yet most effective methods in terms of classification accuracy [23]. It has been considered one of the top 10 methods in data mining for performing recognition tasks [24]. The idea is to classify a test example by the majority label of its k nearest neighbors. It is well suited for multi-class classification problems with a large number of training examples. Moreover, the k-NN classifier makes no assumption about the training data, making it practical for many different application domains. Previous research efforts have shown that an appropriate distance metric can lead to a significant improvement of k-NN classification [16, 18, 20, 21], but the use of advanced distance metrics remains unexplored for cytometry data.

2.1.2. Distance Metric Learning through Maximization of the Jeffrey Divergence

In this study, we learn a Mahalanobis distance metric that preserves the similarity relationships among the data. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ denote the training set containing N cells, where $\mathbf{x}_i$ is the i-th cell in $\mathbb{R}^D$ (where D denotes the number of variables describing $\mathbf{x}_i$) and its corresponding cell population is y_i. Note that "cells" interchangeably refer to the machine learning terminology "instances" of a data set, and "cell populations" refer to the class labels. Formally, the Mahalanobis distance between two cells $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathbb{R}^D$ is defined as:

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)}\,,$$

where $\mathbf{M} \in \mathbb{R}^{D \times D}$ is a positive semidefinite matrix in order to guarantee that $d_{\mathbf{M}}$ satisfies the properties of a (pseudo)metric[1]. Since $\mathbf{M}$ is positive semi-definite, one can factorize (e.g., using eigenvalue decomposition or Cholesky decomposition) $\mathbf{M} = \mathbf{A}\mathbf{A}^\top$, where $\mathbf{A} \in \mathbb{R}^{D \times m}$ with $m = \mathrm{rank}(\mathbf{M})$. By doing

[1] A distance metric d defined on a set $\mathcal{X}$ is a nonnegative symmetric function satisfying the triangle inequality $d(\mathbf{x}_i, \mathbf{x}_j) + d(\mathbf{x}_j, \mathbf{x}_l) \geq d(\mathbf{x}_i, \mathbf{x}_l)$ for any $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l \in \mathcal{X}$.

so, the Mahalanobis distance metric implicitly corresponds to the Euclidean distance metric after applying the linear transformation $\mathbf{x} \mapsto \mathbf{A}^\top \mathbf{x}$, i.e.

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{A}\mathbf{A}^\top (\mathbf{x}_i - \mathbf{x}_j)} = \left\| \mathbf{A}^\top (\mathbf{x}_i - \mathbf{x}_j) \right\| .$$

The goal is to learn a Mahalanobis distance metric from the data (i.e. estimating $\mathbf{M}$), which amounts to learning a linear transformation (i.e. estimating $\mathbf{A}$). The latter has the advantage that it can be optimized easily since the positive semidefiniteness is no longer required. More importantly, when $m < D$, we implicitly reduce the number of variables describing the data. Here, we use the Distance Metric Learning through Maximization of the Jeffrey divergence (DMLMJ) method, recently proposed in [18]. We will briefly describe this method in what follows.

We consider a sample difference vector $(\mathbf{x}_i - \mathbf{x}_j)$, which is called a *positive difference* if $y_i = y_j$ (i.e., two cells are part of the same cell population) and a *negative difference* if $y_i \neq y_j$ (i.e., two cells are part of different populations). Consequently, we define P and Q as the distributions of positive differences and negative differences, respectively. Our goal is to find a linear transformation that maximizes the Jeffrey divergence between these two distributions in the transformed spaces $P_{\mathbf{A}}$ and $Q_{\mathbf{A}}$:

$$\underset{\mathbf{A} \in \mathbb{R}^{D \times m}}{\arg\max} \quad f(\mathbf{A}) = \mathrm{KL}(P_{\mathbf{A}}, Q_{\mathbf{A}}) + \mathrm{KL}(Q_{\mathbf{A}}, P_{\mathbf{A}}) \,, \tag{1}$$

where KL denotes the Kullback-Leibler divergence between two distributions. Note that the Jeffrey divergence is the symmetrized version of the Kullback-Leibler divergence. Increasing this separability criterion leads to a considerable improvement in the performance of the k-NN classifier. According to [18], problem (1) can be formulated as a generalized eigenvalue problem and therefore solved analytically by assuming that P and Q are multivariate Gaussian distributions with zero mean[2]. More specifically, let $\mathbf{\Sigma}_{\mathcal{S}}$ and $\mathbf{\Sigma}_{\mathcal{U}}$ denote the

[2] As noted by Nguyen et al [18], this is not a strong assumption and can easily be satisfied in practice.

covariance matrices of P and Q, respectively, then the solution to problem (1) is a matrix whose columns are m linearly independent eigenvectors of $\mathbf{\Sigma}_{\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{U}}$ corresponding to the m largest values of $\lambda_i + 1/\lambda_i$, where λ_i is the i-th eigenvalue of $\mathbf{\Sigma}_{\mathcal{S}}^{-1}\mathbf{\Sigma}_{\mathcal{U}}$. Note that m is a user-specified parameter, which denotes the desired number of variables in the transformed space. One can tune this parameter using a cross-validation approach. In practice, often a separate validation set is used to measure the performance of a classifier. When no validation set is available, we first create one containing 30% of the training data. DMLMJ is then trained with different numbers of variables. The optimal number of variables is the one that gives the highest score (accuracy or F1 score) on the validation set. At the end, DMLMJ returns a Mahalanobis distance metric that induces a linear transformation of the original data and reduces the number of variables from D to m at the same time.

2.1.3. Transfer Distance Metric Learning

Once the linear transformation $\mathbf{A}$ is determined for a specific sample, we can use it to quantify distances between cells for any related sample that is measured by the same setup. We refer to this setting as transfer distance metric learning. The idea is to learn a distance metric from one task and then apply it in other related tasks due to the lack of knowledge [25, 26], e.g. the unavailability of labeled data. We will denote the transfer distance metric learning through maximization of the Jeffrey divergence method as T-DMLMJ. A schematic overview illustrating the DMLMJ and T-DMLMJ setting can be found in SI Fig. 1.

2.2. Experiments

In order to characterize and validate the potential of DMLMJ for single-cell data, several experiments were conducted for two different cytometry applications, two generated by flow cytometry of synthetic microbial ecosystems and three generated by mass cytometry (CyTOF) for human blood cells. Data were retrieved from experiments that are publicly available. Experiments and

evaluation metrics are reported per dataset. Readers are referred to the original publications for a full overview of data collection and preprocessing. The Euclidean distance metric is used as a benchmark to evaluate the effect of a Mahalanobis distance metric, as determined by DMLMJ.

2.2.1. Application 1: Synthetic Microbial Communities

Dataset 1: Rubbens2017

Data from 20 individual bacterial cultures measured through flow cytometry (FCM) were retrieved from [27] (FlowRepository ID: FR-FCM-ZY6M). In brief, samples were stained with SYBR Green I and measured subsequently. Most bacterial cultures ($n = 17$) were in early-to-mid stationary phase, the other ones ($n = 3$) still were in exponential or linear growth phase. The samples were analyzed on a 3-laser FACSVerse flow cytometer (BD Biosciences), which contained two scatter detectors (forward and side) and eight fluorescence detectors, in which the FITC-detector (527/32 nm) was the targeted detector. Because peak area, height and width signals were captured, the experiment resulted in 30 measured variables in total. All variables were considered in DMLMJ, although only a subset of them contain biologically relevant information [27]. In this way, the functionality of DMLMJ can be evaluated, as we know that a linear transformation should be able to discover this information automatically. A full description of experimental details and preprocessing can be found in [27]. After measurement, samples were denoised in the asinh-transformed bivariate FITC-H – PerCP-Cy5.5-H space, using a robust digital gating strategy [28]. To ensure the quality of the data, the data were additionally filtered using the automated package flowAI (v1.4.4., default settings, target channel = FITC, changepoint detection penalty = 200) [6]. A full list of bacterial species and experimental details can be found in [14, 27].

Dataset 2: Sgier2016

Data were collected from [29] (FlowRepository ID: FR-FCM-ZYLB), in which cyanobacterial and algal cultures were cultured and measured by FCM,

using a Beckman-Coulter Gallios flow cytometer. As these microbial populations exhibit autofluorescent properties, no fluorescence staining was needed. Ten fluorescence and two scatter detectors measured area, height and width signals from the pulse, resulting in 36 variables in total describing the experiment. Sgier et al. removed all cells that were above the signal saturation limit in any of the 12 parameters. We considered only those samples that contained more than 500 cells per replicate. This resulted in 31 individual strains that were used for further analysis.

Experimental Setup

Microbial communities were assembled in different compositions using a data-aggregation step. This means that cells were sampled from bacterial populations that were measured individually and combined into artificial communities, so-called *in silico* communities [14]. The same number of cells ($n = 10,000$ for dataset 1, $n = 1,000$ for dataset 2) were sampled for every population, distributed over the number of technical replicates that were available. Half of the cells were added to a training set and the other half to the test set for every *in silico* community. The complexity of a community can be expressed in terms of the species richness S, denoting the number of phylogenetically distinct microbial populations that were combined in a community. The total number of cells in both training and test set amounts to $S \times n$. The following experiments were conducted for both datasets:

1. Ten *in silico* communities were assembled at random for every increment of S ranging from two to ten. Performance was evaluated in terms of classification accuracy of cells that were part of a held-out test set. In all cases, cells were classified according to their phylogeny using k-NN classification based on the Euclidean distance metric or the Mahalanobis distance metric, as determined by DMLMJ. Experiments were carried out using two different settings, on the raw data and on data for which each variable was transformed by $f(x) = \text{asinh}(x)$, a transformation that is

used as a preprocessing step for the analysis of microbial cytometry data [28, 30].

2. Next, the communities for $S = 10$ were retained, which are called the target communities. For each community, microbial populations were randomly removed one by one, until only two remained. At each step, the resulting community contains one less population compared to the target community. DMLMJ was performed on these subset communities, after which the distance metric was incorporated in the k-NN classifier and cells of the target communities were classified according to this classifier. We call this the partial transfer DMLMJ (partial T-DMLMJ) setting.

3. Finally, a distance metric was determined using microbial populations of which none were present in the target communities for $S = 10$ (i.e., we have 20 in total, so half of the populations are part of a microbial community of interest, the target community, while ten remaining communities were used to only determine a distance metric). This distance metric was used to quantify distances between cells in the target communities, after which cells were classified using k-NN classification. Note that the labels of the target communities and the training set are not the same. This setup is called the transfer DMLMJ (T-DMLMJ) setting.

2.2.2. Application 2: Mass Cytometry

DMLMJ was also evaluated for three mass cytometry (Cytometry by Time-Of-Flight, CyTOF) datasets. All data were transformed using the transformation $f(x) = \text{asinh}(x/5)$, which is a common preprocessing step for the analysis of CyTOF data (see e.g. [31, 32]).

Dataset 1: Levine_13dim

The Levine_13dim-dataset originates from one healthy individual, in which bone marrow mast cells (BMMCs) were analyzed using a 13-color panel. Cell populations were labeled after manual gating using all markers; all markers were used for data analysis. This dataset, as presented by [33] is publicly available

on FlowRepository (ID: FR-FCM-ZZPH). The dataset was originally presented in [31].

Dataset 2: Levine_32dim

The Levine_32dim-dataset originates from two healthy individuals, in which BMMCs were analyzed using a 32-color panel. Cell populations were labeled after manual gating, which was done using 19 out of the 32 surface markers. All markers were used for data analysis. This dataset, as presented by [33], is publicly available on FlowRepository (ID: FR-FCM-ZZPH). This dataset was originally introduced in [10].

Dataset 3: HMIS-2

The HMIS-2 dataset, studying the Human Mucosal Immune System (HMIS), originates from 47 individuals, in which peripheral blood mononuclear cells (PBMCs) were analyzed through CyTOF using a 28-color panel. The dataset was originally published in [19]. This data is publicly available on FlowRepository (ID: FR-FCM-ZYRM). We used the 14 control (CTRL) samples and 14 samples which were identified with Crohn's Disease (CD). Cell labels were used as reported in [34]. In brief, cells were first clustered according to six major immune lineages on the basis of the expression of a corresponding marker. Next, cell populations were automatically determined per lineage using the Gaussian Mean Shift clustering algorithm [35], after which clusters were merged upon high correlation or discarded when they contained too few cells. This resulted in 57 cell types in total. Cell counts per sample and cell type are available in [34].

Experimental Setup

Training and test sets were created in a stratified manner, as the distribution of cell population labels is unbalanced in which instances of the major class heavily outnumber those of the minor class. For the first and second datasets, 20,000 labeled cells per sample were added to the training set and test set,

respectively. For the third dataset 10,000 cells per sample were added to the training and test set respectively. The following experiments were conducted:

1. Single-cell labelling using k-NN classification was conducted, in which the Euclidean distance metric was compared to the Mahalanobis distance metric, as determined by DMLMJ. The hyperparameters were tuned to maximize the macro average F1-score per cell class. The F1-score is calculated as the harmonic average of the precision (which quantifies the number of false positives) and recall (which quantifies the number of false negatives), and lies between zero and one. An F1-score of one resembles perfect cell label classification for all cell populations.
2. The visualization performance of DMLMJ on t-SNE [36] was assessed for the Levine_13dim and Levine_32dim datasets. The test sets from the previous setup were used for this analysis.
3. The effect of an advanced Mahalanobis distance metric was evaluated for the clustering performance of the PhenoGraph algorithm [10]. This was done using the HMIS-2 dataset. We first clustered all individual CD-files, in which distances were quantified using the Euclidean and Mahalanobis distance metrics. The transfer distance metric learning setting was also considered. We calculated a Mahalanobis distance metric for one subject, and applied this metric for the subsequent analysis of all other samples using PhenoGraph. This was done for all possible combinations. The clustering performance was reported using the V-measure [37]. The V-measure is defined as the weighted harmonic mean of completeness (i.e., a cluster is as complete as possible, containing all cells from a specific population) and homogeneity (i.e., a cluster is as pure as possible, containing only those cells that are part of a single population) [38]. Subsequently, all files, both from the CD and control group, were concatenated into one dataframe. PhenoGraph was run on the whole dataframe, after which relative cell fractions per cluster and per sample were tested for statistical differences using the Mann-Whitney-U test. This was done seven

times, once using Euclidean distances and six times using a Mahalanobis distance metric which was determined using a specific sample from the cohort, which was chosen at random.

2.3. Data and Code Availability

DMLMJ was implemented in MATLAB. The Python implementation of PhenoGraph was used with default settings. The implementation of t-SNE (default settings) and performance metrics can be found in the scikit-learn machine learning library (v0.19.1) [39]. All processed datasets, MATLAB code for DMLMJ and Python scripts can be found at `https://github.com/bacnguyencong/CytoDMLMJ`.

3. Results

3.1. Synthetic Microbial Communities

A total of 90 different microbial communities were assembled by using a data-aggregation step, creating *in silico* communities. Ten communities were evenly sampled for every increment of $S = 2, ..., 10$, in which S denotes the total number of microbial populations that were present in a community. Note that these populations were determined beforehand, and upon creation of an *in silico* community, a number of these populations will overlap (i.e., a number of these populations will be present in a number of communities, which is determined at random). The impact of the use of a learned Mahalanobis distance metric on k-NN classification was evaluated in terms of the classification accuracy, which denotes the fraction of correctly labeled cells according to the phylogeny of a single cell. We compared the Mahalanobis distance metric, as determined by DMLMJ, to the Euclidean distance metric in the context of k-NN classification. The accuracy was evaluated using a held-out test set. This was done for two different datasets, the first containing 20 bacterial populations stained with SYBR Green I, the second containing 31 microbial populations (cyanobacteria and algae) with autofluorescent properties. The accuracy increased when DMLMJ was

applied to predict the cell labels of the test data (on average 5.4% for dataset 1, 2.7% for dataset 2) (Fig. 1). An increase in the number of microbial populations resulted in a drop in accuracy.

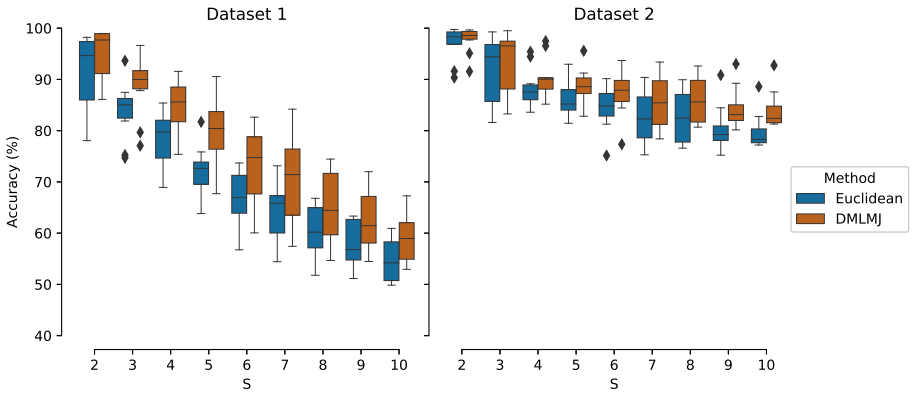


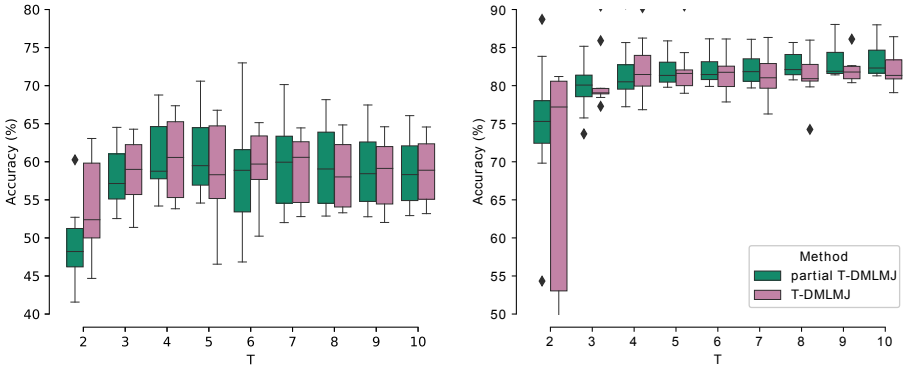
Figure 1: Classification accuracy of k-NN classification for an increasing population richness S with and without the use of DMLMJ. Each boxplot contains the classification accuracy for ten communities. Each box displays the 25% and 75% quartiles of the classification accuracy, of which the whiskers extend the range to maximal 1.5 times the interquartile range. Points that lie outside this range are visualized as outliers.

The same methodology was applied leaving out the asinh transformation as preprocessing step. Applying DMLMJ resulted again in an increased identification of cell phylogeny (SI Fig. 2). Optimizing the distance metric without transforming the data resulted in the best performance for dataset 1 (average accuracy over all communities was 77.7%), while for dataset 2 a combination of asinh and DMLMJ resulted in the best predictions (which resulted in an average accuracy over all communities of 88.6%). We conclude that DMLMJ is able to capture the similarity between examples of the same phylogeny and gave rise to a linear transformation of the data resulting in an improved classification of single cells. Note that preprocessing the data with asinh improved k-NN classification considerably for both datasets when Euclidean distances were used.

Next, the dependency on a specific microbial community of DMLMJ was

evaluated in the so-called transfer learning setting. In other words, we evaluated the capability of an optimized distance metric to quantify distances between cells of communities that have not been seen during the determination of the Mahalanobis distance metric during training. Two types of experiments were conducted, the so-called partial transfer learning setting using DMLMJ (partial T-DMLMJ) and the transfer learning setting using DMLMJ (T-DMLMJ). Both experiments started from the same test sets as created in the previous experiment for $S = 10$, denoted as the *target* communities (we have 10 in total). For the first experiment, one microbial population was randomly eliminated in an incremental way until only two microbial populations were left. A Mahalanobis distance metric was determined using DMLMJ at every step using the populations that were present in the reduced communities. If we let T denote this number, then T quantifies the number of common populations in the reduced and target communities and ranges from nine to two. This means that single cells in the target communities were classified using a k-NN classifier that uses a Mahalanobis distance metric based on a reduced number of bacterial populations. For the T-DMLMJ setting, a Mahalanobis distance metric was determined using microbial populations of which none was part of the target communities. For dataset 1, as the target communities contained ten bacterial populations, the left-out populations were used to determine the distance metric through DMLMJ for $T = 2, ..., 10$ (as there are 20 in total). For dataset 2 these were chosen at random for every community. Results are shown in Fig. 2.

When T was large enough, the accuracy remained stable. The classification accuracy for T-DMLMJ was comparable to (dataset 1) or slightly lower than (dataset 2) than that of the partial T-DMLMJ setting. Results for both partial T-DMLMJ and T-DMLMJ were slightly lower than or comparable to those of the DMLMJ setting (see $T = 10$ in Fig. 2, as in this case the partial DMLMJ setting is the same as for $S = 10$ in Fig. 1). Large variability in performance was reported for $T = 2$ in dataset 2. This indicates that the learned distance metric depends on the community. As T increased, performance increased and variability decreased, indicating that DMLMJ for this setting became indepen-



(a) Dataset 1: *In silico* bacterial communities (b) Dataset 2: *In silico* autofluorescent microbial communities

Figure 2: Classification accuracy of k-NN classification for ten target communities each containing ten microbial populations, in which the distance metric was determined using only a subset of the microbial communities present (partial T-DMLMJ) or a set of microbial communities of which none was part of the target community (T-DMLMJ). Data was first transformed using $f(x) = \text{asinh}(x)$. Each boxplot displays the 25% and 75% quartiles of the classification accuracy, of which the whiskers extend the range to maximal 1.5 times the interquartile range. Points that lie outside this range are visualized as outliers.

dent of the microbial community and thus useful for communities that contain a larger number of microbial populations. These empirical results confirm that DMLMJ results in an optimized distance metric that does not depend on a specific community, and is therefore generalizable to other communities that have been measured by the same experimental setup.

3.2. Application 2: Mass cytometry

DMLMJ was evaluated for three mass cytometry (CyTOF) datasets. Data were first split in a training and test set using stratified sampling, as cells are unevenly distributed over the different cell populations. We determined a Mahalanobis distance metric based on the training sets in function of the average F1-score over all cell populations (i.e., the macro-average). Performances of k-NN classification were reported for the test sets (Fig. 3). Using a Maha-

lanobis distance metric improved the performance of k-NN classification when compared to that of the Euclidean distance metric. The average increase for the Levine_32dim-dataset was 2.6%, for the HMIS-2 dataset 6.3%. The average F1-score dropped for the Levine_13dim-dataset, although the median was higher (91% versus 86.5%). This is due to the fact that most cell populations resulted in a better identification, but a few of them returned low identification scores, resulting in a lower average F1-score.

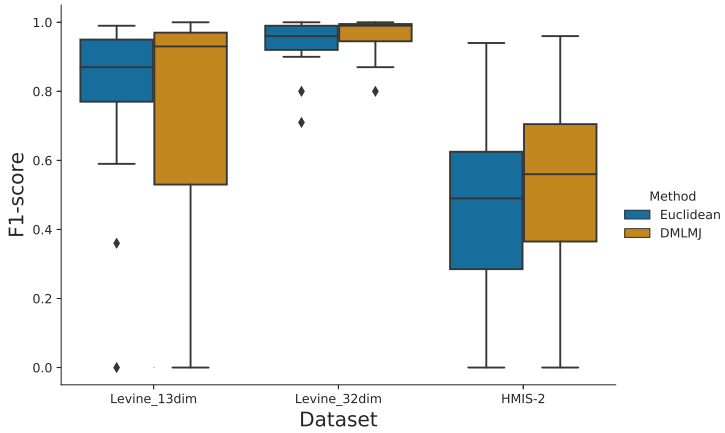


Figure 3: F1-score with and without the use of DMLMJ using k-NN classification of single-cell labels for CyTOF data. Boxplots show the distribution of F1-scores per dataset and per cell population, in which each cell population is represented by a black dot. Each boxplot displays the 25% and 75% quartiles of the F1-score, of which the whiskers extend the range to maximal 1.5 times the interquartile range. Points that lie outside this range are visualized as outliers.

In order to evaluate the effect of a Mahalanobis distance metric on an unsupervised analysis of CyTOF-data, we performed an additional unsupervised analysis of the HMIS-2 dataset. We first analyzed the samples that were diagnosed Crohn's Disease (CD, $n = 14$) using the clustering algorithm PhenoGraph. Each sample was first split in a training and test set. A Mahalanobis distance metric was determined for each sample separately, and used to transform the data of all other samples. In other words, a distance metric, determined using one sample, was used to quantify distances between cells for all other samples (the T-DMLMJ setting). Cells were clustered using PhenoGraph, and cluster-

ing results were evaluated using the V-measure, which lies between zero and one, in which a score of one equals 'perfect' clustering (Fig. 4A). Clustering was compared with samples analyzed using the Euclidean distance metric (Fig. 4A, last column). The clustering performance improved when using a Mahalanobis distance metric, both in the DMLMJ and T-DMLMJ setting (Fig. 4B). The reason for the improvement can be attributed to the number of clusters that are automatically determined by PhenoGraph (Fig. 4C). Although DMLMJ resulted in a lower number of variables compared to the original representation of the data, PhenoGraph will automatically detect more clusters compared to the use of the Euclidean distance metric.

We also assessed whether the use of DMLMJ improves the statistical power upon discriminating between samples which have been diagnosed with CD and a control (CTRL, $n = 14$) group. Samples from both groups were concatenated and cells were clustered using PhenoGraph. We determined the relative cell counts per cluster and sample, and tested for statistical differences using the Mann-Whitney-U test. P-values were adjusted using Benjamini-Hochberg correction. Data was analyzed using the Euclidean distance metric and using the Mahalanobis distance metric based on DMLMJ, which was applied for six different samples chosen at random (SI Figs 3 and 4). While the number of clusters that were retrieved was of the same order (sometimes higher, sometimes lower) compared to the use of the Euclidean distance metric, we always retrieved the same or more clusters that revealed statistically significant differences ($P \leq 0.05$) between the CD and CTRL group.

We also evaluated the use of t-SNE on the test sets of the Levine_13dim-dataset and Levine_32dim-dataset, with and without the use of DMLMJ to visualize the data (SI Figs. 5-7). t-SNE was able to return a clear visualization of most cell populations without the use of DMLMJ, although we note that DMLMJ improved the visualization to some extent. This was for example reflected in the visualization of megakaryocyte and erythroblast cells, which were more separated for the Levine_13dim dataset. In general, we note that separation between large cell populations that were already separated improved

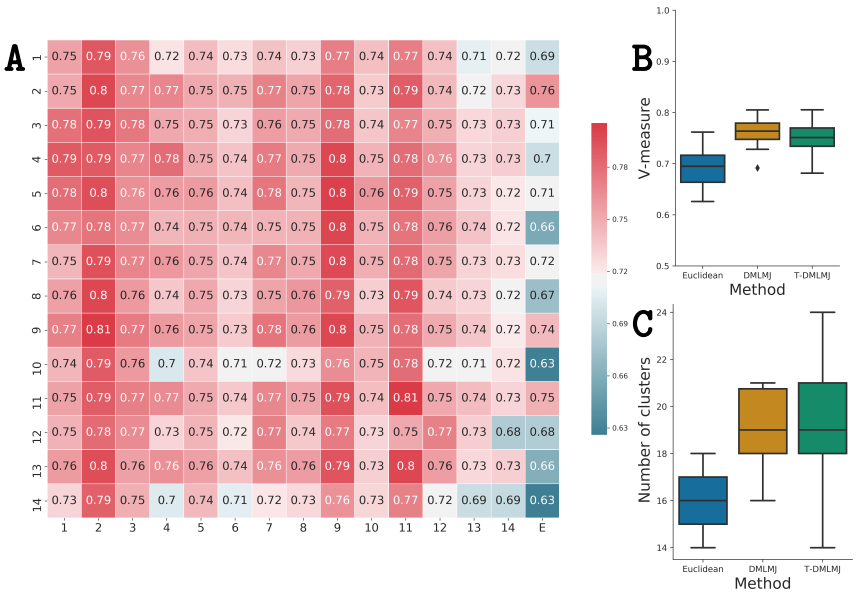


Figure 4: Summary of clustering results using PhenoGraph. **A**: Heatmap displaying V-measure results, in which each element V_{ij} corresponds to an analysis in which sample j was transformed by DMLMJ based on sample i. The column 'E' contains clustering results using the Euclidean distance metric. **B**: Boxplots showing the distribution of of the V-measure, using the Euclidean distance metric, DMLMJ (diagonal elements of the heatmap) and T-DMLMJ (the non-diagonal elements of the heatmap). **C**: Boxplots showing the distribution of the number of clusters that are detected per sample using PhenoGraph, using the Euclidean distance metric, DMLMJ (diagonal elements of the heatmap) and T-DMLMJ (the non-diagonal elements of the heatmap). Each boxplot displays the 25% and 75% quartiles of the V-measure or number of clusters, of which the whiskers extend the range to maximal 1.5 times the interquartile range. Points that lie outside this range are visualized as outliers.

slightly because of DMLMJ.

4. Discussion

More advanced distance metrics have been applied to cytometry data, but only to define distances between samples. Examples of these are the quadratic form (QF) distance [40, 41] and the Earth Mover's distance (EMD) [42]. To

the best of our knowledge, the use of more advanced distance metrics to define distances between single-cell measurements remains relatively unexplored in the field of cytometry. Most algorithms make use of the Euclidean distance metric to define a distance between two cells. A few studies have discussed the impact of alternative distance metrics for automated cell population identification, but only briefly. For example, Van Gassen et al. reported that the Euclidean distance metric gave the best results compared to the Manhattan and Chebyshev distance metrics for the FlowSOM algorithm [9]. Boddy et al. noted that a Mahalanobis distance metric consistently resulted in a 4% increase in terms of the classification accuracy compared to the weighted Euclidean distance metric for the classification of phytoplankton single cells using neural networks [43]. Pouyan & Nourani proposed a more advanced distance metric based on Random Forest graphs, showing that their distance metric improved k-NN classification compared to the use of the Euclidean or Manhattan distance metric [13]. Pouyan & Kostka proposed an extension of the previous model for the analysis of single-cell RNA sequencing data [44]. Kim and colleagues showed that the correlation-based similarity measures improved the automated clustering analysis compared to the distance-based similarity measures for the analysis of single-cell RNA sequencing data [45].

In this paper, we evaluated and quantified the usefulness of a Mahalanobis distance metric for the analysis of data originating from two different cytometry applications. We determined the coefficients of the Mahalanobis distance metric using Distance Metric Learning through Maximization of the Jeffrey divergence (DMLMJ). This method makes use of single-cell labels to determine a Mahalanobis distance metric in a data-driven way. DMLMJ was compared to the use of Euclidean distances for k-NN single-cell classification. We first showed that the performance of a distance-based method such as the k-NN classifier can be improved by employing an appropriate distance metric. Next, we tested whether a Mahalanobis distance metric, which is determined using a specific sample, can be used for the analysis of additional samples that have been measured by the same experimental setup. This property can be relevant

for different cytometry experiments. For instance, in the field of synthetic microbial ecology, researchers have knowledge of bacterial populations that are present in the experiment [46], and are able to measure their individual cytometric profile at the start of their experiment [14]. These data can then be used to perform distance metric learning, from which the resulting distance metric can subsequently be included in the analysis of the dynamics of the respective community. We used all variables for the Rubbens2017 dataset, for which we know that a number of them are redundant. Yet, previous work has shown that correlated variables, due to spillover, can assist in the discrimination between bacterial populations [27]. Therefore, DMLMJ results in a better identification of single cells, as it intrinsically performs a linear transformation of the data, in which only interactions between relevant variables are included. To illustrate this, we have included an example of a metric determined by DMLMJ (SI Fig. 8). When organisms contain autofluorescent properties, which is the case for the Sgier2016 dataset, researchers might not know which detectors to include in their analysis. Distance metric learning can offer a way to decide this automatically. Another important difference between these two datasets is the origin of the studied species. A large part of the studied populations of the latter consists of phytoplankton species, of which the cells are typically larger than those of bacterial populations. Therefore, their expression of scatter and fluorescence signals is also higher. This might explain why the asinh is not necessary to perform DMLMJ for the first dataset, while it is for the second. Note that the identification of microbial populations considerably improved for both datasets when we applied the asinh transformation as a preprocessing step.

Cytometry by Time-Of-Flight (CyTOF) is often applied using different machines and for various subjects. In case a supervised classification method breaks down due to variability between samples, an appropriate distance metric determined by DMLMJ can still be used to improve downstream cluster analysis. We have evaluated this in a clustering setting, in which a Mahalanobis distance metric was determined using one sample, and used for a clustering analysis for all other samples using the PhenoGraph algorithm. We have shown that DMLMJ

is robust for inter-subject variability and increases the cell population identification using PhenoGraph at the same. The reason for this is that, while DMLMJ results in a reduction of the number of variables, the number of clusters that are detected by PhenoGraph increases at the same time. Distance metric learning can therefore provide an alternative way to incorporate pre-existing knowledge. When cell population annotation is available for at least one sample, this information can be included in automated cell annotation techniques to analyze samples that have been studied with the same experimental setup. Naturally, as with any semi-supervised or supervised technique, the performance of distance metric learning depends on the quality of the annotated dataset. We also tested whether the use of DMLMJ resulted in more cell populations that captured statistical differences between a disease and control group analyzed by the same experimental setup. We conclude that in some cases additional cell populations gave rise to statistical differences, but it is not the case for any situation. However, statistical power never degraded. Future research will therefore focus on the evaluation of alternative distance metrics and their impact on downstream statistical analyses.

The cytometry applications studied in this research are related but quite different at the same time. Microbial FCM is characterized by the analysis of small particles, as bacterial cells are often orders of magnitude smaller in both size and volume compared to human eukaryotic cells. In addition, due to the high diversity of microbial communities, no general antibody-based panels have been established for microbial cells [47]. Therefore, one has to rely on general DNA stains, for which multicolor approaches are limited [48]. This results in few targeted detectors, resulting in overlapping cell populations, as the number of bacterial populations is typically larger than the number of differentiating signals. This is why the performance for synthetic microbial communities is lower compared to performances reported for CyTOF data. As cells for the latter are described by a lot more variables, many more cell populations can be separated.

To conclude, the performance of distance-based data analysis techniques

depends on the used distance metric. Distance metric learning provides a solution to improve their performance when some supervised information, in this case single-cell labels, is available. DMLMJ allows to estimate a Mahalanobis distance metric in a data-driven way, which improves the performance of automated cell population identification using both classification and clustering algorithms.

Acknowledgements

B. Nguyen is supported by the Special Research Fund (BOF15/DOS/039) of Ghent University. P. Rubbens is supported by the Special Research Fund (BOFSTA2015000501) of Ghent University. The authors declares that there is no conflict of interest regarding the publication of this article.

References

[1] K. O'Neill, N. Aghaeepour, J. Spidlen, R. Brinkman, Flow cytometry bioinformatics, PLoS Comput Biol 9 (2013) e1003365.

[2] R. R. Brinkman, N. Aghaeepour, G. Finak, R. Gottardo, T. Mosmann, R. H. Scheuermann, Automated analysis of flow cytometry data comes of age, Cytometry A 89 (2016) 13–15.

[3] Y. Saeys, S. Van Gassen, B. Lambrecht, Computational flow cytometry: helping to make sense of high-dimensional immunology data, Nat Rev Immunol 16 (2016) 449–462.

[4] A. Rahim, J. Meskas, S. Drissler, A. Yue, A. Lorenc, A. Laing, N. Saran, J. White, L. Abeler-Drner, A. Hayday, R. R. Brinkman, High throughput automated analysis of big flow cytometry data, Methods 134-135 (2018) 164–176.

[5] G. Finak, J. M. Perez, A. Weng, R. Gottardo, Optimizing transformations for automated, high throughput analysis of flow cytometry data, BMC Bioinformatics 11 (2010) 546.

[6] G. Monaco, H. Chen, M. Poidinger, J. Chen, J. P. de Magalhaes, A. Larbi, Flowai: automatic and interactive anomaly discerning tools for flow cytometry data, Bioinformatics 32 (2016) 2473–2480.

[7] Y. Ge, S. C. Sealfon, Flowpeaks: A fast unsupervised clustering for flow cytometry data via k-means and density peak finding, Bioinformatics 28 (2012) 2052–2058.

[8] E. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, D. Pe'er, viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia, Nat Biotechnol 31 (2013) 545–552.

[9] S. Van Gassen, B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene, Y. Saeys, Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data, Cytometry A 87A (2015) 636–645.

[10] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, G. P. Nolan, Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis, Cell 162 (2015) 184–197.

[11] S. Pyne, X. Hu, K. Wang, E. Rossin, T.-I. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler, P. De Jager, J. P. Mesirov, Automated high-dimensional flow cytometric data analysis, Proc Natl Acad Sci U S A 106 (21) (2009) 8519–8524. doi:{10.1073/pnas.0903028106}.

[12] N. Aghaeepour, R. Nikolic, H. H. Hoos, R. R. Brinkman, Rapid cell population identification in flow cytometry data, Cytometry A 79 A (2011) 6–13.

[13] M. B. Pouyan, M. Nourani, Clustering Single-Cell Expression Data Using, IEEE J Biomed Health Inform 21 (4) (2017) 1172–1181.

[14] P. Rubbens, R. Props, N. Boon, W. Waegeman, Flow cytometric single-cell identification of populations in synthetic bacterial communities, PLoS One 12 (1) (2017a) e0169754.

[15] M. Lux, R. R. Brinkman, C. Chauve, A. Laing, A. Lorenc, L. Abeler-Drner, B. Hammer, flowlearn: fast and precise identification and quality checking of cell populations in flow cytometry, Bioinformatics 15 (2018) 1–9.

[16] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, J Mach Learn Res 10 (2009) 207–244.

[17] A. Bellet, A. Habrard, M. Sebban, Metric learning, Synthesis Lectures on Artificial Intelligence and Machine Learning 9 (2015) 1–151.

[18] B. Nguyen, C. Morell, B. De Baets, Supervised distance metric learning through maximization of the Jeffrey divergence, Pattern Recognit 64 (2017) 215–225.

[19] V. van Unen, N. Li, I. Molendijk, M. Temurhan, T. Höllt, A. E. van der Meulen-de Jong, H. W. Verspaget, M. L. Mearin, C. J. Mulder, J. van Bergen, B. P. Lelieveldt, F. Koning, Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets, Immunity 44 (5) (2016) 1227–1239. doi:10.1016/j.immuni.2016.04.014.

[20] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighbourhood components analysis, in: Adv Neural Inf Process Syst 17, 2005, pp. 513–520.

[21] J. V. Davis, B. Kulis, P. Jain, S. Sra, I. S. Dhillon, Information-theoretic metric learning, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 209–216.

[22] E. P. Xing, M. I. Jordan, S. Russell, A. Ng, Distance metric learning with application to clustering with side-information, in: Adv Neural Inf Process Syst 14, 2002, pp. 505–512.

[23] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans Inf Theory 13 (1967) 21–27.

[24] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems 14 (1) (2008) 1–37.

[25] Y. Zhang, D.-Y. Yeung, Transfer metric learning by learning task relationships, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 1199–1208.

[26] J. Hu, J. Lu, Y. P. Tan, Deep transfer metric learning, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2015, pp. 325–333.

[27] P. Rubbens, R. Props, C. Garcia-Timermans, N. Boon, W. Waegeman, Stripping flow cytometry: How many detectors do we need for bacterial identification?, Cytometry A 91A (2017b) 1184–1191.

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

[29] L. Sgier, R. Freimann, A. Zupanic, A. Kroll, Flow cytometry combined with visne for the analysis of microbial biofilms and detection of microplastics, Nat Commun 7 (2016) 11587.

[30] R. Props, M. L. Schmidt, J. Heyse, H. A. Vanderploeg, N. Boon, V. J. Denef, Flow cytometric monitoring of bacterioplankton phenotypic di-

versity predicts high population-specific feeding rates by invasive dreissenid mussels, Environmental Microbiology 20 (2) (2018) 521–534. doi:10.1111/1462-2920.13953.

[31] S. Bendall, E. Simonds, P. Qiu, E.-a. Amir, P. Krutzik, R. Finck, R. Bruggner, R. Melamed, A. Trejo, O. Ornatsky, R. Balderas, S. Plevritis, K. Sachs, D. Pe'er, S. Tanner, G. Nolan, Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum, Science 332 (6030) (2011) 687–696.

[32] R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, G. P. Nolan, Automated identification of stratifying signatures in cellular subpopulations, Proceedings of the National Academy of Sciences 111 (26) (2014) E2770–E2777. arXiv:1606.01346, doi:10.1073/pnas.1408792111.
URL http://www.pnas.org/lookup/doi/10.1073/pnas.1408792111

[33] L. M. Weber, M. D. Robinson, Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data, Cytometry A 89 (2016) 1084–1096.

[34] T. Abdelaal, V. van Unen, T. Höllt, F. Koning, M. Reinders, A. Mahfouz, Predicting cell types in single cell mass cytometry data, bioRxiv (2018) 316034doi:10.1101/316034.
URL https://www.biorxiv.org/content/early/2018/05/07/316034?{%}3Fcollection=

[35] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.

[36] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J Mach Learn Res 9 (2008) 2579–2605.

[37] N. Aghaeepour, A. H. Khodabakhshi, R. R. Brinkman, An empirical study

of cluster evaluation metrics using flow cytometry data, in: Adv Neural Inf Process Syst 22, no. 2, 2009, pp. 2–5.

[38] A. Rosenberg, J. Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure, Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language 1 (2007) 410–420.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J of Mach Learn Res 12 (2011) 2825–2830.

[40] T. Bernas, E. K. Asem, J. P. Robinson, B. Rajwa, Quadratic form: A robust metric for quantitative comparison of flow cytometric histograms, Cytometry A 73 (2008) 715–726.

[41] D. Y. Orlova, S. Meehan, D. Parks, W. A. Moore, C. Meehan, Q. Zhao, E. E. Ghosn, L. A. Herzenberg, G. Walther, QFMatch: Multidimensional flow and mass cytometry samples alignment, Scientific Reports 8 (1) (2018) 1–14. doi:10.1038/s41598-018-21444-4.
URL http://dx.doi.org/10.1038/s41598-018-21444-4

[42] D. Y. Orlova, N. Zimmerman, S. Meehan, C. Meehan, J. Waters, E. E. Ghosn, A. Filatenkov, G. A. Kolyagin, Y. Gernez, S. Tsuda, W. Moore, R. B. Moss, L. A. Herzenberg, G. Walther, Earth mover's distance (emd): A true metric for comparing biomarker expression levels in cell populations, PLOS ONE 11 (2016) e0151859.

[43] L. Boddy, C. Morris, M. Wilkins, L. Al-Haddad, G. Tarran, R. Jonker, P. Burkill, Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data, Mar Ecol Prog Ser 195 (2000) 47–59.

[44] M. B. Pouyan, D. Kostka, Random forest based similarity learning for single cell RNA sequencing data, Bioinformatics 34 (13) (2018) i79–i88. doi:10.1093/bioinformatics/bty260.

[45] T. Kim, I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, P. Yang, Impact of similarity metrics on single-cell RNA-seq data clustering, Brief Bioinform 00 (August) (2018) 1–11. doi:10.1093/bib/bby076.
URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby076/5077112

[46] K. De Roy, M. Marzorati, P. Van den Abbeele, T. Van de Wiele, N. Boon, Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities, Environ Microbiol 16 (2014) 1472–1481.

[47] C. Koch, S. Müller, Personalized microbiome dynamics Cytometric fingerprints for routine diagnostics, Mol Aspects Med 59 (2018) 123–134. doi:10.1016/j.mam.2017.06.005.
URL https://doi.org/10.1016/j.mam.2017.06.005

[48] B. Buysschaert, B. Byloos, N. Leys, R. Van Houdt, N. Boon, Reevaluating multicolor flow cytometry to assess microbial viability, Appl Microbiol Biotechnol 100 (21) (2016) 9037–9051. doi:{10.1007/s00253-016-7837-5}.