Ming Chuan-Health Tech Jounal, 2010, Vol. 1, No. 1, e7

Predicting Cancer Survival Time Using an Artificial Neural Network with Gene Expression Data

Yen-Chen Chen[1], Wen-Wen Yang[2], Tsung-Chieh Lee[3], Hung-Wen Chiu[1*]

[1]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taiwan
[2]Graduate Institute of Medical Sciences, Taipei Medical University, Taiwan
[3]Department of Biomedical Engineering, Yuanpei University, Taiwan

Received 15 April 2010/Revised 15 May 2010/Accepted 16 June 2010

ABSTRACT

Objectives. Survival analysis is commonly used for analyzing time-to-event data in medical research. This study aimed to determine the usefulness of training artificial neural networks (ANNs) for predicting the survival time in cancer patients using microarray and clinical data. ***Methods***. We analyzed public-domain microarray and clinical data sets with different kinds of cancer. We selected 15–30 genes (with correlation coefficient values of >0.4) as variables to train the ANNs. All models were tested with a testing set to determine their accuracy in predicting the survival time. The network with the highest classification accuracy was used in subsequent experiments. ***Results***. The selection of 15–30 genes as ANN variables allowed well-trained networks to be produced, with correlation coefficients of greater than 0.7. ***Conclusions***. The results showed that the survival times predicted by an ANN using microarray gene expression data are in good agreement with real observations.

INTRODUCTION

Survival analysis is commonly used for analyzing time-to-event data in medical research. The survival time is the length of time that a patient survives after the occurrence of a given event related to a disease, such as the time period from the beginning to the end of a remission period or the time period from the diagnosis of a disease to death. Many cancer studies have considered gene expression and clinical data, such as those related to lymphoma [1–4] and ovarian cancer [5]. Such studies have applied microarray technology to identify specific cancer-related genes that can be used to diagnose and predict the cancer stage.

The artificial neural network (ANN) is a form of artificial intelligence that employs nonlinear mathematical models to simulate the problem-solving process of the human brain. Humans apply knowledge gained from past experience to new problems, and a neural network similarly uses previously solved examples to build a system of “neurons” that makes new classifications, decisions, and predictions. The classification rules are not written into algorithms; instead, the network learns them from the examples that it encounters. An ANN can be constructed from one or more layers of neurons, and many biomedical studies have shown such networks to be good at predicting and classifying clinical outcomes [4]. The present study aimed at training ANNs using microarray and clinical data in order to predict the survival time in cancer patients.

MATERIALS AND METHODS

Data preprocessing. We analyzed three public-domain microarray and clinical data sets that can be downloaded from the Internet. Data preprocessing and normalization were performed using BRB-ArrayTools [6]. The flowchart in Figure 1 shows the methodology applied in this study, and the data sets used are described below.

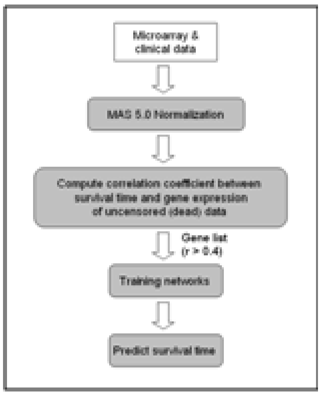

FIG. 1. Flowchart of the normalization and selection of candidate genes from microarray data for training networks.

Diffuse large B-cell lymphoma (DLBCL), the most common lymphoid malignancy in adults, is curable in less than half of patients [7]. Microarray gene expression profiles and survival data were obtained for the study of Shipp et al. [4]. Microarray (Affymetrix Hu6800) and clinical data were obtained from 58 untreated patients who had been diagnosed with DLBCL. The median follow-up time was 43 months (range, <12 to 182 months), during which 27 deaths occurred [4]. The Kaplan-Meier survival curve for the 58 DLBCL patients is shown in Figure 2A.

Follicular lymphoma (FL) is the second most frequent type of non-Hodgkin's lymphoma (NHL), comprising 22% of all NHL cases [8]. Microarray gene expression profiles and survival data were obtained for the study of Dave et al. [1]. The data set consists of 191 untreated patients who had been diagnosed with FL, with 95 deaths occurring during the follow-up. The array chips used were the Affymetrix U133A and U133B platforms. The median age at diagnosis was 51 years (range, 23 to 81 years), and the median follow-up time was 6.6 years (range, <1.0 to 28.2 years). The details of the experiment were reported by Dave et al. [1]. The Kaplan-Meier survival curve for the 191 FL patients is shown in Figure 2B.

Ovarian cancer is a leading cause of cancer death among women in the United States and Western Europe, and has the highest mortality rate of all gynecologic cancers. Microarray gene expression profiles and survival data were obtained from the study of Dressman et al. [5]. Microarray (Affymetrix U133A) and clinical data were obtained at the initial cytoreductive surgery from 119 patients who received platinum-based primary chemotherapy at Duke University Medical Center and the H. Lee Moffitt Cancer Center and Research Institute. The median follow-up time was 21 months (range, <1 to 148 months), during which 69 deaths occurred. The details of the experiment were reported by Dressman et al. [5]. The Kaplan-Meier survival curve for the 119 ovarian cancer patients is shown in Figure 2C.

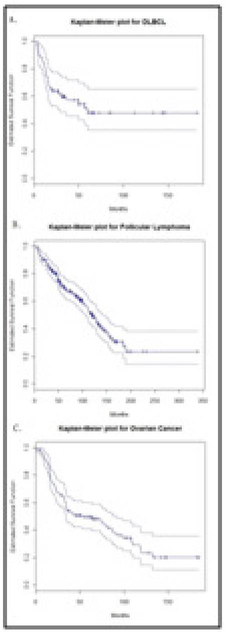

FIG. 2. Kaplan-Meier estimates of the probability of survival (median and 95% CI values) for three kinds of cancer.

Training network. Several types of ANN analysis were performed. The first aim was to determine the optimal ANN architecture. The data set was divided into two separate groups: (1) random selected 90% uncensored patients (the training set) and (2) the other patients (the test set). To train the network, we calculated correlation coefficients between variables (genes) and survival times in the training set, with the variables for which $|r|>0.4$ being used as the inputs for network training. All networks were trained using commercial software (STATISTICA version 8.0). During the supervised training stage, a data set was presented to the ANN along with the correct outputs. The ANN was trained by first randomly initializing the connection weights between the neurons, and then the data were run through the network. Finally, the generated output was compared with the known survival time. The process was repeated, and the network altered the connections weights between neurons until the errors between the generated and real outputs became negligible, at which point the ANN could be used for prediction. Because there is no well-established theoretical method for designing an ideal ANN [9], and the optimal numbers of hidden nodes and iterations are unknown, the best designs are typically determined through trial and error [10]. An optimal network was determined in the present study by constructing and training different ANN architectures comprising 5–30 hidden neurons using the training set. The numbers of iterations and hidden neurons were limited due to the learning algorithm of an ANN being able to overfit the training examples, which would decrease the generalization accuracy. All models were tested with the testing set to determine their accuracy in predicting the survival time. The network with the highest classification accuracy was used in subsequent experiments.

RESULTS

DLBCL. In this experiment there were too many variables (genes) matching the originally chosen criterion (r>0.4), and hence the criterion r>0.5 was applied to choose the model variables. This criterion resulted in an ANN with the following 16 genes as input variables being selected from the data on an Affymetrix Hu6800 microarray chip: D90084_at, D63879_at, U23803_at, HG1879-HT1919_at, D89077_at, M73047_at, U13695_at, M99701_at, S69232_at, U41815_at, D15050_at, X77366_at, L37882_at, X98001_at, U00238_rna1_at, and U68488_at. The training and test results for DLBCL are shown in Figure 3A and 3B, respectively. The optimal ANN architecture MLP16-15-1 was found to be a standard feedforward, fully connected, back-propagation multilayer perceptron. The root mean square error (RMSE) between observed values and the ANN training set was 2.89, and the correlation coefficient was 0.986. The RMSE between observed values and the ANN test set was 2.68, and the correlation coefficient was 0.956.

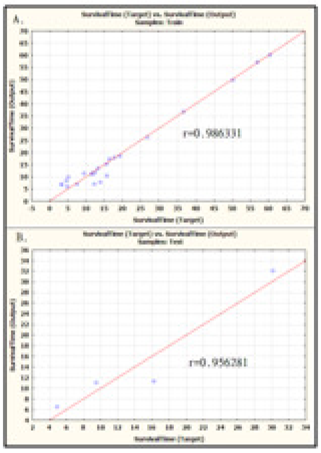

FIG. 3. DLBCL survival time (in months): observation results vs. prediction results for the training (A) and test (B) sets.

FL. An ANN with the following 30 genes as input variables was selected from the data on an Affymetrix U133AB microarray chip using the criterion r>0.4: 242131_at, 242895_x_a, 206854_s_a, 243293_at, 242980_at, 222923_s_a, 215095_at, 236348_at, 236775_s_a, 202979_s_a, 225981_at, 235047_x_a, 212177_at, 201083_s_a, 204732_s_a, 200045_at, 222789_at, 203566_s_a, 214048_at, 243054_at, 240295_at, 224052_at, 217929_s_a, 212381_at, 203970_s_a, 220482_s_a, 221045_s_a, 232932_at, 229086_at, and 243835_at. The training and test results for FL are shown in Figure 4A and 4B, respectively. The optimal ANN architecture MLP30-28-1 was found to be a standard feedforward, fully connected, back-propagation multilayer perceptron. The RMSE between observed values and the ANN training set was 23.61, and the correlation coefficient was 0.886. The RMSE between observed values and the ANN test set was 27.69, and the correlation coefficient was 0.771.

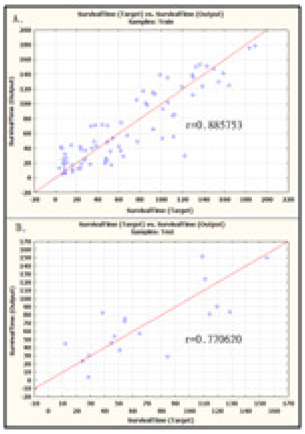

FIG. 4. FL survival time (in months): observation results vs. prediction results for the training (A) and test (B) sets.

Ovarian cancer. An ANN with the following 20 genes as input variables was selected from the data on an Affymetrix U133A microarray chip using the criterion r>0.4: 202322_s_at, 213270_at, 201455_s_at, 204777_s_at, 212483_at, 202350_s_at, 213396_s_at, 211622_s_at, 209251_x_at, 202923_s_at, 209654_at, 205679_x_at, 213646_x_at, 212639_x_at, 211481_at, 213019_at, 200047_s_at, 213976_at, 202314_at, and 204726_at. The training and test results for ovarian cancer are shown in Figure 5A and 5B, respectively. The optimal ANN architecture MLP20-8-1 was found to be a standard feedforward, fully connected, back-propagation multilayer perceptron. The RMSE between observed values and the ANN training set was 5.10, and the correlation coefficient was 0.988. The RMSE between observed values and the ANN test set was 17.23, and the correlation coefficient was 0.868.

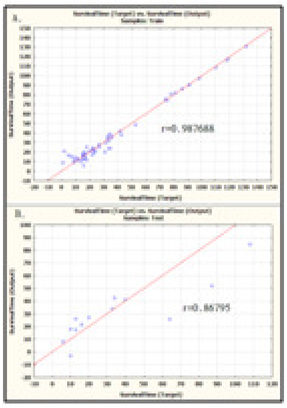

FIG. 5. Ovarian cancer survival time (in months): observation results vs. prediction results for the training (A) and test (B) sets.

DISCUSSION

In the above experiments, the results of which are summarized in Table 1, we observed that selecting 15–30 genes as ANN variables allowed training into good networks. When more than 30 or less than 15 variables were used to train the network, the correlation coefficient was lower than 0.7. The ANN prediction results of the data sets displayed some aberrations, but these might have been attributable to the use of different patient treatment methods.

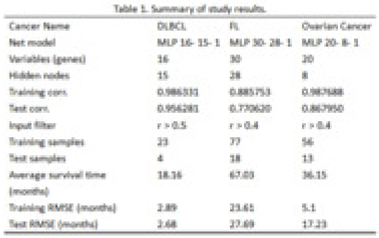

Table 1. Summary of study results.

Cancer Name	DLBCL	FL	Ovarian Cancer
Net model	MLP 16- 15- 1	MLP 30- 28- 1	MLP 20- 8- 1
Variables (genes)	16	30	20
Hidden nodes	15	28	8
Training corr.	0.986531	0.885753	0.987688
Test corr.	0.956281	0.770620	0.867950
Input filter	r > 0.5	r > 0.4	r > 0.4
Training samples	23	77	56
Test samples	4	18	13
Average survival time (months)	18.16	67.03	36.15
Training RMSE (months)	2.89	23.61	5.1
Test RMSE (months)	2.68	27.69	17.23

TABLE 1. Summary of study results.

For the DLCBL data set, when only using 16 genes as variables the survival time predicted by the ANN was strongly correlated with the observed survival time. In addition, genes D63879_at (KIAA0156), HG1879-HT1919_at (ARHQ), U41815_at (NUP98), and X77366_at (TCF11) were cancer-related genes that are reported in OMIM [11] databases. The ANN prediction result for this data set was therefore considered to be good.

For the FL data set, when using 30 genes as variables the survival time predicted by the ANN was correlated with the observed survival time. Moreover, genes 206854_s_at (MAP3K7), 225981_at (DMC1), 235047_x_at (BTBD14B), and 214048_at (MBD4) were cancer-related genes that are reported in OMIM [11] databases. The ANN prediction result for this data set was considered acceptable.

For the ovarian cancer data set, when using 20 genes as variables the survival time predicted by the ANN was correlated with the observed survival time. Moreover, genes 213270_at (MPP2), 200047_s_at (YY1), 213976_at (CIZ1), and 204726_at (CDH13) were cancer-related genes that are reported in OMIM [11] databases. The ANN prediction result for this data set was considered good.

Microarrays are not yet routinely applied not in the diagnosis of clinical patients, and hence the present study was limited to gathering sufficient public-domain data to build and validate the prediction models.

CONCLUSIONS

In conclusion, we have developed ANNs that yielded higher prediction accuracies for survival times using cancer microarray data. It is evident that information related to gene expression levels may have played an important role in cancer prognosis assessment.

FOOTNOTES

* Corresponding author. Mailing address: Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, 250 Wu-Xin St., Taipei, 11031 Taiwan. Phone: 886 2 2736 1661 ext 3347. Fax: 886 2 2739 2914. E-mail: hwchiu@tmu.edu.tw

REFERENCES

1. Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. New Engl J Med 2004; 351: 2159–2169. Available form: http://llmpp.nih.gov/
2. Rosenwald A, et al. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell 2003; 3: 185–197.
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403: 503–511.
4. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002; 8: 68–74. Available from: http://www.genome.wi.mit.edu/science/data
5. Dressman HK, Berchuck A, Chan G, Zhai J, Bild A, Sayer R, et al. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. J Clin Oncol 2007; 25: 517–525. Available from: http://data.cgt.duke.edu/

6. McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics 2002; 18: 1462–1469.
7. Armitage JO, Weisenburger DD. New approach to classifying non-Hodgkin’s lymphomas: clinical features of the major histologic subtypes. Non-Hodgkin's Lymphoma Classification Project. J Clin Oncol 1998; 16: 2780–2795.
8. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001; 7: 673–679.
9. Miller AS, Blott BH, Hames TK. Review of neural network applications in medical imaging and signal processing. Med Biol Eng Comput 1992; 30: 449–464.
10. Penny W, Frost D. Neural networks in clinical medicine. Med Decis Making 1996; 16: 386–398.
11. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. Nucleic Acids Res 2002; 30: 52–55.

Ming Chuan-Health Tech Jounal, 2010, Vol. 1, No. 1, e7