Using a Neural Network with Flow Cytometry Histograms to Recognize Cell Surface Protein Binding Patterns

Eun-Young Kim, M.D.[1], Qing Zeng, Ph.D.[1], James Rawn, M.D.[2], Matthew Wand, Ph.D.[4], Alan J. Young, Ph.D.[2,5], Edgar L. Milford, M.D.[3], Steven J. Mentzer, M.D.[2], Robert A. Greenes, M.D., Ph.D.[1]

[1]Decision Systems Group, [2]Department of Surgery, [3]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, [4]Department of Biostatistics, Harvard School of Public Health, [5]Department of Biology & Microbiology, South Dakota State University

Abstract

Flow cytometric systems are being used increasingly in all branches of biological science including medicine. To develop analytic tools for identifying unknown molecules such as the antibodies that recognize different structure in the identical antigens, we explored use of a neural network in flow cytometry data comparison. Peak locations were extracted from flow cytometry histograms and we used the Marquardt backpropagation neural networks to recognize identical or similar binding patterns between antibodies and antigens based on the peak locations. The neural network showed 93.8% to 99.6% correct classification rates for identical or similar molecules. This suggests that the neural network technique can be useful in flow cytometry histogram data analysis.

Introduction

Flow cytometry is a clinical and experimental technique that quantitatively measures the individual protein molecules on cell surfaces. It has been used not only for quantifying the expression of known molecules, but also for identifying unknown molecules through the comparison with known molecules. However, there is a need to improve flow cytometry analysis, to distinguish the "similarity" of antibodies which recognize the identical surface molecules but different target sites, from those of different antibody subtypes (primary structure).

Some traditional comparison methods have been widely used for flow cytometry data analysis despite their deficiency in assessing non-identical similarity[1, 2]. Recently, application of information theory to cytometric histogram showed improvement over traditional methods, although it still did not differentiate the similarity correlation from other types of correlation[3].

In this study, we explored the application of a Marquardt neural network in flow cytometry data analysis. Peak locations are the key features of a histogram and were extracted using a kernel smoothing technique. The Marquardt neural networks were trained and tested to recognize identical or similar binding patterns between antibodies and antigens based on the peak locations.

Monoclonal antibodies that recognize antigens with identical as well as with different primary structure were used in this study. The results indicated that a neural network can successfully recognize certain molecular binding patterns: (1) "similar, but not identical" – binding between an antibody and different sites (structures) in an antigen; (2) "identical or similar" – binding between an antibody and an antigen regardless of binding sites; (3) "different" – binding between an antibody and different antigens.

Background

Flow cytometry

Flow cytometry is a technique for the automatic and quantitative determination of physical parameters of molecules using fluorescence induced by attached fluorescent dyes[4]. It has been used increasingly in various branches of biomedical research and clinical practice. Major applications of flow cytometry include DNA histogram analysis for ploidy determination and immunofluorescence phenotyping. In addition, monoclonal antibodies have been used for the identification, enumeration, localization, and isolation of individual types of cells from blood or solid tissues. The molecular composition of a cell, reflecting its differentiation state and patterns of molecules expressed selectively by a particular cell type, can serve as a marker for the cell type[4]. Various monoclonal antibodies recognizing molecules on the lymphocytes were used in this study to test the applicability of the neural network on flow cytometry histogram comparison.

Kernel smoothing and SiZer

In order to analyze histograms, we used a new kernel smoothing technique, – SiZer, to smooth the histograms and extract meaningful features. SiZer performs cross-bandwidth (resolution) smoothing and

provides a family of smoothed histograms. It also distinguishes the statistical significant features from noise. The first-derivative characteristics (i.e., curve increasing, decreasing, or neither) are provided in the form of a color-coded map where blue represents curve increasing, red represents curve decreasing, purple represents neither of the previous two, and gray represents that not enough data at this resolution to determine direction. The number of bandwidths used for smoothing is user defined[5-7].

Marquardt backpropagation network

A backpropagation neural network consists of input, hidden, and output layers[8]. Each layer contains at least one node with an output that is simply the sum of its inputs and modified by a nonlinear transfer function. Backpropagation is a training method that uses backward propagation algorithms (from the output layer to the hidden layers, and then to the input layer) to update weights in each layer. In our study, we chose the Marquardt backpropagation algorithm provided in the Matlab software (version 6, The MathWorks, Inc., Natick, MA). The Marquardt backpropagation algorithm is a modification of the Marquardt-Levenberg algorithm into a backpropagation algorithm. A network was known to have converged when the conjugate gradient and variable learning rate algorithms failed to converge[9].

Material

Antibodies used were 86D, 18-106, ST-1, H1-68, A1-107, 17-63, 2-87, 2-128, 2-128-1, 14-24, 14-109, 7C-2/38-65, ST-8, 6-87, 6-99, FW4, and ERD2/114. Peripheral sheep blood lymphocytes (PBLs) of a single animal were used as a test cell population. The molecular "target" of each of these antibodies is shown in Table 1. Some antibodies recognizing the same cell surface proteins (e.g., TCR and CD3) were also included.

Methods

Data acquisition

Reactivity of antibodies to the target molecules was detected by indirect immunofluorescence. This was carried out using FITC-conjugated goat-anti-mouse Immunoglobulin (Ig) diluted 1:10 with Phosphate Buffered Saline. Total of 50,000 cells were analyzed by Coulter Epics XL flow cytometry (Coulter Electronics, Inc., Hialeah, FL). The fluorescence intensity was presented as log-scale histograms containing a total of 1024 channels. A total of 51 flow cytometry histograms were obtained from 17 antibodies and there were 3 replications for each antibody. The flow cytometry histograms were recorded as FCS 2.0 listmode files[10]. The raw histogram data was exported from WinList 4.0 (Verity, Topsham, ME) into ASCII for data analysis.

Table 1. Antibodies, their isotypes, and their target molecules.

Antibody	Isotype	Target molecule
86D	α1	TCR
18-106	α1	CD3
ST-1	α2a	CD5
H1-68	α2b	CD5
A1-107	α2a	CD5
17-63	α2a	CD5
2-87	α1	CD21
2-128	α1	CD21
2-128-1	α1	CD21
14-24	α1	CD21
14-109	α1	CD21
7C-2/38-65	α2a	CD8
ST-8	ν	CD8
6-87	α1	CD8
6-99	α1	CD8
FW4	α1	CD29
ERD2/114	α1	CD29

Feature Extraction

SiZer maps with 5-bandwidth and 11-bandwidth were generated for the histograms. We used these maps to determine the peak locations on a histogram. In a particular bandwidth, a peak was defined as a region consisting of a strip of blue (increasing region) to its left and a strip of red (decreasing region) to its right. Thus, each peak location was described by two parameters: the start and the end location of the peak. A histogram may have multiple peaks, but typically no more than 2 or 3. Empirically, the peaks further to the right are more biologically significant. In this study, we extracted the locations of the two rightmost peaks from each histogram. (When there was only one or no peak, its location was marked with a default number '-1'.) Therefore, a total of 20 and 44 features were extracted from the 5-bandwidth and 11-bandwidth SiZer maps, respectively.

Neural Network Analysis

The neural networks were constructed to compare two histograms and determine if they were identical, similar, or different. The peak location features extracted from the SiZer maps of the two comparing histograms were used as inputs to the neural network. The input vector sizes were 40 and 88, respectively, when 5-bandwidth and 11-bandwidth SiZer map were used.

We trained, validated, and tested the networks for two tasks: (A) to distinguish similar histograms from different histograms, when the histograms are not identical; (B) to distinguish identical/similar histograms from different histograms. We defined identical, similar, and different histograms as follows: Identical histograms: histograms of the same antibody that binds with the same antigen, including absolute and replicate identity (Figure 3); Similar histograms: histograms of different antibodies that bind with the same antigen (Figure 2); Different histograms: histograms of different antibodies that bind with different antigens (Figure 4).

The target of the networks was set to 0 for recognizing histograms as different (different target molecule), and 1 for recognizing histograms as identical/similar or just similar (same target molecule).

Four Marquardt backpropagation networks were constructed using Matlab, corresponding to the two different input vector sizes and two different tasks. They were all two-layer networks: a 40-20-1 network for 40-feature input, and an 88-9-1 network for 88-feature input. A hyperbolic tangent transfer function was used in the hidden layer, and a logistic sigmoid transfer function was used in the output layer (Figure 1).

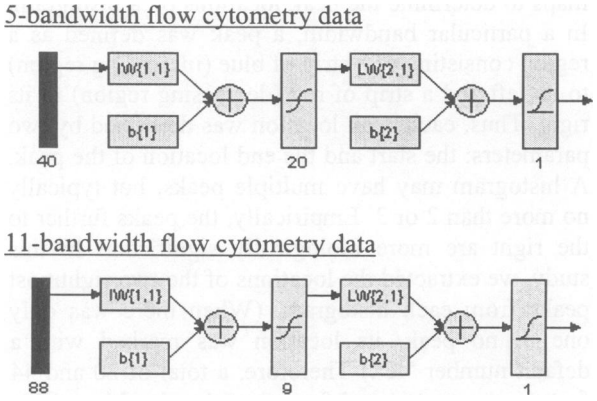

Figure 1. Architecture of the networks (IW: input weight matrix, LW: layer weight matrix, b: bias).

In our sample set, there were 102 unique pairs of identical histograms, 216 unique pairs similar histograms, and 1008 unique pairs different histograms. The 1224 pairs of similar or different histogram were used as the data set for the task (A) as described before; and the 1326 pairs of identical, similar, or different histograms were used as the data set for the task (B). For each network, we assigned a half of the data as a training set, a quarter as a test set and the remaining quarter as a validation set. The validation set was included to prevent the overfitting problem. When errors in the validation set showed an increasing tendency, training was stopped.

Initial weights in the networks were randomly set. For each network, we experimented with five different sets of initial weights and saved the one with the best correct classification. The designated epochs (cycles) were 1000, but most of the networks were trained with less than 1000 epochs (11 to 1000) to avoid overfitting.

In all 4 networks, we used 0.5 as a cut-off value, as the output of the logistic sigmoid function is between 0 and 1. We considered values greater or equal to 0.5 as "identical" or "similar", and less than 0.5 as "different".

Evaluation

We evaluated the neural network outputs in terms of correct classification rate and mean-squared error (MSE). The correct classification rate represented the percentage of correct pattern recognition cases to total cases. The best correct classification rate indicated the highest correct classification rate among the 5 neural networks' correct classification rates according to 5 different initial weights.

Results

Each network was trained and tested with 5 sets of randomly generated initial weights. The mean-squared error (MSE) of the four networks with different initial weights ranged from 0.027 to 0.148 and the correct classification rate ranged from 80.7% to 99.1% (Table 2). The best correct classification rate of each network was over 90%.

Table 2. Range of MSE and correct classification rate (CCR) of the networks based on different initial weights.

	Similar*		Identical/similar**	
	MSE	CCR (%)	MSE	CCR (%)
5 band$^{\Psi}$	0.060 –0.104	85.9 – 92.2	0.042 – 0.127	82.5 – 95.5
11 band$^{\Psi}$	0.031 –0.155	81.4 – 95.8	0.027 – 0.148	80.7 – 99.1

* : Networks that distinguish similar from different
**: Networks that distinguish identical/similar from different
$^{\Psi}$: 5-bandwidth and 11-bandwidth data sets

The correct classification rate and MSE of the four different networks showed little difference, indicating

that the inclusion or exclusion of identity and that the number of features used did not have drastic impact.

All the training sets showed over 95% correct classification rate (95.9 –99.8%), which are better than those of the test (92.2 – 99.1%) and validation (91.2 – 99.7%) sets (Table 3). The network for distinguishing similarity from difference showed more overfitting tendency.

When errors occurred, we observed that the networks tended to classifying identical/similar histograms as different rather than the reverse. We also observed that the networks made more errors in the recognition of similarity than in the recognition of identity/similarity.

Table 3. Best correct classification rate (CCR) and MSE of the selected data sets.

			Train	Test	Validation
S*	5 band$^{\Psi}$	MSE	0.035	0.060	0.072
		CCR (%)	95.9	92.2	91.2
	11 band$^{\Psi}$	MSE	0.026	0.031	0.065
		CCR (%)	98.0	95.8	92.5
ID**	5 band$^{\Psi}$	MSE	0.038	0.051	0.041
		CCR (%)	97.3	94.9	96.1
	11 band$^{\Psi}$	MSE	0.021	0.027	0.023
		CCR (%)	99.8	99.1	99.7

* : Similar from different data set
**: Identical/similar from different data set
$^{\Psi}$: 5-bandwidth and 11-bandwidth data sets

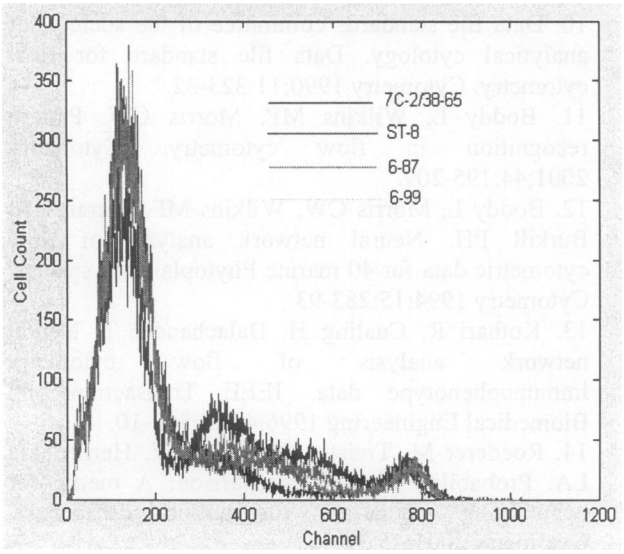

Figure 2. Reactivity of 4 antibodies recognizing the CD8 on a subset of peripheral blood leukocytes.

Histograms of some antibodies that recognize the same antigen had more variance than others and thus were more challenging for the network to handle. The most common error the networks made was to classify the 4 antibodies that all recognize CD8 as different (Figure 2).

Discussion

The application of the neural network technique on flow cytometry data had been successfully used for molecular and cellular classification[11-13]. Previous studies performed classification of individual molecules and did not involve histogram analysis. In this study, we explored the possibility of using neural networks with flow cytometry histograms for molecular identification. Our study differs from previous studies in that it focuses on molecular populations, the characteristics of which are revealed in the form of histograms.

Comparing to other methods that have been used and studied (Kolmogorov-Smirnov, information theory measurement, binning method)[1,3,14], neural network analysis is a learning algorithm and it can memorize and recognize patterns. The advantage of neural networks is their ability to model both linear and non-linear relationship models. Traditional linear models are simply inadequate for non-linear data[8].

The best correct classification rates in the four networks we implemented were consistently over 90%. This demonstrated the feasibility of combining neural network and kernel smoothing technique for flow cytometry histogram analysis.

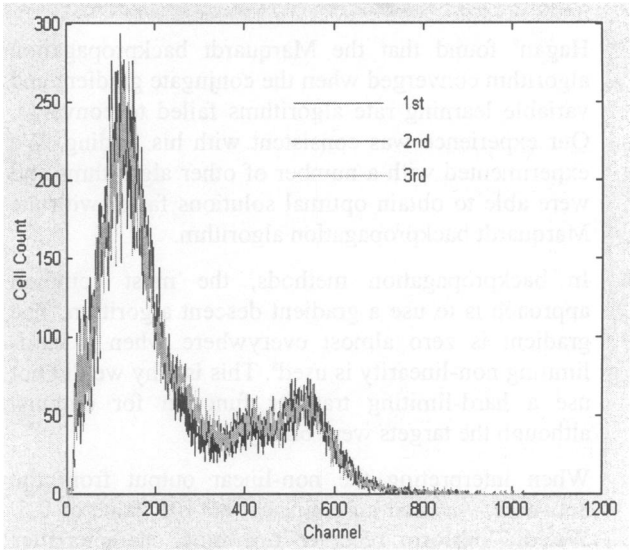

Figure 3. Three replications of 2-128 antibody reactivity.

We picked the more general antibody binding patterns to train the neural network to recognize: identity, similarity, and difference. Such patterns are relatively consistent across antibodies and cell lines. An alternative way was to train the neural network to recognize specific molecules or binding patterns associated with specific molecules. For instance, we could train the networks to classify the histograms by their target molecules such as CD8 or CD3. The resulting network, however, would not be applicable to other molecules or cell lines.

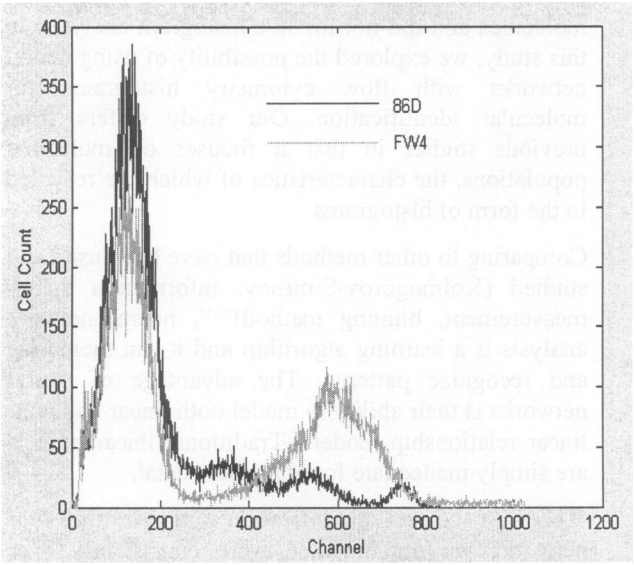

Figure 4. Two different antibody reactions of 86D and FW4 to recognize different target molecules.

Hagan[9] found that the Marquardt backpropagation algorithm converged when the conjugate gradient and variable learning rate algorithms failed to converge. Our experience was consistent with his finding. We experimented with a number of other algorithms and were able to obtain optimal solutions faster with the Marquardt backpropagation algorithm.

In backpropagation methods, the most common approach is to use a gradient descent algorithm. The gradient is zero almost everywhere when a hard-limiting non-linearity is used[8]. This is why we did not use a hard-limiting transfer function for outputs, although the targets were binary.

When interpreting the non-linear output from the networks, we used an empirical cut-off value of 0.5. We did perform receiver operating characteristics curve (ROC) analysis using different cut-off values. The optimal cut-off values to recognize identical/similar vs. different was around 0.5, and The optimal cut-off values to recognize similar vs. different was around 0.3. Due to space limitations, we are not including the detailed results from the ROC analysis in this paper.

References

1. Lampariello F. On the use of the Kolmogorov-Smirnov statistical test for immunofluorescence histogram comparison. Cytometry 2000;39(3):179-88.
2. Bagwell CB, Hudson JL, Irvin GL. Nonparametric flow cytometry analysis. J Histochem Cytochem 1979;27(1):293-6.
3. Zeng Q, Young AJ, Boxwala AA, et al. Molecular identification using flow cytometry histograms and information theory. Proc AMIA Symp 2001:776-80.
4. Zola H. Monoclonal antibodies: A manual of techniques. Boca Raton, FL: CRC press, Inc.; 1987:89-146.
5. Wand MP, Zeng Q, Rawn JD, et al. Analysis of flow cytometry histogram structure using kernel smoothing. Decision Systems Group Technical Report TR-2001-017.
6. Wand MP, Jones MC. Kernel smoothing. London: Chapman and Hall; 1995:1328-45.
7. Chaudhuri P, Marron JS. SiZer for exploration of structures in curves. Journal of the American Statistical Association 1999;94:807-23.
8. Hush DR, Horne BG. Progress in supervised neural networks: What's new since Lippmann? IEEE Signal Processing Magazine 1993;10:8-39.
9. Hagan MT, Menhaj MB. Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks 1994;5(6):989-93.
10. Data file standards committee of the society for analytical cytology. Data file standard for flow cytometry. Cytometry 1990;11:323-32.
11. Boddy L, Wilkins MF, Morris CW. Pattern recognition in flow cytometry. Cytometry 2001;44:195-209.
12. Boddy L, Morris CW, Wilkins MF, Tarran GA, Burkill PH. Neural network analysis of flow cytometric data for 40 marine Phytoplankton species. Cytometry 1994;15:283-93.
13. Kothari R, Cualing H, Balachander T. Neural network analysis of flow cytometry immunophenotype data. IEEE Transactions on Biomedical Engineering 1996;43(8):803-10.
14. Roederer M, Treister A, Moore W, Herzenberg LA. Probability binning comparison: A metric for quantitating univariate distribution differences. Cytometry 2001;45:37-46.