Computer Methods and Programs in Biomedicine 141 (2017) 27–34

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Cancer subtype prediction from a pathway-level perspective by using a support vector machine based on integrated gene expression and protein network

Fei-Hung Hung, Hung-Wen Chiu*

Graduate Institute of Biomedical Informatics, Taipei Medical University, 250 Wu-Hsing Street, Taipei 11031, Taiwan

ARTICLE INFO

Article history:
Received 9 July 2016
Revised 6 January 2017
Accepted 16 January 2017

Keywords:
Cancer subtype
Protein–protein interaction
Gene expression
Signaling pathway
Neuroepithelial tumor
Computational method

ABSTRACT

Background and objective: Distinguishing cancer subtypes is critical for selecting the appropriate treatment strategy. Bioinformatics approaches have gradually taken the place of clinical observations and pathological experiments. However, these approaches are typically only used in gene expression profiling. Previous studies have primarily focused on the gene level or specific diseases, and thus pathway-level factors have not been considered. Therefore, a computational method that integrates gene expression and pathway is necessary.

Methods: This study presented an approach to determine potential fragments of activated pathways around protein networks in different stages of disease. We used a scored equation that integrates genomic and proteomic information and determined the intensity of the pathway link change. A support vector machine (SVM) was used to train and test subtype-predicted models.

Results: The performance of the proposed method was evaluated by calculating prediction accuracy. The average prediction accuracy was 67.64% for three subtypes in tumors of neuroepithelial tissues. The results demonstrate that the proposed method applies fewer features than gene expression methods used to obtain similar results

Conclusions: This study suggests a method to implement a cancer subtype classifier based on an SVM from a pathway-level perspective.

1. Introduction

Gene expressions differ in different diseases, and cells affected by the same disease can have different gene expressions. Furthermore, distinguishing different disease types at a single location is difficult. Previous studies have reported distinguishing different disease types through clinical observations and by determining differences through pathological experiments [1,2]. For example, brain tumors were classified according to the location of the tumor in the brain, type of tissues involved, and the original location of the tumor. However, with advancements in bioinformatics, gene expressions can be determined and disease subtypes can be classified according to accumulated gene expression data. Although some disease information can be derived from gene expressions, data related to pathways underlying cellular mechanisms have not been analyzed extensively. Currently, machine learning methods with gene expressions are primarily used to classify disease subtypes. A signal transduction pathway is a respondent for an extracellular activity and is important for disease subtypes.

In an era of rapid accumulation of genomic and proteomic information, bioinformatics has greatly benefited from advances in computer science and biology laboratory techniques. For example, the Gene Expression Omnibus (GEO) database is one of several public genomic data repositories [3]. Researchers can easily access these online resources through a web-based interface.

MacQueen proposed the k-means algorithm in 1967 [4]. This algorithm is widely used for cluster analysis in data mining. The algorithm first divides observations into k groups. The center of each group is then calculated, and observations are allotted to the nearest center. These steps are repeated until the members and the centers of all groups are stable. Shai et al. applied the k-means algorithm and discovered three clusters: glioblastomas, low-grade astrocytomas, and oligodendrogliomas [5]; they showed that a relatively small number of genes can be used to separate molecular subtypes. Similarly, Tothill et al. used the k-means algorithm to identify six molecular subtypes of serous and endometrioid tumors of the ovary, peritoneum, and fallopian tubes [6].

* Corresponding author.
E-mail address: hwchiu@tmu.edu.tw (H.-W. Chiu).

http://dx.doi.org/10.1016/j.cmpb.2017.01.006
0169-2607/

Fuzzy c-means was developed from c-means by Dunn in 1973 and improved by Bezdek in 1981. In contrast to c-means, in fuzzy c-means, the membership degree of a group is quantified in the 0–1 range, and the result is presented as a probability. Dembélé and Kastner used fuzzy c-means to attribute cluster membership values to yeast genes [7]. Zainuddin proposed an enhanced fuzzy c-means clustering algorithm for heterogeneous cancer classification using four microarray benchmark data sets [8].

Graph theory refers to the study of graphs wherein relationships among objects in graphs are modeled as mathematical structures. Song et al. used graph theory to differentiate meaningful clusters and hierarchies in various real lymphoma data sets [9]. They reported that gene expression patterns of lymphoma samples revealed biologically significant groups of genes in lymphoid malignancies.

An artificial neural network (ANN) is a computational system that simulates neural processes by simulating neuron links. In an ANN, known inputs and outputs are used to facilitate self learning for a model between the inputs and outputs. Selaru et al. developed an ANN using five inflammatory bowel disease-associated neoplasms (IBDN) and 22 sporadic colorectal adenomas (SAC). They correctly diagnosed all 12 examined blinded samples in a test set comprising three IBDNs and nine SACs [10].

Kanth et al. combined clustering and classification to differentiate cancer tissues. The accuracy of their classification of acute myeloid and acute lymphoblastic leukemia was 94.12% [11].

A support vector machine (SVM) is supervised learning model typically used in machine learning. SVMs can be applied to both classification and regression. Cortes and Vapnik proposed the SVM in 1995 [12]. The main function of an SVM is to search for the maximum marginal hyperplane.

Some of the aforementioned methods use only gene expressions or protein–protein interaction (PPI), which does not sufficiently represent the entire pathway mechanism. Furthermore, some methods are limited to specific diseases. An approach that considers both gene- and pathway-level data is required because such a method can approximate the true situation of different disease subtypes.

2. Materials and methods

2.1. Gene expression data

We used gene expression profiles of gliomas, GDS1962, from the GEO (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1962) to compute pathway change. This data set was generated and used to determine the effects of glioma-derived stem cell factor on angiogenesis in the brain by Sun et al. [13] It was used to analyze gliomas of different grades. Data were produced using Affymetrix gene chips. The data set contains 180 patients and 30,744 genes. GDS1962 includes 23 non-tumor, 26 astrocytoma, 81 glioblastoma, and 50 oligodendroglioma samples. In total, 157 patients with cancer were included.

Astrocytomas are a type of brain cancer that originates from star-shaped brain cells (astrocytes). This tumor type does not typically spread to other organs. Astrocytomas are the most common gliomas and primarily affect the brain and sometimes the spinal cord.

Glioblastoma (glioblastoma multiforme) is the most invasive brain cancer. Its symptom aggravation is rapid and causes unconsciousness in patients.

Oligodendrogliomas are a type of glioma that generate at the oligodendrocytes or a glial precursor cell and primarilly affect adults.

The samples from the four aforementioned groups are listed in Supplementary 1. These samples were grouped into three categories: healthy versus astrocytomas, healthy versus glioblastomas, and healthy versus oligodendrogliomas.

2.2. Protein–protein interaction data

To construct a PPI network of homosapiens, the PPI data were downloaded from the Interologous Interaction Database (I2D) [14]. I2D integrates PPI data from various sources. In this study, data from BioGrid [15], DIP [16], the Human Protein Reference Database [17], InnateDB [18], IntAct [19], and the Malaysian Institute for Nuclear Technology Research [20] were retained and other experimental and predicted data were eliminated. Approximately 160,000 non-duplicate interactions were storedin a local MySQL database.

2.3. Equations

To simulate the pathway fragment between normal and developed statuses, a method to calculate the change in each link between two statuses is required.

The gene expression change score (*GCS*) and edge change score (*ECS*) equations were used to obtain activated subnetworks between different statuses [21]. The *GCS* calculates the intensity of gene expression change in each gene or point. The *ECS* was used to quantify the intensity of change in the link or edge between the original and developed statuses.

Student's *t*-test was used for the normal against the developed statuses of each patient. Here a one-sample *t*-test was used because each patient's disease status was relative to the mean of all normal statuses. $p < 0.00001$ of the change in gene expression between the normal and developed statuses was considered significant.

The *GCS* of whole genes were then calculated. Next, *GCS*s were used as inputs to the *ECS* equation.

2.4. Support vector machine

SVMs are typically used for supervised learning and are primarily applied to classification and regression. SVMs are based on hyper planes that maximize the separating margin between several labeled classes. The most widely used SVM program uses LIBSVM by Chang and Lin [22]. LIBSVM is a simple, easy-to-use, and efficient tool.

2.4.1. Data preprocessing

After each edge's *ECS* is calculated to prepare the input matrix, a score is used as the threshold of activated edges or links. Statistical comparisons among features with different scores were created (Fig. 1). The feature number of score 30 was the largest and was therefore used as the threshold. Whole edges or links with a score greater than or equal to 30 were set to 1 and whole edges or links with a score less than 30 were set to 0. Here, three sparse matrices for three subtypes were produced. Each 0 or 1 suggests the feature of one edge or link. For input to LIBSVM, the three subtype spares were merged to a complete sparse matrix. Three subtype sparse matrices had some similar and different features. When merging, the different features among the three subtype sparse matrices were set to 0. This complete sparse matrix (157×1586, Table 1 for partial matrices and Supplementary 1 for complete matrices) could reveal the activated and mute pathway edges.

Here, group 1 was "from normal to astrocytomas," group 2 was "from normal to glioblastoma," and group 3 was "from normal to oligodendrogliomas." The weight parameters of groups 1, 2, and 3 were set to 6, 2, and 3, respectively. The LIBSVM subset.py tool randomly generated 100 training and testing sets. The patients