Chapter 14
Selection of Pathway Markers for Cancer Using Collaborative Binary Multi-Swarm Optimization

Prativa Agarwalla
Heritage Institute of Technology, India

Sumitra Mukhopadhyay
Institute of Radiophysics and Electronics, India

ABSTRACT

Pathway information for cancer detection helps to find co-regulated gene groups whose collective expression is strongly associated with cancer development. In this paper, a collaborative multi-swarm binary particle swarm optimization (MS-BPSO) based gene selection technique is proposed that outperforms to identify the pathway marker genes. We have compared our proposed method with various statistical and pathway based gene selection techniques for different popular cancer datasets as well as a detailed comparative study is illustrated using different meta-heuristic algorithms like binary coded particle swarm optimization (BPSO), binary coded differential evolution (BDE), binary coded artificial bee colony (BABC) and genetic algorithm (GA). Experimental results show that the proposed MS-BPSO based method performs significantly better and the improved multi swarm concept generates a good subset of pathway markers which provides more effective insight to the gene-disease association with high accuracy and reliability.

INTRODUCTION

Genes control the different functioning of a cell, like growth, division, death etc. When the normal profile of a gene is changed or damaged, it causes the abnormal behavior of the cell and we, in generic sense, call it as cancer. Cancer is nothing but out-of-control cell growth due to change in the expression profile of genes. Advancement of microarray technology has made the genomic study more fast and

DOI: 10.4018/978-1-5225-2607-0.ch014

efficient by analysing the expression of thousands of genes in a single chip (Zhang et.al, 2008). But, the huge dimension of the gene expression data leads to statistical and analytical challenges to identify differentially expressed genes in different classes for the study of their effect on diseases. So the selection of the most relevant genes is very essential for the proper medical diagnosis as well as for drug target prediction and in this context, different aspects of big data processing and analysis come into play. Big data analysis is one of the very popular and recent day technologies, used for examining large dataset. It helps to reveal hidden pattern and correlations information of the data which can be used in the field of computational biology to analyze huge biological data for extracting the relevant information and to enrich the knowledge related to the biological system. While exploring small number of significant genes participating in a tumour progression, it has been observed that those genes are functioning similar and work as a group to form a certain cancer. The set of genes having identical biological functioning is known as a pathway. To understand the biological functioning of those groups of genes, involved in tumour progression and the phonotypical changes at the pathway level is a very interesting research topic now a day. But, the dataset contains high amount of noise and overlapping samples which decreases classification accuracy and the identification of dominant genes related to the disease suffers. Proper and efficient identification of those differentially expressed genes at the pathway level is very crucial for the treatment of the prior disease related to the pathway.

Different approaches have been developed by the researchers for finding the pathway marker genes related to different diseases (Khunlertgit, 2013; Ma, 2009; Mandal, 2015; Mukhopadhyay, 2014). The statistical approaches like mean, median are widely applied for this purpose (Guo et.al, 2005). Principal component analysis is used in the literature (Ma et.al, 2009). However, the uses of different statistical tests lack functionality (Mandal et.al, 2015). In another literature, Log-Likelihood Ratio (LLR) is implemented for accurate and reliable classification of cancer through the identification of pathway markers (Su et.al, 2009). Another method named as Condition Responsive Genes (CORGs) is proposed for delivering optimal discriminative power for the disease phenotype (Lee et.al, 2008). A family of bio-inspired algorithms, like particle swarm optimization algorithm (PSO), genetic algorithm (GA), artificial bee colony optimization (ABC), differential evolution (DE), ant colony optimization (ACO) etc. have also been applied by formulating the problem as a global optimization problem (Das et.al, 2008) where the position of a particle is termed as the solution which is associated with some fitness function. Binary version of PSO (Mandal et.al, 2015) has been adopted to find the pathway activity where the mean t-score of the pathways are computed as the metric of evaluation, that is the fitness function has to be maximized (Bandhopadhyay et.al, 2014). Again, Multi-objective particle swarm optimization (MOPSO) has also been designed to select significant pathway features for cancer identification (Mukhopadhyay et.al, 2014). Generally, the performance of meta-heuristic algorithms deteriorates as the dimension of the real world problem increases. So, involving a single group of swarm and using an average fitness value as a solution of the problem may fail to produce more accurate results for the problem of gene selection from the huge microarray data. In case of generalized optimization problems, multi-swarm concept (Niu et.al, 2007) working in a collaborative way produces better result than the conventional single swarm based methods (Eberhart et.al, 1995). In multi-swarm concept, total population is divided into several groups and each group searches for a new promising solution. The interactions between the groups influence the balance between exploration and exploitation and maintain a suitable diversity in the population, reducing the risk of pre-convergence. Therefore, a novel pathway gene selection tech-