Journal of Biomedical Informatics 56 (2015) 1–7

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Cancer adjuvant chemotherapy strategic classification by artificial neural network with gene expression data: An example for non-small cell lung cancer

Yen-Chen Chen, Yo-Cheng Chang, Wan-Chi Ke, Hung-Wen Chiu *

Graduate Institute of Biomedical Informatics, Taipei Medical University, 250 Wu-Hsing Street, Taipei City, Taiwan

ARTICLE INFO

Article history:
Received 25 September 2014
Revised 2 April 2015
Accepted 11 May 2015
Available online 18 May 2015

Keywords:
Microarray
Gene expression
Neural network
Machine learning
Survival analysis
Adjuvant chemotherapy
Outcome prediction
Lung cancer

ABSTRACT

Purpose: Adjuvant chemotherapy (ACT) is used after surgery to prevent recurrence or metastases. However, ACT for non-small cell lung cancer (NSCLC) is still controversial. This study aimed to develop prediction models to distinguish who is suitable for ACT (ACT-benefit) and who should avoid ACT (ACT-futile) in NSCLC.

Methods: We identified the ACT correlated gene signatures and performed several types of ANN algorithms to construct the optimal ANN architecture for ACT benefit classification. Reliability was assessed by cross-data set validation.

Results: We obtained 2 probes (2 genes) with T-stage clinical data combination can get good prediction result. These genes included 208893_s_at (DUSP6) and 204891_s_at (LCK). The 10-fold cross validation classification accuracy was 65.71%. The best result of ANN models is MLP14-8-2 with logistic activation function.

Conclusions: Using gene signature profiles to predict ACT benefit in NSCLC is feasible. The key to this analysis was identifying the pertinent genes and classification. This study maybe helps reduce the ineffective medical practices to avoid the waste of medical resources.

1. Introduction

Non-small cell lung cancer (NSCLC) is the leading cause of cancer deaths in the worldwide [1]. Complete surgical resection with adjuvant chemotherapy (ACT) is the most widely used treatment for NSCLC. Adjuvant chemotherapy is used after surgery to prevent recurrence or metastases. According to guidelines from medical society, cisplatinum based ACT is now recommend a standard treatment in NSCLC. Several large randomized studies have tried to understand the benefit of ACT in NSCLC [2–5]. However, adjuvant chemotherapy for non-small cell lung cancer is still controversial [6]. Especially in the treatment of early-stage (IB) NSCLC is unclear in clinical oncology. Furthermore, several studies report opposite outcomes of adjuvant chemotherapy in NSCLC [4,7]. Because adjuvant chemotherapy has considerable toxicity, each patient needs a prudent assessment of the risks before ACT treatment. Recently, several studies have to investigate the survival benefit of adjuvant chemotherapy in NSCLC with other impact factors such as age, sex, stage and gene signature [8–13]. However, there is no proper to assess adjuvant chemotherapy benefit. In this study, we aim to develop prediction models to distinguish who is suitable for adjuvant chemotherapy and who should avoid adjuvant chemotherapy in NSCLC.

Many studies described to develop classifiers for detection or diagnosis of disease by using machine learning techniques. The artificial neural network (ANN) is a kind of machine learning methods [14]. It works as humans apply knowledge gained from past experience to new problems. ANN takes previously solved examples to build a system of "neurons" that makes new decisions, classifications, and forecasts. The basic ANN receives many inputs (it can be from original data, or from the output of other ANN). Each input is connected to neurons with different weights. The output of the neuron is produced by the activation function. In ANN research field, multi-layer perceptron (MLP) and radial basis function (RBF) are commonly used ANN model types, which have a strong classification ability. It can solve very complex distribution pattern classification problem, but there are some differences in the structure and function. In which, MLP is the most widely used ANN model type for classification and regression in medical research. MLP consists of three parts, including the input layer, hidden layer and output layer. The input layer neurons accept a large

* Corresponding author. Tel.: +886 2 27361661#3347; fax: +886 2 27392914.
E-mail address: hwchiu@tmu.edu.tw (H.-W. Chiu).

http://dx.doi.org/10.1016/j.jbi.2015.05.006
1532-0464/

number of non-linear input. The output layer is the signal analysis result that weighted, analyzed and transmitted via neurons. The hidden layer is composed of many neurons and links all levels between the input layer and output layer. The hidden layer can have multiple layers which typically used one layer in MLP networks. MLP can distinguish data that are not linearly separable.

2. Methods

To assess the prognostic benefit of various parameters with ACT, we analyzed multiple data sets in lung cancer. Previous studies [8–10] are using Cox Proportional Hazards Model to calculate risk scores of gene signature values with survival time as prognostic risk classification threshold. According to the median risk score, NSCLC patients were divided into low risk and high risk group. Then, they observed the survival benefit of ACT/non-ACT patients in these two groups. However, median risk score is a prediction value and has bias error. According bias error to create predictive models will result in greater deviation. Hence, we used the median survival time rather than the median risk score as the risk classification threshold. The flowchart of our study is shown in Fig. 1.

2.1. Gene expression and clinical data collection

In recent years, a large number of genetic data generated due to advances in high-speed gene expression measurement techniques. Many studies have reported the combination with gene expression data and other high-dimensional genomic data for survival analysis [15–20]. In NSCLC studies, Zhu (2010) has provided a large study comparing the complete clinical data (including: age, race, sex, survival time, adjuvant chemotherapy, adjuvant radiation therapy and stages) [8]. We downloaded NSCLC patients' gene expression raw data (CEL files) and clinical data from NCI caArray database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68465) which is a repository of high-throughput gene expression data and hybridization arrays, chips, and microarrays [18]. These NSCLC data were recorded primarily from 4 institutes and represent 442 NSCLC patients. After deleting duplicate data, survival time missing value and ACT information unknown patients, 280 NSCLC patients were included in this study. Of these, 82 attended the University of Michigan Cancer Center (UM), 72 attended the Moffitt Cancer Center (HLM), 84 attended the Memorial Sloan-Kettering Cancer Center (MSKCC), and 42 attended the Dana-Farber Cancer Institute (DFCI). The clinical data included patients' survival times, age, diagnoses, stages, treatment, and smoking history. All gene-expression profiling was performed using HG-U133A Affymetrix microarray chips. Affymetrix U133A array chip are about 22,000 probe sets corresponding to 14,500 well-characterized human genes. A more detailed description of the clinical data can be found in supplement file I.

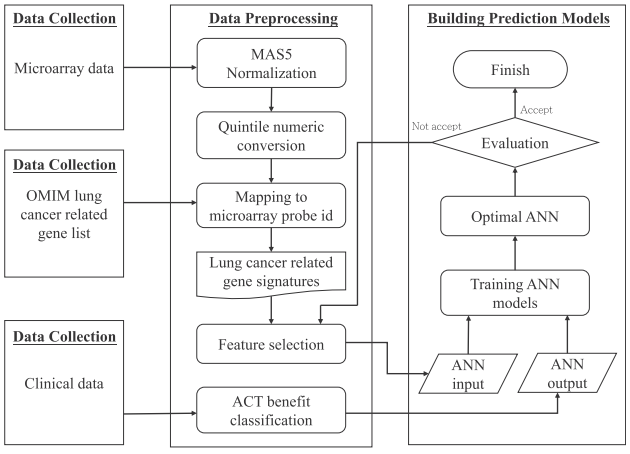

Fig. 1. Flow chart of building artificial neural network prediction model for adjuvant chemotherapy benefit classification in non-small lung cancer.

2.2. Data preprocessing and ACT-Associated Gene Signatures

We collected microarray gene expression data from 4 hospitals, which used the same technology platform, for cross-laboratory data comparisons. Raw data (CEL files) for microarray gene expression profiles were imported into the statistical program R. To avoid the effects from variation in the technology rather than from biological differences between the RNA samples or between the printed probes, we need to do normalization to adjuvant microarray data. The expression variables were normalized and computed using the MAS5 function in the microarray package of R. The MAS5 function includes background adjustment, normalization, and summarization on Affymetrix microarray probe-level data.

Microarray chips contain thousands of gene expression data. This type of high-dimensional data contains many more gene expressions than the number of individuals represented. In addition, the data set contained censored information, which meant that we could not directly establish a gene prediction model. Thus, we needed to reduce the number of variables and find a suitable subset of genes that correlated with survival time as the inputs of a prediction model. The strategy of our approach to filtering the genes included several steps:

2.2.1. Establish lung cancer related gene subset

OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that contain information on all known Mendelian disorders and over 12,000 genes. In our previous study, we collect OMIM database lung cancer-related gene list and mapping to their corresponding microarray probe-id [21]. This can be narrowed down to facilitate the calculation of target genes. The advantage is that the results obtained will easily explain their biological.

2.2.2. Quintile numeric conversion

All gene expression in accordance with its quintile values were converted to class 1–5 (very low, low, normal, high and very high). The purpose is to reduce individual differences in gene expression, and to facilitate the transition to other inspection technologies for future use.

2.2.3. ACT benefit classification

In order to build predictive models to assess which patients appropriate to accept adjuvant chemotherapy, we need to classify all patients into ACT-futile and ACT-benefit groups. Since approximately 50% of these patients did not survive over 40 months, the classification threshold we set at 40 months. All patients can be divided into ACT and observation (OBS) sets according to the treatment method. In ACT set, patients who lived greater than 40 months represent that the patients get significant ACT help. Patients who lived less than 40 months represent that the patients do not get significant ACT help. These patients need to reduce the damage by chemotherapy. In OBS set, patients who lived less than 40 months represent that the patients may need get ACT. Patients who lived greater than 40 months represent that patients have a good prognosis and do not need ACT. As shown in Fig. 2, all patients were divided into two groups according to their survival time and ACT information. Among them, patients who lived less than 40 months with ACT or who lived greater than 40 months without ACT were belonging to the ACT-futile group. And, patients