Chapter 29
Biclustering of DNA Microarray Data:
Theory, Evaluation, and Applications

Alain B. Tchagang
National Research Council, Canada

Fazel Famili
National Research Council, Canada

Youlian Pan
National Research Council, Canada

Ahmed H. Tewfik
University of Minnesota, USA

Panayiotis V. Benos
University of Pittsburgh, USA

ABSTRACT

In this chapter, different methods and applications of biclustering algorithms to DNA microarray data analysis that have been developed in recent years are discussed and compared. Identification of biological significant clusters of genes from microarray experimental data is a very daunting task that emerged, especially with the development of high throughput technologies. Various computational and evaluation methods based on diverse principles were introduced to identify new similarities among genes. Mathematical aspects of the models are highlighted, and applications to solve biological problems are discussed.

INTRODUCTION

Recent developments in genomics and highthrouput technology have shown that biclustering is an emerging and powerful methodology for gene expression data analysis. This is driven by the fact that biclustering performs simultaneous row-column clustering and is able to identify local behaviors of the dataset. When dealing with DNA microarray data, biclustering is capable to find subgroups of genes that are intimately related across subgroups of attributes, *e.g.* experimental conditions, time points, or tissue samples. In other words, by simultaneously clustering the rows and columns of the gene expression matrix, 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.

DOI: 10.4018/978-1-4666-3604-0.ch029

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. In this chapter, we explain the biclustering problem, some of its variations, and the main techniques to solve them. Obviously, given the huge amount of work on this topic, it is impossible to explain or even mention all proposed algorithms. Instead, in this chapter, 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 relevant statistical assumptions so that it is accessible to biologists, statisticians, and computer scientists, and can be of use to those starting to do research on biclustering of microarray data as well as users experienced with this technique. Furthermore, we give more insights regarding the methodologies available for statistical and biological evaluations of the biclusters, and demonstrate the applicability of biclustering algorithms to solve specific problems in computational biology and gene expression data analysis in particular.

This chapter is divided into five sections with several examples at the end. 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 bicluster models interpretations and validations discusses the available procedures to measure the validity of the resulting biclusters. The section on algorithms for biclusters identification presents the implementation of popular algorithms. The section on biological applications shows several examples of biclustering applied to microarray data to answer specific biological questions. Lastly, in the final section, we conclude and provide some insights on future research directions.

BICLUSTERING OF DNA MICROARRAY DATA

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 techniques.

DNA Microarray

Microarrays are solid substrates hosting hundreds 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 during a reverse transcription procedure. The targets reflect the amount of mRNA isolated from a sample obtained under a particular condition. Thus, the amount of fluorescence emitted by each spot is proportional to the amount of mRNA transcribed from corresponding DNA sequence. The microarray is scanned and the resulting image is analyzed using signal and image processing techniques so that the signal from each probe can be quantified into numerical values. Such values represent the expression level of the gene in the given condition (Simon et al., 2003).

Microarrays can be fabricated by depositing cDNAs or previously synthesized oligonucleotides; this approach is usually referred to as printed microarrays. In contrast, *in situ* manufacturing encompasses technologies that synthesize the probes directly on the solid support. Slightly different oligonucleotides array platforms are manufactured by companies such as Affymetrix, Agilent, and NimbleGen. Each technology has its advantages and disadvantages and serves a particular research goal. A good review on DNA microarray technology can be found in (Irizarry et al., 2005).