

Gene Screening Method for Prognostic Prediction Using Projective ART Model

Hiro Takahashi

h021413m@mbbox.nagoya-u.ac.jp

Takeshi Kobayashi

takeshi@nubio.nagoya-u.ac.jp

Hiroyuki Honda

honda@nubio.nagoya-u.ac.jp

Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho,
Chikusa-ku, Nagoya 464-8603, Japan

Keywords: gene expression analysis, class prediction, projective ART

1 Introduction

Recent advances in DNA microarray technologies have made it possible to measure the expression levels of thousands of genes simultaneously. Artificial neural network (ANN), and fuzzy neural network (FNN) combined with SWEEP operator method (FNN-SWEEP method) are useful for constructing cancer class prediction model with high accuracy [1]. However, the gene expression data are easy to include an experimental error. Therefore, it is necessary to selectively find significant genes and also necessary to eliminate nonspecific genes so as to prevent the model from overfitting for learning data before modeling. In this paper, we applied projective adaptive resonance theory (PART) [2] to gene expression data for eliminating nonspecific genes. Furthermore, the genes selected by PART were applied to FNN-SWEEP method for constructing cancer class prediction model. The results in modeling were evaluated by being compared with those without using PART.

2 Method and Results

2.1 Data Processing

In this study, we used gene expression data from a study of Golub *et al.* [4]. These gene expression data consisted of 72 bone marrow sample (47 acute lymphoblastic leukemia (ALL), 25 acute myeloid leukemia (ALL)) were obtained from acute leukemia patients at the time of diagnosis. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 7,129 human genes. For these dataset, we used a standard deviation threshold of 50 for expression units to select the 5,401 most variable genes. Furthermore, these data were separated into modeling dataset consisted of 38 samples (27 ALL, 11 AML) for constructing class prediction model and independent dataset consisted of 34 samples (20 ALL, 14 AML) for evaluating class predictor constructed.

2.2 Evaluation of Extracted Genes by PART

To evaluate extracted genes by PART, we constructed two kinds of FNN class predictors. As one way, we applied PART to modeling dataset and extracted 253 genes from 5,401 genes. Then, class predictor genes were selected by FNN with SWEEP from extracted genes by PART (Predictor 1). As an alternative way, the predictor genes were selected directly from 5,401 genes without PART (Predictor 2). These two predictors were compared. The outline of this paper is shown in Figure 1.

2.3 PART Model

PART was proposed to find projected clusters for data sets in high dimensional spaces. The architecture is based on the well known ART developed by Carpenter and Grossberg [3], and a major modification is provided in order to deal with the inherent sparsity in the full space of the data points from many data-mining applications (Figure 2).

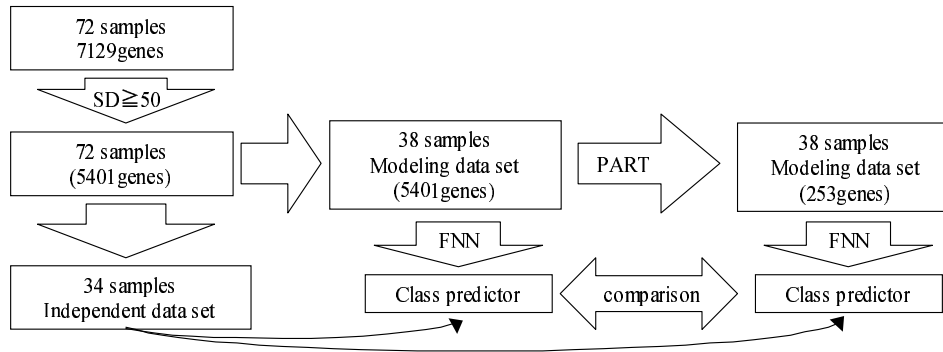


Figure 1: Outline of this paper.

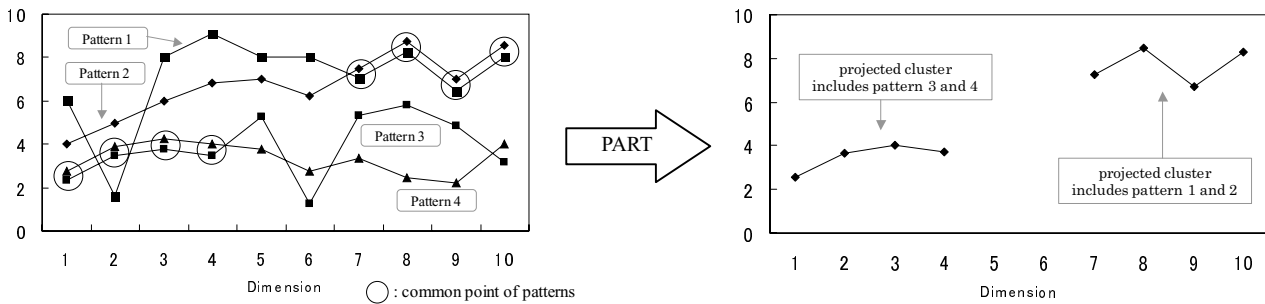


Figure 2: Function of PART clustering.

3 Results and Discussion

We constructed two kinds of FNN class prediction models (with PART screening and without PART). As shown in Table 1, predictor 1 showed 93% correctness ratio for modeling data set and 97% for independent data set. On the other hand, predictor 2 without PART was 93%, 76%, respectively. In both cases, 10 FNN models were constructed for prediction and in predictor 1 only 10 significant genes were used for class prediction. In the conventional method using weight voted method, 50 genes were selected for the same analysis. From these results, this result suggests that PART has a potential to function as a new method of genes screening for class prediction.

Table 1: Correctness ratio of class prediction for two kinds of acute leukemia.

Model	Gene Screening by PART	Discrimination correctness (3-fold cross validation)	
		Modeling Data Set	Independent Data Set
Predictor 1	YES	93%	97%
Predictor 2	NO	93%	76%

References

- [1] Ando, T. *et al.*, Selection of casual gene sets for lymphoma prognostication from expression profiling and construction of prognostic fuzzy neural network models, *Journal of Bioscience and Bioengineering*, 96:161–167, 2003.
- [2] Cao, Y. *et al.*, Projective ART for clustering data sets in high dimensional spaces, *Neural Networks*, 15:105–120, 2002.
- [3] Carpenter, G.A. *et al.*, A massively parallel architecture for a self-organizing neural pattern recognition machine, *Computer Vision, Graphics, and Image Processing*, 37:54–115, 1976.
- [4] Colub, T.R. *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286:531–537, 1999.