Analysis of Flow Cytometry Data by Matrix Relevance Learning Vector Quantization

Michael Biehl[1]*, Kerstin Bunte[2], Petra Schneider[3]

1 Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Groningen, The Netherlands, 2 CITEC, Faculty of Technology, University of Bielefeld, Bielefeld, Germany, 3 CEDAM, School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, United Kingdom

Abstract

Flow cytometry is a widely used technique for the analysis of cell populations in the study and diagnosis of human diseases. It yields large amounts of high-dimensional data, the analysis of which would clearly benefit from efficient computational approaches aiming at automated diagnosis and decision support. This article presents our analysis of flow cytometry data in the framework of the DREAM6/FlowCAP2 Molecular Classification of Acute Myeloid Leukemia (AML) Challenge, 2011. In the challenge, example data was provided for a set of 179 subjects, comprising healthy donors and 23 cases of AML. The participants were asked to provide predictions with respect to the condition of 180 patients in a test set. We extracted feature vectors from the data in terms of single marker statistics, including characteristic moments, median and interquartile range of the observed values. Subsequently, we applied Generalized Matrix Relevance Learning Vector Quantization (GMLVQ), a machine learning technique which extends standard LVQ by an adaptive distance measure. Our method achieved the best possible performance with respect to the diagnoses of test set patients. The extraction of features from the flow cytometry data is outlined in detail, the machine learning approach is discussed and classification results are presented. In addition, we illustrate how GMLVQ can provide deeper insight into the problem by allowing to infer the relevance of specific markers and features for the diagnosis.

Citation: Biehl M, Bunte K, Schneider P (2013) Analysis of Flow Cytometry Data by Matrix Relevance Learning Vector Quantization. PLoS ONE 8(3): e59401. doi:10.1371/journal.pone.0059401

Editor: Avi Ma'ayan, Mount Sinai School of Medicine, United States of America

Received April 5, 2012; **Accepted** February 16, 2013; **Published** March 18, 2013

Funding: PS was supported by the Medical Research Council UK (Strategic Biomarker Grant G0801473). KB gratefully acknowledges funding by DFG (HA 2719/6-1) and the CITEC Centre of Excellence at the University of Bielefeld. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: m.biehl@rug.nl

Introduction

We present in this article our main results obtained in the context of the DREAM6/FlowCAP2 *Molecular Classification of Acute Myeloid Leukemia Challenge 2011* [1–3]. This challenge was organized in a joint effort by the *Dialogue for Reverse Engineering Assessments and Methods* (DREAM) project [3–6] and the *Flow Cytometry: Critical Assessment of Population Identification Methods* (Flow-CAP) initiative [2].

Flow cytometry constitutes a powerful technique which is widely used in medical research and clinical practice for the study and diagnosis of various diseases [7]. Flow cytometry measurements typically yield a quantitative description of several tens or even hundreds of thousands of cells in a given sample. Light scatter and fluorescence properties are used to identify deviations from normal cell size or structure and to quantify functional properties in terms of, e.g., protein marker expressions [7,8]. The amount of available data, its high dimension, and the complexity of the diagnosis tasks trigger a significant interest in systems for automated analysis and decision support.

Along these lines, the DREAM6/FlowCAP2 challenge addressed the analysis of given flow cytometry data, representing peripheral blood and bone marrow samples of, in total, 359 subjects. Some of these corresponded to cases of Acute Myeloid Leukemia (AML) and the ultimate goal was to predict the condition of a number of patients whose diagnosis was unknown to the participants. Hence, the goal of the challenge could be formulated as a machine learning problem: From the given example data with known diagnoses, criteria were to be inferred which then allowed for the classification of the *test samples*.

We extracted feature vectors from the data in terms of a few characteristic quantities, summarizing the statistics of the observed marker values. Predictions were obtained by means of a specific machine learning technique termed Generalized Matrix Relevance Learning Vector Quantization (GMLVQ) [9–11]. This prototype based method extends standard Learning Vector Quantization [12,13] by using *Adaptive Distance Measures in Relevance LVQ*, which motivated the acronym and team name *Admire-LVQ*. In the challenge, our team achieved the best possible performance with respect to the required test set prediction.

In the following section a description of the data set and our analysis is given. Thereafter we present and discuss our main results and the obtained prediction. We conclude with a brief outlook on possible extensions and future work.

Data Set and Analysis

In this section we first describe the extraction of features from the given data. The specific machine learning analysis based on Generalized Matrix Relevance Learning Vector Quantization is outlined. Furthermore, its validation in terms of the given data set is discussed.

The data set provided in the challenge comprised 359 subjects. For each of these, a varying number of cells, on the order of a few thousands, had been analysed by means of flow cytometry, see [2,3] for details. The first 179 subjects served as the training data; label information $S^{\mu} \in \{1,2\}$, $\mu = 1,2,\ldots,179$ was provided, specifying 23 subjects as AML patients ($S^{\mu} = 2$). The remaining 156 subjects are referred to as *healthy donors* ($S^{\mu} = 1$) throughout this contribution. Note that the latter group of subjects includes a number of patients with a diagnose different from AML [14].

The task was to predict the diagnosis with respect to a test set of 180 subjects for which no label information was provided. The total number of AML cases in the test set, 20, was also disclosed to the participants. However, this information was not exploited in our approach. We have analysed the *transformed* and *compensated* flow cytometry data as provided by the organizers of the challenge [2,3]. In our analysis we omitted the non-specific isotope control data representing non-human binding antibodies, which corresponds to *tube 8* in the data set [3].

In clinical practice, a possible workflow is to sort cells according to a small number of *gating* variables in a first step, identifying potentially degenerate or immature cells. Subsequently, the selected cells are analysed according to the remaining markers, aiming at a reliable diagnosis and potential identification of the AML subtype [7,8]. In our approach we follow a simpler, more direct strategy in which we omit cell specific information. After visual inspection in terms of histograms we decided to represent the data by a limited number of statistical characteristics per patient and marker. Moreover, we took into account all markers at once in order to assign each subject to one of the two classes in a single processing step.

Feature Extraction and Normalization

A key step in the design of a classifier in this challenge was the extraction of appropriate features from the provided data. The data corresponding to *tubes 1–7* represents 31 characteristic quantities per cell: the so-called *Forward Scatter* on linear scale (FS Lin), the *Sideward Scatter* on logarithmic scale (SS Log), and 29 fluorescence intensities on logarithmic scale quantifying the expression of various surface proteins. All of these quantities are referred to as *markers* in the following. Table 1 lists the considered markers and the index $j = 1,2\ldots31$ which we refer to in the analysis.

Note that the potential *gating* markers FS Lin, SS Log, and CD45-EDC were provided for all cells in the data set. The other 28 markers were measured in one tube only, representing a subpopulation of cells per subject. We rescaled all markers by the respective largest possible value as to limit all observations to the interval [0,1].

FS Lin can be interpreted as a measure of cell size, while SS Log roughly quantifies intracellular granularity [7]. Note furthermore that the expression of IgG1 was measured by means of four different binding antigens. In our analysis, however, the corresponding values were treated as four independent markers ($j = 4,11,18,25$), formally.

For the purpose of a first, visual inspection, we computed histograms corresponding to the frequency of marker values in the training set. Figures 1 and 2 display histograms of 4 example markers: FS Lin ($j = 1$), SS Log ($j = 2$), CD45-EDC ($j = 3$), and CD10-PC7 ($j = 31$) for one patient per class ($\mu = 100$ and103). The main purpose of Figures 1 and 2 is to illustrate the extraction of feature vectors from the sample data which is described in the following.

For each patient and marker a varying number M of cell measurements, typically a few thousands, were made available. In our analysis, we did not make use of cell specific information, as it is done frequently in terms of a so-called *gating* procedure in clinical practice [7,8]. We extracted information only on the level of single marker statistics over the entire population of cells. A direct classification of histograms using, for instance, entropic distance measures or statistical divergences would be feasible here [15,16]. We resorted, however, to reducing the information to only six quantities per marker which summarize the characteristics of the corresponding histogram. We denote by $v_j^{\mu}(m)$ the value measured for marker $j (1 \leq j \leq 31)$ in individual cell $m (1 \leq m \leq M^{\mu})$ of patient μ. From the available data we determined the following quantities:

$$(a)\,\text{mean} \qquad mean_j^{\mu} = \frac{1}{M^{\mu}} \sum_{m=1}^{M^{\mu}} v_j^{\mu}(m)$$

$$(b)\,\text{standard deviation}$$

$$std_j^{\mu} = \left[\frac{1}{M^{\mu} - 1} \sum_{m=1}^{M^{\mu}} \left(v_j^{\mu}(m) - mean_j^{\mu} \right)^2 \right]^{1/2}$$

$$(c)\,\text{skewness} \qquad skew_j^{\mu} = \frac{1}{M^{\mu}} \sum_{m=1}^{M^{\mu}} \frac{\left(v_j^{\mu}(m) - mean_j^{\mu} \right)^3}{(std_j^{\mu})^3}$$

Table 1. List of the 31 markers used in the analysis.

1	2	3	4	5	6	7	8
FS Lin	SS Log	CD45- EDC	IgG1- FITC	Kappa- FIT	CD7- FITC	CD15- FITC	CD14- FITC
9	10	11	12	13	14	15	16
HLADR- FITC	CD5- FITC	IgG1- PE	Lambda- PE	CD4- PE	CD13- PE	CD11c- PE	CD117- PE
17	18	19	20	21	22	23	24
CD19- PE	IgG1- PC5	CD19- PC5	CD8- PC5	CD16- PC5	CD64- PC5	CD34- PC5	CD3- PC5
25	26	27	28	29	30	31	
IgG1- PC7	CD20- PC7	CD2- PC7	CD56- PC7	CD33- PC7	CD38- PC7	CD10- PC7	

doi:10.1371/journal.pone.0059401.t001

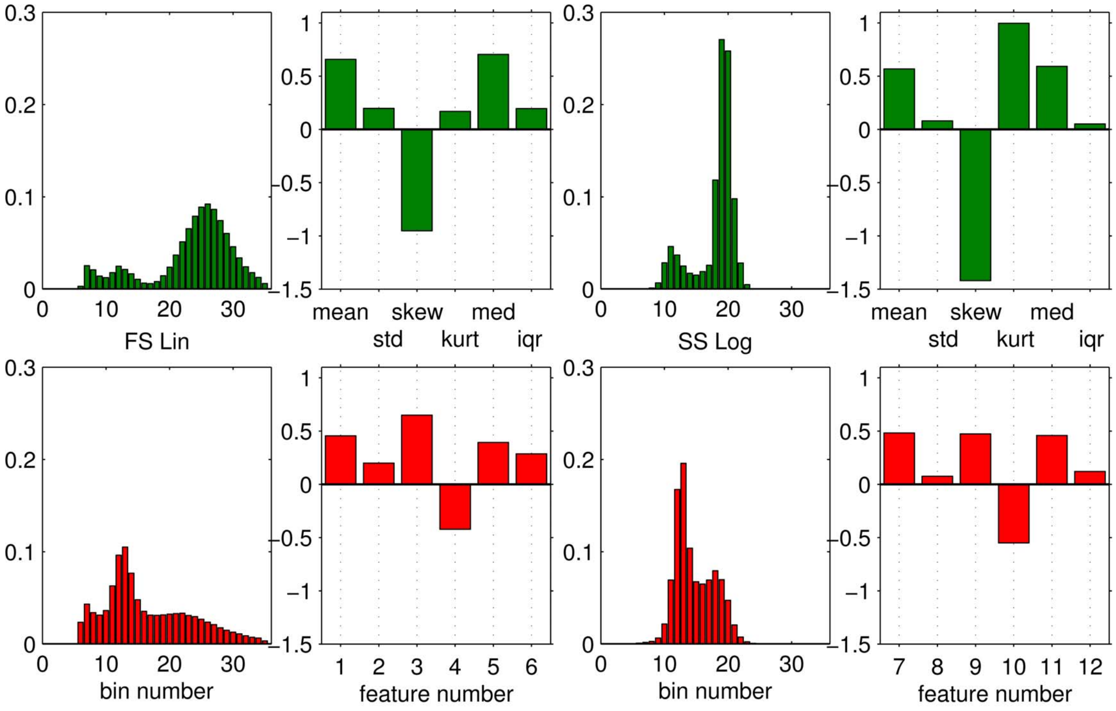

Figure 1. Example histograms and extracted features: FS Lin and SS Log. Histograms and extracted features correspond to one healthy donor (subject $\mu = 100$, upper panels) and one AML patient (subject $\mu = 103$, lower panels), respectively. Histograms display the frequency of a particular marker value for visual inspection. Six features are extracted per patient and marker, corresponding to mean, standard deviation, skewness, kurtosis, median, and interquartile range of the observed frequency of marker values, cf. Eq. (1). Here the first 12 components of the 186-dim. feature vectors are displayed before z-score transformation.
doi:10.1371/journal.pone.0059401.g001

$$(d)\,\text{kurtosis}$$

$$kurt_j^\mu = \frac{1}{M^\mu} \sum_{m=1}^{M^\mu} \frac{\left(v_j^\mu(m) - mean_j^\mu\right)^4}{(std_j^\mu)^4} - 3. \qquad (1)$$

In addition we computed (e) median (med_j^μ) and (f) interquartile range (iqr_j^μ) in the set of observed values $\left\{v_j^\mu(m)\right\}$. The skewness is a measure of the asymmetry, with positive values indicating that more weight is contained in the *left side of the histogram*. The kurtosis quantifies how sharply peaked a histogram is. Note that in the above defined $kurt$, sometimes termed *excess kurtosis* in the literature, a constant 3 is subtracted yielding $kurt = 0$ in case of normal densities.

Hence we obtained, for each patient μ, a set of 6 quantities per marker. A particular subject was subsequently represented by the concatenated vector $\mathbf{x}^\mu \in \mathbb{R}^N$ of $N = 186$ characteristic features. As one example, the skewness of marker 17 (CD19-PE, see Table 1) observed for patient 42 corresponds to component $skew_{17}^{42} = x_{99}^{42}$ of the feature vector $\mathbf{x}^{42}$ since $(17-1)\cdot 6 + 3 = 99$.

The features representing markers 1–3 (FS Lin, SS Log, CD45-EDC) and marker 31 (CD10-PC7) are shown for one example subject from each class in Figures 1 and 2, together with the corresponding histograms.

In the training processes described in the following, we applied an additional z-score transformation: Given a (sub-)set of P training examples we computed for $i = 1,2,\ldots,186$ the quantities

$$\bar{x}_i = \frac{1}{P} \sum_{\mu=1}^{P} x_i^\mu \text{ and } \sigma_i = \left[\frac{1}{P-1} \sum_{\mu=1}^{P} \left(x_i^\mu - \bar{x}_i\right)^2\right]^{1/2}$$

and rescaled all features in training, validation or test data by subtracting the mean $\bar{x}_i$ and subsequently dividing by σ_i. Consequently, the transformed features display zero mean and unit variance in the actual training set. While the transformation did not affect the classification performance, it enhances the interpretability of the results, in particular with respect to the relevance matrix, see below.

Matrix Relevance Learning Vector Quantization

We employed Generalized Matrix Relevance Learning Vector Quantization (GMLVQ) for the analysis of the obtained feature vectors. This highly flexible and powerful variant of LVQ is described in detail in [9–11]. Here we employed the algorithm in its simplest setting with one prototype per class and a single, global relevance matrix as defined below.

The two classes, i.e. healthy donors (class 1) and AML patients (class 2), are represented by the prototype vectors $\mathbf{w}^1, \mathbf{w}^2 \in \mathbb{R}^N$, respectively. Given a particular z-score-transformed feature vector $\mathbf{x} \in \mathbb{R}^N$ representing one of the patients, its distance from the prototypes is determined as

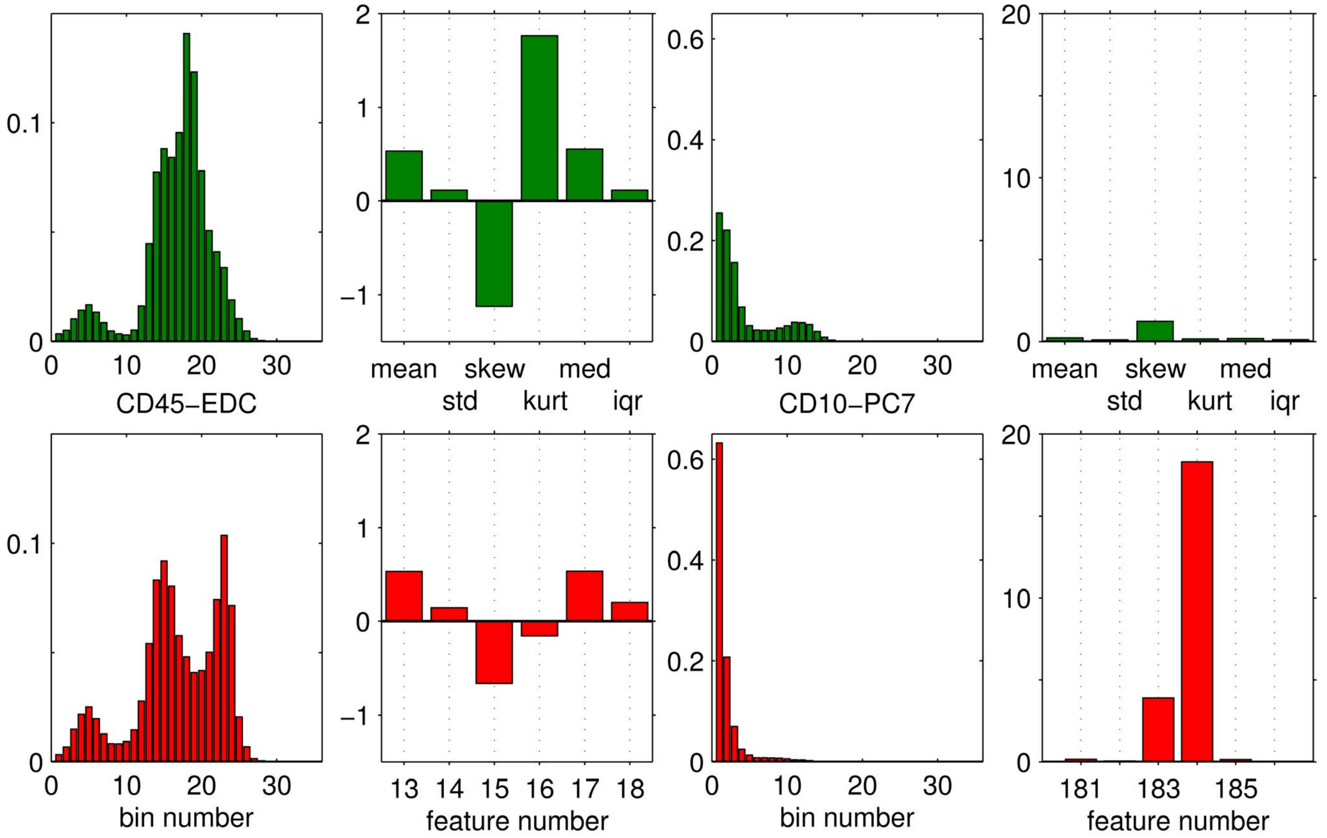

Figure 2. Example histograms and extracted features: CD45-EDC and CD10-PC7. For further description see Figure 1. The quantities displayed here correspond to features 13–18 and 181–186 of the 186-dim. vectors before z-score transformation.
doi:10.1371/journal.pone.0059401.g002

$$d(\mathbf{x},\mathbf{w}^i) = \sum_{j,k=1}^{N} \left(x_j - w^i_j\right) \Lambda_{jk} \left(x_k - w^i_k\right) = (\mathbf{x}-\mathbf{w}^i)^{\mathrm{T}}\,\Lambda(\mathbf{x}-\mathbf{w}^i) \qquad (2)$$

with $\Lambda = \Omega^{\mathrm{T}}\Omega$.

Here Λ and Ω are $N \times N$ matrices and the specific parameterization of the distance guarantees non-negativity of the measure:

$$d(\mathbf{x},\mathbf{w}^i) = \left[\Omega(\mathbf{x}-\mathbf{w}^i)\right]^2 \geq 0. \qquad (3)$$

In a simple *Nearest Prototype Classification* (NPC) scheme, a feature vector $\mathbf{x}$ is assigned to class 1 if $d(\mathbf{x},\mathbf{w}^1) < d(\mathbf{x},\mathbf{w}^2)$ and to class 2, else. While $\mathbf{w}^{1,2}$ serve as typical representatives of the classes, elements Λ_{ij} of the symmetric matrix Λ can be interpreted as to quantify the relevance of a pair of feature dimensions i,j in the classification scheme.

Both, prototypes and relevances, are determined in the same supervised training process. Given a set of P examples $\{\mathbf{x}^\mu, S^\mu\}_{\mu=1}^{P}$ with class labels $S^\mu \in \{1,2\}$, training is guided by the minimization of the cost function [9,17,18]

$$E(\mathbf{w}^1,\mathbf{w}^2,\Omega) = \sum_{\mu=1}^{P} e^\mu(\mathbf{w}^1,\mathbf{w}^2,\Omega), \qquad \text{with } e^\mu = \frac{d(\mathbf{x}^\mu,\mathbf{w}^J) - d(\mathbf{x}^\mu,\mathbf{w}^K)}{d(\mathbf{x}^\mu,\mathbf{w}^J) + d(\mathbf{x}^\mu,\mathbf{w}^K)} \qquad (4)$$

where the index J corresponds to the *correct* prototype with $J = S^\mu$ while $K \neq S^\mu$ identifies the *wrong* prototype. In general, the objective of training can be further specified by introducing a function $\Phi(e^\mu)$ in the cost function, e.g. a *sigmoidal* [17]. Here, we resorted to the simple case $\Phi(x) = x$. Note that the contribution e^μ of a single example to the cost function satisfies $-1 < e^\mu < 1$. It is negative if $\mathbf{x}^\mu$ is classified correctly and its absolute value relates to the *margin* of the classification.

Alternatively we refer to the closely related *score* s which is computed as

$$s(\mathbf{x}) = \frac{1}{2}\left[1 + \frac{d(\mathbf{x},\mathbf{w}^1) - d(\mathbf{x},\mathbf{w}^2)}{d(\mathbf{x},\mathbf{w}^1) + d(\mathbf{x},\mathbf{w}^2)}\right] \text{with } 0 \leq s \leq 1. \qquad (5)$$

A value $s(\mathbf{x}) \approx 0$ indicates that feature vector $\mathbf{x}$ is assigned to class 1, healthy donors, with high certainty. Large values close to $s = 1$ signal confident classification as an AML patient (class 2). The NPC scheme can be reformulated as assigning vector $\mathbf{x}$ to class 1 if $s(\mathbf{x}) \leq 1/2$ and to class 2 else. While the score may serve as a relative measure of certainty in GMLVQ, it should not be interpreted directly as a probability for AML. Note that any monotonically increasing function $f:[0,1]\rightarrow[0,1]$ could be used to transform s without modifying the actual ordering of patients according to s.

We implemented the iterative optimization of E, cf. Eq. (4), by means of a gradient descent procedure with respect to the adaptive quantities $\mathbf{w}^1,\mathbf{w}^2$, and Ω. At iteration step t, updates along the normalized gradients and subsequent normalization of Ω were performed:

$$\mathbf{w}^i(t+1)=\mathbf{w}^i(t)-\alpha_w(t)\,\frac{\partial E}{\partial \mathbf{w}^i}/|\frac{\partial E}{\partial \mathbf{w}^i}|$$

$$\Omega(t+1)=\frac{\widetilde{\Omega}(t+1)}{\left[\sum_{i,j}\widetilde{\Omega}_{ij}^2\right]^{1/2}}\ \text{where}$$

$$\widetilde{\Omega}(t+1)=\Omega(t)-\alpha_\Omega(t)\frac{\partial E}{\partial \Omega}/|\frac{\partial E}{\partial \Omega}| \qquad (6)$$

with the time-dependent step sizes $\alpha_w(t)$ and $\alpha_\Omega(t)$. The full form of the gradient terms is given in [9,11,19], for instance. We employed gradient descent with *waypoint averaging and step size control*, which has been introduced and described in greater detail in [19]: After a gradient step, Eq. (6), the achieved value of the cost function $E(t+1)=E(\mathbf{w}^1(t+1),\mathbf{w}^2(t+1),\Omega(t+1))$ is compared with $\hat{E}(t+1)=E(\bar{\mathbf{w}}^1(t),\bar{\mathbf{w}}^2(t),\bar{\Omega}(t))$ where

$$\bar{\mathbf{w}}^i(t)=\frac{1}{T}\sum_{\tau=1}^{T}\mathbf{w}^i(t-\tau)\ \text{and}\ \bar{\Omega}(t)=\frac{1}{T}\sum_{\tau=1}^{T}\Omega(t-\tau)$$

corresponds to the position in search space, on average over the last T updates. The observation of $\hat{E}(t+1)<E(t+1)$ signals oscillatory behavior of the iteration. In this case, we set $\mathbf{w}^i(t+1)=\bar{\mathbf{w}}^i(t)$ and $\Omega(t+1)=\bar{\Omega}(t)$ and reduced the step sizes by a factor $a<1$: $\alpha_{w,\Omega}(t+1)=a\cdot\alpha_{w,\Omega}(t)$. All results presented here were obtained with parameters $T=3$ and $a=2/3$ in the waypoint averaging scheme. Initial step sizes were $\alpha_w(0)=10/N$ for prototypes and $\alpha_\Omega(0)=100/N$ for matrix updates, respectively. In the problem at hand, the obtained classification scheme and error rates turned out very robust with respect to the choice of these parameters.

For a given training set, we initialized prototypes $\mathbf{w}^{1,2}$ close to the corresponding class conditional means with small random deviations; similarly we chose the initial Ω close to the $(N\times N)$ identity matrix:

$$\mathbf{w}^i(0)=\frac{\sum_{\mu=1}^{P}\delta_{iS^\mu}\,\mathbf{x}^\mu}{\sum_{\mu=1}^{P}\delta_{iS^\mu}}+\mathbf{r},\ \ \Omega(0)=\frac{I_{N\times N}+\Gamma}{\left[\sum_{i,j}(\delta_{ij}+\Gamma_{ij})^2\right]^{1/2}} \qquad (7)$$

where the Kronecker–Delta $\delta_{mn}=1$ if $m=n$ and $\delta_{mn}=0$ if $m\neq n$. The components of $\mathbf{r}\in\mathbb{R}^N$ and all elements of $\Gamma\in\mathbf{R}^{N\times N}$ were generated independently according to a uniform density $U(-0.1,0.1)$. Results were found to depend only very weakly on details of the initialization.

Validation

In order to evaluate the performance of the GMLVQ classifier before applying it to the test set, we employed a validation scheme based on randomized subsets of the available training data. In every run we selected ca. $3/4$ of the data from each class randomly, i.e. 17 of the 23 AML examples and 117 of the 156 healthy donors. These $P=134$ example data were used for training the GMLVQ system while the remaining 45 served as a validation set. The random split of the data was repeated 50 times and, if not stated otherwise, results presented in this section were obtained on average over the validation runs.

Figure 3 displays the averaged error rates of naïve Nearest Prototype Classification in the course of gradient based training. Note that an over-fitting effect was observed: Performing more than ca. 60 training steps decreased the error rates with respect to training examples to very low values. At the same time, however, validation set performance deteriorated. Closer inspection revealed that this effect was essentially due to patient $\mu=116$, listed as a case of AML in the training set. If contained in the validation set, this patient was consistently misclassified by the NPC scheme. On the contrary - if employed for training - the system achieved agreement with the label, eventually, but at the expense of an increased error rate in the validation set.

Based on this observation, we employed an *early stopping* strategy, terminating the training process after 40 gradient steps. When omitting patient $\mu=116$ from the training set or re-labeling the subject as healthy donor, the learning curves converged smoothly and overfitting was not observed anymore. Moreover, we obtained virtually the same classification, i.e. the same order of scores with respect to the test set patients in all these scenarios. The precise numerical results reported in the following section were obtained by means of the early stopping strategy including subject $\mu=116$ labelled as an AML case ($S^{116}=2$).

In addition to the error rates of the naïve NPC scheme we also evaluated the validation set performance in terms of the Receiver Operating Characteristics (ROC) [20]. By introducing a threshold Θ, the GMLVQ scheme can be biased with respect to one of the two classes:

$$\text{assign vector}\,\mathbf{x}\,\text{to class}\begin{cases}1 & \text{if}\, s\leq 1/2-\Theta\\ 2 & \text{else,}\end{cases} \qquad (8)$$

with the score s defined in Eq. (5) For thresholds in the range $-1/2\leq\Theta\leq 1/2$ we computed the corresponding class-wise error rates with respect to the validation set on average over the 50 training runs, yielding the threshold-averaged ROC curves [20] displayed in Figure 4.

The ROC analysis revealed very high sensitivity (true positive rate) and specificity (1 - false positive rate) with respect to the validation set performance, the corresponding Area Under Curve being $AUC=0.9935$ [20]. In addition, removal of patient $\mu=116$ from the data set resulted in an almost perfect ROC with $AUC=0.9996$. Given the close to error–free classification we refrained from employing complementary performance measures such as precision/recall or other characteristics [20]. For the same reason, we did not compare the validation performance of the simple GMLVQ scheme with more sophisticated settings or alternative classifiers.

Results and Discussion

Final results, including the test set scores, were obtained using all 179 training samples for training. In addition, we performed an average over 50 randomized intializations in order to rule out an

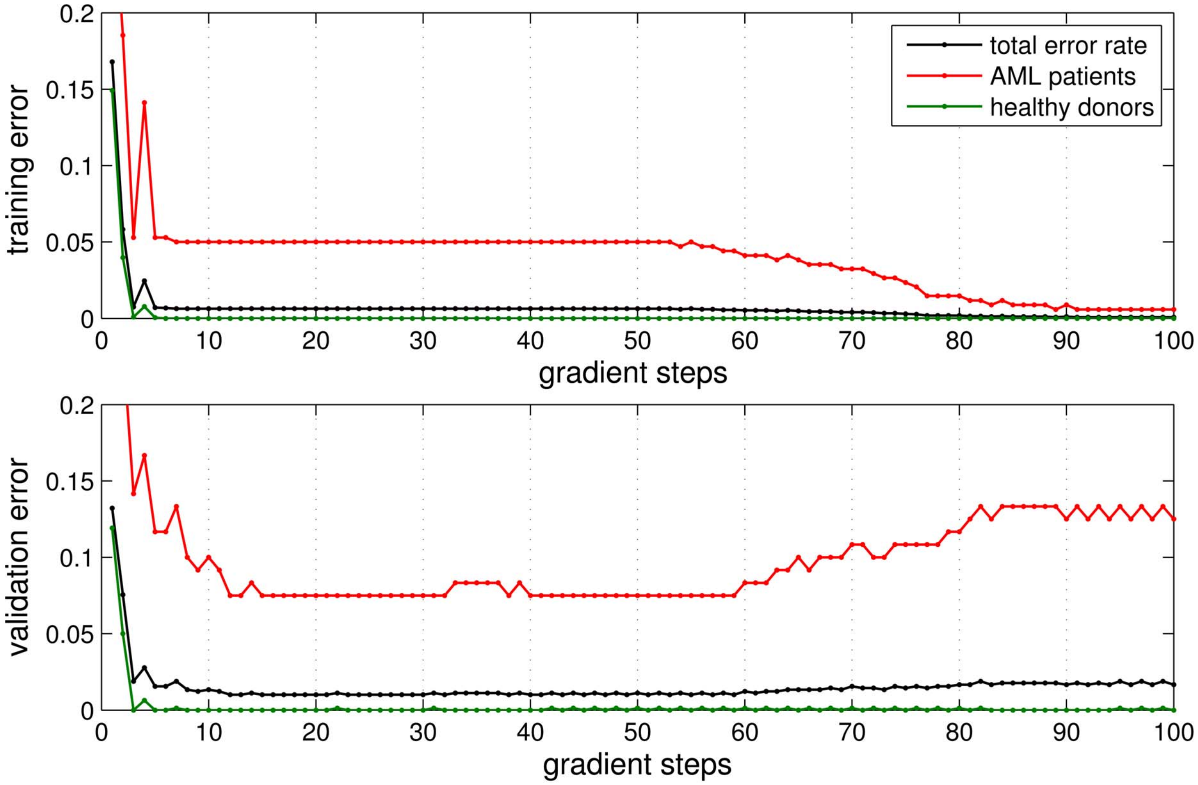

Figure 3. Learning curves in the validation procedure. Class specific and total error rates of Nearest Prototype Classification, corresponding to $\Theta = 0$ in Eq. (8), on average over 50 randomized validation runs. The upper panel corresponds to the performance in the respective training set, the lower panel displays error rates with respect to the validation set. The curves correspond to including patient $\mu = 116$ in training or validation set.
doi:10.1371/journal.pone.0059401.g003

influence of the initial configuration of the GMLVQ system. In each run, 40 gradient steps were performed with waypoint averaging and step size control as described above. The final test set scores were obtained on average over the 50 randomized training runs.

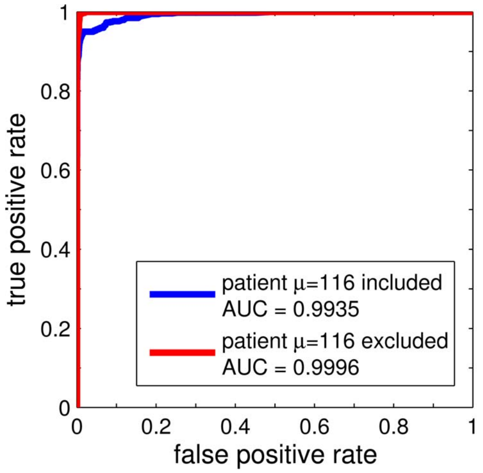

Figure 4. Validation set performance. Threshold-averaged ROC as obtained in the validation runs using labeled data. The curves correspond to using the data set including patient $\mu = 116$ (lower, blue line) and excluding patient $\mu = 116$ from the analysis completely (upper, red line), respectively.
doi:10.1371/journal.pone.0059401.g004

Before discussing the outcome of the GMLVQ training in terms of prototypes and relevances we present the actual test set predictions.

Test Set Prediction

Figure 5 displays the GMLVQ based scores s, cf. Eq. 5, with respect to the 180 test set patients. Values close to $s = 1$ correspond to patients that are identified as AML patients with high certainty, while small $s \approx 0$ correspond to a classification as healthy donor. It would be very interesting to study potential correlations of the scores s with additional information about the patients, e.g. measures of the *severity* of the AML cases. Unfortunately such information was not disclosed and is not available for the given data set.

Although it was known to the participants that the test set contained 20 AML cases, we did not make explicit use of this information. In the GMLVQ training, a threshold value does not have to be specified. The result in terms of scores s and the corresponding ranking of test set patients is independent of the actual number of AML cases. In a practical context, and if a *crisp* classification is the goal, the actual value of Θ should be set according to domain expert (user) preferences concerning the compromise between sensitivity and specificity. The example threshold value marked in the right panel of Figure 5 was chosen a posteriori for illustration purposes only and is neither a result nor a parameter of the training process. With respect to performance in the challenge it is irrelevant.

The comparison with the unknown test set labels after submission of the predictions [3] revealed that the 20 patients with highest GMLVQ score s corresponded precisely to the 20

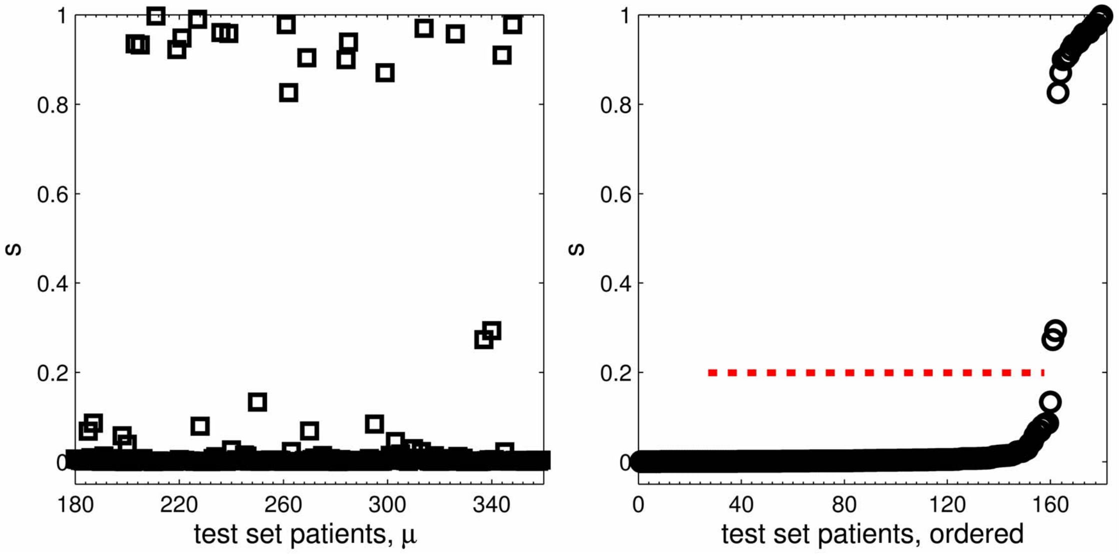

Figure 5. Test set predictions. GMLVQ based score s vs. patient number μ in the test set (left panel) and ordered according to s (right panel). The dotted line marks an example posterior choice of the threshold Θ, cf. Eq. (5), for crisp classification yielding correct prediction of 20 AML patients in the test set.
doi:10.1371/journal.pone.0059401.g005

AML patients in the test set. Hence, we achieved the best possible prediction according to Receiver Operator Characteristics or other evaluation methods like Precision/Recall, which only depend on the order of scores and the corresponding ranking of patients.

The obtained classifier can be illustrated in terms of a two-dimensional visualization: Figure 6 displays the training and test data in terms of projections on the leading eigenvectors of the relevance matrix Λ [21]. Two rather well separated clusters can be identified which reflect the assignment of classes. Note that the training set subject (patient 116) that was consistently misclassified by the NPC scheme is, indeed, located in the cluster representing healthy donors. This relates to the overfitting behavior discussed in greater detail in the previous section.

It is remarkable that error-free classification of the test set data was obtained by a number of teams who extracted different features from the data and used a variety of classification approaches [1]. For example, Vilar et *al.* employed a histogram based classifier in connection with the Kullback-Leibler divergence used as an entropic distance measure [16]. Amar et *al.* also extracted statistical moments from the data, but applied Support Vector Machine Regression, subsequently [22]. Logistic Regression was applied successfully by Manninen et *al.* [23]. Strickert and Seifert based their predictions on a method termed Correlative Matrix Mapping [24]. Using their software library *Jstacs* [25], Keilwagen and Grau built a weighted ensemble of classifiers which also achieved perfect classification.

An additional ranking of the best performing teams was suggested by the organizers in retrospect [1,3,26]. It hinges on interpreting the submitted scores as probabilistic assignments and on the reliability of the test set labels. In our opinion, the suggested posterior ranking according to, e.g., the Pearson correlation between scores s and the test set class labels is questionable, see also the DREAM6 discussion forum at [3].

Characteristics of the GMLVQ Classifier

Apart from yielding the actual classification scheme, the GMLVQ analysis provides insights into the structure of the data which become available by inspection of the prototypes and relevance matrix. The interpretability of the classifier has proven useful in several applications and facilitates discussions with the respective domain experts [27,28].

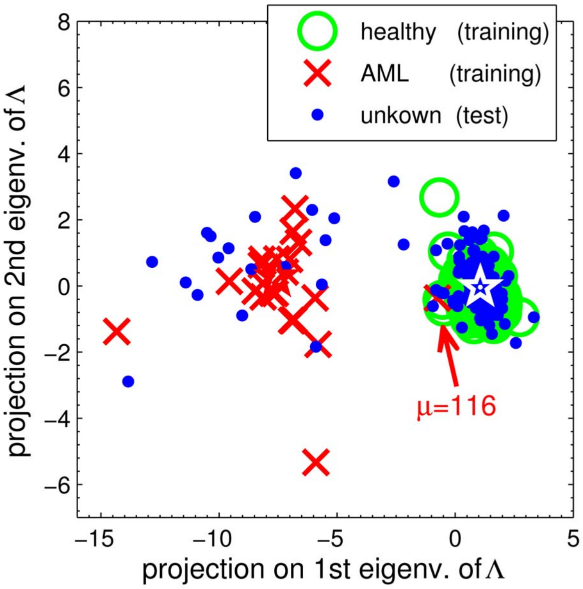

Figure 6. Visualization of the data set as obtained by GMLVQ. Projections of normalized feature vectors on the leading eigenvectors of Λ are displayed. Green circles correspond to healthy donors, red symbols mark AML patients in the training set, while blue dots represent test set data. Stars indicate the positions of the prototypes. The red arrow marks patient $\mu = 116$ in the training set, who is labeled as AML but is misclassified persistently for a large range of thresholds Θ, cf. Eq. (5).
doi:10.1371/journal.pone.0059401.g006

Figure 7 visualizes the difference vector $(\mathbf{w}^2-\mathbf{w}^1)$ of prototypes representing healthy donors (1) and AML patients (2), respectively. For the sake of clarity, we have shown only the 31 components which correspond to the features $mean_j$, cf. Eq. (1 a). A positive difference corresponds to markers which display a greater value in the AML prototype compared to the typical healthy donor in the data set, examples being HLA-DR-FITC ($j=9$), CD117-PE ($j=16$), and CD34-PC5 ($j=23$). Example markers which display reduced values $mean_j$ in AML patients are CD15-FITC ($j=7$), CD16-PC5 ($j=21$), and CD10-PC7 ($j=31$).

In addition we analysed the resulting relevance matrix $\Lambda=\Omega^{\mathrm{T}}\Omega\in\mathbb{R}^{186\times186}$. We focused on the diagonal elements $\Lambda_{ii}=\sum_j\Omega_{ij}^2$ which formally accumulate the importance of feature i for the resulting classification.

The direct interpretation of Λ is simplified if all features assume values on the same order of magnitude. This condition was realized here by the explicit z-score transformation mentioned above. Moreover, it is important to note that, given a particular set of feature vectors $\mathbf{x}^\mu$ and prototypes $\mathbf{w}^j$, a continuum of matrices Ω may exist which yield the same distances $d(\mathbf{x}^\mu,\mathbf{w}^j)$, cf. Eq. (3) and, hence, the same classification scheme with respect to the training data. This ambiguity problem is particularly pronounced for interdependent or highly correlated features in high dimension. Resulting difficulties concerning the interpretation of Λ in terms of feature relevances are discussed in depth in [28]. There, schemes are suggested for posterior regularization which provide unique, interpretable Ω and Λ which we also applied here: Note that arbitrary vectors from the null-space or kernel $ker(C)$ of the matrix.

$$C=\sum_\mu \mathbf{x}^\mu\mathbf{x}^{\mu\mathrm{T}}+\sum_j \mathbf{w}^j\mathbf{w}^{j\mathrm{T}}$$

can be added to the rows of a given Ω_{raw} without changing the GMVLQ cost function (4) and the actual classification of training data. In [28] a column space projection is suggested in order to remove contributions from $ker(C)$ [28]:

$$\Omega=\Omega_{raw}\Psi \quad \text{where} \quad \Psi=I-\sum_{\mathbf{u}^m\in ker(C)}\mathbf{u}^m\mathbf{u}^{m\mathrm{T}} \qquad (9)$$

is constructed from the eigenvectors $\mathbf{u}^m$ of C with eigenvalues zero.

Zero eigenvalues of C reflect the presence of linear dependent or strongly correlated features and the corresponding eigenvectors mark directions in input space in which training samples and prototypes do not vary. In the data considered here, one clearly expects dependencies between related markers, the four versions of IgG1 being an obvious example. In addition, extracted features like std_j and iqr_j or $mean_j$ and med_j should be strongly correlated.

For the following discussion we determined Ω by means of a posterior column space projection (9) with Ψ retaining only the leading eigendirections of C with eigenvalues $\geq 10^{-12}$. Thereafter, the matrix was normalized again to satisfy $\sum_{ij}\Omega_{ij}^2=1$ and we computed the regularized $\Lambda=\Omega^T\Omega$.

It is remarkable that, in the given problem, this posterior regularization has very little influence on the test set classification. In particular, the ordering of test set scores obtained with the regularized system is the same as the one presented in the previous section. This suggests that the correlations and dependencies observed in the training set are already representative for the entire data. In [28] example problems are presented where the posterior regularization has a non-trivial effect also on test set performance.

Figure 8 displays the diagonal entries of Λ for all 186 features. After regularization, the heuristic interpretation of Λ_{ii} as the relevance or significance of feature x^j in the classification is justified [28]. The figure displays the features in groups of six, corresponding to the 31 markers, cf. Table 1.

A relatively small number of markers appears to contribute the most significant features: FS-Lin (1), SS-Log (2), CD15-FITC (7), CD117-PE (16), CD16-PC5 (21), CD34-PC5 (23), and CD10-PC7 (31). A more detailed discussion of the obtained Λ_{ii} provides further, valuable information: For instance, the histogram shape as measured by skewness and kurtosis appears to be of minor importance with respect to marker 16 (CD117-PE), while measures of the corresponding histogram width (std, iqr) seem to

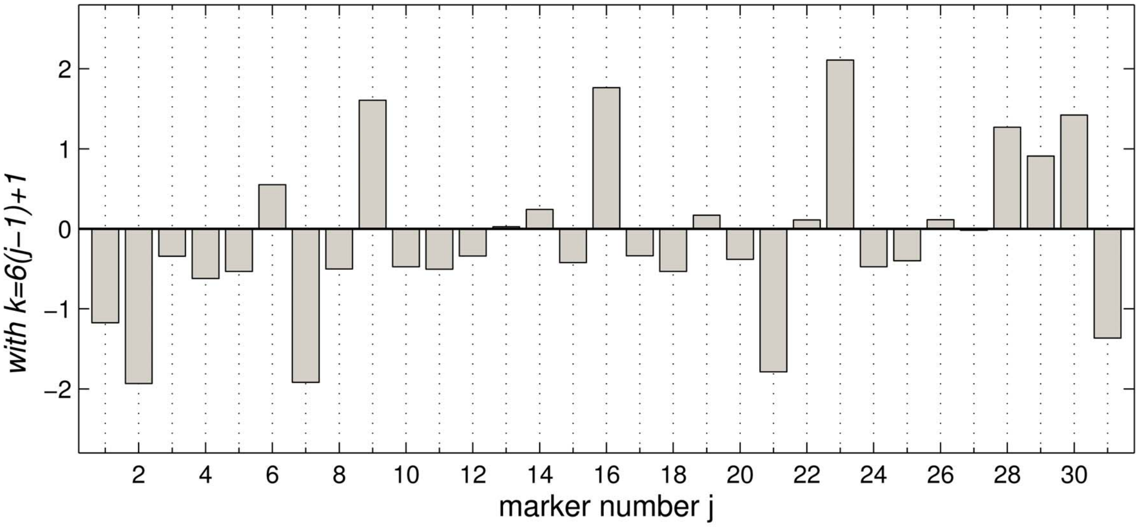

Figure 7. GMVLQ prototypes. Components of the difference vector $(\mathbf{w}^2-\mathbf{w}^1)$ corresponding to the feature $mean_j$, cf. Eq. (1), as represented by the AML prototype $\mathbf{w}_2$ and healthy donor prototype $\mathbf{w}^1$. Positive bars indicate that $mean_j$ is typically greater in AML patients than in healthy donors.
doi:10.1371/journal.pone.0059401.g007

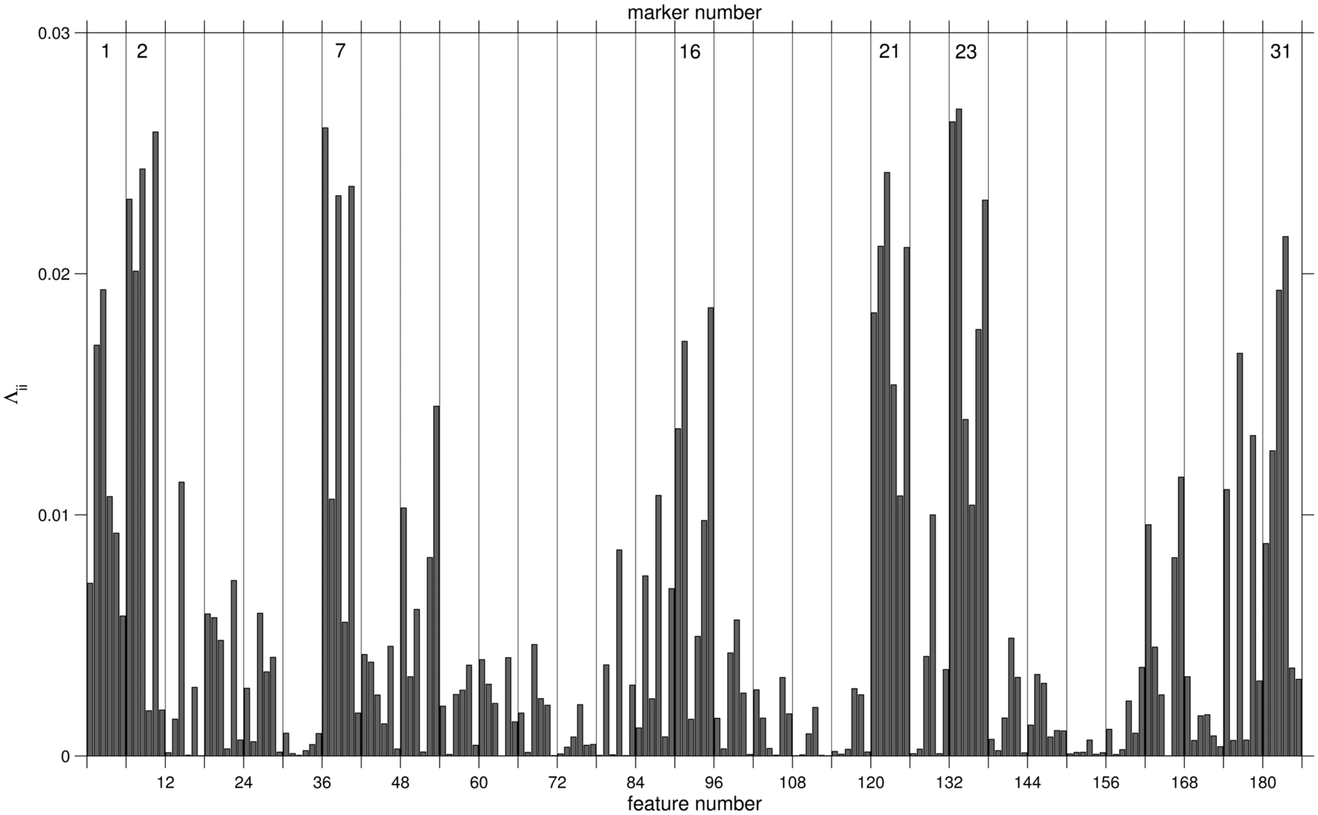

Figure 8. Relevance profile. Diagonal relevances Λ_{ii} of features $i=1,2,\ldots,186$. Vertical grid lines separate the groups of 6 quantities corresponding to each of the 31 markers, cf. Table 1. Marker numbers are given explicitly for 7 highly relevant ones.
doi:10.1371/journal.pone.0059401.g008

represent significant differences between AML patients and healthy donors. On the contrary, for CD10-PC7 (marker 31) skewness and kurtosis carry most discriminative power.

While several of the above mentioned markers have been discussed as relevant in the context of AML in the literature, see e.g. [7,28,29,30], their expression characteristics can vary a lot with the actual AML variant. For instance, both, HLA-DR positive and HLA-DR negative types of AML exist [31], the same applies to several other markers.

Due to the limited size of the data set and because information about AML subtypes was not disclosed, one should not over-interpret the results presented above. It is very likely that our findings in terms of relevances and prototypes are highly specific for the provided data set which seems to represent particular types of AML only. Nevertheless, our results demonstrate the interpretability of the GMLVQ approach and illustrate how the method could be used for efficient biomarker selection in collaboration with domain experts.

Obviously, the outcome and interpretation of relevance parameters depends on the precise form of the distance measure, Eq. (3), or more generally, on the parameterization of the classifier. For instance, systems with diagonal matrix Λ could only take into account single features and would disregard the discriminative power of particular pairs of features. Accordingly, featurs which display low relevance in our scheme might become significant in more complex classifiers. Nevertheless we believe that our method provides valuable insight into the discriminative power of features and pairs of features. The following simple experiment further illustrates this claim: We ranked features according to the corresponding Λ_{ii} and restricted the obtained GMLVQ classifier to the use of only 18 features for classification. All other features were omitted when evaluating distances and scores, cf. Eqs. (3,5), no re-training of the system was performed. The restricted classifier was evaluated in terms of its test set ROC. Close to perfect test set classification with an AUC ≈ 1 was retained when using only the leading 18 features which all are derived from the above mentioned 7 markers. It is interesting to note that also the following two subsets of 18 features, i.e. with relevance ranks 19–36 and 37–54, yielded excellent test set performance. Figure 9 shows how the resulting AUC decreases for subsequent subsets of 18 features with decreasing relevance. Performance deteriorated when subsets of features with very low relevance were used, resulting in essentially random class assignments with AUC ≈ 0.5.

A more reliable determination of discriminative markers, and even more so, the selection of a minimal set of features for correct classification would require systematic validation studies including the re-training of the GMLVQ system on the respective feature sets. Due to the limitations of the data provided in the challenge we postponed this line of research to forthcoming studies.

Outlook

More challenging data sets will have to be inspected to further demonstrate the usefulness of the approach in the analysis of flow cytometry data. This should, of course, include the systematic comparison with other methods. A comparison of various classifiers in the context of the FlowCAP2/DREAM6 challenge can be found in [1].

The identification of leukemia subtypes in a larger study population requires the introduction of several prototypes representing the class of AML patients. The extension of GMLVQ

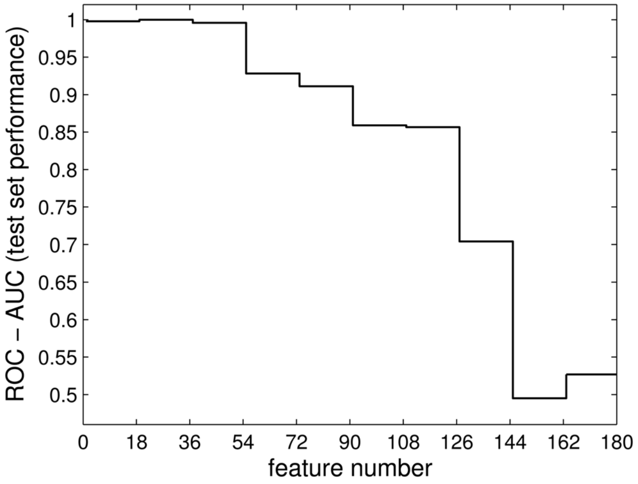

Figure 9. Test set performance (AUC of ROC) for the GMLVQ system restricted to subsequent subsets of 18 features which are ordered according to diagonal relevance Λ_{ii}. The AUC deteriorates from values close to 1 for highly relevant features to AUC ≈ 0.5 when using 18 features of low relevance.
doi:10.1371/journal.pone.0059401.g009

in terms of localized distance measures [9,11] appears also promising in this context.

The reliable identification of feature relevances for marker selection should also be based on larger, more representative data sets. For a successful application of GMLVQ for bio-marker selection in the context of tumor classification see [27]. The application of multi-class, potentially localized, GMLVQ will open new routes to the identification of discriminative markers in the differential diagnosis of AML subtypes. In forthcoming studies, the consideration of histogram specific distance measures will also be studied along the lines of [15].

The analysis presented here was based on the entire cell population of a given subject. More general problems, including the above mentioned identification of AML subtypes, might require an analysis on the level of individual cells. We intend to consider the development of prototype based automated gating procedures in forthcoming projects.

Available Software

The specific Matlab code used to generate our contribution to the DREAM6/FlowCAP2 *Molecular Classification of Acute Myeloid Leukemia Challenge 2011* is publicly available at http://www.the-dream-project.org/story/code [3].

A Matlab toolbox *Relevance and Matrix adaptation in Learning Vector Quantization*, including GMLVQ and important variants, is made available at http://matlabserver.cs.rug.nl/gmlvqweb/web/[32].

Acknowledgments

We would like to thank all members and supporters of the DREAM project and the FlowCAP initiative for the organization of this highly interesting challenge. We furthermore thank Wade T. Rogers for permission to use the data for this publication. We are indebted to Marc Strickert, Michael Seifert, Jens Keilwagen, and Jan Grau for drawing our attention to the challenge and for useful discussions. M.B. thanks Carmen Wesemeier for providing a brief introduction to the clinical practice of flow cytometry.

Author Contributions

Development and implementation of algorithms: MB KB PS. Analyzed the data: MB KB PS. Wrote the paper: MB KB PS.

References

1. Aghaeepour N, Finak G, The FlowCap Consortium, The Dream Consortium, Hoos H, et al. (2013) Critical assesment of automated flow cytometry analysis techniques. Nature Methods, advance online publication, 10 February 2013. URL http://www.nature.com/nmeth/journal/vaop/ncurrent/abs/nmeth.2365.html. Accessed 2013 Feb 19.
2. The FLowCAP initiative (2011) FlowCAP - Flow Cytometry: Critical Assessment of Population Identification Methods. The FlowCAP website. URL http://flowcap.flowsite.org. Accessed 2013 Feb 19.
3. The DREAM Project: Dialogue for Reverse Engineering Assessments and Methods (2011) DREAM6/FlowCAP2 molecular classification of Acute Myeloid Leukaemia. The DREAM project website. URL http://www.the-dream-project.org/challenges. Accessed 2013 Feb 19.
4. Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: The dream of high-throughput pathway inference. In: Stolovitzky G, Califano A, editors, Annals of the New York Academy of Sciences. volume 1115, 11–22.
5. Stolovitzky G, Prill R, Califano A (2009) Lessons from the dream2 challenges. In: Stolovitzky G, Kahlem P, Califano A, editors, Annals of the New York Academy of Sciences. volume 1158, 159–195.
6. Prill R, Marbach D, Saez-Rodriguez J, Sorger P, Alexopoulos L, et al. (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PLoS One 5: e9202.
7. Nguyen D, Diamond L, Braylan P (2007) Flow Cytometry in Hematopathology: A Visual Approach to Data Analysis and Interpretation (Current Clinical Pathology). Humana Press.
8. Gorczyca W, Weisberger J, Emmons F (2005) Atlas of Differential Diagnosis in Neoplastic Hematology. Taylor & Francis.
9. Schneider P, Biehl M, Hammer B (2009) Adaptive relevance matrices in learning vector quantization. Neural Computation 21: 3532–3561.
10. Schneider P, Bunte K, Stiekema H, Hammer B, Villmann T, et al. (2010) Regularization in matrix relevance learning. IEEE Transactions on Neural Networks 21: 831–840.
11. Bunte K, Schneider P, Hammer B, Schleif FM, Villmann T, et al. (2012) Limited rank matrix learning, discriminative dimension reduction and visualization. Neural Networks 26: 159–173.
12. Kohonen T (1997) Self-Organizing Maps. Berlin, Heidelberg: Springer, second edition.
13. Biehl M, Ghosh A, Hammer B (2007) Dynamics and generalization ability of LVQ algorithms. Journal of Machine Learning Research 8: 323–360.
14. Rogers W (2011) Description of the acute myeloid leukemia (AML) data set. Conference presentation, 6th Annual DREAM conference. The Recomb-DREAM2011 website. URL http://www.recomb-dream2011.org. Accessed 2013 Feb 19.
15. Mwebaze E, Schneider P, Schleif FM, Aduwo J, Quinn J, et al. (2011) Divergence based classification in Learning Vector Quantization. Neurocomputing 74: 1249–1435.
16. Vilar J (2011) Multidimensional entropies for diagnosing Acute Myeloid Leukemia from patient samples using flow cytometry data. Conference presentation, 6th Annual DREAM conference. The Recomb-DREAM2011 website. URL http://www.recomb-dream2011.org. Accessed 2013 Feb 19.
17. Sato A, Yamada K (1996) Generalized Learning Vector Quantization. In: Touretzky DS, Mozer MC, Hasselmo ME, editors, Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference. Cambridge, MA, USA: MIT Press, 423–429.
18. Hammer B, Villmann T (2002) Generalized relevance learning vector quantization. Neural Networks 15: 1059–1068.
19. Papari G, Bunte K, Biehl M (2011) Waypoint averaging and step size control in learning by gradient descent. In: Schleif FM, Villmann T, editors, Proc. Mittweida Workshop on Computational Intelligence, MiWoCi 2011. volume 6/2012 of Machine Learning Reports, 16–26.
20. Fawcett T (2006) An introduction to ROC analysis. Patt Rec Lett 27: 861–874.
21. Biehl M, Bunte K, Schleif F, Schneider P, Villmann T (2012) Large margin linear discriminative visualization by matrix relevance learning. In: The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8.
22. Amar D, Orenstein Y, Zeira R, Shamir R (2011) Regression based classification of AML flow cytometry profiles. Conference presentation, 6th Annual DREAM conference. The Recomb- DREAM2011 website. URL http://www.recomb-dream2011.org. Accessed 2013 Feb 19.
23. Manninen T, Huttunen H, Ruusuvuori P, Nykter M (2011) Logistic regression for AML prediction. Conference presentation, 6th Annual DREAM conference. The Recomb-DREAM2011 website. URL http://www.recomb-dream2011.org. Accessed 2013 Feb 19.

24. Strickert M (2011) Correlative matrix mapping. The Machine Learning Open Source Software (MLOSS) website. URL http://mloss.org/software/view/293/ . Accessed 2013 Feb 19.
25. Grau J, Keilwagen J, Gohr A, Haldemann B, Posch S, et al. (2012) Jstacs – a Java framework for statistical analysis and classification of biological sequences. JMLR 13: 1967–1971.
26. Norel R (2011) Scoring the classification of AML challenge. Conference presentation, 6th Annual DREAM conference. The Recomb-DREAM2011 website. URL http://www.recomb-dream2011.org. Accessed 2013 Feb 19.
27. Arlt W, Biehl M, Taylor A, Hahner S, Libé R, et al. (2011) Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors. Journal of Clinical Endocrinology & Metabolism 96: 3775–3784.
28. Strickert M, Hammer B, Villmann T, Biehl M (2013) Regularization and improved interpretation of linear data mappings and adaptive distance measures. In: Proc. of the 2013 IEEE Symposium on Computational Intelligence (IEEE SSCI). IEEE, p. 8 pages. In press.
29. Raspadori D, Damiani D, Lenoci M, Rondelli D, Testoni N, et al. (2001) CD56 antigenic expression in acute myeloid leukemia identifies patients with poor clinical prognosis. Leukemia 15: 1161–1164.
30. Feller N, Schuurhuis G, van der Pol M, Westra G, Weijers G, et al. (2003) High percentage of CD34-positive cells in autologous AML peripheral blood stem cell products reflects inadequate in vivo purging and low chemotherapeutic toxicity in a subgroup of patients with poor clinical outcome. Leukemia 17: 68–75.
31. Promsuwicha O, Auewarakul C (2009) Positive and negative predictive values of HLA-DR and CD34 of acute promyelocytic leukemia and other types of acute myeloid leukemia with recurrent chromosomal translocations. Asian Pac J Allergy Immunol 27: 209–216.
32. Biehl M, Schneider P, Bunte K (2013) Relevance and Matrix Adaptation in Learning Vector Quantization (website). URL http://matlabserver.cs.rug.nl/gmlvqweb/web. Accessed 2013 Feb 19.