Chapter 8
An Ultra-Fast Method for Clustering of Big Genomic Data

Billel Kenidra
National Superior Institute of Computer Science (ESI), Constantine, Algeria

Mohamed Benmohammed
Lire Laboratory, University of Constantine-2, Constantine, Algeria

ABSTRACT

The clustering process is used to identify cancer subtypes based on gene expression and DNA methylation datasets, since cancer subtype information is critically important for understanding tumor heterogeneity, detecting previously unknown clusters of biological samples, which are usually associated with unknown types of cancer will, in turn, gives way to prescribe more effective treatments for patients. This is because cancer has varying subtypes which often respond disparately to the same treatment. While the DNA methylation database is extremely large-scale datasets, running time still remains a major challenge. Actually, traditional clustering algorithms are too slow to handle biological high-dimensional datasets, they usually require large amounts of computational time. The proposed clustering algorithm extraordinarily overcomes all others in terms of running time, it is able to rapidly identify a set of biologically relevant clusters in large-scale DNA methylation datasets, its superiority over the others has been demonstrated regarding its relative speed.

1. BACKGROUND

1.1. Cancer and Bioinformatics

Cancer is a genetic disease, caused by changes in genes that control the way how our cells function, especially how they grow and divide. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place. However, when cancer develops, this ordered process breaks down. As cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not

DOI: 10.4018/979-8-3693-3026-5.ch008

needed. These extra cells begin to divide without stopping and spread into surrounding tissues called tumors (Nielsen et al., 2010). Cancer is a complicated disease with complex treatments, because different causes of cancer will lead to different prognosis and need different treatments. Whereas, the same treatment for different patients with the same cancer may lead to different results (Xu et al., 2015).

The application of computer technology to the management of molecular biology is known as bioinformatics. The ultimate goal of bioinformatics is to better understand a living cell and how it functions at the molecular level using computational tools. Starting by storing and mining raw genomic data, and going into analyzing and interpreting relations found within data, then deducing information and discovering meaningful knowledge thereof, this knowledge is crucial for making the right decision on diagnosis and prognosis, as well as being able to generate new insights and provide a global perspective about the cell, aiming at exploring the genetic relationships of deadly diseases.

1.2. Gene Expression Data

It would be difficult for the biologist to discover the knowledge that resides in many DNA, RNA and protein data with traditional methods. Information technology can be very useful. Microarray gene expression studies have been actively pursued to extract significant biological knowledge hidden under a large volume of gene expression profiles accumulated by microarray experiments. Particularly interesting attention has been given to a variety of data mining systems for the discovery of genetic function, diagnosis of the disease (Saiki et al., 2008), pathway analysis (Werner, 2008), identification of pharmaceutical targets (Corn et al., 2007) and so on.

DNA methylation, occurring in the context of a CpG dinucleotide, is a kind of epigenetic modification that is critical to gene regulations and genomic functions, it can be inherited through cell division (Chandrasekhar et al., 2011). Some special methylation patterns are found in many genetic diseases including various types of cancer (Macgregor & Squire, 2002). The strong correlation between cancer and DNA methylation had been demonstrated in many researches (Kulis & Esteller, 2010). Indeed, the analysis of DNA methylation datasets is proven to be very interesting for the distinction between the cancerous genes and the normal ones. It becomes a powerful tool in cancer diagnosis, treatment, and prognostication.

While cancer cells have more genetic changes than normal cells, regulation of the biological process can occur by controlling mRNA gene expression, gene expressions provide information about the way how the cell reacts to a particular condition. DNA microarray technology has been developed to establish the levels of gene expression in various tissue types. This can be particularly useful in cancer studies, where mutations occur in malignant samples, as mutations can amplify or turn off gene expression.

DNA microarray technology has revolutionized the monitoring of the expression levels of thousands of genes. Finding the hidden patterns in this huge volume of gene expression data requires efficient computational methods and thus offers a huge opportunity for understanding the genomic functions. However, the large number of genes and the complexity of biological relationships greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements.

DNA microarray is generally a glass or plastic substrate, or silicon chip, onto which tens of thousands of DNA molecules are deposited in a regular grid-like pattern. Each grid spot corresponds to a DNA sequence of a specific gene. The idea of a microarray is to detect the presence and abundance of specific DNA molecules (targets) in biological samples of interests. cDNAs are obtained and labeled