Discrimination of malignant neutrophils of chronic myelogenous leukemia from normal neutrophils by support vector machine

Wanmao Ni a, Xiangmin Tong a,*, Wenbin Qian a, Jie Jin a, Hongchan Zhao b

a Institute of Hematology, The First Affiliated Hospital, College of Medicine, Zhejiang University, 79# Qingchun Road, Hangzhou 310003, PR China
b The Department of Clinic Laboratory, The First People’s Hospital of Hangzhou, Hangzhou 310006, PR China

1. Introduction

Flow cytometry has been applied in immunophenotyping of malignant hematological diseases for several decades and is helpful in the diagnosis of acute leukemia, chronic lymphocytic leukemia, myelodysplastic syndrome, lymphoma and others. In general, malignant cells were mainly characterized by aberrant antigen expression or asynchronous antigen expression according to their maturation stage in clinical flow cytometry analyses. However, mature neutrophils from CML patients have similar antigen expression patterns compared to normal neutrophils and have low probability of aberrant expression of antigens such as CD56 [1], CD2, CD5 and CD7 [2]. This similarity made it difficult to differentiate mature neutrophils of CML patients from normal mature neutrophils by routine flow cytometry. The Bethesda group also suggested that flow cytometry was not indicated in the differential diagnosis of mature neutrophilia in the absence of blasts [3]. Therefore, polymerase chain reaction (PCR) analysis of the BCR/ABL fusion gene is used as an alternative to flow cytometry in the diagnosis of CML clinically. The use of flow cytometry to discriminate between mature neutrophils from patients with chronic myelogenous leukemia and normal neutrophils (such as those from patients with leukemoid reaction, benign hematological diseases, infection and non-myeloid leukemia) is challenging. Routine clinical grade flow cytometers can provide simultaneous detection of eight or more measurements, while only two parameters are analyzed simultaneously in clinical sequential two-dimensional analysis. The treatment of data points as high-dimensional objects has become more common in applied genomics and proteomics. In this way, simultaneous analysis of multiple parameters takes full advantage of flow cytometry and can be useful in classification of different cell populations [3–6].

Support vector machines (SVM) are supervised algorithms used to learn from samples (training group) and assign labels to unknown objects (test group) [7]. These can be used to analyze multiple parameters simultaneously and to classify two sets of data in an n-dimensional space. Each cell in a flow cytometry dataset has six or more characteristics and can therefore be treated as an n-dimensional vector. The multi-dimensional advantage of flow cytometry can be better utilized with SVM. From the first introduction of this methodology in 1995 [8], many algorithm updates and extensions have been made [9–11,20,21]. LIBLINEAR is one of the most popular SVM libraries developed by Fan et al. which is suitable for linear classification of large sets of sparse data with a huge number of instances and features identified on the basis of LIBSVM [11].

In this study, a four-color panel consisting of anti-CD45, anti-CD65s, anti-CD15 and anti-CD11b, was used to obtain high-dimensional datasets that achieved routine acute leukemia immunophenotyping. We applied LIBLINEAR to learn from data-sets consisting of mature neutrophils of CML patients and normal mature neutrophils. A predictive model was then built on these training data and used for classifying the origin of mature neutrophils. To calculate classification accuracy, we used a
confusion matrix to compare predicted classification with clinical diagnosis, which was confirmed by real-time PCR of the BCR/ABL fusion gene, immunophenotyping and chromosome morphology.

2. Materials and methods

2.1. Case selection

Immunophenotyping data for nine patients with CML and nine healthy donors were selected from our archives (2006–2008) as the training group for data training and model building. Subsequently, 67 patients with different diagnoses (excluding those with a diagnosis of acute myeloid leukemia) were randomly selected into the test group after analysis (see Section 2.2), regardless of their mature neutrophil counts. CML was diagnosed as follows: (1) Mature neutrophilia and basophilia with/without blasts in bone marrow smears. (2) Bone marrow histocytometry showed low neutrophil alkaline phosphatase (NAP) score. (3) Positive for the BCR/ABL fusion gene by real-time PCR. 4. Cytogenetic analysis showed t(9;22). (5) Routine blood routine tests showed white blood cells were higher than 10×10^9/L and most of them were neutrophils. Patients with normal mature neutrophils were confirmed according to the following criteria: (1) No myelodysplasia and percentage of blast cells were lower than 5% in bone marrow smears. (2) Negative for the BCR/ABL fusion gene by real-time PCR. (3) Normal karyotype in cytogenetic analysis.

2.2. Data retrieval

Raw flow cytometry data were generated using a FACSCalibur flow cytometer and CellQuest Pro software was used for data analysis (Becton Dickinson). A four-color staining tube containing anti-CD65 fluorescein isothiocyanate (FITC; clone: VIM2, AN DER GRUB), anti-CD15 phycoerythrin (PE; clone: VIMC6, INVITROGEN), anti-CD11b allophycocyanin (APC; clone: VIM12, Invitrogen), and anti-CD45 peridinin chlorophyll protein-cyanin 5.5 (PerCP-cy5.5; clone:VIM12, Becton Dickinson), was included in our routine immunophenotyping panel. Briefly, the antibodies were incubated with samples at room temperature for 15 min. Samples were lysed with ammonium chloride lysis solution for 10 min and centrifuged at 1500 rpm for 5 min. The supernatant was discarded and samples were then ready for detection.

The LIBLINEAR software toolkit processes a specific data file format containing information for one cell in each line, comprising a numeric category label and an n-dimensional instance. This file format is known as a label-instance data file in this study. These files were prepared as follows: First, neutrophils were selected by manually gating on the CD45 versus SSC dot plot. Data for neutrophils were then reprojected onto the CD65s versus CD15 or the CD15 versus CD11b dot plot and exported as FCS (Flow Cytometry Standard) 2.0 data files by CellQuest Pro. Finally, a library from the MATLAB website (http://www.mathworks.com/matlabcentral/fileexchange/8430-flow-cytometry-data-reader-and-visualization) was used to read the information for each cell from the FCS files according to the FCS 2.0 standard and to generate label-instance data files. In 18 label-instance data files for training, the label was set as 0 for normal neutrophils and 1 for those from CML patients. These files were merged into a single large label-instance data file for training. However, for each label-instance data file in the test group, labels were set as 0 by default and the data file was kept separately for individual predictions. That is, mature neutrophils from each patient in the test group were regarded as normal by default.

2.3. Data analysis

The label-instance data files were analyzed in three steps on the basis of the method described by Fan et al. [11]. First, nine label-instance data files from mature neutrophils of CML patients and nine files from normal neutrophils were rescaled with the svm-scale.exe command toolkit from the LIBSVM software (command line: svm-scale –s range1 trainingdata > training-data.scale). These data were trained as the prediction model with the train.exe of the LIBLINEAR software with the parameter epsilon set as 0.001 (command line: train –e 0.001 trainingdata.scale training.model).

Second, label-instance data files of the test group were rescaled with previous generated training data range (command line: svm-scale.exe –r range1 testdata) and predicted probability of our hypothesis of normal neutrophils by the trained prediction model (command line: predict testdata.scale training.model testdata. result). Here, we introduced a heuristic cut-off probability of 50% for discriminating mature neutrophils of CML patients from normal neutrophils. That is, mature neutrophils were considered as normal if the predicted probability of being normal mature neutrophils was higher than 50%, or were deemed to be mature neutrophils from CML patients if that probability was lower than 50%.

Finally, these probabilities of being a normal mature neutrophil and the diagnosis from all patients were collected and then stored in a data file with assigned labels, which were set as 0 for normal mature neutrophils and 1 for mature neutrophils from CML patients. This data file was processed by Receiver Operating Characteristic (ROC) analysis [12] for determination of the optimal cut-off probability to distinguish normal from CML for best specificity and sensitivity.

3. Results

3.1. Patient characteristics

The training group (n=18) comprised 14 males and four females, with an average age of 45 years at the time of bone marrow biopsy (range 21–75 years). Mean bone neutrophil counts were 9.73 (range, 4.1–20.6) ×10^9/L for healthy donors and 77.07 (range, 19.7–171.9) ×10^9/L for CML patients. The test group (n=67) comprised 40 males and 27 females, with an average age of 44 years at the time of bone marrow biopsy (range, 10–78 years). The mean peripheral neutrophil counts were 42.87 (range, 0.2–603.72) ×10^9/L. Clinically, in the test group, 24 patients were diagnosed with CML and 43 patients were confirmed with normal neutrophils according to criteria outlined in Section 2.

3.2. Differences in predicted probability of being normal mature neutrophils

In the test group, the predicted probability of being normal mature neutrophils was 17.22 ± 22.85% for patients with CML and 82.77 ± 15.89% for patients with normal mature neutrophils. There was a statistically significant difference between the predicted probability for the CML patients and normal subjects (P < 0.05, Fig. 1). The cut-off value was 51.79%, which was calculated by ROC curve analysis.

3.3. ROC curve analysis of probability of being normal mature neutrophils

To determine the specificity and sensitivity of our method for discrimination of different sources of mature neutrophils, ROC
Fig. 1. Scatter plot of predicted probability of being normal mature neutrophils from different groups. The cut-off value of predicted probabilities for discrimination was optimized to 51.79% (see Section 3.3). At this value, mature neutrophils of CML patients could be accurately distinguished from their normal counterparts.

Fig. 2. ROC curve analysis of predicted probabilities of being normal mature neutrophils. The circle near upper-left corner represents the cut-off point (51.79%) of probability nearest to the coordinate point (0.1) and the highest specificity and sensitivity could be achieved.

curve analysis was conducted. A cut-off value for the optimal specificity and sensitivity of the predicted probability was also defined. According to the results, the area under curve (AUC) (0.97), Youden’s Index (0.91) [13] and discrimination power (3.4) [10,11] provided evidence that support vector machines offer high specificity and sensitivity in the classification of mature neutrophils. The cut-off point with the optimal specificity and sensitivity for the predicted probability of being normal mature neutrophils was determined at 51.79% on the ROC curve (Fig. 2). With this cut-off value, the sensitivity and specificity of the classifier were 95.80% (95% confidence interval: 87.80%–100.00%) and 95.30% (95% confidence interval: 89.10%–100.00%) respectively. Generally, accuracy of 95.50% and a mis-classification rate of 4.50% were achieved in discrimination between mature neutrophils from CML patients and their normal counterparts following the application of our built model and support vector machine. The results of our study can be presented as a confusion matrix (Table 1), which contains information about actual and predicted classifications performed by a classification system.

3.4. Differences in the predicted probability of being normal mature neutrophils for the chronic phase (CP) and blast crisis (BC) of CML

Of 24 patients diagnosed with CML in the test group, 10 were classified as CML-BC and 14 as CML-CP. Probability of being normal mature neutrophils predicted by our model were 12.21 ± 15.69 for patients with CML-BP and 20.80 ± 26.83 for patients with CML-CP. There was no statistically significant difference between these groups of patients (P=0.213), which indicates that mature neutrophils in the chronic and blast phases of CML were both neoplastic and of the same origin.

4. Discussion

This study illustrates a novel use of flow cytometry to discriminate between the mature neutrophils of CML patients from their normal counterparts. In this method, a prediction model was built on datasets from diagnosed CML patients and health donors. This model was then used to predict the accuracy of hypothesis that all neutrophils were normal, based on multicolor flow cytometry datasets obtained from different patient groups. In these processes, all datasets were treated as high-dimensional objects, which took full advantage of the multi-dimensional power of flow cytometry to differentiate between cell populations more effectively in a process analogous to that of face recognition. Moreover, only four antibodies were used for discrimination.

SVMs are supervised algorithms used to learn from samples and assign labels to unknown objects. This approach allows the analysis of multiple parameters simultaneously. Various derivations of SVM were developed and successfully applied in the classification of tumors [14–21] over the past decade. However, these applications were based mainly on microarray data or information from other types of gene technology. Recently, several SVM applications based on flow cytometry have been published. Toedling et al. automated the monitoring of minimal residual leukemic cells and achieved high specificity and sensitivity [4]. Pedreira et al. described the automated analysis of peripheral blood lymphocyte subsets with SVM [6]. These studies showed good discrimination power in multi-dimensional flow cytometry datasets. However, the cells analyzed in these studies exhibited significantly different antigen expression levels, which could also be easily distinguished by traditional two-dimensional iteration methods.

In our study, mature neutrophils from CML patients and their normal counterparts were all CD45 +, CD65sstr, CD15sstr and CD11bstr, which made it hard to discriminate between these two groups using the classical two-dimensional flow cytometric analyses. To date, there have been no reports of the application of SVM for the classification of cells with similar antigen expression profiles. LIBLINEAR had been successfully applied for text classification of

Table 1

<table>
<thead>
<tr>
<th>Predicted probability of being normal</th>
<th>Clinical diagnosis</th>
<th>Row sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 51.79%</td>
<td>Normal</td>
<td>CML</td>
</tr>
<tr>
<td>&gt; 51.79%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column sum</td>
<td></td>
<td>67</td>
</tr>
</tbody>
</table>
more than one million features with good performance [11], although there are no reports of the application of this tool to flow cytometry datasets which contain tens of thousands of instances with six or more features. Therefore, we applied a support vector machine and a four-color panel for the detection of CD45, CD65s, CD15, CD11b, in the discrimination of mature neutrophils from CML patients and normal neutrophils. The cut-off predicted probability of being normal mature neutrophils was determined by ROC analysis [12].

After building a prediction model using malignant neutrophils from several cases of diagnosed CML patients and healthy donors, flow cytometry data for analysis was presumed to represent normal mature neutrophils by default. In this context, the probability of being normal mature neutrophils was predicted by the prediction model. The cut-off predicted probability for discrimination between different sources of mature neutrophils was adjusted to 51.79% after ROC curve analysis. With this cut-off predicted probability, the sensitivity and specificity of this technique reached 95.80% and 95.30%, respectively. These results may provide superior ability to distinguish mature neutrophils of CML patients from their normal counterparts, even if there are no statistically significant differences between their antigen expression levels. Moreover, no statistically significant differences in the predicted probability of being normal mature neutrophils were found between mature neutrophils from patients in the chronic phase and blast crisis of CML, which indicates that the prediction results are stable regardless of the phase of CML.

Real-time PCR, chromosome smears and in situ fluorescent hybridization are the main techniques used for the diagnosis of CML clinically. Although aberrant antigen expression in CML might be detected by flow cytometric analysis, there are almost no significant differences between mature neutrophils from CML patients and their normal counterparts. LIBLINEAR compares objects in multiple dimensions and fully utilizes the advantage of multicolor flow cytometry. Therefore, differences between these two cell populations are magnified and identified easily. Only four antibodies were used in this study, but the simplicity of this panel did not decrease the capacity to discriminate different sources of mature neutrophils.

The approach described here simplifies the process of identification of mature neutrophils from CML patients with a four-color panel and LIBLINEAR software. This approach may be a highly informative method for the diagnosis of CML, although the analysis of more clinical samples and additional studies will be required to appraise the potential value of this technique. On this basis, it may be used for monitoring disease in CML patients in the future.

Summary

We distinguished between malignant neutrophils in chronic myelogenous leukemia and normal neutrophils using a simple four-color panel and LIBLINEAR software.

Conflict of interest statement

On behalf of Wannao Ni, Xiangmin Tong, Wenbin Qian, Jie Jin and Hongchan Zhao, it is declared that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled “Discrimination of malignant neutrophils of chronic myelogenous leukemia from normal neutrophils by support vector machine”.

Acknowledgments

This work was supported by Grants from the National Natural Science Foundation of China (81172250, 2012C13021-3), and Zhejiang Provincial Healthy (2011ZDA008, 2011KYA15, 2012RCA 017 and 20120533Q02).

References