Computers in Biology and Medicine 43 (2013) 1192–1195

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

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

ARTICLE INFO

Article history:
Received 18 March 2013
Accepted 9 June 2013

Keywords:
Flow cytometry
Immunophenotyping
LIBLINEAR
Support vector machine
Mature neutrophils

ABSTRACT

Malignant neutrophils of chronic myelogenous leukemia (CML) have similar antigen expression patterns compared to their normal counterparts, thus making the cells difficult to distinguish by clinical flow cytometry. In this study, we applied the support vector machine method to build a malignant neutrophil prediction model based on nine CML patients and nine healthy donors. This approach effectively differentiated between malignant and normal neutrophils with high specificity and sensitivity (≤95.80% and ≤95.30%, respectively). This approach may broaden the application of flow cytometry for differentiation between CML and normal neutrophils and become an important diagnostic tool in CML.

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 datasets 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

Abbreviations: SVM, support vector machine; CML, chronic myelogenous leukemia; FCS, flow cytometry standard; PCR, polymerase chain reaction; ROC, receiver operating characteristic; BCR, breakpoint cluster region; ABL, Abelson murine leukemia

* Corresponding author. Tel.: +86 571 87236575; fax: +86 571 87236702.
E-mail address: tongxiangmin@163.com (X. Tong).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.06.004

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 2.3), 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 histochemistry 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∗10E9/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-CD65s 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:2D1, Becton Dickinson), was included in our routine immunophenotyping panel. Briefly, the antibodies were incubated with samples at room temperature for 15 min. Samples were lyzed 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 > trainingdata.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 neutrophils 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 blood neutrophil counts were 9.73 (range, 4.1–20.6) $\times 10^9$/L for healthy donors and 77.07 (range, 19.7–171.9) $\times 10^9$/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) $\times 10^9$/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 $17.22 \pm 22.85\%$ for patients with CML and $82.77 \pm 15.89\%$ 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