Chapter 9
Subspace Clustering of DNA Microarray Data: Theory, Evaluation, and Applications

Alain B. Tchagang
National Research Council of Canada, Canada

Fazel Famili
National Research Council of Canada, Canada

Youlian Pan
National Research Council of Canada, Canada

ABSTRACT

Identification of biological significant subspace clusters (biclusters and triclusters) of genes from microarray experimental data is a very daunting task that emerged, especially with the development of high throughput technologies. Several methods and applications of subspace clustering (biclustering and triclustering) in DNA microarray data analysis have been developed in recent years. Various computational and evaluation methods based on diverse principles were introduced to identify new similarities among genes. This review discusses and compares these methods, highlights their mathematical principles, and provides insight into the applications to solve biological problems.

1. INTRODUCTION

Recent developments in genomics and highthrouput technology have shown that subspace clustering (biclustering and triclustering) is an emerging and powerful methodology for gene expression data analysis. This is driven by the fact that subspace clustering is able to identify local behaviors of the dataset. When dealing with DNA microarray data, biclustering for example is capable to find subgroups of genes that are intimately related across subgroups of attributes, *e.g.* experimental conditions, time points, or tissue samples. On the other hand, triclustering identifies subgroups of genes that are coherent on subsets of samples along segments of time series. In other words, by simultaneously clustering the gene expression

DOI: 10.4018/978-1-5225-8903-7.ch009

matrix in each one of its dimensions, one can identify candidate subsets of attributes that are associated with specific biological functions, in which only a subset of genes potentially plays a role. Biological analysis and experimentation could then confirm the significance of the candidate subsets.

Since the introduction of biclustering algorithms in DNA microarray data analysis in 2000 by Cheng and Church, biclustering has received a great deal of attention. Thousands of research papers have been published, presenting new algorithms or improvements to solve this biological data mining problem more efficiently. Similarly, with the advent of three-dimensional DNA microarray data, triclustering algorithms have emerged as a natural extension of biclustering algorithms in 3D space, same as biclustering was a natural extension of the classical full space clustering in 2D space. In this paper, we explain the biclustering and the triclustering problems, some of their variations, and the main techniques to solve them. Obviously, given the huge amount of work on these topics, it is impossible to cover all proposed algorithms. Instead, in this article, we attempt to give a comprehensive survey of the most influential algorithms and results. It begins with a description of the biological problem motivating the underlying methodology. At each step, an attempt is made to describe both the relevant biological and statistical assumptions so that it is accessible to biologists, statisticians, and computer scientists, and can be of use to graduate students and to those starting to do research on biclustering and triclustering of microarray data as well as users experienced with this technique. Furthermore, we provide insights regarding the methodologies available for statistical and biological evaluations of the subspace clusters, and demonstrate the applicability of biclustering and triclustering algorithms to solve specific problems in computational biology and gene expression data analysis in particular.

This review is divided into seven sections with several examples at the end. The first two sections give a quick overview of DNA microarray technologies and DNA microarray data preprocessing and normalization. The section on clustering versus biclustering and triclustering algorithms describes the main differences between full space clustering, biclustering and triclustering algorithms. The section on biclustering of DNA microarray data introduces the application of biclustering to microarray data, illustrating the practical aspects of these techniques. The section on triclustering algorithms presents triclustering as a natural extension of biclustering algorithms and methodologies for extracting triclusters from 3D microarray data. The section on statistical and biological evaluations of clusters presents statistical tools and biological knowledge for the evaluation and biological significance of biclusters and triclusters. The section on biological applications shows several examples of biclustering and triclustering applied to microarray data to answer specific biological questions. Lastly, in the final section, we conclude and provide some insights for future research directions.

2. DNA MICROARRAY

Quantitative gene expression measurements using microarrays were first performed by Schena et al. (1995) on 45 *Arabidopsis thaliana* genes and shortly after, on thousands of genes or even a whole genome (DeRisi et al., 1996; DeRisi et al., 1997). Since that time, various methods for the analysis of such data have been developed. This includes the biclustering and triclustering techniques.

Microarrays are solid substrates hosting thousands of single stranded DNAs with a specific sequence, which are found on localized features arranged in grids. These molecules, called probes, hybridize with single stranded cDNA molecules, named targets, which have been labeled with fluorescence during a reverse transcription procedure. The targets reflect the amount of mRNA isolated from a sample obtained