Gene Clustering in Time Series Microarray Analysis

Camelia Chira*, Javier Sedano, José R. Villar, Carlos Prieto, and Emilio Corchado

[1] Instituto Tecnológico de Castilla y León, Burgos, Spain
{camelia.chira,javier.sedano}@itcl.es
[2] University of Oviedo, Gijón, Spain
villarjose@uniovi.es,
[3] Instituto de Biotecnología de León, León, Spain
carlos.prieto@inbiotec.es
[4] Universidad de Salamanca, Salamanca, Spain
escorchado@usal.es

Abstract. A challenging task in time series microarray data analysis is to identify co-expressed groups of genes from a large input space. The overall objective of this study is to obtain knowledge about the most important genes and clusters related to production and growth rate in a real-world microarray data analysis task. Various measures are engaged to evaluate the importance of each gene and to group genes based on their correlation with the output and each other. Some strategies for grouping and selecting genes are integrated resulting in several models tested for real biological data. All proposed models are tested on a real microarray data analysis problem and the results obtained are throughtly presented as well as interpreted from a biological perspective.

1 Introduction

Microarray data analysis (MDA) deals with a large number of features (genes) and needs efficient tools and techniques for the identification and classification of information [1–3]. The number of samples usually available is very low mainly due to the cost associated. This issue combined with the high dimensionality of the feature space make the task of extracting significant knowledge from microarray data an extremely difficult one. Time course (TC) microarray analysis [4–7] aims to find the best gene subset that promotes a certain variable or event when subsequent samples are taken from the same biological data at a certain time rate.

In TC MDA, the overall objective is to provide groups of genes meaningfully correlated and a ranking for each group in some well specified conditions. This paper focuses on a particular TC MDA problem with some specific requirements received from the biological experts. The input data consists of time series

* Corresponding author.

samples which contain the expression levels of 8848 genes of a certain bacteria measured at 12 time points, each with 3 replicates. Three output values are available for each sample as follows: (i) the *production* - a real value indicating a production level in the studied bacteria, (ii) the *production growth* - a boolean value indicating if production is produced or not in the current sample, and (iii) the *growth rate* - a real value representing the level of growth in the bacteria. The objective of the problem is to select and group those genes which are the most relevant and related with the changes in the production and growth.

A related problem is that of gene expression classification where each sample has a corresponding output class and the aim is to find the most relevant subset of genes able to correctly classify new samples. A typical approach is to apply a gene selection method in order to reduce dimensionality and then engage a classifier system to evaluate the accuracy of the classification based on the selected genes [8–10]. In the MDA problem considered in this paper, the output does not represent the class label for a sample and the aim is not to classify samples but rather to group them in a meaningful way.

This paper presents framework for MDA that can be used in the case of time course (TC) analysis [4–7] with certain restrictions: groups of genes have to be identified so that genes in the same group are related with each other and with some production or growth output (corresponding to each sample). The proposed generic model to address this task includes three main steps as follows: (i) gene sorting according to some information measures and correlation with the output, (ii) formation of groups using a Markov Blanket (MB) approach, and (iii) validation of groups based on rate of change from one time point to another. Several algorithms result from this model according to the measures (information based or statistic) used in grouping the genes and the strategy selected for the validation step. All resulted methods are applied for a real MDA problem and the experiments performed are discussed.

The paper is structured as follows: information and statistical measures commonly used in microarray analysis are briefly reviewed, a model for gene clustering and selection based on infomation theory measures and new proposed similarity measures are presented, and experiments and results obtained for several model variants of the proposed approach are discussed.

2 Relevant Information Measures for Gene Ranking and Selection

Measures coming from information theory are useful in several fields and often engaged in feature selection [11]. Let X be a random variable and $p(x)$ the probability distribution of X. The entropy $H(X) = -\int p(x) \cdot \log(p(x))dx$ is a measure of the information the feature supports. Similarly, $H(Y|X)$ denots the entropy of a feature y provided the feature x.

The *mutual information* between two features x and y (denoted by $I(X, Y)$) is defined by means of their probability distribution, as stated in Eq. 1. Higher values of the mutual information between two features correspond to higher

degrees of relevance between the two features. For our MDA problem, a naive way to select genes would be to calculate the mututal information between each gene and the output and then sort them in descending order. However, this approach would only consider the individual gene contribution and correlation with output.

$$I(X,Y) = \int\int p(x,y) \cdot \log(\frac{p(x,y)}{p(x) \cdot p(y)}) dxdy \tag{1}$$

For feature selection, these measures lack the ability of choosing independent features, particularly in high dimensional datasets. Let us consider two dependent features: if one of them has a high information measure then the second one does too. This results in a disadvantage of the above information metrics that lead to the proposal of other information measures described below.

The *Information Correlation Coefficient* (ICC) measures how independent two features are from each other (see Eq. 2). The higher the value the more relevant the relationship is. This measure is reflexive, symmetric and monotonic.

$$ICC(X,Y) = \frac{I(X,Y)}{H(Y|X)} \tag{2}$$

If $ICC(X,Y) = 1$ then the two variables X and Y are strictly dependent whereas a value of 0 indicates that they are completely irrelevant to each other.

The *Pearson's Correlation Coefficient* (PCC) measures the correlation between two features using statistics [12]. Let Y be the output feature and X is a feature from the input space. Let (x,y) be a pair of values of features X and Y, respectively. The PCC is calculated using Eq. 3.

$$PCC(X,Y) = \frac{\sum (x - \hat{x}) \cdot (y - \hat{y})}{\sqrt{\sum (x - \hat{x})^2 \sum (y - \hat{y})^2}} \tag{3}$$

3 Proposed Methods for Gene Clustering and Validation

The model proposed in this paper to approach the given MDA problem is based on information theory measures engaged to facilitate clustering and a new proposed measure mainly used in cluster validation. Genes are first grouped using different information and statistical measures in connection with the *Markov blanket* concept. In a second phase, groups are validated using a new measure for evaluating the rate of change in time series.

3.1 Gene Clustering

The formation of gene groups has to take into account the degree of relevance between each gene and the output as well as similar changes in gene expression levels in the time series. The phase of gene clustering in the proposed model addresses this problem with an emphasis on the first objective. The gene-output

and gene-gene relevant degrees are computed using information correlation measures. Two such measures are considered in this study as follows: *Information Correlation Coefficient (ICC)* and *Pearson Correlation Coefficient (PCC)*.

As described in the previous section, ICC (see Eq. 2) measures the relevance between two variables based on mutual information and joint entropy. ICC takes values between 0 and 1. The higher the ICC value the more relevant the relationship between the two variables is. For instance, if ICC(X,Y) = 1 then X and Y are strictly dependant and relevant. A correlation degree can be expressed by stating that X is relevant to Y with degree ICC(X,Y).

On the other hand, PCC (see Eq. 3) is a statistical measure of the strength of the association between the two variables. PCC values range from -1 to +1. Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases the other decreases (and viceversa).

The basic procedure for gene clustering follows some ideas described in [8]. An ensemble gene selection by grouping (EGSG) method has been proposed in [8] for classification tasks in MDA. In the EGSG method, genes are first clustered by approximate MB and then ensemble classifiers applied. In this study, we adapt the first step from EGSG in order to group similar genes based on the correlation with the production and growth output. Furthermore, clusters are validated continously during their formation using a newly introduced rate of change measure (detailed in the next subsection).

In the clustering phase, genes are first ranked according to the *Correlation Measure (CM)* with the output. Both ICC and PCC are considered as CM in different combinations for experiments. Groups of genes are formed starting from the highest-ranked gene so that genes in each group are correlated with each other and with the output based on the MB (Markov Blanket) strategy. The CM is used in determining if one gene is the approximate MB of another gene. The first gene added to a group is called the center of that group. A new gene g is accepted in an existing group if the center of the group is the approximate MB of g (otherwise, gene g forms a new group becoming the center of that group). The number of groups emerges from this schema and does not have to be a-priori known.

The main steps of the gene clustering phase are as follows:

A. For each gene $g_i, i = 1 \dots M$ calculate $CM(g_i, y)$ based on the formula of ICC (Eq. 2) / PCC (Eq. 3).
B. Sort the gene set according to the calculated CM value (starting with the highest value, meaning the most relevant genes to the output y are first in the list). Let S be the sorted gene set.
C. Initialize the number of groups $k = 1$. Initialize the first group G_k with the top ranked gene from S: $G_k = \{S[1]\}$.
D. For each gene $g_i, i = 2 \dots M$ do
   – *D.1. Grouping phase:*
   –– *D.1.1* If none of the centers of any group is the approximate MB of gene g_i then create a new group for g_i: $k = k + 1$; $G_k = \{g_i\}$.

–– *D.1.2* If there is a group G_h such that $G_h[1]$ is the approximate MB of gene g_i then add g_i to group G_h: $G_h = G_h \cup \{g_i\}$.
– *D.2. Validation phase:* Check if the membership of gene g_i to the chosen group is validated from a rate of change similarity perspective. The strategy used for validation is detailed in the next subsection.

E. The final number of groups is k and the resulting groups are $G_1 \dots G_k$.

The first phase results in k groups of genes clustered based on the correlation with the production output and each other.

3.2 Validation of Clusters

After a gene is added to an existing cluster (*grouping phase* - step D.1 described in the previous subsection), the updated cluster is validated by checking if the new gene has the same dynamics with the genes already present in the group (step D.2 - *validation phase*). If the new gene does not actually fit with the cluster then it is moved to a special group of *'unclustered'* genes (denoted by G_0).

The *Rate of Change Similarity (RCS)* measure is proposed to evaluate the similarity of the dynamics between two genes. RCS is defined as the number of significant changes that co-occur in two gene expression profiles. A percentage of the gene span is considered as a parameter to assess if a change is significant or not.

Let the span of a gene or of the output (production or growth) be the full extend of the variable, that is, the difference between its maximum and minimum values. Let τ be the parameter indicating a predefined percentage of the span. A significant change of a variable a (denoted by $\phi_i(a)$ at sample i in the dataset) is considered to occur when the difference of two consecutive values of that variable is higher than the product $\tau \cdot span$ (see Eq. 4). Then, given two variables a and b, the RCS is calculated as stated in Eq. 5.

$$\phi_i(a) = \begin{cases} 1, |a_i - a_{i-1}| > \tau \cdot span(a) \\ 0, otherwise \end{cases} \tag{4}$$

$$RCS(a,b) = \frac{\sum_{i=2}^{N} \phi_i(a) \cdot max_{j \in \{i-1,i,i+1\}} \phi_j(b)}{\sum_{i=2}^{N} \phi_i(a)} \tag{5}$$

It should be noted that $i = 2$ or $i = N$ represent special situations for Eq. 5. In these extreme situations, the strategy for finding a maximum ϕ value for the second variable b has to be changed from considering three possible rates of change to the only two actually available. In this way, when $i = 2$ the second term in the sum is $max_{j \in \{i,i+1\}} \phi_j(b)$ while for $i = N$ the $max_{j \in \{i-1,i\}} \phi_j(b)$ is considered.

The online validation phase (carried out once a gene is selected to be added to an existing cluster) determines if the new gene has a similar RCS to the output as the most representative gene in the cluster and further between each other. A parameter called δ is used to decide if the RCS for two different pairs

of genes (x_1, y_1) and (x_2, y_2) is similar. If $|RCS(x_1, y_1) - RCS(x_2, y_2)| < \delta$ then they are considered similar. The validation phase checks the difference between the RCS of the center of group and the output with both RCS of the new gene and the center of group and the RCS of the new gene and the output. As already mentioned, if the new gene does not pass the validation step, it is added to a special group G_0 of unclustered genes.

The clustering and validation phase results in k meaningful groups of genes as well as a special group G_0 which contains those genes that have no similarity to any cluster.

3.3 Summary of Proposed Methods

Several variants of the proposed model can be specified according to different measures chosen for clustering and validation phases. In order to allow an extensive analysis, we have selected three different model variants with or without validation and based on ICC, PCC or RCS in different combinations.

The following variants of the model have been selected for the study presented here: (i) *ICC_MB* - genes are sorted and MB clustered based on ICC, (ii) *ICC_MB_PCC* - genes are sorted and MB clustered based on ICC; validation of clusters is based on PCC, and (iii) *ICC_MB_RCS* - genes are sorted and MB clustered based on ICC; validation of clusters is based on RCS.

4 Computational Experiments

The dataset consists of 36 (3x12) samples and 8848 genes. A normalization and an optional discretization step was applied to the dataset. Some experiments use a discretization phase for the data which is applied after normalization. This phase means that the gene expression values are discretized so that insignificant changes are ignored. For this discretization phase, a parameter called *d_step* is used to decide a significant change.

4.1 Normalization

The normalization process was performed with the limma package [13]. Median and none background correction methods were applied for all results reported in this paper. Method none computes M and A values without normalization so the corrected intensities are equal to foreground intensities. On the other hand, method median substracts the weighted median of background intensities from the M-values for each microarray.

4.2 Experiments Setup

Experiments consider the input dataset as follows: the mean value of the 3 samples at each time point providing a dataset of 12 samples with 12 set of outputs (called *Mean12*). With each resulted dataset, the correlation with one of the

three available outputs i.e. the production growth output (*Bool*), the production rate (*RealProd*) and the growth output (*RealGrowth*) can be considered in the experiments.

Therefore, experiments and results are grouped in the following categories: *Mean12_Bool*, *Mean12_RealProd* and *Mean12_RealGrowth*. For each experiment category, all three model variants described in the previous section (i.e. *ICC_MB*, *ICC_MB_PCC* and *ICC_MB_RCS*) have been applied and the obtained results are discussed in the following subsections.

The possible parameters in each method include *d_step* (used in the discretization phase), τ and δ (both used in the validation phase). Based on many experiments performed and the results obtained, we selected the following values for each parameter to discuss the results in this paper: $d\_step \in \{0, 0.001, 0.01, 0.1\}$ ($d\_step = 0$ corresponds to no discretization), $\tau \in \{0.01, 0.1, 0.5\}$ and $\delta \in \{0.005, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9\}$.

4.3 Results

Considering the mean value over 3 replicas and the boolean production output, ICC_MB groups all genes in the same cluster except when discretization with step 0.1 is used in which case two clusters are obtained: one with 8830 genes and the second one with 18 genes. ICC_MB_PCC and ICC_MB_RCS further produce a group G_0 for which the size depends on the δ value.

The results obtained considering the real value of production as output are overall better compared to those obtained for the production growth boolean value.

Without discretization, ICC_MB puts all the genes in the same group (similarly with ICC_MB for Mean12_Bool). However, when $d\_step = 0.1$, ICC_MB reports 7 clusters where majority of genes is in the first cluster and the other 6 groups are formed by fewer genes. The result is still poor as the size of one group is too large compared to the rest. This is emphasized by the validation phase (particularly of ICC_MB_RCS) which results in a group G_0 containing many uncorrectly clustered genes from the big size cluster.

To be more specific, ICC_MB_PCC produces a G_0 group which contains from 0 to 8604 genes depending on the value of δ. Again, for $d\_step = 0.1$ best results are obtained: 7 clusters and G_0 with 1590 genes.

ICC_MB_RCS also gives up to 7 clusters depending on parameters τ and δ used in the RCS measure and the cluster validation. Furthemore, the discretization step highly influences the results. When no discretization is used, all genes go in one cluster and the size of G_0 increases with lower values of δ. For discretization step lower than 0.1, two clusters are formed: one with very high number of genes (from 8837 to 8754) and the other with very low number of genes (from 11 to 94). For $d\_step = 0.1$ genes are grouped in 7 clusters and a G_0 group for which the size depends on δ (see Table 1).

From Table 1, it can be seen that for $\delta = 0.005$ the size of G0 is rather large at over 7500 genes whereas at δ 0.5 and 0.9 all clusters are validated by RCS

Table 1. ICC_MB_RCS ($d\_step = 0.1$) results for Mean12_RealProd with different τ and δ values.

τ	δ	Clusters							G_0
		G_1	G_2	G_3	G_4	G_5	G_6	G_7	
0.01	0.005	941	14	110	5	1	17	8	7752
0.01	**0.1**	**4361**	**16**	**357**	**5**	**1**	**17**	**8**	**4083**
0.01	0.5	8207	215	395	5	1	17	8	0
0.1	0.005	1130	73	97	1	1	5	8	7533
0.1	0.1	6832	195	273	5	1	7	8	1527
0.1	0.9	8207	215	395	5	1	17	8	0

regardless the value of τ. The most balanced results are obtained for $\tau = 0.01$ and $\delta = 0.1$.

Considering the relation of gene values with the growth value output results in more clusters of genes in all methods compared to the Mean12_RealProd where the real production value was considered.

Without discretization, ICC_MB puts all the genes in the same group (same as for Mean12_RealProd). When $d\_step = 0.1$, ICC_MB reports 31 clusters (as opposed to 7 clusters for Mean12_RealProd). The majority of genes go in the first cluster and the other groups are formed by fewer genes (similar behavior with ICC_MB for Mean12_RealProd).

ICC_MB_PCC obtains similar results with ICC_MB except that it also produces the G_0 group which contains from 0 to 8584 genes depending on the value of δ. Again, for $d\_step = 0.1$ best results seem to be obtained: 31 clusters and G_0 with 631 genes at $\delta = 0.25$.

ICC_MB_RCS reports 1 to 31 clusters depending more on the discretization step rather than on the τ parameter (used in the RCS measure) and δ (used in the cluster validation). When no discretization is used, all genes go in one cluster. For discretization step of 0.0001, three clusters are formed: one with very high number of genes (8846) and the other two groups with one gene each. For discretization step of 0.01, 12 clusters are formed: one with very high number of genes (8818) and the other 11 groups having 1 to 8 genes each. For discretization step of 0.1, 31 clusters are formed: one with high number of genes (7731) and the other 30 groups having among 1 and 317 genes each.

4.4 Discussion and Biological Perspective on the Results

Experiments have shown that the model variant ICC_MB is not able to provide any clustering in most scenarios considered. The inclusion of a validation phase (based on either RCS or PCC) is crucial in obtaining a more reliable clustering result starting from ICC_MB. Figure 1 emphasizes the different results obtained by ICC_MB compared to the ICC_MB_PCC where a validation phase is included and also the difference in results between a validation based on PCC and the other based on RCS. A triangular matrix is created as follows: for each pair of genes (g_i, g_j) associate value 0 (corresponding to white color) if none of the

two methods grouped genes g_i and g_j together, value 0.5 (corresponding to grey color) if only one of the methods put the two genes in the same group and a value of 1 (corresponding to black color) if both methods produced the same grouping result. Clearly, the RCS provides different kind of groups by checking the rate of change in the gene expression values.

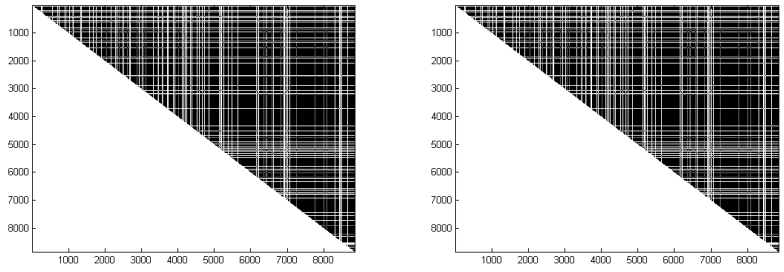

Fig. 1. Comparison of model variants with and without validation: (left) ICC_MB vs. ICC_MB_PCC, and (right) ICC_MB_RCS vs ICC_MB_PCC.

Although the computational results are encouraging, their biological utility is limited due to the big size of resulting groups and the lack of co-expression between the genes of each group. It is known that functionally related genes tend to have similar expression values [14] and hence, the possibility of obtaining groups with a common expression profile is of great interest because it enhances the biological significance. However, it is important to emphasize that gene ranking and selection measures help to identify genes that are involved in the production and growth processes. Therefore, the combination of co-expression and gene ranking approaches could be beneficial because (i) the size of groups is reduced based on a co-expression measure, (ii) genes are ranked based on the growth and production values and (iii) the biological significance is improved based on the assumption in which related biological processes have similar expression patterns.

5 Conclusions and Future Work

The task of gene clustering and selection in connection with a real-world time series microarray problem has been investigated. Several methods based on information theory methods are developed and analysed. Experiments show a poor performance of measures such as ICC in the ability to meaningfully cluster the genes in the considered dataset. However, the importance of validation by similarity measures is clearly emphasized through the comparisons performed.

Future work focuses on development and investigation of methods able to provide gene groups based on the distance between gene expression levels and the correlation with the output.

Acknowledgments. This research has been partially supported through the projects of the Junta de Castilla y Leon CCTT/10/BU/0002 and the projects from Spanish Ministry of Science and Innovation PID 560300-2009-11 and TIN2011-24302.

References

1. Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.
2. Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, and Victor Robles. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006.
3. Chien-Pang Lee and Yungho Leu. A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing*, 11:208–213, 2011.
4. Shyamal D. Peddada, Edward K. Lobenhofer, Leping Li, Cynthia A. Afshari, Clarice R. Weinberg, and David M. Umbach. Gene selection and clustering for time-course and doseresponse microarray experiments using order-restricted inference. *Bioinformatics*, 19(7):834–841, 2003.
5. Jason Ernst and Ziv Bar-Joseph. Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7(1):191, 2006.
6. John D. Storey, Wenzhong Xiao, Jeffrey T. Leek, Ronald G. Tompkins, and Ronald W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.
7. Tianqing Liu, Nan Lin, Ningzhong Shi, and Baoxue Zhang. Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments. *BMC Bioinformatics*, 10(1):146, 2009.
8. Huawen Liu, Lei Liu, and Huijie Zhang. Ensemble gene selection by grouping for microarray data classification. *Journal of Biomedical Informatics 43 (2010) 81–87*, 43:81–87, 2010.
9. Ying Lu and Jiawei Han. Cancer classification using gene expression data. *Information Systems*, 28(4):243 – 268, 2003.
10. Yu Wang, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus F.X. Mayer, and Hans W. Mewes. Gene selection from microarray data for cancer classification—a machine learning approach. *Computational Biology and Chemistry*, 29:37–46, 2005.
11. Hanchuan Peng, Fuhui Long, and Chri s Di ng. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Learning*, 27(8):1226–1238, 2005.
12. Sorana-Daniela Bolboaca and Lorentz Jantschi. Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, (9):179–200, 2006.
13. G.K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–73, 2003.
14. Carlos Prieto, Alberto Risueno, Celia Fontanillo, and Javier De Las Rivas. Human gene coexpression landscape: Confident network derived from tissue transcriptomic profiles. *PLoS ONE*, 3(12):e3911, 12 2008.