Chapter 6
Observer-Biased Analysis of Gene Expression Profiles

Paulo Fazendeiro
Instituto de Telecomunicações (IT), Portugal

José Valente de Oliveira
University of Algarve, Portugal

ABSTRACT

Microarray generated gene expression data are characterized by their volume and by the intrinsic background noise. The main task of revealing patterns in gene expression data is typically carried out using clustering analysis, with soft clustering leading the more promising candidate methods. In this chapter, Fuzzy C-Means with a variable Focal Point (FCMFP) is exploited as the first stage in gene expression data analysis. FCMFP is inspired by the observation that the visual perception of a group of similar objects is (highly) dependent on the observer position. This metaphor is used to provide a new analysis insight, with different levels of granularity, over a gene expression dataset.

INTRODUCTION

A gene usually corresponds to a sequence used in the production of a specific protein or ribonucleic acid (RNA) molecule. It is a region of deoxyribonucleic acid (DNA) that controls a hereditary characteristic. A gene carries biological information in a form that must be copied and transmitted from each cell to all its progeny. Each gene has a fixed location on its chromosome and helps to specify a trait. Defective genes may cause diseases hence they need to be identified. Despite some evidences pointing that microarray technology is slowly being phased out in favor of several next-generation sequencing methods (Ozsolak & Milos, 2011; Wang, Gerstein, & Snyder, 2009) DNA microarrays are commonly being used in first-tier clinical testing (Riggs, 2014) and still are essential tools for various genomic studies, e.g. (Belfield, 2014; Sanmann, 2013; Nylund, 2013). This technique is providing a wealth of data on global patterns of gene expression. Currently, efforts are being made to describe and understand the global view of these patterns, *i.e.*, trying to uncover the hidden structures in gene expression data.

Gene expression refers to transcription levels of genes. The expression level refers to the amount of messenger RNA (mRNA) in a gene, which is the

DOI: 10.4018/978-1-4666-6611-5.ch006

transcription of an activated gene that is later translated into a protein. A wide range of approaches are being use to measure gene expression levels. These methods, which fall under the category of microarrays technology, include cDNA microarray (Schena et al., 1996a; Schena et al., 1996b) and oligonucleotide microarray (Fodor et al., 1993; Lipshutz et al., 2000). Gene expression profiling can also be performed using serial analysis of gene expression (SAGE) (Velculescu et al., 1997) and reverse transcription-polymerase chain reaction (RT-PCR) (Somogyi et al., 1995).

The analysis of microarrays generated data remains a quite challenging task. According to (Simon, 2008) gene expression profiling offers both a great opportunity for new kinds of investigation and great risk of error because it provides a high-dimensional read-out for each specimen assayed. The datasets are typically large with large background noise, cf. (Chu et al., 1998). The yeast cell cycle dataset analysed in this chapter is one relevant example of such datasets.

Clustering is usually the first step in gene expression data analysis (Jiang, Tang & Zhang, 2004). Apart from gene expression, clustering plays a major role in data mining applications such as information retrieval and text mining, web analysis, scientific data exploration, spatial database applications, CRM and marketing, image processing and recognition systems, medical diagnostics and computational biology, just to mention a few (de Oliveira & Pedrycz, 2007; Soowhan, Lee & Pedrycz, 2009; Ming, Kiong & Soong, 2011; Zhang & Lu, 2010; Chaira, 2011).

In the past, various techniques have been used for gene expression microarray data analysis such as K-means (Tavazoie et al., 1999; Richards et al., 2008), hierarchical clustering (Eisen et al., 1998; Lein et al., 2007; Finak et al., 2008), self-organizing maps (Tamayo et al., 1999; Ghouila et al. 2009), graph-theoretic approaches (Amir & Zohar, 1999; Huttenhower et al., 2007), and fuzzy c-means (FCM) (Futschik & Carlisle, 2005; Dembélé & Kastner; 2003). A recent review of these and other techniques can be found in (Pirim et al. 2012).

Most of the clustering methods used in gene expression analysis fall in the category of the hard clustering methods. One gene belongs to exactly one cluster. These methods implicitly assume that the clusters are well separated, which is hardly the case in gene expression data, with several biological studies reporting no clear boundaries between clusters. Moreover, hard clustering methods appear to detect clusters even in gene expression randomised data, cf. (Futschik, & Carlisle, 2005). By allowing one gene to belong, with different degree of membership, to more than one cluster, soft clustering allows to identify meaningful, biologically relevant, clusters. In this chapter we further exploited soft clustering for gene expression analysis by applying the fuzzy C-means with a variable focal point (FCMFP) algorithm (Fazendeiro & de Oliveira, 2014; Fazendeiro & de Oliveira, 2008) to the analysis of gene expression data of yeast cell cycle.

FCMFP is inspired in the following everyday live observation: The position at which the observer is located relatively to a set of objects determines how the observer perceives these objects. Suppose that the observer is located far away from the objects. In this case, and due to the observation distance objects tend to be undistinguishable, that is, objects tend to be seen as a single cluster. As the observer gets closer and closer to the objects the differences between them tend to become clearer and clearer. The initial single cluster tends to split in a number of clusters which is becoming higher and higher and eventually becomes equal to the number of objects. The authors have integrated this metaphor into the popular FCM algorithm. This is accomplished by incorporating a focal point and a zoom factor into the original FCM objective function. The focal point represents the point where the observer is located relatively to the objects to be clustered. The zoom factor acts