journal homepage: www.intl.elsevierhealth.com/journals/cmpb

TC-VGC: A Tumor Classification System using Variations in Genes' Correlation

Eunji Shin[a], *Youngmi Yoon*[b], *Jaegyoon Ahn*[a], *Sanghyun Park*[a,*]

[a] *Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea*
[b] *Division of Information Technology, Gachon University of Medicine and Science, 534-2 Yonsu-dong, Yonsu-gu, Inchon 534-2, South Korea*

ARTICLE INFO

Article history:
Received 24 June 2010
Received in revised form
11 January 2011
Accepted 7 March 2011

Keywords:
Tumor classification
Microarray data analysis
Gene–gene correlation
Cancer specific genetic network

ABSTRACT

Classification analysis of microarray data is widely used to reveal biological features and to diagnose various diseases, including cancers. Most existing approaches improve the performance of learning models by removing most irrelevant and redundant genes from the data. They select the marker genes which are expressed differently in normal and tumor tissues. These techniques ignore the importance of the complex functional-dependencies between genes. In this paper, we propose a new method for cancer classification which uses distinguished variations of gene–gene correlation in two sample groups. The cancer specific genetic network composed of these gene pairs contains many literature-curated prostate cancer genes, and we were successful in identifying new candidate prostate cancer genes inferred by them. Furthermore, this method achieved a high accuracy with a small number of genes in cancer classification.

1. Introduction

DNA microarray technologies make it possible to simultaneously monitor the expression levels of thousands of genes [1]. The large amount of data generated by microarray experiments has stimulated the development of many computational methods to study different biological processes at the gene expression level. These microarray techniques are expected to result in precise cancer detection and classification.

The major difficulty of microarray data analysis is the large number of genes compared to the limited number of samples in a typical experiment. Furthermore, many of the genes are 'noise' genes that are not relevant in differentiating between normal and tumor samples. One of the major challenges in designing an accurate classifier using microarray data is identifying the optimal subset of relevant genes. This is known as gene selection and corresponds to feature selection in the field of pattern classification.

Existing gene selection methods just select the marker genes which are differentially expressed in normal and tumor tissues. They do not consider the gene–gene correlations between two groups of samples. However, variations in gene–gene correlations can be a good guide for making a cancer diagnosis. For example, when a transcription factor activates or represses two genes, A and B, simultaneously, the expression levels of A and B reveal that they have high correlation. If gene B is affected by a specific disease, the transcription factor continues to activate or repress gene A, while it can no longer influence gene B.

Most of microarray studies often focus on the identification of differentially expressed genes or the construction of prediction rule. However, gene itself which does not appear to be oncogenic could be highly relevant to be cancer specific when they are considered with others. In this paper,

* *Corresponding author*. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park).
0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.03.002

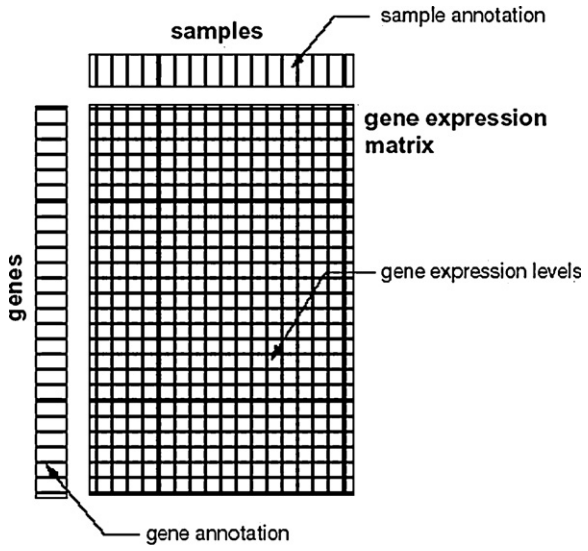

Fig. 1 – Microarray data set.

we are proposing the TC-VGC (Tumor Classification using Variations in Genes' Correlation) that considers complex functional-dependencies between genes using the correlation coefficients of gene pairs. This is a new cancer classification method which uses substantially smaller number of genes and retains a high predictive accuracy. Given a new patient's sample, the method predicts the class of the sample using variations in gene correlations, with high accuracy. Furthermore, the cancer specific genetic network composed of these gene pairs contains many literature-curated prostate cancer genes, and we were successful in identifying new candidate prostate cancer genes inferred by them.

2. Related works

2.1. *Microarray data set*

The gene expression data which is generated from microarrays is organized as matrices. Rows represent genes and columns represent various samples such as normal or tumor tissues. Values in each cell represent the expression level of the particular gene in the sample. Fig. 1 shows an example of a gene expression matrix [2].

2.2. *Existing cancer classification algorithms based on microarray*

Machine-learning methods are used to classify and cluster data. These methods are especially useful in cancer diagnosis and detection [3–5]. The current trend is to apply a computational approach to traditional biological research, which is then used to understand biological processes.

Many researchers have applied a machine-learning approach to microarray data analysis. Examples of these techniques include the SVM (Support Vector Machine), k-NN (K-Nearest Neighbors), Random Forest, and Bayesian networks [6]. Pirooznia et al. recently introduced the commonly used classification methods, and applied these methods to publicly available datasets [7]. Results revealed that SVM classification and RBF Neural Nets had the best accuracy. Wang et al. demonstrated that feature subset selection algorithms, namely wrappers, filters and CFS (Correlation-based Feature Selection), can be very useful in extracting relevant information in microarray data analysis [8]. They exhibited that the filters and CFS are recommended for fast analysis of data. However, for better classification accuracy and fewer genes that could be further used for a cancer diagnosis toolkit, the wrapper approaches are more recommended.

SVM is becoming increasingly popular classifiers for many data, including microarrays. Guyon et al. proposed an SVM-RFE (Support Vector Machine Recursive Feature Elimination) algorithm to recursively classify the samples using SVM and to select genes according to their weights in the SVM classifiers [9]. Duan et al. proposed MSVM-RFE [10]. This technique trains multiple linear SVMs on subsamples of training data and computes the feature ranking score of SVM-RFE. The performance of MSVM-RFE is better than that of SVM-RFE and fewer genes were needed in MSVM-RFE than in SVM-RFE.

Pan et al. proposed a comprehensive KNN/LSVM classification approach [11]. This technique used a combination of a k-NN majority voting approach and a local Support Vector Machine approach which makes optimal decisions at the local level. The goal of this technique is to classify cases based on a local approach without the time burden which is usually necessary to run traditional algorithms. Diaz-Uriarte and Alvarez de Andres investigated the use of a Random Forest for classification of microarray data including multi-class problems [12]. Random Forest is a classification algorithm that is well suited for microarray data even when most predictive variables are noise and the amount of input data is large.

The main challenge in classifying gene expression data is to solve the curse of dimensionality problem. In addition, irrelevant and redundant features increase the search space and make patterns more difficult to detect and make it more difficult to capture rules necessary for prediction or classification. Albrecht has proven that selecting the best feature subset is an NP-complete problem [13]. To overcome these problems, feature selection is an indispensable task in classification to identify a smaller subset of relevant genes for building robust learning models. Gheyas and Smith proposed a hybrid method for feature subset selection consisting of a combination of simulated annealing and a genetic algorithm and compare its performance to a variety of other greedy and stochastic search algorithms [14]. Jaeger et al. reduced the number of relevant genes by eliminating highly correlated ones [15]. Kim et al. attempted to use several methods for extracting informative features and combining classifiers learned from the negatively or complementarily correlated features [68]. Berretta et al. suggests feature set model to extract differentially expressed genes [16]. The genes that must be in a different state in any pair of samples with different labels and in the same state in any pair of samples with the same label are chosen to be a feature set. Dougherty and Brun considered optimal feature sets in the framework of a model in which the features are grouped in such a way that intra-group correlation is substantial whereas inter-group correlation is minimal [17].

Principle components analysis (PCA), a mathematical algorithm that reduces the dimensionality of the data while