Chapter 4
A Novel Deep Learning Method for Identification of Cancer Genes From Gene Expression Dataset

Pyingkodi Maran

https://orcid.org/0000-0002-5247-1870

Kongu Engineering College, India & Anna University, India

Shanthi S.

Kongu Engineering College, India

Thenmozhi K.

https://orcid.org/0000-0003-0934-2552

Selvam College of Technology, India

Hemalatha D.

Kongu Engineering College, India

Nanthini K.

Kongu Engineering College, India

ABSTRACT

Computational biology is the research area that contributes to the analysis of biological information. The selection of the subset of cancer-related genes is one amongst the foremost promising clinical research of gene expression data. Since a gene can take the role of various biological pathways that in turn can be active only under specific experimental conditions, the stacked denoising auto-encoder(SDAE) and the genetic algorithm were combined to perform biclustering of cancer genes from huge dimensional microarray gene expression data. The Genetic-SDAE proved superior to recently proposed biclustering methods and better to determine the maximum similarity of a set of biclusters of gene expression data with lower MSR and higher gene variance. This work also assesses the results with respect to the discovered genes and spot that the extracted set of biclusters are supported by biological evidence, such as enrichment of gene functions and biological processes.

INTRODUCTION

Bioinformatics is a multidisciplinary subject, related to area as diverse as Computer Science, Mathematics, Biology, Statistics and Information Technology. Cancer is featured by an irregular, unmanageable growth that may destroy and cause the neighboring healthy body tissues. In the past, cancer classification by medical practitioners and radiologists was based on clinical and morphological features and had limited diagnostic ability. It deals with different kinds of biological data. The dimension and complexity of raw gene expression data creates challenging data analysis and data management problems. The fundamental goal of microarray gene expression data analysis is to find the behaviourial patterns of genes.

Computational molecular biology deals with different kinds of biological data. Gene expression data is one among them. Hence Gene expression data are the basic data used in this paper. Gene expression is the process by which the information encoded in a gene is changed into an observable phenotype (protein). It is the degree to which a gene is active in certain tissues of the body, measured by the amount of Messenger Ribonucleic Acid (mRNA) in the tissue. Individual genes can be switched on (apply their effects) or switched off according to the needs and situations of the cell at a particular instance. Thus, abnormalities or deviations of gene expression may result in the death of cells, or their uncontrolled growth, such as cancer (Subramanian 2010).

Gene Expression Data

The gene expression matrix is a processed data obtained after the normalization. Each row in the matrix corresponds to a particular gene and each column could either correspond to an experimental condition or a specific time point at which expression of the genes has been measured (Tiwari *et al.* 2012). The expression level for a gene across different experimental conditions is cumulatively called the gene expression profile, and the expression level of each gene under an experimental condition is cumulatively called the sample expression profile (Androulakis *et al.* 2007). An expression profile of an experimental condition or a gene is thought of as a vector and can be represented in vector space. For example, an expression profile of a gene can be considered as a vector in n dimensional space where n is the number of conditions, and an expression profile of a condition with m genes can be considered as a vector in m dimensional space where m is the number of genes. Figure 1 shows the gene expression matrix A with m genes across n conditions is considered to be an $m \times n$ matrix. Each element a_{ij} of this matrix represents the expression level of a gene i under a specific condition j, and is represented by a real number.

SIGNIFICANCE OF CANCER DIAGNOSIS USING MICROARRAY

Current cancer classification includes more than 200 types of cancer (American Cancer Society). For any patient to receive proper therapy, the clinician must identify as accurately as possible the cancer type. Although analysis of morphologic characteristics of biopsy specimens is still the standard diagnostic method, it gives very limited information and clearly misses much important cancer aspects such as rate of proliferation, capacity for invasion and metastases, and development of resistance mechanisms to certain treatment agents. To appropriately classify cancer subtypes, therefore, molecular diagnostic methods are needed. The classical molecular methods look for the DNA, RNA or protein of a defined marker that is correlated with a specific type of tumor and may or may not give biological information

14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/novel-deep-learning-method-identification/342522

Related Content