Virtual Error: A New Measure for Evolutionary Biclustering

Beatriz Pontes[1], Federico Divina[2], Raúl Giráldez[2], and Jesús S. Aguilar–Ruiz[2]

[1] Department of Computer Science, University of Seville
Avenida Reina Mercedes s/n, 41012 Sevilla, Spain
bepontes@lsi.us.es
[2] School of Engineering, Pablo de Olavide University
Ctra. de Utrera, km. 1, 41013, Sevilla, Spain
{fdivina,giraldez,aguilar}@upo.es

Abstract. Many heuristics used for finding biclusters in microarray data use the mean squared residue as a way of evaluating the quality of biclusters. This has led to the discovery of interesting biclusters. Recently it has been proven that the mean squared residue may fail to identify some interesting biclusters. This motivates us to introduce a new measure, called *Virtual Error*, for assessing the quality of biclusters in microarray data. In order to test the validity of the proposed measure, we include it within an evolutionary algorithm. Experimental results show that the use of this novel measure is effective for finding interesting biclusters, which could not have been discovered with the use of the mean squared residue.

1 Introduction

Nowadays, technological advances offer the possibility of completely sequentialize the genome of some living species. This constitutes a great source of information which needs to be analyzed. Microarray techniques allow us to study genomes on their own or also to combine some of them in order to extract relational knowledge [12].

Microarray data are usually transformed into a numerical matrix which could then be analyzed. There exist various techniques to extract relevant information from a microarray, depending on the specific application in study. These techniques include clustering methods [4], where the goal is to cluster together genes that have a similar behaviour under all the experimental conditions. This grouping is carried out by means of any specific algorithm or mathematical formula based on genes similarity over all conditions [13]. It may be interesting, however, to analyze whether several genes in a microarray show the same behaviour under a subset of the experimental conditions. This has motivated a recent line of research named biclustering. Biclustering techniques aim at individuating subsets of genes that present the same behaviour under a subset of experimental conditions. This problem has been proven to be even much more complex than clustering [8].

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 217–226, 2007.

Biclustering was first applied to genomic data in [6], where the authors present a greedy search method for finding biclusters. In the same work, a measure for assessing the quality of biclusters, named *Mean Squared Residue* (MSR), is proposed. This measure has been used by many researches who have proposed different heuristics for biclustering biological data, e.g., [2,14]. Some other authors have established a search model to detect significant biclusters, without using a specific formula to optimize [11]. For a review of different biclustering techniques, we refer the reader to [9,10]. Among the used techniques, it is interesting to emphasize the application of evolutionary computation to the problem of finding biclusters in microarray data [5,8]. In these work the search was biased towards biclusters with low MSR.

The use of MSR to guide the search for biclusters in microarray data has led to the discovery of interesting biclusters. However, it has been proven that MSR may fail to recognize some interesting biclusters as quality biclusters [1]. This motivates us to introduce a new measure, called *Virtual Error* (VE), for assessing the quality of biclusters in microarray data. In order to evaluate the validity of the proposed measure, we include it within an evolutionary algorithm (EA). In a previous version of this EA, the search was guided mainly by the MSR. Experimental results show that the so modified EA is capable of finding interesting biclusters, which could not have been discovered with the use of MSR.

This paper is organized as follows: in Section 2 we present the motivations for this paper, we therefore describe the quality measure we propose in Section 3. Section 4 describes the used EA and how VE has been included into the algorithm, while some experimental results are shown in Section 5. Finally, Section 6 summarizes the main conclusions.

2 Motivation

One of the most used quality measures for biclusters is the *Mean Squared Residue*, MSR [6]. MSR tries to evaluate the coherence of the genes and conditions of a bicluster $\mathcal{B}$ consisting of I rows and J columns. MSR is defined as:

$$\mathtt{MSR}(\mathcal{B}) = \frac{1}{I \cdot J} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \tag{1}$$

where b_{ij}, b_{iJ}, b_{Ij} and b_{IJ} represent the element in the ith row and jth column, the row and column means, and the mean of the submatrix, respectively. If the gene expression levels fluctuate in unison under the conditions contained in a bicluster $\mathcal{B}$, then $\mathtt{MSR}(\mathcal{B}) = 0$. In general, the lower the MSR, the stronger the coherence exhibited by the bicluster, hence the better the quality. It follows that a trivial or constant bicluster where there is no fluctuation is characterized by a very low value of MSR. In order to reject these kind of biclusters, most heuristics combine the MSR with some other measures, e.g., the row variance [8,6].

As demonstrated in [1], MSR may not be the optimal measure for assessing the quality of some kinds of biclusters. In this work, the author makes a further study on the main characteristics inherent to biclusters, extracting from them