Max-variance Clustering and Biclustering of Microarray Data

C. Cano[1], S. Blanco[2], F. García[1], A. Blanco[1]
[1]Department of Computer Science and A. I. University of Granada.
{ccano, fernan, armando}@decsai.ugr.es
[2]Department of Electronic, Computer Science and Automatic Engineering. University of Huelva.
sergio.blanco@diesia.uhu.es

Abstract

Microarray technology allows to measure the expression of thousands of genes simultaneously, and under tens of specific conditions. Clustering and Biclustering are the main tools to analyze gene expression data, since they reveal genes with the same behavior across samples. In this paper we present three novel approaches for Clustering and Biclustering based on Estimation of Distribution Algorithms (EDA) and Principal Components Analysis. The goal is to find non-exclusive (potentially overlapping) groups of genes with similar behavior and maximum between-sample variance. We tested the proposed methods on two real datasets, outperforming previous results in terms of quality and size of revealed patterns.

Keywords: Microarray, Clustering, Biclustering.

1 Introduction

Microarray technology makes use of the sequence resources created by the genome projects to monitor the expression of thousands of genes in particular cell samples, times and conditions.

A microarray is a slide on which single-stranded DNA molecules are attached at fixed locations called spots. There can be thousands of spots on a single microarray, each one containing a huge number of identical DNA molecules which identify one gene. The hybridization experiment consists on dying the total mRNA from a cell sample with a fluorescent label and washing it over the microarray. These labeled gene products hybridize to their complementary sequences in the spots. The fluorescence emitted from each spot when the microarray is excited by a laser, allows to measure the amount of sample bounded to the DNA of the spot.

Gene expression profiles are usually presented in a matrix $A_{n\times m}$ where rows represent genes, columns represent different cell situations under study, and each element a_{ij} indicates the expression level of gene i under condition j. Thereby microarray data provide a global picture of the cell activities and open the way to a high-level understanding of its behavior.

Previous work on Clustering

The potential of clustering to reveal patterns in microarray data was shown by Eisen *et al.* [7], who applied hierarchical clustering to identify functional groups of genes. After that, many other approaches have been proposed to cluster microarray data (for a review see [10]). Clustering methods group genes which have similar expression levels across conditions (co-expressed genes). Genes in the same cluster respond similarly in different circumstances, so they are likely to share a common function.

However, traditional approaches group genes

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 690–697
Torremolinos (Málaga), June 22–27, 2008

into mutually exclusive clusters, whereas in the real biological system a gene may play multiple roles in different biological processes. To address this, several methods have been proposed ([9], [6]). Our work is largely motivated by the algorithm Gene Shaving [9], which has become one of the most widely-extended algorithms for the analysis of microarray data. Gene Shaving searches for potentially-overlapping coherent clusters with high variance across samples (i.e. high row variance). Therefore, provided clusters become very useful for identifying distinct types of samples and studying the biological processes which may cause these differences in the behavior of the genes.

However, there can be tens of heterogeneous conditions in one expression matrix, so looking for genes behaving similarly in all the conditions can lead us to miss relevant and interesting patterns. Biclustering has emerged as a suitable method to solve this limitation as it allows to identify groups of genes co-expressed under subsets of samples.

Previous work on Biclustering

Let $R = \{1..n\}$ and $C = \{1..m\}$ be the set of rows and columns of $A_{n\times m}$. A bicluster B is defined as a submatrix $B = A_{IJ}$ of $A_{n\times m}$, with values correlated according to a certain criterion. Particularly, we are interested in coherent groups of genes which behave very differently across conditions, following the model proposed by Hastie *et al.* [9]. Finding maximal size biclusters in a matrix is a NP-complete problem [15], hence almost all the existing algorithms use heuristics approaches ([12, 16, 4], for a review see [13]).

The only previous work which address the identification of coherent biclusters with maximum variance is the work by Aguilar and Divina [1]. They propose a Genetic Algorithm (GA) for identifying non-overlapping coherent biclusters with maximum variance across conditions. However, their quality measure requires the establishment of a threshold (upper bound) δ for the coherence of the bicluster which requires some prior knowledge which depends on the dataset.

The aim of this paper is to present three new algorithms (one for Clustering and two for Biclustering) for identifying potentially-overlapping groups of genes with the maximum variance across samples. Section 2 describes the first proposed approach: Gene-&-Sample Shaving, which uses Principal Component Analysis to identify biclusters, extending the Gene Shaving algorithm proposed by Hastie *et al.* [9]. In section 3 we propose a novel framework for clustering and biclustering using Estimation of Distribution Algorithms: EDA-Clustering and EDA-Biclustering. Section 4 shows the results obtained by the proposed approaches on two real microarray datasets. Finally, section 5 presents conclusions and future work.

2 Identifying max-variance patterns with Principal Components

2.1 Initial Approach: Gene Shaving

Rather than simply looking for genes with similar expression patterns, Gene Shaving searches for coherent clusters with high variance across samples [9]. The algorithm takes the expression matrix $A_{n\times m}$ and the number of desired clusters M as input. Let S_k be a cluster of k genes and

$$\overline{a_{S_k}} = (\frac{1}{k}\sum_{i\in S_k} a_{i1}, \frac{1}{k}\sum_{i\in S_k} a_{i2}, ..., \frac{1}{k}\sum_{i\in S_k} a_{im}) \quad (1)$$

be the collection of m column averages of the expression values for this cluster.

For every cluster size k, the algorithm seeks a cluster S_k having the highest variance of the column averages, i.e. $\arg\max var(\overline{a_{S_k}})$. For obtaining this cluster, Gene Shaving generates a sequence of nested clusters:

$$S_n \supset ... \supset S_{k_i} \supset S_{k_j} \supset ... \supset S_1 \quad (2)$$

of decreasing size, starting with $k = n$, the total number of genes, and finishing with $k = 1$ gene. At each stage the largest principal component of each cluster of genes is computed. This *eigen-gene* is the normalized linear combination of genes with the largest

variance across the samples. Then we discard a fraction ($\alpha \in [0,1]$) of the genes having lowest correlation (lowest absolute inner-product) with this *eigen-gene*, obtaining the next nested cluster. The process is repeated until we get a cluster with one gene.

Once the nested sequence of clusters has been completed, the algorithm selects one cluster from the sequence. This selection is done by calculating, in analogy with ANOVA (Analysis of Variance), the following measures of variance for each cluster S_k:

$$\begin{array}{l} V_W = \frac{1}{m}\sum_{j=1}^{m}[\frac{1}{k}\sum_{i\in S_k}(a_{ij}-\overline{a}_j)^2] \\ V_B = \frac{1}{m}\sum_{j=1}^{m}(\overline{a}_j-\overline{a})^2 \\ V_T = \frac{1}{k\times m}\sum_{i\in S_k}\sum_{j=1}^{m}(a_{ij}-\overline{a})^2 = V_W + V_B \end{array} \quad (3)$$

Where $\overline{a}_j = \frac{1}{k}\sum_{i\in S_k} a_{ij}$ in all the expressions above.

The *Within Variance* (V_W) measures the variability between the genes of the cluster (cohesion of the cluster). The *Between Variance* (V_B) is the variance of the mean gene of the cluster (variance across samples). To minimize the V_W and maximize the V_B, the *percentage of variance explained* (R^2) is computed:

$$R^2 = 100\frac{V_B}{V_T} = \frac{\frac{V_B}{V_W}}{1+\frac{V_B}{V_W}} \quad (4)$$

So large R^2 values imply high values of V_B and low values for V_W. To know whether a value of R^2 for a given cluster S_k is larger than we would expect by chance, i.e. if the rows and columns of A were independent, Hastie *et al.* proposed the GAP measure [9].

Let D_k be the R^2 measure for S_k, and A^{*b} a permuted data matrix, obtained by randomly permuting the elements of each row of A. If we form B such matrices, we can define GAP as the function:

$$GAP(S_k) = D_k - \overline{D}_k^* \quad (5)$$

Where $\overline{D}_k^*$ is the R^2 meanvalue for S_k in the B randomly permuted matrices: $A^{*1}, ..., A^{*B}$. Therefore, a large GAP value for S_k reveals a relevant (non-spurious) pattern.

After selecting one cluster from the sequence, A is orthogonalized with respect to the mean of the selected cluster, promoting new patterns to be revealed in further iterations.

2.2 Biclustering based on Principal Components: Gene-&-Sample Shaving

Gene-&-Sample Shaving is a novel approach we propose for identifying biclusters with max-variance patterns based on Principal Components.

The main idea of the algorithm is to compute the leading Principal Component (PC) for both the rows and the columns of the expression matrix. The genes with the lowest correlation with the *eigen-gene* (PC of the genes in A) are removed in order to keep those genes with maximum variance across samples (by analogy with the Gene Shaving algorithm). However, we also remove those samples which present the highest correlation with the *eigen-sample* (PC of the samples in A), hence keeping the samples with minimum variance over the genes. The latter is what we call *Sample Shaving*. Combining the removal of genes and samples (Gene Shaving and Sample Shaving) we get coherent biclusters in which genes exhibit very different behavior across samples.

Due to the fact that expression matrices present much more genes than samples (typically thousands of genes and tens of samples), the best results are obtained by firstly removing genes and then removing conditions from the obtained clusters of genes. Thus, we apply Gene Shaving to obtain a sequence of nested clusters of genes (eq. 2) and then we perform a Sample Shaving to the clusters of this sequence to convert them into biclusters. The best bicluster found in this process (the one with the highest GAP value) is returned.

After obtaining one bicluster, the original expression matrix is orthogonalized with respect to the mean of the bicluster. This is done to promote new signals to be revealed in fur-

ther iterations, allowing potential overlapping among signals.

3 Identifying max-variance patterns with Estimation of Distribution Algorithms

3.1 EDA-Clustering

As it can be noted, the generation of the sequence of nested clusters in Gene Shaving (eq. 2) is strongly driven by the variance of the genes across conditions, as genes are shaved-off depending on their correlation with the leading principal component. However, we are also interested in obtaining high-coherence clusters and this criterion is not directly used for obtaining the clusters sequence. It is only considered at the end of the process, when the GAP statistic for every cluster of the sequence is computed in order to select one of them.

The Gene Shaving process can be seen as a multiple-step *Feature Subset Selection* (*FSS*) problem in which, given a set of genes S_k with $k \in [2, n]$, we want to select a subset with $k \times (1-\alpha)$ genes: $S_{k\times(1-\alpha)} \subset S_k$, which maximizes a given criterion. Gene Shaving uses the variance of the cluster mean as optimization criterion. We consider that maximizing *GAP* function instead of between-sample variance provides overall better results. Therefore, we address the *FSS* problem of finding clusters with high values for the *GAP* function with Evolutionary Algorithms (EA) and particularly with Estimation of Distribution Algorithms (EDA), which have been proven to have an excellent performance on highly complex optimization problems [11].

Estimation of Distribution Algorithms (EDAs) are a set of Evolutionary Algorithms mainly characterized by the use of explicit probability models to recover the information of the selected individuals and to sample new solutions [11]. In EDAs, there are neither crossover nor mutation operators. Instead, a probabilistic model is inferred from selected individuals of the current generation, and the new population of individuals is sampled from the estimated distribution (see Figure 1).

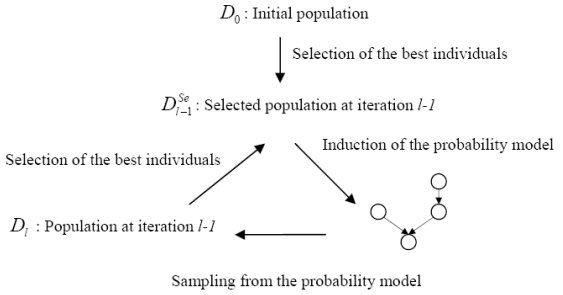


Figure 1: Scheme of an EDA algorithm.

We implemented the Univariate Marginal Distribution Algorithm (UMDA) by Muhlenbein [14], which considers the variables (i.e. genes) to be independent. To represent the solutions, each individual is coded as a binary string of length n representing whether each gene is selected in the cluster or not. The univariate distribution for every gene can be computed as the fraction of individuals from the selected population D^{Se}_{l-1} (Figure 1) for which that gene is selected.

We applied the UMDA algorithm to solve the FSS problem following two different schemes:

- Single-step FSS: only one execution of the EDA which takes the whole matrix as input and directly generates a cluster maximizing GAP.
- Multiple-step FSS: we can generate a nested sequence of clusters (eq. 2) by selecting/discarding, step by step, a fraction of the remaining genes with an EDA guided with the GAP function.

Both approaches provide good results as shown in Section 4.1 [1]. To evaluate EDA performance in comparison with other Evolutionary Algorithms, we also implemented a generational Genetic Algorithm (GA) with elitism to perform the selection of genes into clusters

[1]The best results are obtained with Baker's stochastic universal sampling selection, $|D^{Se}_{l-1}| = |D_l|/2$, $|D_l| = k \cdot \alpha/15 + 20$ (see eq. 2), $\sharp Iterations = 150$ for the multiple-step approach and $|D_l| = 200$, $\sharp Iterations = 200$ for the single-step approach.

maximizing GAP (GA-Clustering) [2]. EDAs showed overall better results than GA for this problem as can be seen in Section 4.1.

3.2 EDA-Biclustering

We can extend the FSS problem we presented in the previous section to the problem of biclustering. To represent the solutions, each individual is now coded as a binary string of length $n + m$ representing whether each gene or sample is selected in the bicluster. We applied an UMDA algorithm to find the bicluster with the highest GAP value following the one-step scheme previously detailed [3]. Results are shown in section 4.

4 Experimental Results and Discussion

We test the proposed methods on two real microarray datasets: the yeast cell cycle microarray data by Cho *et al.* [5] and the human lymphoma data by Alizadeh *et al.* [2]. We also implemented the Gene Shaving algorithm and run it on the two datasets to perform a comparison with the proposed methods [4]. We focus the comparison on the quality (GAP) and size of obtained patterns.

We also analyze the biological significance of these revealed patterns. For this end, we use the Gene Ontology [8] and GO Term Finder [3] to retrieve the most significant biological processes associated to each cluster and bicluster, extracting relevant and significant insights from expression data. Particularly, we compute the statistical significance of every GO biological process in every group of genes, calculating the associated p-value by using the hypergeometric distribution and the Bonferroni multiple-hypothesis correction [3].

[2]The best results are obtained with Baker's stochastic universal sampling selection, a crossover operator which maintains the common values of both parents and a *BitFlip* mutation operator. $Population\_size = k \cdot \alpha/15 + 20$. Stop condition: $k \cdot \alpha \cdot 12 + 500$ calls to the fitness function reached. $P_{crossover} = 0.9$. $P_{mutation} = 0.2$.

[3]In this case, $|D_l| = 300$, $\sharp Iterations = 300$.

[4]Best results shown are obtained for $\alpha = 0.1$.

4.1 Yeast dataset

Yeast dataset contains the expression levels of 2879 yeast *Saccharomyces cerevisiae* genes under 17 cell cycle conditions, covering approximately two full cell cycles [5].

Results comparison. Table 1 shows average GAP and size (number of genes) for 100 clusters obtained in ten executions of each algorithm: Gene Shaving, GA-Clustering and EDA-Clustering (with both multiple-step and single-step schemes).

Table 1: Clustering results in yeast dataset. Shown average GAP and size (with standard deviations in parenthesis) of obtained clusters.

Algorithm	No. genes	GAP
Gene Shaving	13.26 (10.33)	61.89 (23.87)
GA-Clustering	14.56 (4.01)	79.92 (3.8)
EDA-Clustering (multiple-step)	15.3 (6.4)	81.87 (4.8)
EDA-Clustering (single-step)	35.53 (10.1)	72.64 (4.6)

GA-Clustering and EDA-Clustering (both single-step and multiple-step schemes) show higher average GAP and size than Gene Shaving. Based on a two tailed t-test, we check that the improvements in terms of GAP are statistically significant ($p - value < 0.05$) for GA-Clustering and EDA-Clustering (multiple-step) with respect to the results obtained by Gene Shaving. EDA-Clustering (single-step) obtains good GAP values and the largest clusters (the difference in terms of cluster size is significant: $p - value < 0.05$ with respect to any other method). Figure 2 shows the GAP value and size of a subset of the clusters obtained with each algorithm.

The proposed Biclustering algorithms outperform the above clustering results, as can be seen in Table 2. Gene-&-Sample Shaving obtains biclusters with higher GAP than Gene Shaving ($p - value < 0.01$), thus obtaining more detailed patterns of better quality. EDA Biclustering also outperforms both EDA-Clustering algorithms in terms of GAP ($p - values < 0.01$). Indeed, EDA Biclustering shows the best performance among all the tested methods for this dataset.

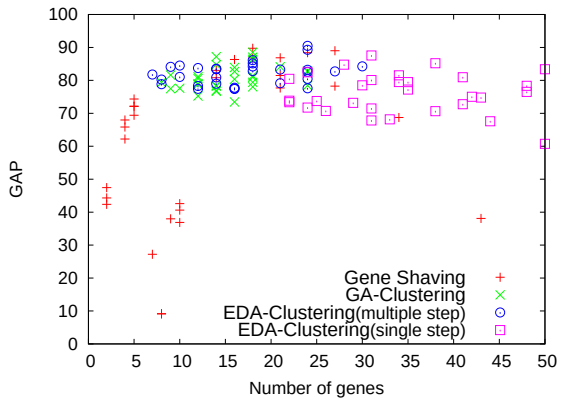


Figure 2: Scatter plot representing the GAP and size of the clusters obtained with each algorithm. For clarity, only the results of the first three executions (i.e. the first 30 clusters) are shown for each algorithm.

Table 2: Biclustering results in yeast dataset. Shown average GAP and size (with standard deviations in parenthesis) of obtained biclusters.

Algorithm	No.genes	No.cols	GAP
Gene & Sample Shaving	11.5(7.5)	4.7(2.7)	86.6(8.4)
EDA Bic.	25.1(4.4)	6.7(2.6)	88.8(4.1)

Biological interpretation of the results Using GO Term Finder we can assign the GO term with the lowest p-value to every obtained group of genes. Significant biological signals are revealed when we consider high-GAP and low-p-value clusters/biclusters (see Figure 3). Likewise we can validate our algorithms and interpret the results to extract new and reliable biological knowledge. For example, Figure 3(a) confirms the correspondence between the biological process *DNA metabolism*, which is the one with the lowest p-value for this cluster, and the expression behavior of the genes belonging to the cluster, which are over-expressed in samples 2-3 and 10-12. These samples are associated to the S phase of cell cycle, in which DNA replication takes place [5].

Figure 4 shows the gene expression patterns from two biclusters obtained with Gene-&-Sample Shaving and EDA Biclustering whose genes are significantly associated to *DNA metabolism* and *mitotic cell cycle*, respectively. Since biclustering algorithms do not require the genes to behave the same under all the samples, obtained patterns show higher quality and involve more genes than clustering results.

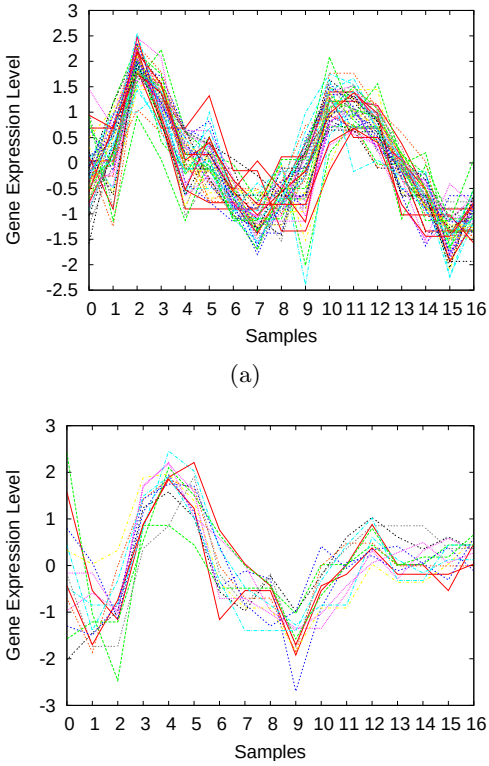
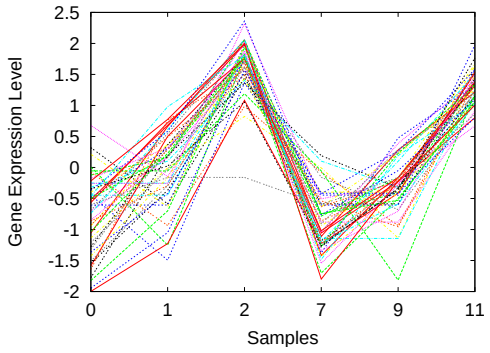


Figure 3: Clusters obtained with EDA-Clustering. 3(a) Single-step EDA-Clustering. GO term: DNA metabolism (P-value: $1.8x10^{-13}$). GAP:83.38 . Size:50 genes. 3(b) Multiple-step EDA-Clustering. GO term: Sulfur metabolism (P-value: $7.2x10^{-15}$). GAP:83.4 . Size:14 genes.

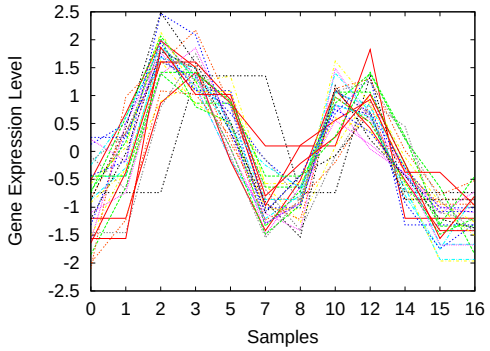
4.2 Lymphoma dataset

Lymphoma dataset contains the expression levels of 4026 genes under 96 human tissue samples, which are classified into 9 types of lymphoma and healthy tissues [2].

Biclustering algorithms are specially suitable to analyze this dataset since it contains a large number of heterogeneous conditions and there might be genes showing similar behavior in only a subset of the samples, i.e. some types of lymphomas. Thus, we focus on the proposed biclustering algorithms to carry out the analysis of this dataset.



(a)



(b)

Figure 4: 4(a) Bicluster significantly associated to *DNA metabolism* obtained with Gene-&-Sample Shaving. P-value: $2x10^{-10}$. GAP: 92.9 . Size: 48 genes, 6 samples. 4(b) Bicluster significantly associated to *mitotic cell cycle*, obtained with EDA Biclustering. P-value: $5.35x10^{-5}$. GAP: 86.32 . Size: 30 genes, 12 samples.

Results comparison Table 3 shows average GAP and size of 500 biclusters obtained in ten executions of the algorithms Gene-&-Sample Shaving and EDA Biclustering. We also show Gene Shaving results for comparative purposes. The proposed biclustering algorithms outperforms Gene Shaving results in terms of GAP ($p - value < 0.05$), with similar number of genes per group. They also show very low GAP dispersion which proves the robustness of these methods.

Biological interpretation of the results Genes from the human Lymphoma dataset are poorly annotated in GO. This lack of knowledge makes the obtained groups of genes poor significant from a biological point of view. However, this dataset contains different types of samples we know *a priori* [2]. Therefore, we may also consider whether the obtained biclusters help to discriminate the different types of samples.

Table 3: Biclustering results in Lymphoma dataset. Shown average GAP and size (with standard deviations in parenthesis) of obtained biclusters. Gene Shaving results are also shown for comparative purposes.

Algorithm	No.genes	No.cols	GAP
Gene Shaving	13.3(96.6)	96	52.1(17.3)
Gene & Sample Shaving	10.9(7.3)	14.9(14.2)	83.9(6.9)
EDA Bic.	30.2(6.6)	17.9(4.5)	68.6(8.3)

In order to determine if our results fit this classification, we compute the statistical significance of each type of condition in every bicluster. This leads us to extract promising conclusions. For example, Figure 5 shows the expression levels of genes in a bicluster obtained with EDA Biclustering which significantly represents conditions from *Chronic lymphocytic leukaemia (CLL)*, with a corrected p-value of $1.4x10^{-05}$. These expression profiles show that the genes belonging to this bicluster are under-expressed in the samples associated to this type of leukaemia (samples numbered from 83 to 94) and the same genes are over-expressed in conditions representing the other types of tissues of the dataset. Therefore, this bicluster represents genes whose behavior helps to discriminate CLL samples from the other healthy and cancerous tissues under consideration.

5 Conclusion

We have presented three new clustering and biclustering methods for identifying max-variance patterns in microarray data. The proposed approaches use Principal Components and Estimation of Distribution algorithms for maximizing the GAP measure defined in the Gene Shaving algorithm. Experimental results demonstrate that the proposed approaches EDA-Clustering, EDA-Biclustering and Gene-&-Sample Shaving outperform Gene Shaving in terms of quality

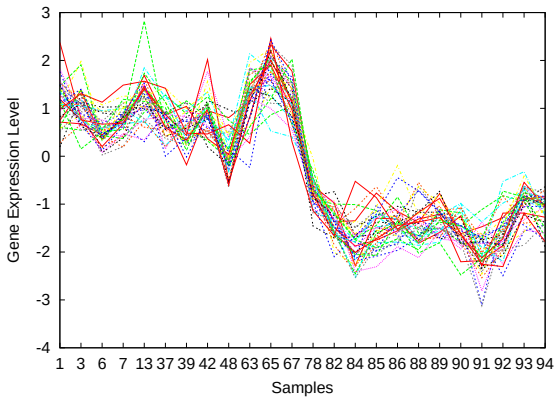


Figure 5: Bicluster obtained with EDA Biclustering on the lymphoma dataset. The bicluster contains 10 out of the 11 samples representing Chronic lymphocytic leukaemia (CLL) (samples from 84 to 94) (corrected p-value $1.4x10^{-5}$). GAP: 90.22 . Size: 39 genes, 24 samples.

and size of obtained patterns. Moreover, we validated the results from a biological point of view using the Gene Ontology.

Further work is needed to integrate information from different biological data sources, such as gene expression matrices, biological ontologies, biomedical literature and transcription factors binding sites.

Acknowledgements

This work has been carried out as part of projects TIC-640 of J.A. Sevilla and TIN2006-13177 of DGICT. Madrid.

References

[1] Aguilar-Ruiz, J.S., Divina, F. (2006) Biclustering Expression Data with Evolutionary Computation, *IEEE Trans Knowledge and Data Eng*, **18(5)**,590-602.

[2] Alizadeh, A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*,**403**, 503-510.

[3] Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., Sherlock, G. (2004) GO::TermFinder, *Bioinformatics*, **20**, 3710-3715.

[4] Cano, C., Adarve, L., Lopez, J., Blanco, A. (2007) Possibilistic Approach for Biclustering Microarray Data. *Comput Biol Med, Elsevier* , **37 (10)**, 1426-1436.

[5] Cho, R.J. et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol Cell.*, **2(1)**, 65-73.

[6] Dembelé, D., Kastner, D. (2001) Fuzzy c-means method for clustering Microarray data. *Bioinformatics*, **19**, 973-980.

[7] Eisen, M., *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS. USA*, **95**, 14863-14868.

[8] Gene Ontology Consortium (2004) *Nucleic Acids Res.*, **32**, D258-D261.

[9] Hastie, T. *et al.* (2000) Gene shaving as a method for identifying distinct sets of genes with similar expression, *Genome Biol*, **1**, 1-21.

[10] Jiang, D., Tang, C., Zhang, A. (2004) Cluster analysis for gene expression data: A survey. *IEEE Trans Knowledge and Data Eng*, **16(11)**, 1370-1386

[11] Larrañaga, P., Lozano, J.A. (2001) Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation. Kluwer Academic Publishers.

[12] Lazzeroni, L., Owen, A. (2002) Plaid models for gene expression data. *Statistica Sinica*, **12**, 61-86.

[13] Madeira, S., Olivera, A. (2004) Biclustering Algorithms for Biological Data Analysis: A survey, *IEEE/ACM Trans Comput Biol Bioinform*, **1(1)**, 24-45.

[14] Muhlenbein, H. (1998) The equation for response to selection and its use for prediction. *Evol Comput*, **5**, 303-346.

[15] Peeters, R. (2003) The Maximum Edge Biclique Problem is NP-Complete. *Discrete Applied Math.*, **131(3)** 651-654

[16] Tanay, A., Sharan, R., Shamir R. (2002) Discovering statistically significant biclusters in gene expression data, *Bioinformatics*, **18**, S136-S144.