Applied Soft Computing 62 (2018) 230–250

Contents lists available at ScienceDirect

Applied Soft Computing

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

Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach

Prativa Agarwalla [a,*], Sumitra Mukhopadhyay [b,*]

[a] Heritage Institute of Technology, Kolkata, India
[b] Institute of Radiophysics and Electronics, Kolkata, India

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form 13 August 2017
Accepted 11 October 2017

Keywords:
DNA microarray
Biological pathway
Feature selection
Particle swarm optimization (PSO)
Artificial bee colony (ABC)

ABSTRACT

Background: Understanding of molecular mechanism, lying beneath the carcinogenic expression, is very essential for early and accurate detection of the disease. It predicts various types of irregularities and results in effective drug selection for the treatment. Pathway information plays an important role in mapping of genotype information to phenotype parameters. It helps to find co-regulated gene groups whose collective expression is strongly associated with the cancer development.
Method: In this paper, we have proposed a bi-stage hierarchical swarm based gene selection technique which combines two methods, proposed in this paper for the first time. First one is a multi-fitness discrete particle swarm optimization (MFDPSO) based feature selection procedure, having multiple fitness functions. This technique uses multi-filtering based gene selection procedure. On top of it, a new blended Laplacian artificial bee colony algorithm (BLABC) is proposed and it is used for automatic clustering of the selected genes obtained from the first procedure. We have performed 10 times 10-fold cross validation and compared our proposed method with various statistical and swarm based gene selection techniques for different popular cancer datasets.
Result: Experimental results show that the proposed method as a whole performs significantly well. The MFDPSO based system in combination with BLABC generates a good subset of pathway markers which provides more effective insight into the gene-disease association with high accuracy and reliability.

1. Introduction

The fundamental cause of cancer [1] involves abnormal growth of cells and the underlying factor is the certain changes at the genetic expression level. Gene controls the functioning of a living cell. When the normal expression profile of gene changes, it causes some abnormalities. However, a series of pathway activities related to biological process collectively leads to the growth of cancer. So it becomes very essential to analyze the molecular mechanism and genomic expression profile of cancer. The efficient mapping of genomic information to the phenotype parameter can help in better understanding in the progression and early detection of the disease. Moreover, analysis about the changes in gene expression helps in proper cancer identification and class prediction that is necessary for the drug selection in the course of treatment. The variation in the expression profile of genes can be visualized by the recent advancement of microarray technology [2] which is nothing but the expression level of thousands of genes in a single chip. To prepare the gene expression data, samples are collected from patients having different classes of disease and then through the hybridization procedure the changes in expression level are examined. The schematic diagram of generating microarray gene expression is shown in Fig. 1 where the dataset is formed for two classes of disease. Now, to identify differentially expressed genes for different classes and to study their effects on diseases, we need to investigate gene expression dataset. This leads to statistical and analytical challenges because of its huge dimension. Again, the availability of larger number of genes compared to the number of samples in the dataset can cause the overfitting of classification model and the genetic heterogeneity across patients weakens the discriminating power of individual gene [3]. Also the presence of noisy genes makes it very challenging to understand the nature of gene regulation in the cancer samples compared to the normal cell. The removal of uninformative genes not only reduces the processing time but also diminishes the interference of noisy or unwanted information leading to the incorrect classification of data. The procedure of relevant gene (feature) selection from the gene expression profile and the expected outcome is described by a schematic diagram, given

* Corresponding authors.
E-mail address: sumitra.mu@gmail.com (S. Mukhopadhyay).

https://doi.org/10.1016/j.asoc.2017.10.024
1568-4946/

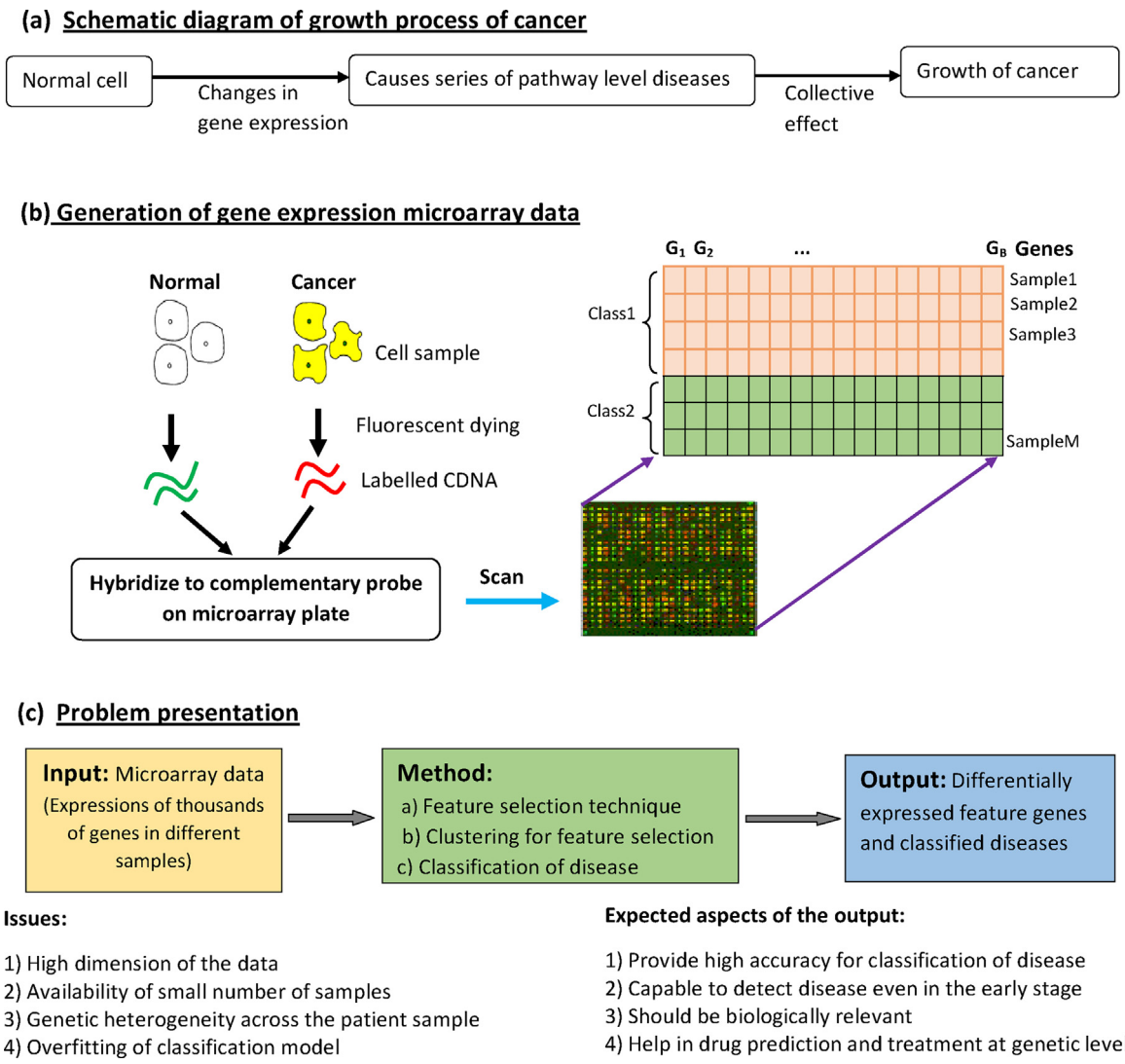

Fig. 1. General problem presentation of gene marker selection for cancer growth.

in Fig. 1. To solve the above stated issues, different methodologies are approached in different literatures [4–11] for the feature selection and classification of cancer, but the reproducibility of results is a challenging task as in most of the cases the resultant genes are varying. Also, the selected genes which are providing good classification result may not always be biologically relevant to cancer progression. So, the genes used for cancer prediction can lead to a false discovery of disease which could be vital for the patient. It would be better to enrich with additional biological knowledge rather concentrating only on high classification accuracy.

While exploring the features of significant non-redundant genes participating in a tumor progression, it has been observed that those genes are functioning in a co-regulated manner and work as a group. They confer a selective growth advantage during the development of certain cancer. So, those differentially expressed genes having a similar biological contribution to the progression of tumor are to be identified and their phenotypical changes at the pathway level [12,1] for all the patients are to be studied for the treatment of prior related pathway diseases. So shifting to pathway based study can help in better understanding of prior disease related information and how a disease occurs altogether. Several publicly available databases are developed for accessing the pathway information and detailed information about the interaction of genes and their regulatory pathways [13,14]. To use those pathway-based marker genes in disease classification, initially we need a way to infer the activity of a given biological pathway on the gene expression and then the differential genes (features) significantly associated with the disease outcomes are to be identified for the efficient classification of samples.

Rather, computing the statistical significance of each and individual genes, we propose an efficient swarm based stochastic method for the selection of pathway marker genes. The marker genes are selected from functional genomic expression data incorporating different parametric and non-parametric statistical techniques that are able to reflect the heterogeneity in the gene expression level. The proposed method consists of two basic stages: first a multi-fitness discrete particle swarm optimization (MFDPSO) technique in combination with multiple statistical filters is used for the initial selection of genes from biomedical databases. It reduces the dimensionality of the microarray data as well as the combined effect of different filters provides a better differential gene subset. Reduced feature set obtained in the first stage is used in the second stage where distinct gene expression levels in each class are identified using a newly proposed blended Laplacian artificial bee colony (BLABC) based automatic inter-class clustering technique. The resultant genes can be used for classification. It solves the problem of overfitting of the classifiers and reduces the false discovery rate of the disease identification.