Selection of Causal Gene Sets from Gene Expression Profiles Using GeneFis®, New Software Based on FNN

Hiroyuki Honda honda@nubio.nagoya-u.ac.jp
Takeshi Kobayashi takeshi@nubio.nagoya-u.ac.jp

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

Keywords: GeneFIS, DNA microarray, fuzzy neural network, noninferior model

1 Introduction

Microarray data were useful in disease diagnosis and prognosis. Most approaches to the computational analysis of gene expression data are functionally significant classification of genes [1, 6, 7]. Fuzzy Neural Network (FNN) is one of the advanced ANN models. FNN system can automatically select the causal gene set consisting of several genes for prediction of the disease diagnosis and prognosis and the constructed prediction models showed more than 90% accuracy from cDNA microarray [2] or oligonucleotide microarrays [3] for diffuse large B cell lymphoma (DLBCL) patients. In the present paper, we introduce the customized software, GeneFIS® (Fuzzy Inference System for Gene expression analysis), which is incorporated in FNN modeling for prognostic prediction from gene expression data. The majoritarian decision using multiple noninferior models can be also provided as an optional function. Here, we analyzed here the outcome prediction of 220 DLBCL patients with high heterogeneity using GeneFIS®.

2 Materials and Methods

Contents of GeneFIS® (Fuzzy Inference System for Gene expression analysis) GeneFIS® provided the SWEEP operator method for ranking of candidate genes and selection of partner gene or combination against arbitrary genes, noninferior FNN modelings for outcome prediction (Fig. 1), the extraction of the explicit IF-THEN rule and majoritarian decision using multiple noninferior model as optional fuction. The Software is available from Mutsui Knowledge Industry Co. (Tokyo, Japan).

Data preprocessing Transcriptional profiling data obtained from "Lymphochip" DNA microarrays [6] were used in the present study. A total of 7,384 gene expression data from 220 DLBCL patients with long-term clinical follow-up were available. We divided the 220 patients into 2 groups: those who were alive 4 years after the beginning of anthracycline-based chemotherapy (group 1; n = 102).

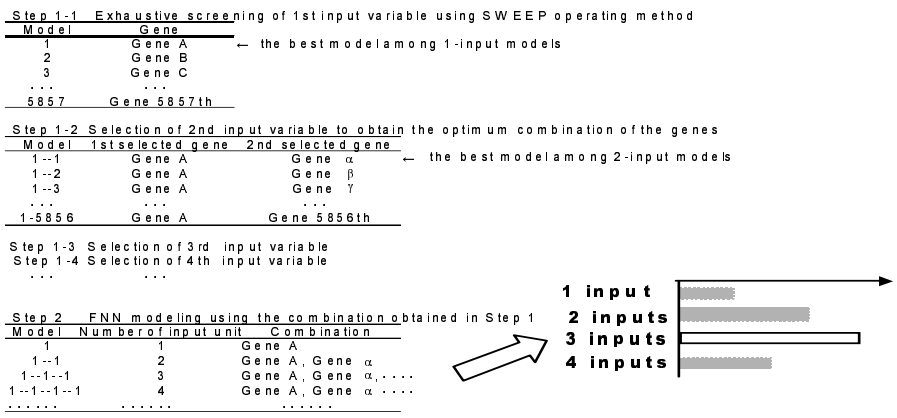

Figure 1: Modeling schema of GeneFIS®.

3 Results and Discussion

Using the SWEEP operator method, we calculated the prediction accuracy of a 1-input FNN model, and ranked 7384 genes using microarray data. The partner genes of the first-ranked gene, *AA805575,*

were selected in order to obtain the best combinations with this gene. The FNN model automatically selected 4 genes, EST *AA805575*, the genes for sentrin/SUMO-specific protease 3, butyrophilin (subfamily 3, member A2) and tumor necrosis factor (ligand) superfamily (member 10), and the accuracy of the constructed FNN model was 73.4%, whereas that of a Cox proportional-hazards model with 17 genes was 68.5% [6].

Kaplan-Meier survival analysis was performed based on the majoritarian decision of 9 FNN models. The results clearly divided the patients into 2 groups ($p<0.001$, Fig. 3). Patients predicted to survive by the FNN models had significantly longer survival time (4-year overall survival [OS], 91%) than those predicted not to survive (4-year OS, 10%).

Using the FNN models, relationships between input genes and clinical outcome were investigated. One of the attractive features of FNN models is that causal relationships can be described explicitly as IF-THEN rules [4, 5]. As shown in Fig.3, patients who satisfied any 2 rules (e.g., high expression of MAX dimerization with high expression of unknown A, or high expression of MAX dimerization with low expression of unknown B) were predicted to have a poor outcome in the FNN model. One third of all non-surviving patients (41/118) have satisfied any 2 rules.

In conclusion, we have established a software, GeneFIS®, for analysis of gene expression developed a method using multiple noninferior FNN models, and have achieved highly accurate prediction outcome of the heterogeneous disease DLBCL. The present paper is the first one describing the multiple noninferior FNN modeling system.

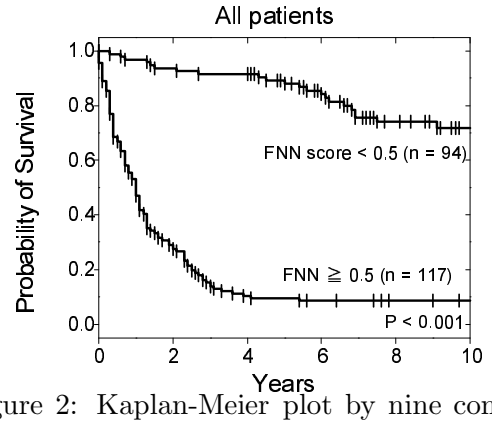

Figure 2: Kaplan-Meier plot by nine combinations.

cluster A			Unknown A			
			L		H	
			nuclear receptor subfamily 3			
			L	H	L	H
MAX dimerization	Unknown B	L	7, 34, 51, 57, 76, 88, 98, 100, 109, 124, 130, 177, 238, 281 286, 320, 324	**199**	3, 24, 63, 68, 95, 97, 227, 322, **48, 66, 117, 230**	
		H	**62, 141, 419**	86, 399, 432	**263**	
	H	L	**134, 244, 262, 266, 285**	**189**	**30, 64, 103, 166, 191, 203, 212**	**5, 8, 11, 12, 13, 15, 61, 89, 101, 107, 110, 120, 144, 152, 163, 179, 198, 205, 206, 249, 255, 313, 317, 415**
		H	297, 433, **391**	118, 136, 186, 248, 294, 298, 394	**300, 404**	**78, 411**

Figure 3: Relationship between expression of 4 genes selected and clinical outcome of DLBCL patients. Dark gray areas represent the poorest prognosis. Bold-type underlined numbers indicate the fatal patients.

References

[1] Alzadeh, A.A. *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000. `http://genome-www5.stanford.edu/MicroArray/SMD/`

[2] Ando, T. *et al.*, Fuzzy Neural Network applied to gene expression profiling for prognosis of diffuse large B-cell lymphoma, *Jpn. J. Cancer. Res.*, 93:1207–1212, 2002.

[3] Ando, T. *et al.*, Selection of causal gene sets for lymphoma prognostication from expression profiling and construction of prognostic fuzzy neural network models, *J. Biosci. Bioeng.*, in press.

[4] Horikawa, S. *et al.*, A study on fuzzy modeling using fuzzy neural networks, *Proc. Int. Fuzzy Eng. Symp., '91*, 562–573, 1991.

[5] Noguchi, H. *et al.*, Fuzzy neural network-based prediction of the motif for MHC class II binding peptides, *J. Biosci. Bioeng.*, 92:227–231, 2001.

[6] Rosenwald, R. *et al.*, The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, *N. Engl. J. Med.*, 346:1937–47, 2002. `http://llmpp.nih.gov/DLBCL`

[7] Shipp, M.A. *et al.*, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nat. Med.*, 8:68–74, 2002. `http://www.genome.wi.mit.edu/MPR/lymphoma`