34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Shape-Influenced Clustering of Dynamic Patterns of Gene Profiles

Georgia Skreti, Ekaterini S Bei, Michalis Zervakis, *Member, IEEE*

Abstract—Statistical evaluation of temporal gene expression profiles plays an important role in particular biological processes and conditions. We introduce a clustering method for this purpose, which is based on the expression patterns but is also influenced by temporal changes. We compare the results of our platform with methods based on expression or the rank of temporal changes. The proposed platform is illustrated with a temporal gene expression dataset comprised of primary human chondrocytes and mesenchymal stem cells (MSCs). We derived three clusters in each cell type and compared the content of these classes in terms of temporal changes, which can support biological performance. For statistical evaluation we introduce a validity measure that takes under consideration these temporal changes and we also perform an enrichment analysis of three central genes in each cluster. Even though we can detect certain statistical similarities, these might be due to different biological processes. Our proposed platform contributes to both the statistical and biological validation of temporal profiles.

I. INTRODUCTION

The interest on the temporal dynamics of gene expression profiles increases dramatically with the development of methods to deal with large-scale genomic data. The genetic progression of an organism or a disease enables the study of complex biological problems and facilitates the evaluation of therapeutic protocols based on the consideration of the dynamics of molecular mechanisms and drug response. To that respect, the consideration of the temporal profile of genes in a microarray experiment becomes of particular importance and several research attempts have been developed aiming to capture the temporal dynamics of gene expression. In particular, the development of gene-clustering algorithms that also detect temporal profiles is becoming increasingly important. Statistical bootstrap methods have been developed for assigning genes to candidate profiles [1]. Ranked-based considerations of the temporal profile have been proposed in [2], where the time instances are ranked on the basis of the corresponding expression values and are used to describe the temporal profile. Furthermore, biclustering methods have been developed to discover local expression patterns that are consistent in a subgroup of conditions or time instances [3]. Most of the developed methods consider clustering criteria based on some form of coded-shape behavior as opposed to the traditional algorithms that are heavily based on gene-expression patterns.

In ranked-based methods the candidate profiles capture only the shape by characterizing both the type of temporal progression (increase or decrease) and the points of maximum or minimum expression. In biclustering algorithms, the temporal behavior of genes is encoded by symbols and the coherent expression patterns of subgroups of genes in specific time instances (biclusters) are characterized as strings of symbols.

In many biological processes, however, not only the temporal behavior but also the expression profile itself is important for grouping genes as significant for a particular process or condition. In clustering of gene profiles therefore, both the temporal shape and the expression profile must be considered. In our study we attempt to develop tools appropriate for such purposes of clustering based on a dual criterion. More specifically, we develop a criterion based on expression profile, which is also influenced by the shape trend. Following the clustering process, we address two issues related to the comparison of partitions. The first relates to the derivation of correspondences in the two partitions, while the second addresses the validity criterion for comparing the compactness of clusters in these partitions. The purpose of this study is to develop tools for clustering and evaluation of cluster quality based on two criteria, expression profile and coded-shape values. Overall, it addresses the following three issues: 1) Clustering of genes based on their expression profile but also influenced by the temporal shape of this profile; 2) Similarity Criterion for matching similar clusters across partitions based on shape; and 3) Validity index of a partition that captures the temporal shape of gene profiles.

In Section II we provide the notation used and present our clustering approach in Section III, whereas the proposed validity index is introduced in Section IV. The results are presented and discussed in Section V.

II. EXPRESSION PROFILE AND CODED-SHAPE OF TEMPORAL BEHAVIOR

For each gene, the expression pattern in time is encoded in a N dimensional vector $\underline{x}$, where N represents the extent of the time course of interest. The graphical representation of expression values $x_i,\ i = 1,\ldots,N$ throughout the time course of interest defines the temporal profile of each gene. Furthermore, the difference of expressions between two consecutive time points defines the shape of each gene. We utilize a coded form of the shape that encodes, in a ternary form, the changes from one to the next time point. Let

*Research supported by "YPERThEN" project, which is funded by the EU and funds from Greece and Cyprus, and by "OASYS" project funded by the NSRF 2007-13 of the Greek Ministry of Development.

G. Skreti, E. S. Bei, and M. Zervakis are with the Department of Electronic and Computer Engineering, TUC, Chania 73100, Greece, (gskreti@isc.tuc.gr, abei@isc.tuc.gr, michalis@display.tuc.gr).

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

$\delta_i = x_{i+1} - x_i$, $i = 1,...,N-1$. Then, by binning the difference values in the set {-1, 0, 1} via a threshold Δ, we create the coded shape vector $\underline{v}$ of dimensionality N-1. Thus, in our case a coded-shape string is composed of digits that can be used in an arithmetic computations rather than symbols.

Suppose we aim at C clusters S_t, $t = 1,...,C$, with the mean vector of cluster S_t be denoted as $\underline{\mu}_t$. The coded-shape of this vector is also derived by the coding operator *c{.}* and is denoted as $m_t = c\{\mu_t\}$. Let a member of S_t with expression pattern $\underline{x}_j \in S_t$, which happens to have a coded-shape pattern $\underline{v}_j$. The size N-1 of the shape patterns, as well as the ternary discretization of values imply the existence of only a limited number of code vectors $M = 3^{N-1}$. Thus, each cluster may contain a number of coded-shape patterns, numbered from *1,..,M*.

III. Clustering Methodology

The proposed methodology for clustering expression profiles with the influence of their shape profiles is based on a modification of the self-organizing map (SOM) methodology composed of three parts. The first step derives a number of clusters based on the SOM organization of the expression vectors into *Q*x*Q* nodes. Then the nodal weight-patterns of SOM are organized into C clusters based on a K-means iterative approach with a criterion that is influenced by shape. Finally, each sample is assigned to one of C classes based on the distance of its expression vector from the mean of each cluster. Our contribution to this methodology is the criterion used in the second step of organizing the SOM nodes. More specifically, the first step derives a number (*Q*x*Q*) of clusters based on the expression profiles. We proceed in a second grouping that organizes nodes more closely, deriving a small number of clusters. At this point we propose to use information about the temporal trend in order to favor re-grouping of clusters with similar performance. The overall clustering approach is summarized in the following.

<u>First step: Self-Organizing Feature Map</u>

Repeat SOM iterations until convergence, when the absolute squared weight changes is smaller than 0.02 over 2500 epochs. In our application we use 10x10=100 nodes and an extra stopping criterion restricts the number of repetitions to 500 the number of nodes, equal to 50000.

<u>Second step: K-means Clustering of SOM Nodes</u>

The K-means algorithm aims to further organize the groups formed by SOM into C clusters based on the minimization of a distance criterion between the samples (SOM nodes) and the cluster means, so as to minimize the overall variance of each cluster. The K-means scheme operates on an iterative form based on the expectation-maximization (EM) solution. The proposed criterion in this step is defined as the overall l_2 norm of expression-vector differences:

$$Q = \sum_{t=1}^{C} \sum_{w_j \in S_t} a_j \left\| \underline{w}_j - \underline{\mu}_t \right\|_2^2$$

where $\underline{w}_j$ is a sample (expression vector of a SOM node) in the class *t* with class mean $\underline{\mu}_t$ and α_j is a weighting factor depending on the coded shape of $\underline{w}_j$ compared to that of the class mean $\underline{\mu}_t$.

More specifically, $a_j = \dfrac{1}{1+e^{-6(r_j-1)}} + 1$ is the logistic function with range [1, 2] evaluated for r_j. This factor reflects the coded difference $r_j = \left\| \underline{v}_j - \underline{m}_t \right\|_1$ as the l_1 norm of the coded-shape differences corresponding to the sample and the class mean, where $\underline{v}_j = c\{\underline{w}_j\}$ and $\underline{m}_t = c\{\underline{\mu}_t\}$ are the corresponding coded values. Notice that r_j ranges in [0, 2], with 0 and 2 being the cases of no difference and max difference in all digits of the code vector. In case of no difference, the shape factor α_j becomes 1, whereas in max difference in all digits this factor becomes 2, thus increasing the expression-based distance. Considering this criterion, the EM optimization proceeds in discrete steps towards the evaluation of new cluster means and the re-assignment of samples in classes. At each step, for certain samples assigned in a cluster, the sample mean is re-computed as:

$$\underline{\mu}_t = \frac{1}{|S_t|} \sum_{w_j \in S_t} a_j \underline{w}_j$$

Subsequently, samples are evaluated and re-assigned based on the minimum distance from the class means, i.e. $\min_t \ \hat{a}_{j,t} \left\| \underline{w}_j - \underline{\hat{\mu}}_t \right\|_2^2$, with the re-computed shape factors $\hat{a}_{j,t}$ for the sample *j* and the tested mean vector of the class *t*.

<u>Third step: Assignment of initial expression values into clusters</u>

Notice that the second step organizes the nodes of the SOM network; the initial expression vectors still need to be assigned in one of the C clusters. This assignment is performed based on the minimum l2 norm of the difference between each expression vector $\underline{x}_i$ and the class means $\underline{\mu}_t$ computed from the 2nd step.

IV. Criteria for Partition Evaluation

In computational biology there is an often need to compare two (or more) partitions. In particular, we need to find correspondences between the partitions but also need to compare the quality of each partition in terms of compactness and discrimination of its clusters. For the first aspect we need to compare one cluster of the first partition with all clusters of the other partition. The second task needs to consider the distribution of patterns within and across clusters in each partition. For our computational needs, we have two possible quantities available, i.e the sample values and the probabilities of samples. In our application we consider samples coded by their shape codes as ternary strings of size N-1 numbered from 1 to M=3^{N-1}.

A. *Cluster Matching based on shape profile*

Each cluster C_i has been assigned a number L_i of samples, each with a corresponding coded-shape vector. The distribution of these shape vectors results in a code histogram for numbered codes *1,…,M* with probabilities *p[1],…,p[M]*. The cluster similarity metric can be based on the difference of code histograms. In the matching process, each cluster of one partition is mapped to the best matching cluster of the other partition. Thus, let two clusters one with numbered-code probabilities *p[1],…,p[M]* and the other with *q[1],…,q[M]*. The similarity of the two clusters is the summed absolute distance of the two probabilities:

$$D_c = \sum_{i=1}^{M} \left| \, p[i] - q[i] \, \right|$$

B. *Validity Index based on shape profile*

The proposed index is based on the compactness of each code histogram as well as the distance between pairs of code histograms. Due to the probability distribution or weighting form of the histogram the larger histogram bins play much more important role in comparison than smaller bins. Consequently, we now rank the probabilities in descending order to obtain the ranked probabilities $p_1,…,p_M$. Ties in ranking are resolved in favor of the minimum Hamming distance from the previous string.

Suppose we have two clusters C_1 and C_2 of sizes L_1 and L_2, respectively. The first has ranked probabilities p_i corresponding to strings s_i, where the second has q_i corresponding to strings t_i, *i=1,…,M*, respectively. In order to built an inner-cluster compactness index we will use the Hamming distance between two strings in descending ranked order, weighted by the probability of the second string. This represents the distance between two strings obtained from the most to least significant ones. The within-cluster distance signified by the most significant string is zero. Considering the next-ranked string, the codeword distance introduced is the difference of codes in as many cases as signified by the probability of this string. In general, for the *i*th-ranked string, the distance introduced can be computed from the codeword distance from the previous string weighted by the probability of the *i*th string. Thus, for within cluster distance we have:

$$Q(C_1) = \sum_{i=1}^{M-1} d(s_i, s_{i+1}) p_{i+1} \quad \text{and} \quad Q(C_2) = \sum_{i=1}^{M-1} d(t_i, t_{i+1}) q_{i+1}$$

where *d(.,.)* signifies the Hamming distance between two strings.

In order now to introduce across-cluster distances, we first consider the differences of clusters for each string. Thus, for each numbered code we calculate $\textit{Numbered}$ probabilities $r[i] = | \, p[i] - q[i] \, |$ and derive the associated $\textit{Ranked}$ probabilities $r_1, r_2, \ldots, r_M$.

We have now defined a histogram of the difference or probability intersection of the two clusters C_1 and C_2, whose compactness index can be defined as before:

$$Q(C_1, C_2) = \sum_{i=1}^{M-1} d(z_i, z_{i+1}) r_{i+1}$$

These indexes reflect the distance of digitized codewords signified by their probabilities. Their values are within a range *[1, K]* scaled by *(1-r_1)*. The minimum value of the index is zero and is obtained when the histogram has only one point with probability one. The maximum value is attained when the histogram involves two codewords with probability ½ each and distance equal to the maximum value of K.

Thus, the ratio of within cluster to across-cluster metrics signifies a relevant validity index for each partition P of C_P clusters referred to as Ranked Shape Index (RSI), which accounts for shape similarities, as:

$$RSI\{P\} = \frac{1}{C_p} \sum_{i=1}^{C_p} \max_{j \neq i} \left\{ \frac{Q(C_i) + Q(C_j)}{Q(C_i, C_j)} \right\},$$

$$i, j = 1, \ldots, C_p.$$

V. Examples

Experimental setup

We applied our platform to the dataset introduced by Bernstein et al. [4] that comprised of pellet culture-conditioned human primary chondrocytes, and human bone marrow-derived MSCs [4]. Their gene-expression profiles were analyzed and compared at 4 different days.

Implication of Validity Indices

The appropriate number of clusters has been considered both manually and using the concept of validity indices. In manual consideration we examined the shape of temporal profiles and also the properties of clusters formed in terms of expression values and temporal changes of their samples. We concluded that the problem under consideration induces 3 types of gene performances (clusters), which are illustrated by their temporal expression profiles in Figure 1. Segment a) illustrates our proposed mixed clustering scheme, whereas section b) demonstrates the expression-based clustering, for the cases of MSCs (upper part) and chondrocytes (lower part). The clustering in this figure presents the prominent two shapes (expressing the largest probabilities) of each cluster of the proposed approach influenced by coded shape, but it is indicative of the performance trends observed in the databases.

The traditional validity indices, like Davis-Bouldin (DB) or Dunn's index, fail to provide a consistent argument regarding the preferred number of clusters. In particular, the DB index that somewhat resembles the philosophy of RSI at the level of expressions, achieves the lowest values for 2 clusters in the proposed mixed clustering scheme for both cell types. For the ranked-based clustering scheme, it fluctuates and is minimized for quite large numbers of clusters. Finally, for the clustering scheme based on expression values, the DB index is minimized for 3 and 4 classes in the case of MSCs and chondrocytes, respectively.

In contrast, the proposed RSI index shows robust performance and indicates an optimal number of 3 clusters in most approaches considered. More specifically, for clustering based on expression values and the proposed scheme of combined-purpose clustering influenced by shape, the RSI index is minimized at 3 classes for both cell type databases. The situation is slightly different for the ranked-based clustering, as discussed in the following, with the RSI obtaining smaller values at 2 classes. In this case, the RSI index is minimized at a large number of classes (equal to 8), in which case the prominent eight classes in terms of coded-shape patterns appear separately in a single class. The RSI index is illustrated in Table 1, along with the corresponding DB index (in parentheses).

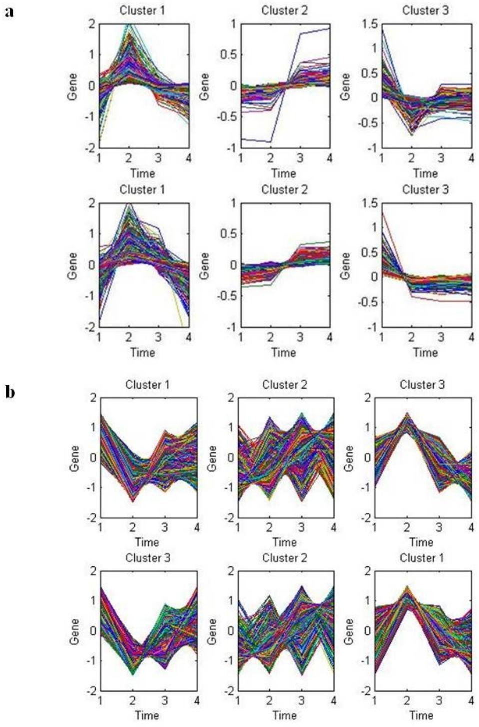

Figure 1. Expression profiles in three classes for the cases of MSCs (upper part) and Chondrocytes (lower part); a) proposed mixed-based and b) expression-based clustering

It is worth mentioning the RSI index favors the combined clustering scheme, whereas the DB index favors the traditional clustering based on expression values. By means of the prominent shapes appearing in the clusters of the proposed methodology in Figure 1, we can also visually indicate that the proposed method results in more tight clusters than the traditional expression-based clustering and this is better reflected on the RSI than the DB index. In this figure we plot the expression pattern over time (4 days) for the two most often appearing trends in each cluster. We present the graphs for the proposed clustering scheme and the traditional approach based on expression. The latter (Figure 1b) operates on the basis of a symmetric distance function (*l2* norm) so that it is forced to include symmetric patterns as in the case of cluster 2. On the other hand, the proposed approach (Figure 1a) also considers the temporal trend, so that it only favors similar-shape performance over time. The issue of symmetry is also encoded into the traditional validity indices, such as the DB, which favors the latter scheme without considering the biological difference implied by opposite temporal trends that result in symmetric distances from the class centroid.

***RSI* Index for each Clustering Methodology**	**Number of Clusters**		
Combined clustering	*2 clusters*	*3 clusters*	*4 clusters*
MSCs	1.1546 (1.1839)	0.9118 (1.3646)	1.3065 (1.2334)
Chondrocytes	1.0058 (0.9997)	0.8906 (1.0553)	1.1141 (1.1794)
Rank-based clustering	*2 clusters*	*3 clusters*	*4 clusters*
MSCs	1.0648 (1.7126)	1.5489 (1.9712)	1.6220 (1.7621)
Chondrocytes	0.9066 (1.3281)	0.9104 (2.1713)	1.3674 (1.5124)
Expression clustering	*2 clusters*	*3 clusters*	*4 clusters*
MSCs	1.2793 (1.1818)	1.1119 (0.9257)	1.4488 (1.4667)
Chondrocytes	1.1602 (0.9906)	1.1089 (0.9780)	1.3014 (0.9065)

Table 1. RSI index for various approaches and cell types; the corresponding DB index appears in parentheses

Overall, our study shows good similarity of classes in MSCs and chondrocytes data, which verifies the results of the original study [4]. However, the cross-class samples need further consideration, since the intermingling of classes is quite heavy, beyond the tolerance of errors due to measurement noise. In a further attempt to compare the two cell types on deeper biological bases, we compare three central genes of MSCs and chondrocytes, in each pair of matched classes, through a gene enrichment analysis. The matched clusters of the two cell types express similar temporal trends, but the biological processes responsible for this performance is different in many aspects, requiring deeper biological interpretation.

Concluding, our proposed platform comprised of a clustering method, similarity criterion and validity index, all based on temporal changes. It provides a consistent tool that facilitates the statistical validation but also the biological evaluation of temporal profiles.

REFERENCES

[1] S. Peddada, E. Lobenhofer, L. Li, C. Afshari, C. Weinberg, and D.M. Umbach, "Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference", *Bioinformatics*, vol. 19, pp. 834–841, 2003.

[2] S-G. Yi, Y-J. Joo and T. Park, "Rank Based Clustering Analysis for the Time-course Microarray Data", *Bioinformatics and Comp. Biology*, vol. 7, no.1, pp. 75-91, 2009.

[3] A. Carreiro, O. Anunciaᴄ̧, J. Carriᴄ̧, S. Madeira, "Prognostic Prediction through Biclustering-Based Classification of Clinical Gene Expression Time Series", *Journal of Integrative Bioinformatics*, vol. 8, no. 3, 2011.

[4] P. Bernstein, C. Sticht, A. Jacobi, C. Liebers, S. Manthey, M. Stiehler, "Expression pattern differences between osteoarthritic chondrocytes and mesenchymal stem cells during chondrogenic differentiation," *Osteoarthritis Cartilage*, vol. 18, no. 12, pp. 1596-607, 2010.