Biclustering of Gene Expression Data Based on Local Nearness

Jesus S. Aguilar-Ruiz*, Domingo Savio Rodriguez*
Dan A. Simovici**

*BIGS BioInformatics Group Seville, University of Seville, Spain
dsavio@lsi.us.es,
**Univ. of Massachusetts Boston, Massachusetts 02125, USA
dsim@cs.umb.edu

Abstract. The analysis of gene expression data in DNA chips is an important tool used in genomic research whose main objectives range from the study of the functionality of specific genes and their participation in biological process to the reconstruction of diseases's conditions and their subsequent prognosis. Gene expression data are arranged in matrices where each gene corresponds to one row and every column represents one specific experimental condition. The biclustering techniques have the purpose of finding subsets of genes that show similar activity patterns under a subset of conditions. Our approach consists of a biclustering algorithm based on local nearness. The algorithm searches for biclusters in a greedy fashion, starting with two–genes biclusters and including as much as possible depending on a distance threshold which guarantees the similarity of gene behaviors.

1 Introduction

The DNA Microarray technology represents a great opportunity of studying the genomic information as a whole, so we can analyze the relations among thousands of genes simultaneously. The experiments carried out on genes under different conditions produce the expression levels of their transcribed mRNA and this information is stored in DNA chips.

A *bicluster* is a subset of genes that show similar activity patterns under a subset of conditions. The research on biclustering started in 1972 with Hartigan's work, in which the way of dividing a matrix in sub–matrices with the minimum variance was studied (Hartigan *et al*., 1972). In that approach the perfect bicluster was the submatrix formed by constant values, i.e., with variance equal to zero. Hartigan's algorithm, named *direct clustering*, divides the data matrix into a certain number of biclusters, with the minimum variance value, so the fact of finding a number of sub-matrices equal to the number of elements of the matrix is avoided. Another way of searching biclusters is to measure the coherence between their genes and conditions. Cheng & Church (Cheng *et al*., 2000) introduced a measure, the *mean squared residue* (MSR), that computes the similarity among the expression values within the bicluster. The ideas of Cheng and Church were further developed by Yang (Yang *et al*., 2002, 2003) who dealt with missing values in the matrices. As a result of this approach an algorithm named