GA–based Approach to Discover Meaningful Biclusters

Jesús S. Aguilar-Ruiz
BioInformatics Group
University of Seville, Spain
aguilar@lsi.us.es

Federico Divina
Computational Linguistics and AI section
University of Tilburg, The Netherlands
f.divina@uvt.nl

Categories and Subject Descriptors

I.2.1 [**Applications and Expert Systems**]: Medicine and Science; I.5.1 [**Clustering**]

General Terms

Algorithms

1. INTRODUCTION

Since the development of the microarray technique in 1995, the interest in extracting meaningful biological knowledge from gene expression data has experimented an enormous increase. In particular, data mining researchers have developed ad–hoc techniques for many task, such as clustering or classification.

Clustering has been applied to gene expression data, which usually refers to conditions or patients, although genes can also be grouped in order to search for functional similarities. However, relevant genes are not necessarily related to every condition, or in other words, there are genes that can be relevant for a subset of conditions. From this point of view, clustering can not only be addressed in one dimension (over genes or conditions), but also in the two dimensions simultaneously. This approach, named *biclustering*, identify groups of genes that show "similar" level expression under a specific subset of experimental conditions.

A bicluster is defined on a gene–expression matrix. Let $G = \{g_1, \ldots, g_N\}$ be a set of genes and $C = \{c_1, \ldots, c_M\}$ a set of conditions. The data can be viewed as an $N \times M$ expression matrix EM. EM is a matrix of real numbers, with possible null values, where each entry e_{ij} corresponds to the logarithm of the relative abundance of the mRNA of a gene g_i under a specific condition c_j. Thus a bicluster can be seen as a sub-matrix IJ of EM.

Cheng and Church [1] proposed the biclustering of gene-expression matrices, introducing the *residue* ($r_{ij} = e_{ij} - e_{iJ} - e_{Ij} + e_{IJ}$) of an element in the bicluster and the *mean squared residue* ($r_{IJ} = \frac{\sum_{i \in I, j \in J} r_{ij}^2}{|I| \cdot |J|}$) of a sub-matrix. In the above formulas e_{iJ}, e_{Ij} and e_{IJ} are the mean of the ith row, of the jth column and of the sub-matrix IJ identifying the bicluster, respectively. The residue is an indicator of the degree of coherence of an element with respect to the remaining ones in the bicluster, given the tendency of the relevant gene and the relevant condition. The lower the residue, the stronger the coherence. In addition, they adjusted that measure to reject trivial biclusters by means of the *row variance* ($var_{I,J} = \frac{\sum_{i \in I, j \in J} (e_{ij} - e_{iJ})^2}{|I| \cdot |J|}$). Biclusters with high row variance contains genes that have large change in expression values over different conditions.

If a bicluster has a mean squared residue lower than a value δ it is called a δ-bicluster.

In this work, we address the biclustering problem with evolutionary computation (EC), which has been proven to have an excellent performance on highly complex optimization problems.

GECCO'05, June 25–29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

2. THE ALGORITHM

The algorithm we propose, called `SEBI`, adopts a sequential covering strategy: a genetic algorithm, called `EBI` (Evolutionary BIclustering), is called several times, until an *end condition* is met. `EBI` takes as input the expression matrix and the δ value and returns either a δ-bicluster or nothing. In the former case, the returned bicluster is stored in a list called LB, and `EBI` is called again. The end condition is also met when `EBI` is called a maximum number of times nb. When the end condition is met, the list LB is returned.

After a bicluster is returned, weights associated with the expression matrix are adjusted. This operation is performed in order to bias the search towards biclusters that do not overlap with already found biclusters. The weight of an element depends on the number of biclusters in LB containing the element. The more biclusters cover an element, the higher the weight of the element will be. The aim of `EBI` is to find δ–biclusters with maximum volume, with a relatively high row variance, and minimizing the effect of overlapping among biclusters.

The initial population consists of biclusters containing only one element of the expression matrix. These biclusters have the property of having a mean squared residue equal to 0. Tournament selection is used for selecting parents. Selected pairs of parents are recombined with a crossover operator with a given probability p_c (default value 0.9), and the resulting offspring is mutated with a probability p_m (default value 0.1). Three crossover operators can be applied with equal probability: one-point, two-point and uniform crossover. Three mutation operators are used, a standard mutation operator, a mutation operator that adds a row and a mutation operator that adds a column to the bicluster.

Each individual of the population encodes one bicluster. Biclusters are encoded by means of binary strings of length $N + M$, where N and M are the number of rows (genes) and of columns (conditions) of the expression matrix, re-

spectively. Each of the first N bits of the binary string is related to the rows, in the order in which the bits appear in the string. In the same way, the remaining M bits are related to the columns. If a bit is set to 1, it means that the relative row or column belongs to the encoded bicluster; otherwise it does not. It is worth to note that given the large value of N (several thousands), the search space size for the evolutionary algorithm is huge, and therefore more emphasis has to be placed on the performance of genetic operators.

The fitness function rewards individuals encoding biclusters with low mean squared residue, with high volume and row variance and covering elements of the expression matrix that are not covered by biclusters found by previous executions of `EBI`. The final objective of the `EBI` is to minimize the fitness.

3. EXPERIMENTAL RESULTS

In order to assess the goodness of the proposed algorithm for finding biclusters in expression data we performed experiments on a well known dataset, the *Embryonal tumors of the central nervous system* dataset. The expression matrix contained in this dataset consists of 60 conditions and 7129 genes. This dataset was used to explore heterogeneity in response to treatment of medulloblastomas [2]. The dataset distinguishes between patients who are alive (39) after treatment compared with those who succumbed to their disease (21). In these experiments, the value of δ was set to 1800.

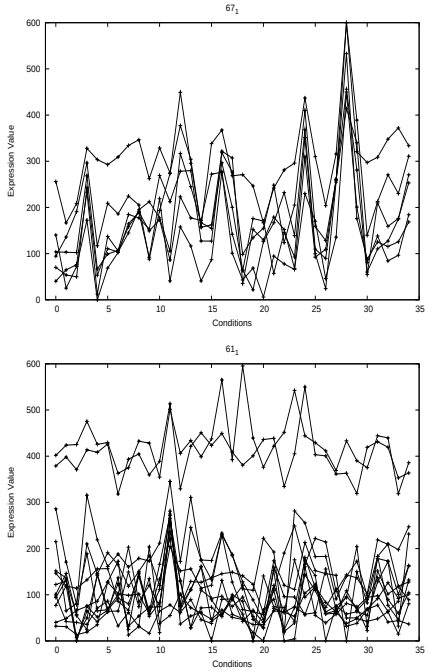

Figure 1: Two examples of bicluster found on the embryonal dataset.

Two examples of bicluster found on the embryonal datasets are shown in figure 1. Bicluster 61_1 is particularly interesting because even though the magnitude of the expression levels of the genes under the given conditions are not close, the patterns they exhibit are very alike. This type of pattern is named *shifting pattern*. Bicluster 67_1 shows, instead, genes that show strikingly up and down regulations under the same set of conditions, and with magnitude of the expression levels of the genes close to each other.

Given the biological interest of the results, we have tested our biclusters with information gathered from public databases [3] (EMBL, The European Molecular Biology Laboratory). Many of the genes found in biclusters belong to the same gene network, in which relationships (edges in the graph) come from co–expression, neighborhood, homology or text mining analysis. For example, the bicluster labeled 61_1 has 13 mRNA sequences, in which there are 8 genes (SCARB2, ACTG2, NR2C1, ATP2A2, CACNA2D1, SRP9, RRM and COMT). This example is very interesting as there are only four subgraphs. Three of them are controlled by the genes SCARB2, NR2C1 and SRP9, respectively. However, the fourth subgraphs contains the other five genes. Such five genes are not connected to each other, but there exist other genes that interconnect one to another. These new "bridge" genes have been found in neighborhood, co–expression or in PubMed with text mining methods. ATP2A2 has been associated to skin disorders, but not the others. However, this relationship might discover interesting interactions during the development of the disease.

4. CONCLUSIONS

In this paper we have introduced an algorithm based on genetic algorithms, called `SEBI`, for finding biclusters on gene expression data. The experimental results show that our approach produces very interesting biclusters, being able to extract shifting patterns from data. In addition, the subsets of genes from some biclusters have been biologically validated, by showing the relationships among genes in biclusters. Other methodologies, like text mining, can help at discovering other genes not present in biclusters that can relate some of them included in biclusters, giving some clues about the possible implications of new genes in biological processes, as regulatory networks or pathways.

In short, `SEBI` is successful in finding set of genes that show strikingly similar up–regulations and down–regulations under a set of conditions.

Acknowledgment

The research was partly supported by the Spanish Research Agency CICYT under grant TIN2004–00159 and Junta de Andalucia (III Research Program).

5. REFERENCES

[1] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Itelligent Systems for Molecular Biology (ISMB'00)*, pages 93–103, 2000.

[2] S. L. Pomeroy and et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.

[3] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 1(31):258–61, 2003.