Artificial Immune System for Classification of Gene Expression Data

Shin Ando[1] and Hitoshi Iba[2]

[1] Dept. of Electronics, School of Engineering
University of Tokyo
ando@miv.t.u-tokyo.ac.jp
[2] Dept. of Frontier Informatics, School of Frontier Science
University of Tokyo
iba@miv.t.u-tokyo.ac.jp

Abstract. DNA microarray experiments generate thousands of gene expression measurement simultaneously. Analyzing the difference of gene expression in cell and tissue samples is useful in diagnosis of disease. This paper presents an Artificial Immune System for classifying microarray-monitored data. The system evolutionarily selects important features and optimizes their weights to derive classification rules. This system was applied to two datasets of cancerous cells and tissues. The primary result found few classification rules which correctly classified all the test samples and gave some interesting implications for feature selection.

1 Introduction

The analysis of human gene expression is an important topic in bioinformatics. The microarrays and DNA chips can measure the expression profile of thousands of genes simultaneously. Genes are expressed differently depending on its environment, such as their affiliate organs and external stimulation. Many ongoing researches try to extract information from the difference in expression profile given the stimulation or environmental change. Some of the experiments have shown promising result in diagnosis of cancer [16, 17, 4]. Our focus is on classifying gene expression data in [16] and [17]. These data are publicly available and has been applied by other classification methods.

This paper describes the implementation of Artificial Immune System (AIS) [2] for classification. The AIS simulates the human immune system, which is a complex network structure, which responds to an almost unlimited multitude of foreign pathogens. It is considered to be potent in intelligent computing applications such as detection, pattern recognition, and classification.

This implementation of AIS defines classification rules as hyperplanes, which divide the domain of sample vectors. The experiments with genomic expression data showed how the border lines are captured and how the features are selected. Compared to other classifier methods, AIS show reduced complexity of the rules, while equaling or improving on the accuracy of prediction.

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1926–1937, 2003.

2 Classification of Microarray Data

Important aspects in gene expression classification are selection of informative genes (feature selection), and optimization of strength (weight) of each gene, and generalization.

Many existing works use ranking methods to select features as a kind of dimensionality reduction on the data. Many studies [7, 15, 16], uses correlation metrics G_i (i) to rank the feature genes. Subset of genes with highest correlations is chosen as classifier genes. In (i), μ and σ are the mean and standard deviation for the expression levels of gene i, in class A or B samples. [5] compares several ranking methods and results when combined with several machine learning methods.

$$G_i = (\mu_A - \mu_B) / (\sigma_A + \sigma_B) \qquad \text{(i)}$$

For optimizing weights, machine learning methods such as weighted vote cast [16], Bayesian Network, Neural Network, RBF Network [7], Support Vector Machine [6, 15], have been applied to such data.

2.1 Classification of ALL/AML in Acute Leukemia

In [16], cancerous cells collected from 72 patients of acute leukemia are monitored over 7109 genes. Discovery and prediction of cancer class is important, as clinical outcome vary depending on its class. Each sample belongs to either ALL or AML cancer classes. Two independent data sets, training data set (38 cell samples, 27 ALL and 11 AML) to learn the cancer classes and test data set (34 samples, 20 ALL and 14 AML) to evaluate its prediction were provided in [16]. The reliable diagnoses of the samples were made by combination of clinical tests.

2.2 Colon Cancer Diagnosis

In [17], using Affymetrix oligonucleotide arrays, expression levels of 40 tumor and 22 normal colon tissues are measured for 6500 human genes. Among these genes, 2000 with the highest minimal intensity across the tissues are selected for classification purposes. Since no training data set or test dataset were classified, we randomly chosen 38 training samples and 24 test samples.

Table 1 shows comparison of performance in ALL/AML classification by various machine learning techniques. The results are cited from [6, 7, 15, 16]. Each method were trained and tested under following conditions. Weighted Vote Cast (WVC) [16] selected 50 genes with high correlation as classifier genes. Each sample is classified by the sum of each gene’s correlation value. Since weights are not learned, training samples are misclassified.

Neural Network (NN) creates different classifier in every run. The success rate shown in Table 1 is the best in 10 runs, while the average and worst success rate of test data was 4.9 and 9 respectively. Bayesian Network[7] uses 4 genes with higher correlations, though how the features were chosen is not clearly stated. Support Vector Machine (SVM) is a combination of linear modeling and instance–based learning. [15] uses 50, 100, and 200 genes of high correlation. [6] uses Recursive Feature

Elimination to select informative genes. Result in Table 1 is obtained when 8, 16, 32 genes were chosen.Our implementation of AIS selects combination of genes and weights in evolutionary recombination process. The result in Table 1 is the best of 20 runs, while average number of misclassification was 0.85. More detailed result will be given in later section.

Table 1. Comparison of machine learning techniques in Leukemia data set

ALL/AML	WVC[16]	BN[7]	NN[7]	SVM[15]	SVM[6]	AIS
Training	2/38	0/38	0/38	0/38	0/38	0/38
Test	5/34	2/34	1/34	2-4/34	0/34	0/34

3 Features of Immune System

The capabilities of natural immune system, which are to recognize, destroy, and remember almost unlimited multitude of foreign pathogens, have drawn increasing interest of researchers over the past few years. Application of AIS includes fields of computer security, pattern recognition, and classification.

The natural immune system responds to and removes intruders such as bacteria, viruses, fungi, and parasites. Substances that are capable of invoking specific immune responses are referred to as antigens (Ag).

Immune system learns the features of antigens and remembers successful responses to use against invasions by similar pathogens in the future. These characteristics are achieved by a class of white blood cells called lymphocytes, whose function is to detect antigens and assist in their elimination. The receptors on the surface of a lymphocyte bind with specific epitopes on the surfaces of antigens. These proteins related to immune system are called antibodies (Ab).

Immune system can maintain a diverse repertoire of receptors to capture various antigens, because the DNA strings which codes the receptors are subject to high probability crossover and mutation, and new receptors are constantly created.

Lymphocytes are subject to two types of selection process. Negative selection, which takes place in thymus, operates by killing all antibodies that binds to any self-protein in its maturing process. The clonal selection takes place in the bone marrow. Lymphocyte which binds to a pathogen is stimulated to copy themselves. The copy process is subject to a high probability of errors, i.e., hypermutation. The combination of mutation and selection amounts to an evolutionary algorithm that produces lymphocytes that become increasingly specific to invading pathogens.

During the primary response to a new pathogen, the organism will experience an infection while the immune system learns to recognize the epitope by evolutionary process. The memory of successful receptors is maintained to allow much quicker secondary response when same or similar pathogens invade thereafter.

There are several theories of how immune memory is maintained. The AIS in this paper stores successful antibodies in permanent memory cells to store adapted results.

4 Implementation of Artificial Immune System

In the application of artificial immune system to ALL/AML classification problem, the following analogy applies. The ALL training data sets correspond to Ag, and AML training data sets to the self-proteins. Classification rules represent Ab, which captures training samples(Ag or self-proteins) when its profile satisfies the conditions of the rule. A population of Ab goes through a cycle of invasion by Ag and selective reproduction. As successful Abs are converted into memory cells, ALL/AML class is learned. Above analogy can be applied to Colon cancer data by replacing ALL with the tumor tissues and AML with normal tissue.

4.1 Rule Encoding

Classification rules are linear separators, or hyper-planes, as shown in (ii). Vector *x* =(x_1, x_2, …, x_i, … x_n) represents gene expression levels of a sample, and vector *w*=(w_1, w_2, …, w_i, …, w_n) represents the weight of each gene. A hyperplane *W*(*x*)=0 can separate the domain and the samples into two classes. *W* determines the class of sample *x*; if *W*(*x*) is larger than or equal to 0, sample *x* is classifies as ALL. If it is smaller than 0, the sample is classified as AML.

$$W(x) \geq 0 \ \left\{ \ W(x) = w^T \cdot x \right\} \tag{ii}$$

Encoded rules are shown in (iii). It represents a vector *w*, where each loci consists of a pointer to a gene and weight value of that gene. It corresponds to a vector where unspecified gene weights are supplemented with 0. It is similar to messyGA[3] encoding of vector *w*.

$$(X_{123}, 0.5)\ (X_i, w_i)\ (\ldots)\ (\ldots)\ (\ldots)\ (\textit{gene index}, \textit{weight}) \tag{iii}$$

4.2 Initialization

Initially, rules are created by sequential creation of locus. For each locus, a gene and a weight value is chosen randomly. With probability P_i, next locus is created. Thus, average lengths of the initial rules are $\Sigma_n nP_i n$. Empirically, the number of initial rules should be in the same order as the number of genes to ensure sufficient building blocks for classification rules.

4.3 Negative Selection

All newly created rules will first go through negative selection. Each rule is met with set of AML training samples, x_i(i=1,2,…,N_{AML}), as self-proteins. If a rule binds with any of the samples(satisfy (iv)0, it is terminated. The new rules are created until N_{Ab} antibodies pass the negative selection. These rules constitute population of antibodies Ab_i(i=1,2,…,N_{Ab}).

$$\prod \delta(W(x)) \neq 1 \begin{Bmatrix} \delta(t) = 0 \mid t > 0 \\ \quad = 1 \mid t < 0 \end{Bmatrix} \tag{iv}$$

4.4 Memory Cell Conversion

The antibodies who endures the negative selection are met with invading antigens, $Ag_i(i=1,2,\ldots,N_{ALL})$, or ALL training samples. Antibodies which can capture many antigens are converted into memory cells $M_i(i=1,2,\ldots,N_{Mem})$.

Ab_i are converted to memory cell in following conditions. A set of antigens captured by Ab_i or a memory cell M_i is represented by $C(Ab_i)$, $C(M_i)$.

- M_i is removed if $C(M_i) \subset C(Ab_i)$.
- Ab_i is converted to M_{N+1} if $C(Abi) \not\subset C(M_1) \cup C(M_2) \cup \ldots \cup C(M_N)$.
- Ab_i is converted to M_{N+1}, if $C(Mi)=C(Abi)$

4.5 Clonal Selection

The memory cells and Abs which bind with Ags go through clonal selection for reproduction. This process is a cycle described as follows:

- First, an Ag is selected from the list of captured Ags. The probability of selection is proportional to $S(Ag_i)$, the concentration of Ag_i, which is initially 1.
- Randomly select two antibodies Ab_{p1} and Ab_{p2} from all the antibodies bound with the antigen.
- Ab_{p1} and Ab_{p2} are crossed over with probability P_c to produce offspring Ab_{c1} and Ab_{c2}.

  The crossover operation is defined as cut and splice operation.

  Crossover is followed by hypermutation, which is a series of copy mutation applied to w_{c1}, w_{c2}, and their copied offspring. There are several types of mutation. Locus deletion, deletes randomly selected locus. Locus addition, adds newly created locus to the antibody. Weight mutation changes the weight value of randomly chosen locus.
- With probability P_m, newly created antibody creates another mutated copy. Copy operation is repeated for $\Sigma n P_m^n$ times on average.
- Parents are selected from memory cells as well. Same crossover and hypermutation process is applied.
- The copied antigens go through negative selection as previously described. The reproduction processes are repeated until N_{Ab} antigens pass the negative selection.
- Finally, the score of each Ag is updated by (v). T is the score of Ag, s is the number of Ab bound to an Ag, and N is the total number of Ab. β is an empirically determined constant, 1.44 in this study. The concentration of Ag converges to 1 with appropriate β.

$$T' = \beta^{T-s/N} \qquad \text{(v)}$$

The process goes back to Negative Selection to start a new cycle.

4.6 Generalization

In a single run, many rules with same set of captured antigens, C(Mem), are stored as memory cells. After the run is terminated, one memory cell is chosen to classify the test samples. A memory cell with largest margin *M* (vi), is chosen.

$$M = \min_i \left| W(x_i) / W(\tilde{x}_i) \right| \qquad \text{(vi)}$$

x_i are the ALL/AML sample vectors and $\tilde{x}_i$ is the median of the samples. We try to maximize generality by choosing a hyperplane whose margin to nearest sample vector is the largest.

4.7 Summary of Experiment

AIS repeats the cycle as previously described. Flow of the system is shown in Fig. 1. Each run was terminated after N_c(=50) cycles. The AIS runs on parameters shown in Table 2. The results were robust to minor tuning of these parameters.

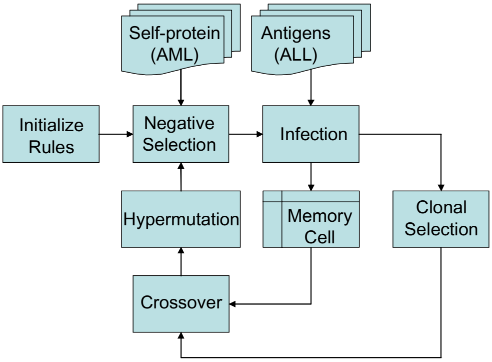

Fig. 1. Cycle of artificial immune system

Table 2. Data attributes and AIS parameters

Leukemia	$N_G = 7109$	$N_{AML} = 11$	$N_{ALL} = 27$		
Colon cancer	$N_G = 2001$	$N_{norm} = 13$	$N_{tumor} = 25$		
AIS parameters	$N_{Ab} = 7{,}000$	$N_c = 50$	$P_i = 0.5$	$P_c = 0.9$	$P_m = 0.6$

4.8 Results

AIS was applied to Leukemia dataset and Colon cancer dataset. Its performance was measured by average and standard deviation of 20 runs. In all runs, training data set was correctly classified. Table 3 shows average and standard deviation of the number of false positives (misclassified AML test data/ misclassified normal tissue) and false negative (misclassified ALL test data / misclassified tumor tissue) prediction on test samples for both dataset.

Table 3. The number of misclassified samples in test data set of Leukemia and Colon cancer

	#FN(Average/Standard Dev.)	#FP(Average/Standard Dev.)
Leukemia	0.3 / 0.47	0.55 / 0.51
Colon Caner	0.75 / 0.44	0.7 / 0.47

Fig. 2 and Fig. 3 show the learning process of AIS in a typical run for Leukemia dataset. Fig. 2 shows the number of Ag captured by best and worst Abs. The average number of captured Ag is also shown. It shows training samples are learned by 20th cycle. It can be read from the graph that cycles afterward are spent to derive more general rules.

Fig. 3 shows the concentration of Ags at each cycle. Most Ags converge to 1, while few are slower to converge. These samples imply the borderline of the classes. Samples near the classification border are prone to misclassification by untrained Abs, thus slower to converge.

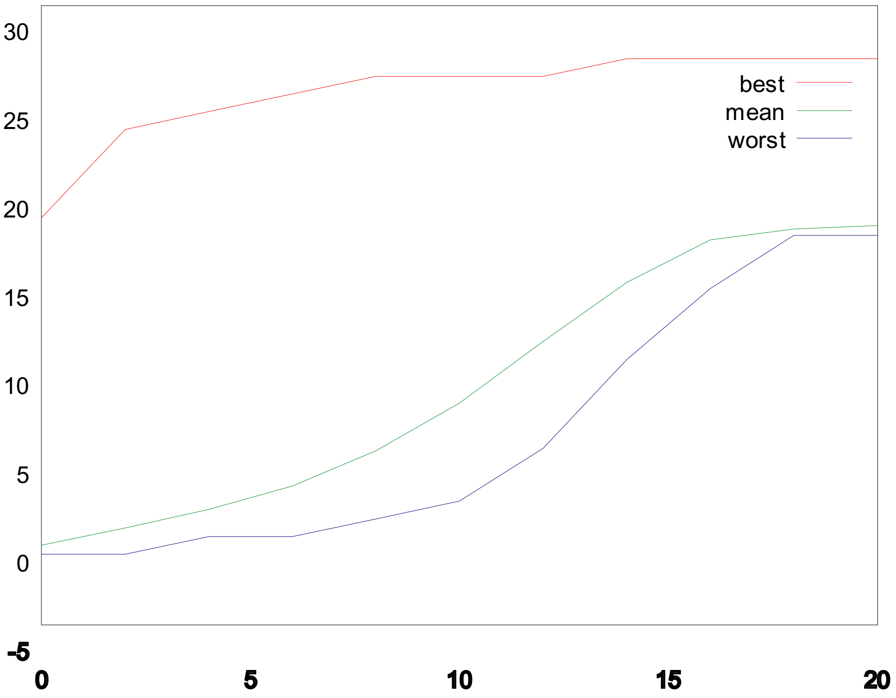

Fig. 2. Number of antigens caught by the best and worst antibodies

Empirically, the results were successful when the grasp of border is clear, i.e. all but few samples have converged. The termination criteria (number of cycle) were determined so that sufficient convergence was achieved by the end of iteration.

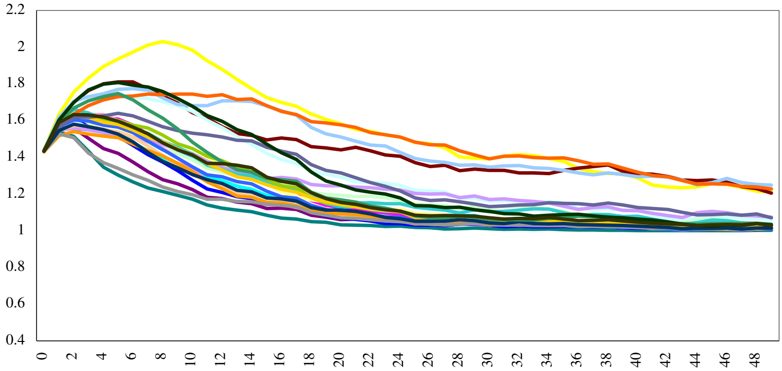

Fig. 3. Transition of antigen scores

5 Analyses of Selected Features

Fig. 4 and Fig. 5 shows some of the classification rules which correctly classified the test samples of Leukemia dataset and Colon cancer dataset respectively.

A) $1.21896X_{3675}$ + $-1.5858X_{4474}$ + $1.46134X_{1540}$ + $-1.19885X_{2105}$ + $1.84803X_{757}$ + $1.82983X_{4038}$

B) $-1.4577X_{1385}$ + $-1.57815X_{4363}$ + $1.31819X_{2317}$+ $1.75329X_{2328}$

C) $1.00809X_{1904}$ + $1.7706X_{6244}$ + $-1.41034X_{4526}$ + $-1.14542X_{759}$ + $1.94696X_{2723}$ + $-1.34382X_{4875}$

D) $1.26745X_{3110}$ + $1.43941X_{1190}$ + 1.97632 + $1.74422X_{5519}$ + $1.79449X_{6874}$ + $-1.44577X_{6022}$

Fig. 4. Examples of ALL/AML classifier rules

A) $-1.678X_{1997}$ $-1.55458X_{1516}$ $1.41783X_{1297}$ $-1.85391X_{696}$ $1.69254X_{416}$ $-1.39212X_{620}$ $-1.19309X_{1672}$ $-1.95164X_{1310}$ $1.95381X_{49}$ $1.64378X_{134}$

B) $-1.25102X_{1997} + 1.77743X_{138} + -1.42182X_{1770} + 1.3154X_{49} + 1.63866X_{183} + 1.17681X_{94}$

Fig. 5. Examples of colon cancer classifier rules

Each rule was different for each run. In this section, we further analyze the selected genes. Some of the genes appear repeatedly in the classification rules. Such genes have relatively high correlations (i) as shown in Table 4. These genes are fairly informative in terms of ALL/AML classification. Fig. 6 shows the expression level of those genes, and how the test data sets can be clustered with features in Table 4. The figure was created with average linkage clustering by Eisen's clustering tool and viewer [10]. It can separate ALL/AML samples with the exception of one sample.

Table 4. Correlation value of the classifier genes (GAN: Gene Accession Number)

Gene	X_{757}	X_{1238}	X_{4038}	X_{2328}	X_{1683}	X_{6022}	X_{4363}
GAN	D88270	L07633	X03934	M89957	M11722	L00634_s	X62654
Gene name	VPREB1	PSME1	CD3D	CD79B	DNTT	FNTA	CD63
Correlation	0.838	0.592	0.647	0.766	0.816	-0.837	-0.834

Many of these genes have relations to Leukemic disease which can be confirmed by biological literature. For example, CD79b is one of the surface marker molecule which could provide important additional information in leukemia cell analysis [8]. CD3D is involved in abnormal location of the genes often observed in acute leukemia [14].

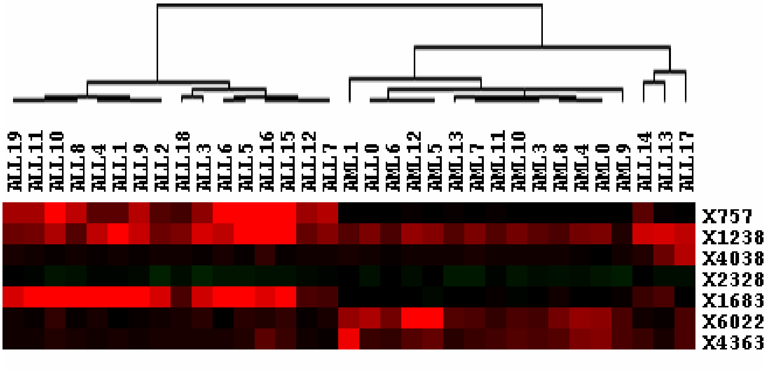

Fig. 6. Expression levels of informative genes and clustering based on those genes

The following section analyzes featured genes in rule A (Fig. 4). Each gene does not always have high correlation value as can be seen in Table 5.

Table 5. Correlation of classifier genes in rule A

Gene	X3675	X4474	X1540	X2105	X757	X4038
Weight	1.21896	-1.5858	1.46134	-1.19885	1.84803	1.82983
GAN	U73682	X69699	L38696	M62762	D88270	X03934
Correlation	0.151	0.371	0.418	-1.02	0.838	0.647

Fig. 7 shows the expression levels of feature genes in rule A. It implies that majority of the samples can be classified by few ALL(X_{2105}, X_{757}) and AML(X_{4038}) classifier genes. These classifier genes have relatively high correlation value.

Some of the samples (AML1, 6, 10, ALL11, 17, 18, 19) in Fig. 7 seem indistinguishable by those classifier genes. Functions of supplementary genes(X_{3675}, X_{4474}, X_{1540}) become evident when these samples are looked at especially.

Fig. 8 shows normalized expression levels of selected samples. In this selected group, X_{3675} and X_{4474} are highly correlated to ALL and AML respectively, while the classifier genes(X_{2105}, X_{757}, X_{4038}) became irrelevant to sample class.

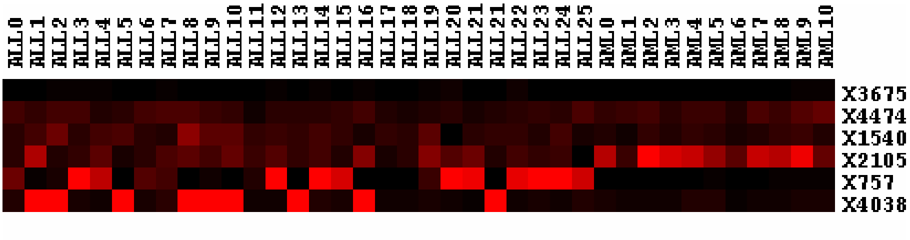

Fig. 7. Expression level of classifier genes in rule A

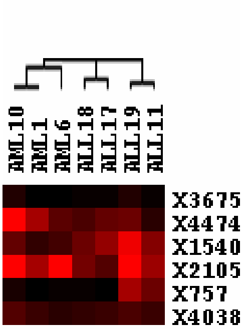

Fig. 8. Selected samples

6 Conclusion

This paper presented Artificial Immune System to classify gene expression of cancerous tissues. Despite sparseness in training data, the accuracy of prediction was satisfactory, as test data were correctly classified 8 out of 20 times in ALL/AML classification. The colon cancer classification is known to be much harder. For comparison, we implemented classification algorithms using popular machine learning methods. Table 6 shows the error rates of the algorithms when applied to the colon cancer data, using 50, 100, and 200 genes selected by correlation metric (i) as features. SVM used linear kernel and error margin of 0.001. NN was implemented as a 3-layer perceptron.

In the experiment with Colon Cancer dataset, the approximate amount of time required for each of the algorithms are: 50 min. for SVM, 3 hours for NN, and 6 hours for AIS. It is hard to compare the efficiency of the algorithms with regards to accuracy of the prediction and amount of computation required. The computational cost may depend on the conditions of the data and the algorithms, e.g. the number of iterations in SVM increase exponentially to the number of training samples, where as the AIS population must increase linear to the number of genes. Considering that preparation of the data, i.e., collecting samples from patients and performing the Microarray tests, takes months, it can be said that all algorithms require considerably small amount of time.

Table 6. Comparison of performance in colon cancer dataset

# of Missclassification	SVM	NN	AIS
Test Data	3/24	9/24	1.45/24

We presume that AIS was more effective than other methods in regards to the feature selection. AIS evolutionarily chooses informative genes whilst optimizing its weight as well. Though gene subset chosen by AIS for classification differs in each run, the genes with strong correlation are chosen frequently as Table 4 shows. Similar results can be obtained by application of Genetic Algorithm [9].

These genes with strong correlation, either selected by frequency or correlation, may not contain enough information to be a sufficient subset for classification, when many co-regulated genes are selected in the subset. Many genes are predicted to be co-expressed and those genes are expected to have similar rankings.

On the other hand, the result in Fig. 8 implies that selection of complementary genes, which are not necessary highly correlated, can be useful in classification. It might be suggested that the choice of feature gene subsets should be based not only on single ranking method, but also on redundancy and mutual information between the genes. Changing of ranking objective, when one feature is removed as a ranking criterion, has been suggested in [13], while [1937] states that performance of machine learning is naïve to choice of features. AIS selects features in the learning process and it is interesting that it can choose primary and complementary feature genes by evolutionary process.

Monitoring the convergence of the clonal selection suggested new termination criteria, as results improved when all but few genes converged. The analysis and quantitative implementation of such criteria is underway. As future work to improve classification capability, use of effective kernel functions, and expressing relations between the genes, such as combining antibodies with AND/OR functions should be addressed.

References

1. A. Ben-Dor, N. Friedman, Z. Yakini, Class discovery in gene expression data, Proc. of the 5th Annual International Conference on Computational Molecular Biology, 31–38, 2001.
2. D. Dasgupta. Artificial Immune Systems and Their Applications. Springer, 1999.
3. D. Goldberg, B. Korb and K. Deb, Messy Genetic Algorithms: Motivation, Analysis and First Results, Complex Systems, 3:493–530, 1989
4. Donna K. Slonim, Pablo Tamayo, Jill P. Mesirov, Todd R. Golub, Eric S. Lander, Class Prediction and Discovery Using Gene Expression Data, Proc. of the 4th Annual International Conference on Computational Molecular Biology(RECOMB), 263–272, 2000.
5. H. Liu, J. Li, L. Wong, A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns, in Proceeding of Genome Informatics Workshop, 2002
6. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning Vol. 46 Issue 1–3, pp. 389–422, 2002
7. K.B. Hwang, D.Y. Cho, S.W. Wook Park, S.D. Kim, and B.Y. Zhang, Applying Machine Learning Techniques to Analysis of Gene Expression Data: Cancer Diagnosis, in Proceedings of the First Conference on Critical Assessment of Microarray Data Analysis, CAMDA2000.
8. Knapp W, Strobl H, Majdic O., Flow cytometric analysis of cell-surface and intracellular antigens in leukemia diagnosis. Cytometry 1994 Dec 15;18(4):187–98
9. L. Li, C. R. Weinberg, T. A. Darden, L. G. Pedersen, Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, Bioinformatics, Vol. 17, No. 12, pp. 1131–1142, 2001
10. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Science, 85:14863–14868, 1998.
11. P. Baldi and A. Long, A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes, Bioinformatics, 17:509--519, 2001.
12. P.J. Park, M. Pagano, and M. Bonetti, A nonparametric scoring algorithm for identifying informative genes from microarry data, PSB2001, 6:52–63, 2001.
13. R. Kohavi and G. H. John, Wrappers for Feature Subset Selection, Artificial Intelligence, vol. 97, 1–2, pp. 273–324, 1997
14. Rowley JD, Diaz MO, Espinosa R 3rd, Patel YD, van Melle E, Ziemin S, Taillon-Miller P, Lichter P, Evans GA, Kersey JH, et al., Mapping chromosome band 11q23 in human acute leukemia with biotinylated probes: identification of 11q23 translocation breakpoints with a yeast artificial chromosome., Proc Natl Acad Sci U S A 1990 Dec;87(23):9358–62
15. T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 2001
16. T.R. Golub, D.K. Slonim, P. Tamayo, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531–537, 1999.
17. U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. Cell Biology, 96:6745–6750, 1999.