Computers in Biology and Medicine 48 (2014) 1–7

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Risk classification of cancer survival using ANN with gene expression data from multiple laboratories

Yen-Chen Chen, 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 15 October 2013
Accepted 13 February 2014

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

ABSTRACT

Numerous cancer studies have combined gene expression experiments and clinical survival data to predict the prognosis of patients of specific gene types. However, most results of these studies were data dependent and were not suitable for other data sets. This study performed cross-laboratory validations for the cancer patient data from 4 hospitals. We investigated the feasibility of survival risk predictions using high-throughput gene expression data and clinical data. We analyzed multiple data sets for prognostic applications in lung cancer diagnosis. After building tens of thousands of various ANN architectures using the training data, five survival-time correlated genes were identified from 4 microarray gene expression data sets by examining the correlation between gene signatures and patient survival time. The experimental results showed that gene expression data can be used for valid predictions of cancer patient survival classification with an overall accuracy of 83.0% based on survival time trusted data. The results show the prediction model yielded excellent predictions given that patients in the high-risk group obtained a lower median overall survival compared with low-risk patients (log-rank test *P*-value < 0.00001). This study provides a foundation for further clinical studies and research into other types of cancer. We hope these findings will improve the prognostic methods of cancer patients.

1. Introduction

Lung cancer is the leading cause of cancer death worldwide [1], with approximately 85–90% of all lung cancer cases being non-small cell lung cancer (NSCLC). Currently, complete surgical resection is the most widely used treatment for NSCLC; however, this approach is unsuitable for all patients. Surgical resection for NSCLC is associated with a relatively poor prognosis. Thus, patients must be carefully selected for surgical resection, and alternative treatments should always be considered [2]. Therefore, prognosis prediction plays an important role in cancer patient classification.

The prognosis of patients having specific types of cancer is assessed by survival analysis according to relevant clinical or laboratory data. Survival analysis is commonly used in medical research to analyze time-to-event data. “Survival time”, is the main outcome variable of survival analysis, often used to analyze a sample of a length of time observed such as the beginning to the end of the time from diagnosis to death. The traditional method of deciding which patients were suitable for surgery was to establish a risk score to separate patients into 2 groups: low risk and high risk [3,9–12]. A patient with a high risk score should rather receive replacement therapy or interrupted treatment. Recent cancer survival studies have analyzed gene expression and clinical data [3–8] to attempt to use genetic information to predict the prognosis of patients in which, lung cancer related studies have large high-throughput gene expression data sets with clinical information.

High-throughput gene expression technologies (i.e., microarray and next-generation sequencing) help clarify the relationship between genes and disease. Numerous studies on the use of microarray in cancer classification have shown the effectiveness of this technology. Several studies have conducted large-scale microarray experiments with data from NSCLC patients, and have attempted to use gene signatures to classify patients into subgroups with differing survival outcomes [3,9–12]. However, the results of such studies have been unsatisfactory. Possible reasons for the poor results include the use of small samples, multiple tumor types, and different parameters. Furthermore, the use of various preprocessing steps in microarray experiments causes dissimilar results [13]. No standard method currently exists for the processing or analysis of cross-laboratory data. Thus, a robust data set is necessary for such research. We collected a large microarray data set from NSCLC-related studies, and used common protocols for data analysis to establish prediction models.

Over the past decade, machine-learning technology has been widely used in the analysis of high-throughput gene expression data [14–16]. These studies have included a range of methods

* Corresponding author. Tel.: +886 2 27361661x3347; fax: +886 2 27392914.
E-mail addresses: hwchiu@tmu.edu.tw, chiuxp@gmail.com (H.-W. Chiu).

http://dx.doi.org/10.1016/j.compbiomed.2014.02.006
0010-4825 © 2014 Elsevier Ltd. All rights reserved.

including support vector machine (SVM) [17], genetic algorithms [18,19], *K*-nearest neighbors [20,21], decision trees [3,12,22], artificial neural network (ANN) [23–25], and clustering [11]. The majority of studies have focused on recognizing differentially expressed genes between control and experimental groups. The cluster of genes is identified and then connected to a known functional genomics approach. The present study assessed the prediction of survival among cancer patients, using microarray data and clinical data obtained from multiple laboratories.

2. Methods

We analyzed multiple data sets to assess the prognostic value of various parameters in lung cancer. Previous studies [3,9,12,26] have used median risk score, which is often calculated by using the Cox Proportional Hazards Model, as the risk classification threshold, however it requires very complex calculation and did not apply for ANN training. Hence, we used the median survival time (36 months) rather than the median risk score as the risk classification threshold. The flowchart of our study is shown in Fig. 1.

2.1. Data collection

Numerous studies have reported the use of gene expression data and other high-dimensional genomic data for survival prediction [3,9–12,26]. We downloaded NSCLC patients' gene expression raw data (CEL files) and clinical data downloaded from NCI caArray database (https://array.nci.nih.gov/caarray/project/jacob-00182) which is a repository of high-throughput gene expression data and hybridization arrays, chips, and microarrays. [27]. These NSCLC data were recorded primarily from 4 institutes and represent 440 NSCLC patients. Of these, 177 attended the University of Michigan Cancer Center (UM), 77 attended the Moffitt Cancer Center (HLM), 104 attended the Memorial Sloan-Kettering Cancer Center (MSKCC), and 82 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 chips. The clinical outcome information includes data on age, race, sex, survival time, adjuvant chemotherapy, adjuvant radiation therapy and stages. Cases with survival time missing value are excluded in this study. A more detailed description of the clinical data can be found in Supplementary material I.

2.2. Data preprocessing and survival-associated gene signatures

The microarray data from 4 units used the same technology platform, enabling 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:

Step1. 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. We collect OMIM database lung cancer-related gene list and mapping to their corresponding microarray probe-id. This can be narrowed down to facilitate the calculation of target genes. The advantage is that the results obtained will easily explain the biological meaning of target genes. In addition there are other ways that we use to establish a subset of genes, including the calculation methods such as the correlation coefficient, RankProd and Principal component analysis (PCA). However, the results obtained are not good. By using only the first gene screening strategy mentioned above, we can get the best results.

Step2. Quintile numeric conversion: all the gene expression, in accordance with its converted quintile values of 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.

Step3. Establishment of a subset of patients without chemotherapy, radiation therapy: all patients will be classified according to their treatment, in which no adjuvant chemotherapy and adjuvant radiotherapy group is for feature selection. This is to avoid the effects of the deviation caused misjudgment by different treatment.

Step4. Feature selection: to select which gene signatures to use in building the prediction model, we calculated chi-square test value between variables (genes) and survival time for each data set. We identified the top 10 rank survival correlated gene signatures for each data set, and used the subset gene signatures as the predictive model variables. Calculate the chi square value for gene expression and patient survival time, and select the top 10 rank genes as the machine learning target genes. Clinical data such as sex, age, T_stage and N_stage were also considered as ANN variables.

Step5. Building prediction models: the ten genes which were identified in the previous process, were used to train an ANN-Network.

2.3. Building prediction models

We used the most popular machine-learning method, Artificial Neural Network (ANN), to model NSCLC survival. The ANN method has been shown to improve the accuracy of prediction for cancer survival outcomes [23]. We performed several types of ANN algorithms to identify the optimal ANN architecture. Reliability was assessed by cross-data set validation. The data sets were randomly selected so that one source provided the training set and the others provided the test set. All networks were trained using commercial software (STATISTICA version 8.0). During the supervised training stage, a data set is presented to the ANN with the correct outputs available (risk classification). Risk was classified according to the median overall survival time. Patients could not be divided into high- and low-risk groups directly according to their gene expression values, so patients who lived longer than the median survival time (36 months) were assigned to the low-risk group, and all other patients were assigned to the high-risk group. Thereafter, the ANN was used for prediction. Because no perfect method exists for designing an ideal ANN and the optimal number of hidden nodes and iterations are unknown, the best design is usually determined by trial and error [28]. To identify the optimal model, we constructed and trained tens of thousands of various