Chem-Bio Informatics Journal, Vol.10, pp.13-23(2010)

Gene expression informatics with an automatic histogram-type membership function for non-uniform data

Akito Daiba[1,2*], Satoru Ito[3], Tsutomu Takeuchi[4], Masafumi Yohda[1]

[1]*Department of Biotechnology and Life Science, Graduate School of Technology, Tokyo University of Agriculture and Technology, Koganei, Tokyo 184-8588, Japan*
[2]*AP Solutions Consulting, Accelrys K.K., Nishishinbashi, Minatoku, Tokyo, 105-0003, Japan*
[3]*Scientific information Department, Fujirebio Inc., FR Bldg., 62-5 Nihonbashi-Hamacho 2-chome, Chuo-ku, Tokyo, 103-0007, Japan*
[4] *Division of Rheumatology/Clinical Immunology Department of Internal Medicine, School of Medicine, Keio University, 35 Shinanomachi, Shinjyuku-ku, Tokyo, 160-8582, Japan*

**E-mail: adaiba@accelrys.com*

(Received November 12, 2009; accepted January 26, 2010; published online February 6, 2010)

Abstract

The non-uniformity of gene expression data is one of the factors that make gene expression analysis difficult. Gene expression data often do not follow a normal distribution but rather various distributions within each group. Thus, it is impossible to apply basic statistical techniques such as the t-test. In this study, we have developed an analysis method for gene expression data obtained by microarrays using a fuzzy logic algorithm with original membership functions. The method automatically evaluates the data from a histogram of gene expression information for a patient group. Using this method, we predicted the efficacy of an anti-TNF-α treatment for rheumatoid arthritis. We created a prediction model for the effects of 14 weeks of anti-TNF-α treatment based on the gene expression data from the peripheral blood of rheumatoid arthritis patients before the treatment. The model had a predictive success of 89% in the model-establishing data group, 94% in the training group, and 89% in the validation group. The results suggest that the method presented here could be an extremely effective tool for gene expression analysis.

Key Words: Microarray, Gene Expression, Prediction of therapeutic efficacy, Rheumatoid arthritis, Fuzzy Logic

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

Microarray technology has become popular, and gene expression analyses by microarray are widely used in the medical research field [1]. However, it is frequently difficult to deduce reasonable conclusions from the large amount of data produced by microarrays, and the limited number of clinical samples makes this type of research difficult. Therefore, many applicable techniques based on computer technologies for numerical analysis and modeling have been researched and developed [2][3].

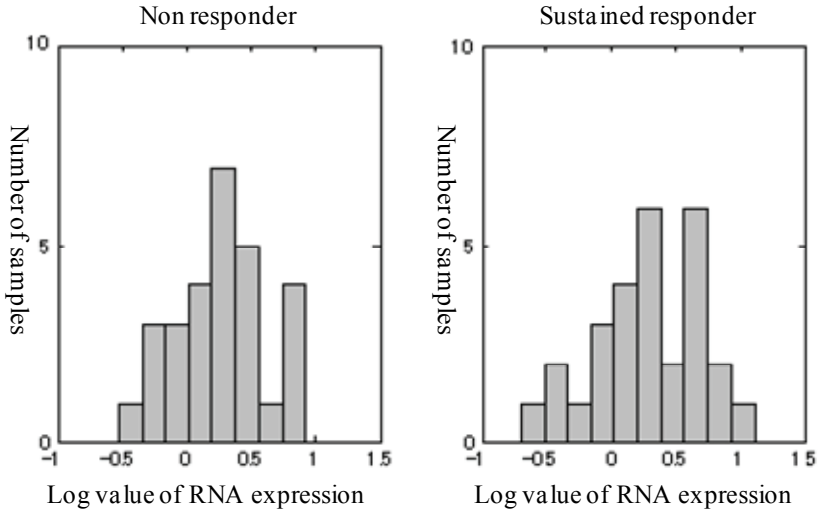

Figure 1. Histograms of 16S ribosomal RNA expression data by microarray
The histogram of the NR (non-responder) group and SR (sustained-responder) group is shown to demonstrate the appearance of gene expression shape. Each histogram has some peaks, but the shapes are different.

In general, gene expression patterns often show various non-uniform distributions, even in a patient group from the same clinical setting. For instance, for the prediction of therapeutic efficacy, most gene expression patterns are not normally distributed, even in the same patient group. Therefore, it is almost impossible to analyze them by a basic statistical technique such as the t-test, which requires that data follow a normal distribution (Fig. 1). Non-uniform shape of the histogram can be observed even in the data of RT-PCR in other study[4]. This non-uniformity is likely to have numerous causes. The age and sex of the patient, the duration of the disorder, the treatment history, the state of the disease progression, and several other factors could strongly influence gene expression, and there may be additional as yet undiscovered factors. If it were possible to partition all of these patient factors, the effects of the treatment might be determined easily based on the gene expression patterns. This scenario, however, is almost impossible because there are so many uncontrollable and uncertain factors, and the availability of clinical samples is often limited. Thus, the clear classification of samples into the groups with different character is very difficult, and methods are needed to analyze the data with various distributions among the limited samples. Moreover, data processing methods, such as those for noise reduction and normalization, are needed for microarray data analyses because numerous factors contribute to the error margins between experiments [5][6][7].

In the present study, we faced an additional problem in predicting the therapeutic efficacy based

on microarray data. Our samples were mRNA from peripheral whole blood cells obtained before the drug administration (anti-TNF-α agent, infliximab). As they were not taken from the tissues affected by the disease, the expression patterns were not directly related to the disease.

After the collection of the efficacy data for the treatment of the rheumatoid arthritis (RA) patients, we attempted to develop a therapeutic efficacy prediction method based on mRNA expression in the blood samples before the treatment. However, we could not obtain any significantly meaningful results from the analysis using basic statistics, hierarchical clustering [8], a discrimination analysis by the Mahalanobis distance method [9], Support Vector Machine (SVM), an artificial neural network (ANN), or Self-Organizing Maps (SOM) (data not shown). We grouped the samples based on the results of the therapy and examined the expression differences of some of the genes within each group. Because there were some differences between the sample groups, it was impossible for the current techniques to classify the expression data per results of the therapy, as two or more peaks coexisted in each distribution. Next, we decided to use fuzzy logic with our original membership function. Fuzzy logic, with an ability to evaluate complex input data, is an algorithm often used in machine learning and in many gene expression studies[10][11][12][13][14]. Our original membership function was created using two kinds of data histograms: one from the sample group that responded positively to therapy and one from the sample group that did not. This function, reflecting patient's gene expression patterns, can evaluate gene expression data. We introduce this data analysis method by demonstrating its application to the prediction of the effect of an anti-TNF-α drug on RA patients. This research project was a collaboration organized by Dr. Takeuchi, with seven university hospitals.

2. Materials and Methods

2.1 Data Set

The gene expression analysis was performed with the Genomessage v2 microarray (registered in GEO, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL5460), which was designed and developed by Japan Genome Solutions, Inc. (JGS). This Stanford-type microarray is printed with genes related to the immune response. The blood samples were obtained from patients with informed consent. The purification of the mRNA and conversion to amplified RNA (aRNA) were performed according to the procedure provided by JGS. Gene expression was measured by the two-color method on the microarray [15]. The clinical data for the RA patients who underwent anti-TNF-α treatment were the ACR (American College of Rheumatology) scores at 14 weeks after the first treatment. We divided the patients according to their ACR score. One group was the NR (non-responder) group, which did not show any positive response to therapy and had ACR scores between 0 and 20. The other group was the SR (sustained-responder) group, which showed a positive response, with ACR scores between 50 and 70. All of the samples were categorized into the two groups. Fifty-five samples were analyzed in total, and the samples were extracted at random using Matlab's rand function from each group to generate three sample groups that were used for modeling (10 NR samples, 9 SR samples, 19 samples total), training (9 NR samples, 9 SR samples, 18 samples total), and validation (9 NR samples, 9 SR samples, 18 samples total) (Fig. 2). The modeling and training sample groups were used for selecting the membership function group. Seven additional samples obtained after the model construction were used to confirm the accuracy of the model.

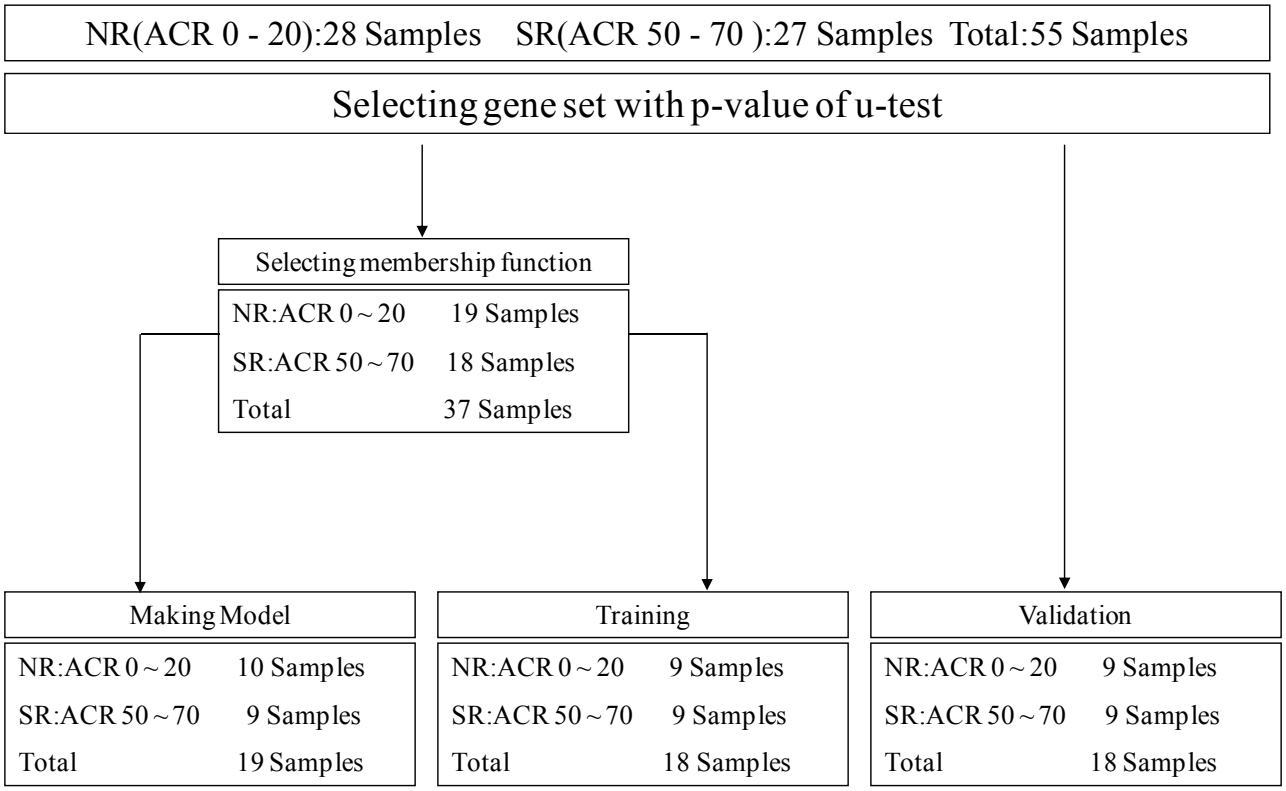

Figure 2. Classification of the samples
Samples were selected from the pool of SR and NR samples at random.
Similar numbers of SR and NR samples were chosen.

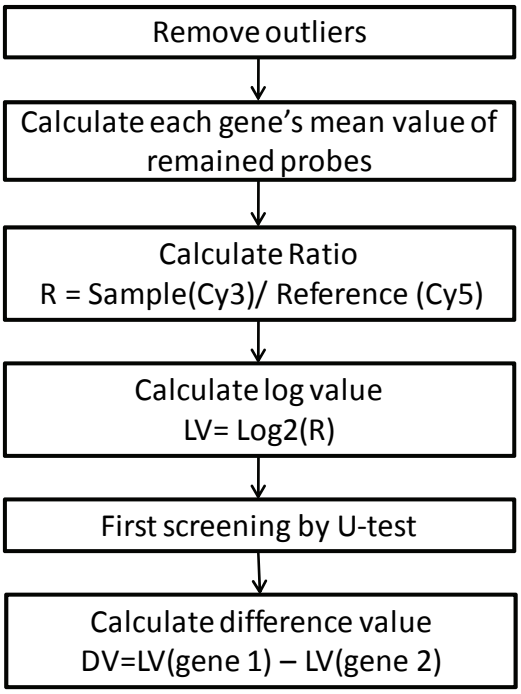

Figure 3. Data preprocessing flow chart
This diagram shows the data preprocessing workflow.

2.2 Data preprocessing

It is important to correct the errors caused by noise, hybridization conditions, fluorophores, etc. during the analysis of microarray gene expression data [16][17]. The data were corrected using the methods developed in this study. The preprocessing workflow is shown in Figure 3.

Genomessage v2 has four discrete location probes for each gene on the microarray. To reduce the error range, the Smirnov-Grubbs test was used to exclude outliers. The outliers in the data from the four probes of same genes were excluded by the method, and the representative value was calculated from the mean of the remaining values. Then, for each gene, the ratio of the sample (Cy3 signal mean value) and reference (Cy5 signal mean value) was calculated and the log2 value determined. In this study, the differences between the two log-transformed gene expression ratios were used as parameters. However, many hours of computing time would be needed to calculate the differences for all the genes. To overcome this obstacle, first screening for the entire gene was performed using the u-test to shorten the computing time. The value differences among the remaining genes were calculated and used for the model construction.

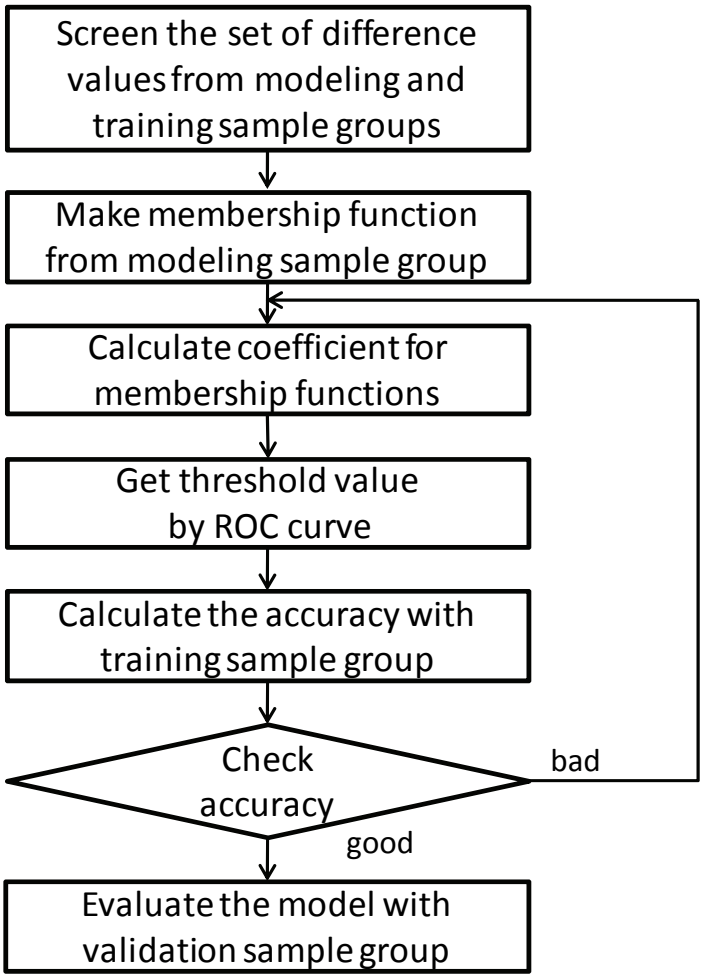

Figure 4. Predicting model development flow
The model was developed through this workflow. Modeling, training, and evaluating were done using different sample groups.

2.3 Algorithms

The prediction model was developed via described workflow (Fig. 4). The screening method has five procedures: (i) to make two histograms that are SR-histogram from SR samples and NR-histogram form NR samples by each difference value set, (ii) to evaluate each sample with SR-histogram and NR-histogram, which are derived from values in Y-axis of SR and NR histogram corresponding to the difference value in X-axis, (iii) to compare the output derived from both histograms and select the result with the bigger value. If the bigger value result is from SR histogram, then, prognosis is SR, Otherwise, if the bigger value result is from NR histogram, prognosis is NR, (iv) to calculate prognosis accuracy which is conducted from modeling samples and training samples in steps (ii) and (iii), and verify the treatment results, (v) to select difference value set which have the highest prognosis accuracy. In this study, remain 30genes, 25 difference value sets.

The next step was making original membership functions for use with the fuzzy logic algorithm. Fuzzy logic is an algorithm that calculates the evaluation value from the input data values. This evaluation function is called the membership function. These functions are used to construct trigonometric functions, wave functions, and other numerical functions. One or more membership functions can be combined. We had to design a more flexible membership function and develop an original function using Matlab (The MathWorks, Inc.). This function requires as the input the difference value between two log-transformed gene expression ratios [18], and the function then determines how close the input value is to the distributions of the SR group or the NR group by the numerical value. The membership function is derived from the histogram of the data from the NR or SR modeling sample group. The output value, evaluated by this membership function, is the value of the height of the bin corresponding to the given difference value. There are two membership functions for each difference value: the SR membership functions (SRMF) were made from the histogram of the SR samples, and the NR membership functions (NRMF) were made from the histogram of the NR samples. Each membership function has same number of bin. In the formula mentioned below, ‘x’ is difference value of each sample, ‘i’ is index of membership functions, and ‘n’ is number of bin. In this study, there were 25 SRMF and 25 NRMF, and the each membership functions had 15 bins.

$$SRMF_{i(xi)} = find_{(xi)}\{SRbin_{i1}, SRbin_{i2}, \cdots, SRbin_{in}\} \quad (1)$$

$$NRMF_{i(xi)} = find_{(xi)}\{NRbin_{i1}, NRbin_{i2}, \cdots, NRbin_{in}\} \quad (2)$$

This method can detect the change in two genes. Additionally, because it deals only with the differences between expression values, reduction of the noise influence can be expected.

Research to predict therapeutic efficacy requires a large amount of gene expression data and multiple membership functions. We also set the coefficient to the output value of each membership function. This coefficient was selected based on each membership function’s classification ability, and multiplied with the output value of a membership function of either SR or NR, using the larger of the two values.

$$Ev = \sum_{i=1}^{n} Ci\left(SRMF_{i(xi)} \mid NRMF_{i(xi)}\right) \quad (3)$$

The final evaluation value was sum of all of the values, and we call this the total score of the output value of the membership functions. We assured the predictive accuracy of the membership

function by using the data group for the model construction. The threshold of the final evaluation value to classify the NR and SR groups was decided based on the Receiver Operating Characteristic curve (ROC curve) analysis of the training data. Points of true positive rate and specificity in each threshold are plotted and a curve is interpolated as seen in the Fig. 5.

The point nearest to the left upper corner of ROC curve chart represents good balance of true positive rate and specificity, thus point No.9 chosen as optimal threshold value.

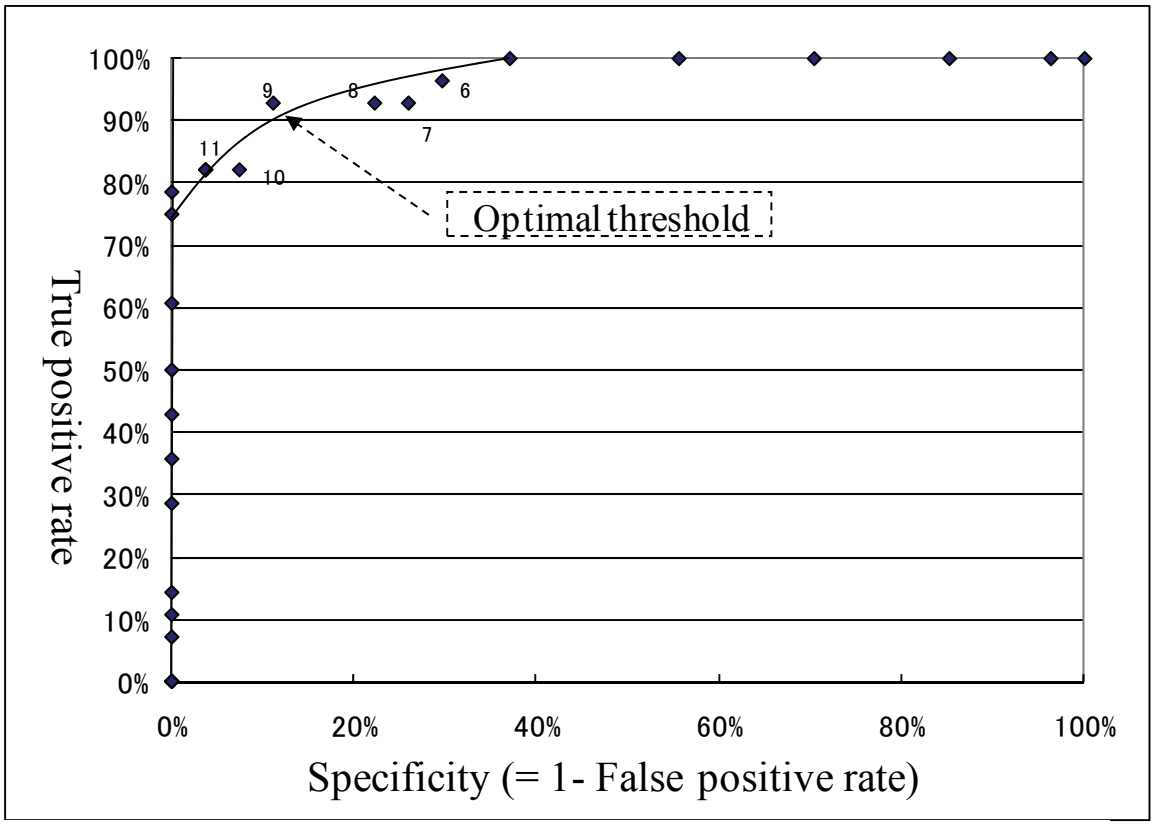

Figure 5. Receiver Operating Characteristic curve
The numbers next to individual points are the threshold values of prediction from the fuzzy logic output.

2.4 Validation

We examined the predictive accuracy in each dataset: the modeling group (19 samples), the training group (18 samples), and the validation group (18 samples). We further confirmed the predictive accuracy with an additional seven samples that were obtained after the model building and first round of predictions.

3. Results

3.1 Gene screening

Table 1 shows the list of genes used to create the membership functions for the prediction of the therapeutic efficacy. First, we screened the expression data of all of the genes by u-test, including genes not related to rheumatism. We expected that the results of the unrelated genes could become

parameters to correct the error between the experiments. The Hepatitis B Virus (HBV) gene, which is not related to any human host system, was used as a negative control for the amount of difference between gene expression patterns. We calculated the difference of the expression values between two genes, and selected the 25 pairs by screening set of difference values. Among the 30 genes for the model construction shown in Table 1, some genes were used for several membership functions.

Table 1. The list of genes used for creating the membership functions.

JGSID	GenBank #	NM_accession	Gene name
3	BC008342	-	cytochrome oxidase 3
4	AF347014	-	cytochrome oxidase 2
24	M12963	NM_000667	alcohol dehydrogenase 1A class I
49	X00955	NM_001643	Apolipoprotein AII (APOA2)
59	M25627	NM_145740	liver glutathione S-transferase subunit 1; glutathione S-transferase A1 (GSTA1)
67	M18907	NM_017460	cytochrome P450 IIIA4 (nifedipine oxidase) (CYP3A4)
120	D15057	NM_001344	defender against cell death 1 (DAD-1)
134	U94354	XM_166539	lunatic fringe; fringe protein (LFNG)
155	D00630	-	HBV surface (subtype adr genomic DNA)
183	M28212	NM_002869	GTP-binding protein (RAB6)
198	S69232	NM_004453	electron transfer flavoprotein-ubiquinone oxidoreductase
237	M26062	NM_000878	interleukin 2 receptor beta chain (p70-75) (IL2RB)
292	D38122	NM_000639	Fas Ligand; tumor necrosis factor (ligand) superfamily, member 6 (TNFSF6)
321	U29656	-	DR-nm23
322	M28882	NM_006500	MUC18 glycoprotein; melanoma cell adhesion molecule (MCAM)
1044	M92642	NM_001856	Alpha-1 type XVI collagen (COL16A1)
1059	L11285	NM_030662	ERK activator kinase (MEK2), dual specificity mitogen-activated protein kinase kinase 2
1064	-	NM_021138	TNF receptor-associated factor 2 (TRAF2)
1093	D86955	NM_004591	CC chemokine LARC precursor, small inducible cytokine subfamily A (Cys-Cys) member 20 (SCYA20); chemokine (C-C motif) ligand 20 (CCL20)
1155	J00123	NM_006211	proenkephalin (PENK) A precursor, exon 3
1171	M34057	NM_000627	latent transforming growth factor beta binding protein 1 (LTBP1)
1178	U86214	NM_032974	caspase 10, Fas-associated death domain protein interleukin-1b-converting enzyme 2 (FLICE2)
1194	L15344	NM_175852	interleukin 14 (high molecular weight B cell growth factor; taxilin)
1198	AF077866	NM_003486	E16 cationic amino acid transporter; solute carrier family 7
1274	AF159442	NM_020360	phospholipid scramblase 3 (PLSCR3)
1322	M15841	NM_003092	U2 small nuclear RNA-associated B,, antigen
1386	BC028078	NM_001337	CX3C chemokine receptor-1 (CX3CR1)
1510	M86511	NM_000591	CD14 (monocyte antigen)
1580	L05144	NM_002591	PCK1; phosphoenolpyruvate carboxykinase (PCK1)
1656	AF000380	NM_000803	folate receptor beta (fetal) (FOLR2)

3.2 Predictions with the fuzzy logic model

Twenty-five sets of membership functions were created from the sample data to generate the model and tune the training. The shapes of membership functions for seven sets are shown in Figure 6 as examples. The output value of these membership functions was evaluated according to the classification performance of the NR group and SR group in the sample data and training data. The evaluation results were weighted to the output value of each membership function, and the cumulative result was used to calculate the total score, which was assumed to be the efficacy prediction.

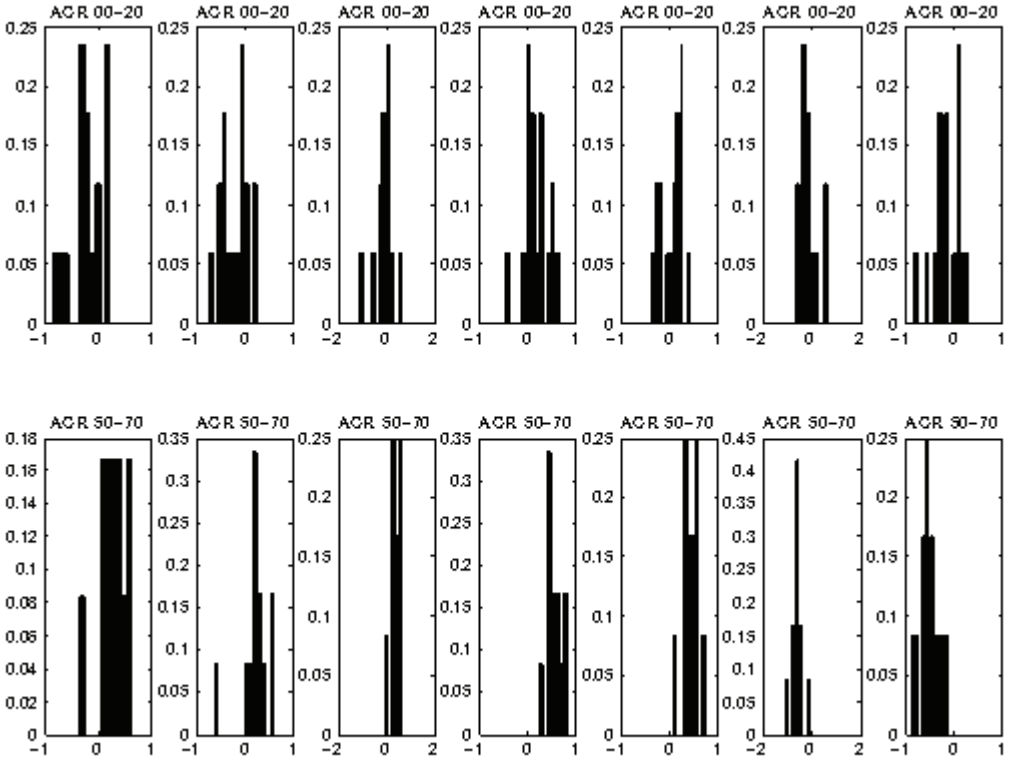

Figure 6. Evaluation by the membership functions
The upper histograms are the NR groups and the lower histograms are the SR groups.
The x-axis is the difference value between two logged gene expression ratios (as input value of membership function). The y-axis is output value of the membership function.

3.3 The predictive accuracy of the fuzzy logic model

The predictive accuracy of each group upon the treatment for RA was as follows (each p-value was determined by Chi-squared test): 100% for the samples used to generate the model (19 samples, $p < 0.0001$), 94% for the samples used to train the model (19 samples, $p<0.0001$), and 89% for the samples used to validate the model (18 samples, $p<0.005$). We also predicted the treatment outcome for an additional seven samples obtained after the model was generated and found a predictive accuracy of 86%. Thus, an accuracy of over 86% was obtained for all of the data groups.

4. Discussion

At first, we were not able to deduce an effective result without using the difference of the two gene expression datasets. Then, we tried to construct a model using fuzzy logic and an automatic membership function generation technique for individual gene expression information. An accuracy exceeding 86% was achieved by calculating the difference between two log-transformed gene expression ratios rather than the difference in the absolute values between two genes. The gene expression measurements were expressed as the difference in value between two genes to reduce the influence of error between the experiments. We think that the mutual effects of two combined

techniques provided this result. The first technique was the construction of the parameters of the membership function, and the other was the use of the fuzzy logic algorithm to analyze the expression pattern of the genes based on the difference between two log-transformed gene expression ratios. This combination of methods contains the possibility of reducing misclassifications due to measurement error. This technique can be applied also to the data of other microarray. However, to make the histogram with this technique, reasonable amount of samples is needed.

In gene expression analyses by microarray, a single experiment provides a vast dataset for the corresponding sample, and it is assumed that the significant differences between target sample groups are few. It is thought that differences due to the positions of probes on the chips and irregularities of the assay may be factors affecting the error margins. Moreover, when gene expression data are gathered from clinical samples, there are numerous sources of error and it is more difficult to control the surrounding conditions. These multiple factors in the error margin include differences between the samples, experimental technique differences, different experimental dates, differences in the environmental conditions, and differences in the batches of manufactured reagents. To exclude the influence of these sources of error as much as possible, an advanced technology for the removal and normalization of noise is needed. We propose that the technique presented here could reduce the influence of noise and abnormal data and improve the accuracy of data comparisons between samples. The measurement of the change of two gene sets is determined by using the difference value between two log-transformed gene expression ratios as a parameter. Moreover, the parameter that shows the change of the living body, which cannot be detected by the fluctuation of only one gene, may be counted as a parameter. This method is significant because there are few studies in the literature that have searched for biomarkers using the ratio of two kinds of measurements with various metrics. However, other studies of rheumatoid arthritis have detected 2 of the 30 genes used in the present study: CX3CR1 was listed in Yano's study[19], and IL2RB was listed in Han's study[20]. This overlap suggests that the 30 genes used in this paper are likely to be related to rheumatoid arthritis. In conclusion, our new method of fuzzy logic is effective in the analysis of gene expression data. Furthermore, we have found that the use of difference values between two gene expression datasets and the technique for the automatic generation of membership functions made from histograms are both effective. We will continue to apply the techniques used in this study to other types of data to develop more efficient numerical analysis techniques.

References

[1] Gerhold, D., R. Jensen, and S. Gullans, Better therapeutics through microarrays, *Nature Genetics*, **32**, 547-551(2002).

[2] Dembele, D. and P. Kastner, Fuzzy C-means method for clustering microarray data, *Bioinformatics*, **19**(8), 973-980(2003).

[3] Gordon, G., et al., Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma, *Cancer Research*, **62**(17), 4963-4967(2002).

[4] TAKATA, R., et al., Predicting response to methotrexate, vinblastine, doxorubicin, and cisplatin Neoadjuvant chemotherapy for bladder cancers through genome-wide gene expression profiling, *Clinical cancer research*, **11**(7), 2625-2636(2005).

[5] Zhang, Q., et al., Which to use?-microarray data analysis in input and output data processing, *Chem-Bio Informatics Journal*, **4**, 56-72(2004).

[6] Takahashi, K., Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification, *Chem-Bio Informatics Journal*, **3**(1), 30-45(2003).
[7] Chuaqui, R., et al., Post-analysis follow-up and validation of microarray experiments, *Nature Genetics*, **32**, 509-514(2002).
[8] Sultan, M., et al., Binary tree-structured vector quantization approach to clustering and visualizing microarray data, *Bioinformatics*, **18**, 111-119(2002).
[9] Daiba, A., et al., A low-density cDNA microarray with a unique reference RNA: pattern recognition analysis for IFN efficacy prediction to HCV as a model, *Biochemical and Biophysical Research Communications*, **315**(4), 1088-1096(2004).
[10] Ando, T., et al., Fuzzy Neural Network Applied to Gene Expression Profiling for Predicting the Prognosis of Diffuse Large B-cell Lymphoma, *Cancer Science*, **93**(11), 1207-1212(2002).
[11] Ando, T., et al., Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling, *Cancer Science*, **94**(10), 906-913(2003).
[12] Ressom, H., R. Reynolds, and R. Varghese, Increasing the efficiency of fuzzy logic-based gene expression data analysis, *Physiological Genomics*, **13**(2), 107-117(2003).
[13] Belacel, N., et al., Fuzzy J-Means and VNS methods for clustering genes from microarray data, *Bioinformatics*, **20**(11), 1690-1701(2004).
[14] Glez-Pe a, D., et al., DFP: a Bioconductor package for fuzzy profile identification and gene reduction of microarray data, *BMC bioinformatics*, **10**(1), 37(2009).
[15] Sekiguchi, N., et al., Messenger ribonucleic acid expression profile in peripheral blood cells from RA patients following treatment with an anti-TNF-alpha monoclonal antibody, infliximab, *Rheumatology (Oxford, England)*, **47**(6), 780-788(2008).
[16] Workman, C., et al., A new non-linear normalization method for reducing variability in DNA microarry experiments, *Genome Biology*, **3**(9), 0048.0001-0048.0016(2002).
[17] Qin, L. and K. Kerr, Empirical evaluation of data transformations and ranking statistics for microarray analysis, *Nucleic Acids Research*, **32**(18), 5471-5479(2004).
[18] Daibata, M., et al., Differential gene-expression profiling in the leukemia cell lines derived from indolent and aggressive phases of CD56+ T-cell large granular lymphocyte leukemia, *International journal of cancer*, **108**(6), 845-851(2004).
[19] Yano, R., et al., Recruitment of CD16+ monocytes into synovial tissues is mediated by fractalkine and CX3CR1 in rheumatoid arthritis patients, *Acta medica Okayama*, **61**(2), 89-98(2007).
[20] Han, T., et al., TRAF1 polymorphisms associated with rheumatoid arthritis susceptibility in Asians and in Caucasians, *Arthritis and rheumatism*, **60**(9), 2577-2584(2009).