Mining High Quality Clusters of SAGE Data

George Tzanis
Aristotle University of Thessaloniki
Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
+302310998433
gtzanis@csd.auth.gr

Ioannis Vlahavas
Aristotle University of Thessaloniki
Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
+302310998145
vlahavas@csd.auth.gr

ABSTRACT

Serial Analysis of Gene Expression (SAGE) is a method that allows the quantitative and simultaneous analysis of the whole gene function of a cell. One of the advantages of this method is that the experimenter does not have to select a priori the mRNA sequences that will be counted in a sample. This makes SAGE a powerful tool for analyzing gene expression and studying various diseases, such as cancer. An important concern in cancer studies is the discovery of the differences between healthy and cancerous samples and the accurate separation of these two groups of samples. However, the high dimensionality of the data, the multiple cell sources (i.e. bulk and cell line) and the multiple cancer subtypes make very difficult the effective clustering of SAGE libraries. Furthermore, the various sources of noise pose an extra challenge to data miners. For all these reasons we propose an approach that involves the discretization of the data, the selection of the most prominent gene tags and the use of a clustering algorithm in order to obtain more compact and reliable clusters that can assist cancer profiling. We experimented with two families of clustering algorithms, partitional and hierarchical, and we utilized various cluster validity criteria in order to evaluate the resulted clustering structures. The experimental results have shown that our approach provides more interesting clustering structures.

1. INTRODUCTION

The last decades the progress in the fields of biology and computer science is quite remarkable. This progress has lead to the introduction of new technologies as well as the improvement of existing ones that made possible the conduction of many and often large-scale experiments. As a consequent large amounts of data, that sometimes are highly complex, are produced in dramatically less time. Biologists demand the tools to organize, maintain, and analyze these data. The fields of data mining and machine learning provide biologists, as well as experts from other areas, a powerful set of tools to analyze new data types in order to extract various types of knowledge fast, accurately and reliably. The use of these tools in biology promises to discover new knowledge and enlighten the molecular and cellular details that govern life.

VLDB '07, September 23-28, 2007, Vienna, Austria.

Every living organism depends on the activities of a complex family of molecules, namely proteins. They are the main structural and functional units of an organism's cell. Two other important molecules, DNA and RNA have the role to carry the genetic information of the organisms. The genetic information that is coded in DNA flows towards the proteins via the processes of transcription and translation. In particular, DNA is transcribed into mRNA (messenger RNA) and then mRNA is translated into proteins. Each organism contains a number of genes that are coding segments of DNA that code the synthesis of an mRNA or protein molecule. Although every cell in an organism contains the same set of genes, two cells may have very different properties and functions. This is due to the differences in abundance of proteins. The abundance of a protein is partly determined by the levels of mRNA which in turn are determined by the levels of the expression of the corresponding gene. Changes in gene expression underlie many biological phenomena.

As implied in the previous paragraph the study of gene expression levels may guide to very important findings. One of the basic aims of gene expression data analysis is to discover differences between the gene expression profiles of diseased and healthy tissues. The last years, two major families of techniques that permit to measure gene expression levels have emerged. The first family consists of techniques based on sequencing (including ESTs and SAGE [1, 21] and the second family consists of techniques based on hybridization procedures (DNA arrays [20]). In this paper we will focus on the first family and especially in SAGE.

SAGE (Serial Analysis of Gene Expression) is a method invented in John Hopkins, Baltimore, USA, in order to provide the quantitative and simultaneous analysis of the whole gene function of a cell [21]. The method works as shown in Figure 1. All the mRNA transcripts of a cell are collected and a short sequence of about ten nucleotides (RNA and DNA are sequences of smaller molecules called nucleotides) called tag is extracted from each transcript. The tags are linked together in a single chain and they are sequenced. Then, the frequency of each tag is counted, so that the relative levels of the corresponding mRNAs and consequently the gene expression levels are determined. The set of all tag counts in a single sample is called a SAGE library.

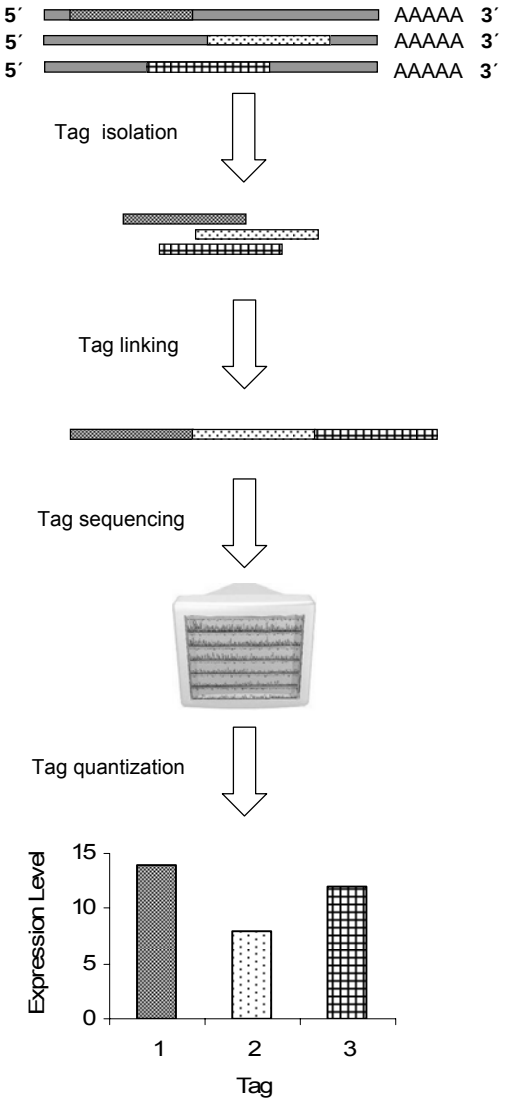

Figure 1. Serial analysis of gene expression.

1.1 Motivation

SAGE is an expensive technique compared to DNA arrays. This disadvantage is the reason that there are not very much publicly available data and consequently there are not too many studies focused on SAGE data analysis. Recently, in 2004 and 2005, the two ECML/PKDD Discovery Challenge Workshops gave a boost on SAGE data mining by providing two SAGE datasets for study. However, although some approaches have been proposed, there are many directions for improvement. An important goal of gene expression data mining is to discover the differences between diseased and healthy tissues. The existence of many different cancer types and different cell sources (bulk and cell line) makes the accurate discrimination of cancerous from normal samples quite difficult. These difficulties expose a great challenge to data mining community and demand improved clustering methods for grouping cancerous samples together and separating them from normal ones. This challenge motivated our work and guided us to an effort to define a better approach that provides improved clustering structures. Moreover, the main advantage of the SAGE method is that the experimenter does not have to select the mRNA sequences that will be counted in a sample. This is quite important, since the appropriate sequences for studying various diseases such as cancer may not be known in advance. In contrast, in DNA array methods, the experimenter by selecting the mRNA sequences introduces a bias to the experiment. This advantage of SAGE makes it a fairly promising method, especially for cancer studies as the one presented in this paper.

1.2 Contribution

We propose an approach for selecting the most prominent features (gene tags) in order to use them and cluster the SAGE libraries according to their cell state (cancerous or normal). The aim of this approach is to discover more compact clusters that group cancerous samples away of the normal ones, in order to assist cancer profiling. On this basis, we provide a thorough study on clustering of SAGE libraries using partitional as well as hierarchical clustering algorithms. We evaluate cluster validity using external criteria based on partitions that are built according to our prior knowledge about the partitioning of the data. As shown by the experimental results, our approach provides better clustering structures, which separate more clearly the cancerous from the normal samples. Moreover, the significant reduction of the feature space that is achieved in most cases leads to considerable improvement of efficiency in terms of time and memory usage.

The paper is outlined as follows. In the next section we provide a concise review of the relative literature. In section three we give a detailed description of our approach. Next, in section four, we present the datasets we used and define our experimental setup. Then, we present our results in section five and finally, we conclude in section six.

2. RELATED WORK

Most of the clustering studies of gene expression focus on clusters of genes and not on clusters of samples as is the case in our study. Moreover, due to the limited availability of SAGE datasets the most studies are based on microarray data [4, 6, 17]. However, the last years a number of studies on clustering SAGE libraries have been presented. Other data mining tasks, like classification and association analysis have recently been applied to SAGE data.

In [8] supervised and unsupervised learning methods were utilized for the analysis of SAGE data. In particular decision trees (C4.5) and support vector machines were used to classify the data according to cell state (normal or cancerous) and tissue type (colon, brain, ovary, etc.). Furthermore, the authors studied hierarchical clustering for identifying different subclasses of tumors and normal tissues.

Hierarchical clustering methods were also applied in [18] for detecting similarities and dissimilarities among different types of cancer. The authors utilized various preprocessing techniques, including error removal, normalization, missing tag imputation, and subspace selection based on Wilcoxon test. Their results shown that the SAGE libraries are grouped according to tissue type and cell state and revealed a possible relation between brain and ovarian cancer.

An approach for cluster analysis of SAGE data using Poisson statistics has been presented in [5]. They proposed two Poisson-based distances and experimented with simulated and experimental mouse retina data. Their results saw that the Poisson-based distances are appropriate and reliable for analyzing SAGE in comparison to other commonly used distances or similarity measures such as Pearson correlation or Euclidean distance.

Another approach for cluster analysis of SAGE data using Poisson statistics as well as self-organizing feature maps has been presented in [22]. The authors proposed two novel clustering algorithms, PoissonS and PoissonHC, for SAGE data analysis that are based on the adaptation and improvement of self-organizing maps and hierarchical clustering techniques. They have tested the proposed algorithms on synthetic and experimental SAGE data. Their results indicate that, the two algorithms and a hybrid approach based on the combination of these two algorithms, offer significant improvements in pattern discovery and visualization for SAGE data.

Rioult [19] presented an approach for mining strong emerging patterns from wide (too many features) SAGE datasets. The approach is based on the transposition of data matrix and the use of Galois connection for inferring the closed sets of the original matrix and finally deriving the strong emerging patterns.

Becquet et al. [3] applied association rules mining to SAGE data. In this study various discretization methods for transforming the data matrix to a Boolean context are described. Depending on the discretization algorithm used, different properties of the data were highlighted. In all cases the extracted collections of rules indicated a very strong co-regulation of mRNA encoding ribosomal proteins. Frequent closed itemset mining algorithms, have been used for SAGE data analysis [10]. In [13] a method to extract all the δ-strong characterization rules from SAGE data is proposed. The authors discuss the potential impact of these rules to characterize cancer versus no cancer biological situations.

The effect of dimensionality reduction methods for supervised learning is studied in [2]. They compare filtering and wrapper methods. They concluded that typical filtering approaches negatively impact the predictive accuracy of classifiers as well as that many groups of genes that are not differentially expressed may contribute critically to classification.

Martinez et al. [16] study the effect of data cleaning and discuss the process of the attribution of a tag to a gene. They use principal components analysis and hierarchical clustering methods. They conclude that the comparisons of cancers from various tissue types is a particularly difficult task, as tissue samples cluster according to tissue origin and not according to cell state (normal or cancerous).

In [14] an analysis of SAGE data using various feature ranking, classification, and error estimation methods is presented. Their results show that the support vector machine with the stump kernel performs well on SAGE data. They have also concluded that a feature set of about 500 to 1000 features is adequate for predicting the cancerous state of a sample.

A hybrid system based on genetic algorithms and artificial neural networks was proposed for classifying SAGE data [7]. The system works by selecting a compact set of genes predicting cancerous and normal cell states.

3. OUR APPROACH

In this section we provide a detailed description of the proposed approach. Before presenting the basic steps of this approach (discretization, feature-gene selection and clustering) we will describe the structure of the input data.

The data are structured in a gene expression matrix *A*. The columns of the matrix represent the tags of the genes and the rows represent the different samples (SAGE libraries). The intersection of the i^{th} row with the j^{th} column, namely the element a_{ij}, is the gene expression level for the gene *j* in the sample *i*.

3.1 Discretization

In this step the data are discretized. The discretization procedure followed in our approach is used in order to detect the strong *under-expressions* (expressions of genes that are significantly lower than the mean gene expression) or *over-expressions* (expressions of genes that are significantly higher than the mean gene expression) of genes. The intuition behind this procedure is that the extreme expressions of genes, namely the over-expressions and under-expressions, may carry important information, which can be utilized for enhancing clustering. Before the discretization step it is not necessary to apply any specific normalization process. The SAGE libraries that are used in our study contain almost the same number of tags. One of the main advantages of the SAGE technique is that, since it relies only on a sampling process, it is "self-normalized" and therefore a SAGE library can directly be compared to other SAGE libraries [9].

The discretization process works as follows. For the data matrix *A* we calculate a 99% confidence interval for the expression levels of each gene. So, for each gene *j* we get a confidence interval [left(*j*), right(*j*)]. Then we create a new matrix A', where $a' \in \{-1,0,1\}$. These values are assigned as follows:

- $a'_{ij} = -1,$ if $a_{ij} <$ left(*j*) AND $a_{ij} \neq 0$
- $a'_{ij} = 0,$ if $a_{ij} \in$ [left(*j*), right(*j*)]
- $a'_{ij} = +1,$ if $a_{ij} >$ right(*j*)

Assigning the value of -1 to a'_{ij}, means that gene *j* is significantly under-expressed in the sample *i*. Similarly, an assignment of +1 to a'_{ij}, means that gene *j* is significantly over-expressed in the sample *i*. Finally, a value of zero assigned to a'_{ij} means that there is not a significant under-expression or over-expression.

The term "$a_{ij} \neq 0$" is used in order not to consider zero values as under-expressions. The rationale behind this is twofold. First, a zero value means that a tag is not found in a sample, so the corresponding gene is not just under-expressed, but it is not expressed at all. As mentioned in [18], biochemists believe that the vast majority of genes in the human genome are only expressed in one tissue type, and only some "housekeeping genes" are expressed in all cells. According to this consideration, it is very probable that a gene with zero expression level in a particular sample is not expressed in the tissue type from which

the sample was taken. So it would be inaccurate if we considered it as an under-expression. Moreover, there are too many zeros in gene expression matrices. If we consider zeros as under-expressions, the valuable under expressions will be lost among the vast majority of inaccurate zero under-expressions.

Matrix A' is the input to the next step that involves selecting the relevant genes for clustering.

3.2 Gene Selection and Clustering

In this step we use the discretized gene expression matrix A' in order to select the most prominent tags for clustering SAGE libraries according to their cell state (cancerous or normal). The criterion of the selection is the *homogeneity* of the over-expressions and under-expressions of each gene. For each gene j we calculate its homogeneity H_j as follows:

$$H_j = \left| \frac{\sum_i a'_{ij}}{\sum_i \left| a'_{ij} \right|} \right|.$$

The nominator of the above fraction is the sum of the values of column (gene) j in the discretized matrix and the denominator is the sum of the absolute values of column j in the discretized matrix. The following examples make more clear how the homogeneity is calculated. Let us consider two genes. The first one has 5 values equal to -1 and 20 values equal to 1, whereas the second one has 11 values equal to -1 and 9 values equal to 1. The zero values, which represent the absence of an over-expression or under-expression, are ignored. The homogeneity of the first gene is equal to $|15/25| = 0.6$ and the homogeneity of the second gene is equal to $|-2/20| = 0.1$.

We believe that the genes with higher homogeneity are more useful for clustering. Although this is a reasonable idea and is, more or less, confirmed by our experiments, we do not propose a rule such as "*select the genes that have homogeneity larger than a value h*". Instead of this, we propose the use of a criterion like *silhouette value* in order to evaluate the clustering structures obtained using genes with different values of homogeneity. In particular, in our approach we divide the range of homogeneity into a number of intervals (i.e. [0, 0.1), [0.1, 0.2), …, [0.9, 1]). For each interval we select only the genes that have a homogeneity value inside this interval. Then we apply a clustering algorithm on the initial data matrix A (not the discretized one) using only the selected genes and we get a clustering structure. We may get more than one clustering structure for an interval by applying many algorithms or the same algorithm with different parameters. This procedure is repeated for all intervals. Then, we evaluate each clustering structure and select the one with the highest mean silhouette value.

4. EXPERIMENTAL SETUP

In this section we define our experimental setup. First, we present the datasets that we experimented with. Then we present the clustering algorithms that were utilized in our experiments and finally, we describe the method we used in order to evaluate our results.

4.1 Datasets

We have used two SAGE datasets in our study. The first one consists of 74 samples (SAGE libraries) and 822 features (tags). The second one consists of 90 SAGE libraries and 27679 tags. From now on we will refer to these datasets as the 74x822 dataset and the 90x27679 dataset respectively. Both datasets have been provided by Dr Olivier Gandrillon's team (Centre de Génétique Moléculaire et Cellulaire de Lyon, France) and have been studied and presented at the ECML/PKDD Discovery Challenge Workshops in 2004 and 2005. The SAGE libraries contained in these datasets are publicly available in the SAGEmap website (http://www.ncbi.nlm.nih.gov/SAGE/index.cgi) and have been prepared as of December 2002 [9]. They are collected from various human tissue types (colon, brain, ovary, etc.) and are labeled according to their cell state that is either normal or cancerous.

4.2 Clustering Algorithms

We used the following two clustering algorithms implemented in MATLAB (http://www.mathworks.com):

- *k-means*. A very popular partitional algorithm [15]. The squared Euclidean distance was used as a distance measure. In order to avoid local minima we set the "replicates" parameter to 3, so that the clustering is repeated three times, each with a new set of initial centroids. We experimented with k (the number of clusters) ranging from 2 to 20.
- *Hierarchical algorithm*. This algorithm is an agglomerative hierarchical clustering algorithm and represents another commonly used family of clustering algorithms. We used the Euclidean distance for measuring distances between samples and the ward method (minimum variance algorithm) for linking clusters together.

4.3 Cluster Validity

The cluster validity criteria can be grouped in three basic categories [12]:

- *External criteria*. The results of a clustering algorithm are evaluated on the basis of a pre-specified structure.
- *Internal criteria*. The results of clustering are evaluated using quantities and features that are inherent to the data (e.g. the proximity matrix).
- *Relative criteria*. In this case the clustering results are compared to other clustering structures resulting by the same algorithm but with different input parameter values.

In our setup we utilize some external criteria to evaluate our clustering results. In particular, we use two pre-specified structures-partitions of the SAGE libraries. The first one is based on the cell state of the sample (cancerous or normal) and the second one is based on the cell source of the sample (bulk or cell line). In the following lines we describe in more detail how we used these external criteria.

In order to compare a clustering structure C of the data with a given partition P of the data we define the following terms:

- SS is the number of pairs of samples that both belong to the same cluster of the clustering structure C and to the same group of partition P.

- *SD* is the number of pairs of samples that both belong to the same cluster of the clustering structure *C* and to different groups of partition *P*.
- *DS* is the number of pairs of samples that both belong to different clusters of the clustering structure *C* and to the same group of partition *P*.
- *DD* is the number of pairs of samples that both belong to different clusters of the clustering structure *C* and to different group of partition *P*.

The total number of pairs of samples is: $SS + SD + DS + DD = \frac{n(n-1)}{2}$, where *n* is the total number of samples in data.

We used the following indices to measure the degree of similarity between *C* and *P*:

- *Jaccard Coefficient*: $J = \frac{SS}{SS + SD + DS}$,
- *Folkes and Mallows Index*: $FM = \sqrt{\frac{SS}{SS + SD} \cdot \frac{SS}{SS + DS}}$.

In the next section we will present only the results of the Folkes and Mallows index, since the results are almost identical for both indices.

5. RESULTS

In this section we present the results of the experiments that were conducted according to the setup described in the previous section.

As described in the experimental setup we applied *k*-means with *k* varying from 2 up to 20 on the 74x822 dataset. Each row of Table 1 presents the best clustering structure -according to the mean silhouette value- among the structures that were obtained using only the genes with a homogeneity value inside the interval of the first column. All the intervals except the last one have the same width. The last one expands from 0.7 to 1, because there are not any genes with a homogeneity value inside the interval [0.7, 0.9). For the most of the intervals the clustering structure with two clusters has the highest mean silhouette value. This is in agreement with the intrinsic characteristics of the data, since there two partitions either if we consider a partitioning based on cell state (cancerous and normal) or based on cell source (bulk and cell line). Among all the intervals the best clustering structure was the one of the [0.6, 0.7).

Table 1. The best clustering structures of k-means on the 74x822 dataset.

Homogeneity	Number of Clusters	Mean Silhouette
[0, 0.1)	2	0.5795
[0.1, 0.2)	2	0.4819
[0.2, 0.3)	20	0.4015
[0.3, 0.4)	2	0.4619
[0.4, 0.5)	3	0.6852
[0.5, 0.6)	2	0.9755
[0.6, 0.7)	2	0.9766
[0.7, 1]	3	0.9448

Figure 2 presents the evaluation of the clustering structure that contains two clusters and is obtained using k-means. The y-axis represents the Folkes and Mallows index, whereas the x-axis represents the homogeneity interval that was used in order to select the genes. As shown in the figure, the clustering structure conforms better to the cell state partition than to the cell source partition. The highest value of the index is obtained using the [0.6, 0.7) interval. This agrees with the results of the mean silhouette value and indicates that the approach we follow in order to select the best interval is accurate.

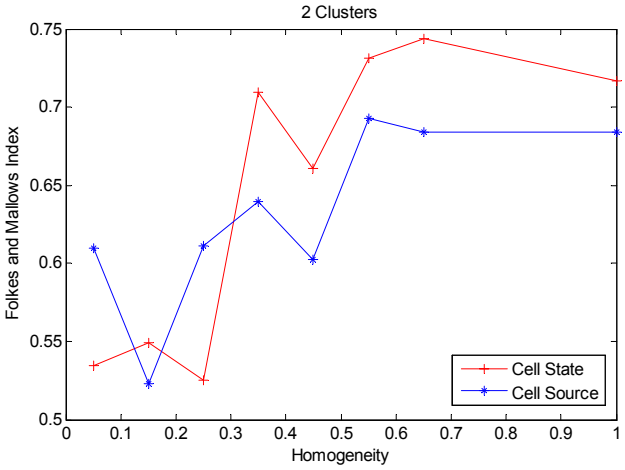

Figure 2. Evaluation of the structure with 2 clusters obtained using k-means on the 74x822 dataset.

In Figure 3 the results obtained using all the features-genes are compared to the results obtained using the selected genes with a homogeneity value inside the interval [0.5, 0.7). The validation is based on the cell state partition. The interval of [0.6, 0.7) that was found to present the best clustering structure for two clusters contains only two genes. Although the results obtained for this interval are very good, two genes are very few for finding compact clusters based on cell state. So we merged this interval with its adjacent [0.5, 0.6) interval that is the second one in the ranking according to mean silhouette value.

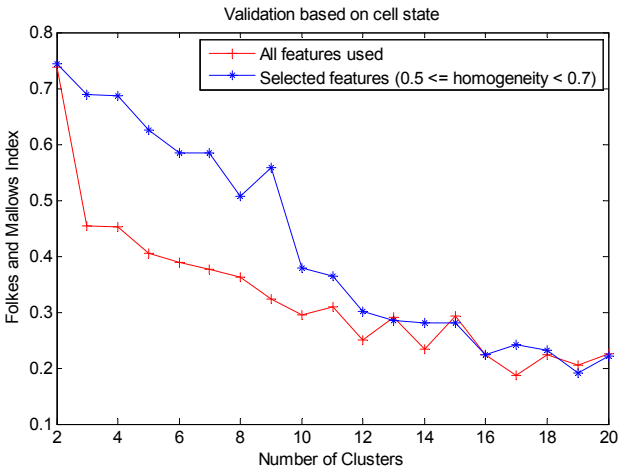

Figure 3. Comparison of the clustering results using k-means on the 74x822 dataset based on cell state partition.

As shown in Figure 3 the clustering structures of the approach with the selected genes are conforming better to the cell state partition for almost all values of *k*. This means that the selection of genes helps to the uncovering of the information that is relevant to discriminate cancerous from normal samples.

In Figure 4 the results obtained using all the genes are also compared to the results obtained using the selected genes with a homogeneity value inside the interval [0.5, 0.7). The validation is based on the cell source partition this time. As before the clustering structures of the approach with the selected genes are conforming better to the cell source partition for almost all values of *k*. This means that the selection of genes helps to the discrimination of bulk from cell line samples, too.

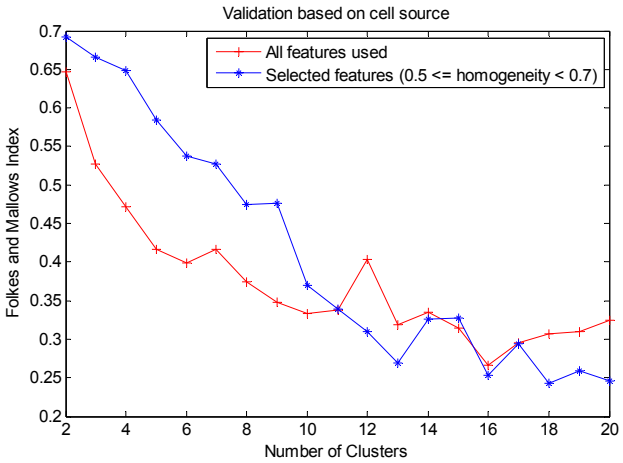

Figure 4. Comparison of clustering results using k-means on the 74x822 dataset based on cell source partition.

Figures 5 and 6 present comparisons of the results of both partitions (cell state and cell source) for the approach with the selected features and the approach with all features. As shown in the figures after feature selection the obtained clustering structures conform slightly better to the cell state partition than to the cell source partition.

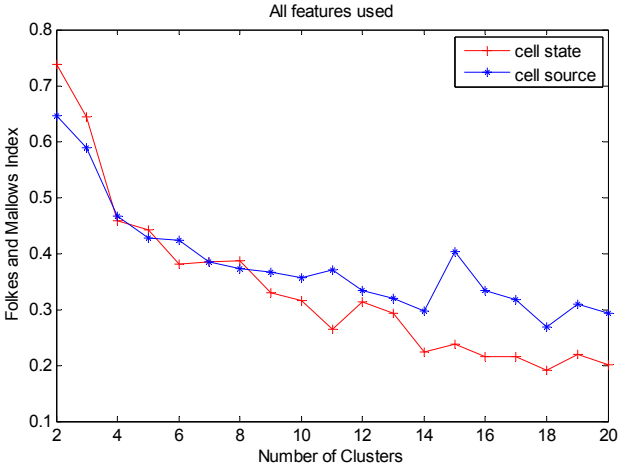

Figure 5. Comparison of clustering results using k-means and all the features on the 74x822 dataset based on both partitions.

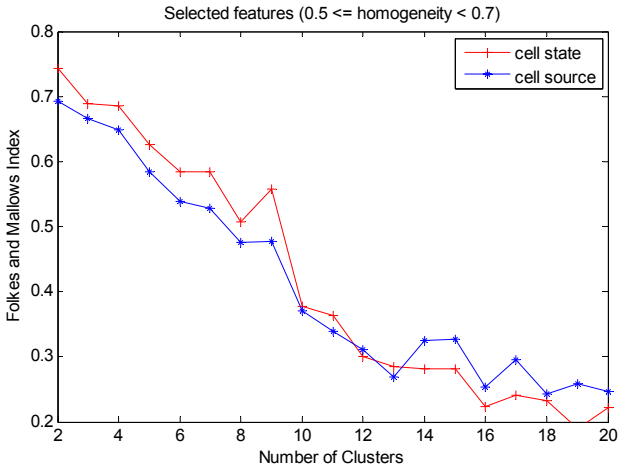

Figure 6. Comparison of clustering results using k-means and selected features on the 74x822 dataset based on both partitions.

In the following lines we present the results obtained using *k*-means with *k* varying from 2 up to 20 on the 90x27679 dataset. As in Table 1, each row of Table 2 presents the best clustering structure -according to the mean silhouette value- among the structures that were obtained using only the genes with a homogeneity value inside the interval of the first column. Again, for the most of the intervals the clustering structure with two clusters has the highest mean silhouette value. So, for this dataset also there is an agreement between the intrinsic characteristics of the data (the existing partitions) and the discovered number of clusters. Among all the intervals the best clustering structure was the one of the [0.9, 1] interval.

Table 2. The best clustering structures of k-means on the 90x27679 dataset.

Homogeneity	Number of Clusters	Mean Silhouette
[0, 0.1)	18	0.2409
[0.1, 0.2)	3	0.5181
[0.2, 0.3)	2	0.4765
[0.3, 0.4)	2	0.4606
[0.4, 0.5)	3	0.8603
[0.5, 0.6)	2	0.8471
[0.6, 0.7)	2	0.9654
[0.7, 0.8)	2	0.9530
[0.7, 0.9)	2	0.8420
[0.9, 1]	2	0.9841

Figure 7 presents the evaluation of the clustering structure that contains two clusters depending on the homogeneity interval that was used in order to select the genes. As shown in the figure, the clustering structure conforms better to the cell state partition than to the cell source partition. The highest value of the index is obtained for the [0.7, 0.8) as well as the [0.9, 1] interval. This agrees with the results of the mean silhouette value, where the [0.9, 1] interval achieved the highest value. The [0.7, 0.8) has also a very high silhouette value.

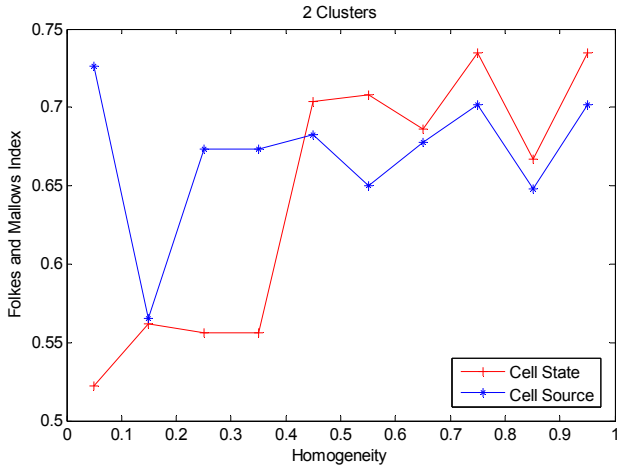

Figure 7. Evaluation of the structure with 2 clusters obtained using k-means on the 90x27679 dataset.

In Figure 8 the results obtained using all the genes are compared to the results obtained using the selected genes with a homogeneity value inside the interval [0.9, 1]. As shown in the figure the clustering structures of the approach with the selected genes are conforming better to the cell state partition for all values of *k* (for 2 and 3 clusters they perform the same). As with the other dataset this indicates that the selection of genes helps to the uncovering of the information that is relevant to discriminate cancerous from normal samples.

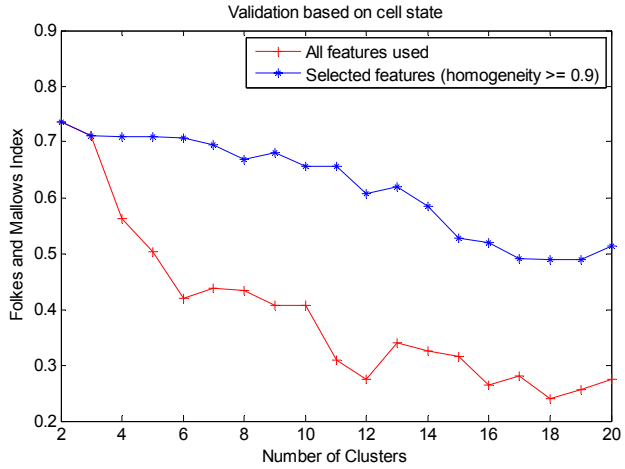

Figure 8. Comparison of the clustering results using k-means on the 90x27679 dataset based on cell state partition.

In Figure 9 the results obtained using all the genes are also compared to the results obtained using the selected genes with a homogeneity value inside the interval [0.9, 1]. The validation is based on the cell source partition. In this case also, the clustering structures of the approach with the selected genes are conforming better to the cell source partition for almost all values of *k*.

By observing the graphs of both Figures 8 and 9 we see that only for the clustering structures of 2 and 3 clusters the results are the same for both approaches. This means that the selection of the gene tags does not lead to improved clustering structures. However, the quality of clusters does not reduce, but remains the same. If we consider the important improvement in all the remaining clustering structure we can undauntedly conclude that the quality of the clustering structures obtained by our approach is considerably improved.

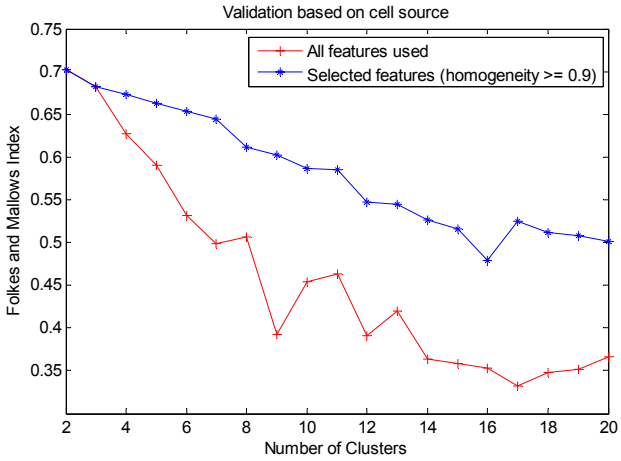

Figure 9. Comparison of clustering results using k-means on the 90x27679 dataset based on cell source partition.

Figures 10 and 11 present comparisons of the results of both partitions (cell state and cell source) for the approach with the selected features and the approach with all features. With this dataset the improvement of the discrimination between different cell states is quite visible. As shown in Figure 10 the clustering structures got using all features conform better to the cell source partition than to the cell state partition, when more than three clusters are obtained except with 2 and 3 clusters. However, in Figure 11 we see that after feature selection the obtained clustering structures conform pretty better to the cell state partition than to the cell source partition.

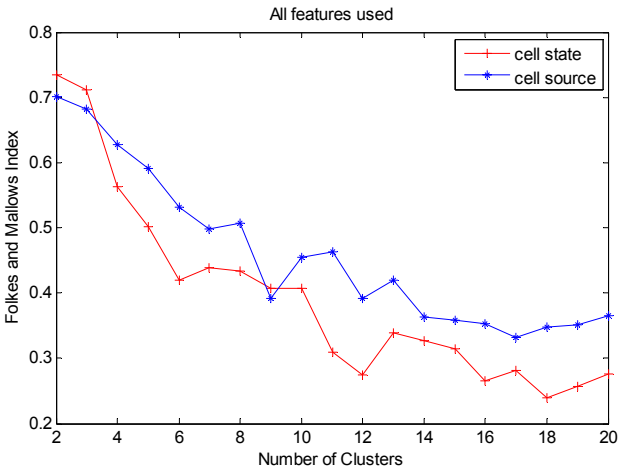

Figure 10. Comparison of clustering results using k-means and all the features on the 90x27679 dataset based on both partitions.

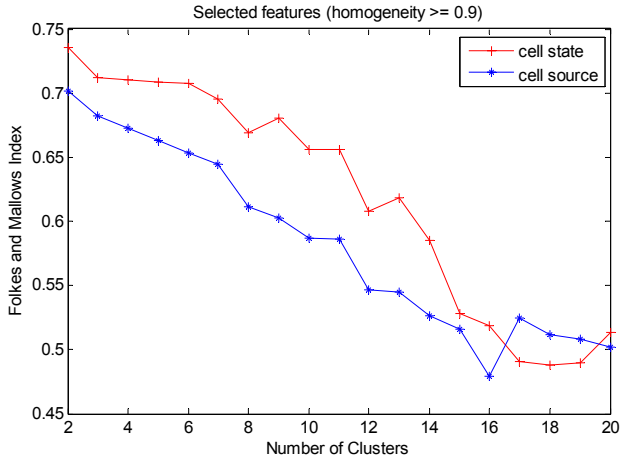

Figure 11. Comparison of clustering results using k-means and selected features on the 90x27679 dataset based on both partitions.

As described in the experimental setup we also used a hierarchical clustering algorithm. We used the distance as a criterion in order to obtain specific numbers of clusters from the hierarchy of clusters. As with *k*-means the number of clusters varies from 2 up to 20. The results of this algorithm are quite similar to those of *k*-means and the observations made for the results of *k*-means are also valid for this algorithm. For this reason we will not present extensive experimental results, but we will focus on the most interesting findings on the 74x822 dataset.

In Figure 12, *k*-means and the hierarchical algorithm are compared based on cell state partition using all the features. As shown in the figure, the clustering structures of *k*-means for 2 and 3 clusters conform fairly better to the cell state partition than the hierarchical algorithm's structures. For the rest clustering structures there are not important differences.

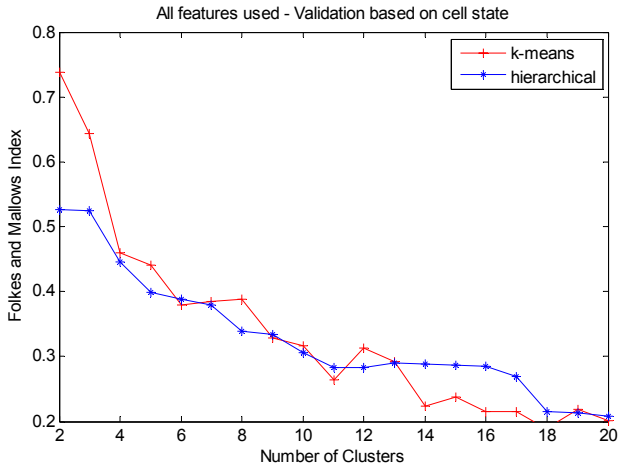

Figure 12. Comparison of k-means and hierarchical algorithm on the 74x822 dataset based on cell state partition.

The same observation could be made for the 2 and 3 clusters of Figure 13, where *k*-means and the hierarchical algorithm are compared based on cell source partition. In this case for 3 to 9 clusters the hierarchical clustering algorithm performs better.

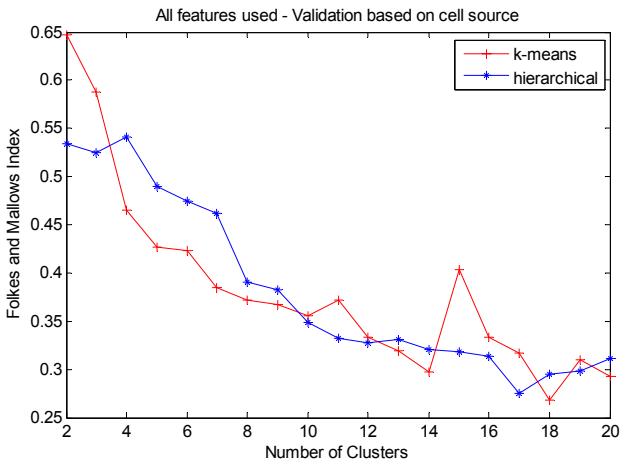

Figure 13. Comparison of k-means and hierarchical algorithm on the 74x822 dataset based on cell source partition.

Figures 14 and 15 present a comparison of k-means and hierarchical algorithm for 2 clusters with various homogeneity intervals for selecting the genes. In Figure 14 the comparison is based on cell state, whereas in Figure 15 is based on cell source. In both cases the two algorithms perform identically for homogeneity values over 0.5. This observation is quite important since the most prominent genes as indicated by the experiments tend to have values of homogeneity greater than 0.5 and near to 1. In contrast, as shown in Figures 12 and 13 there is a great difference between *k*-means and the hierarchical algorithm for the clustering structures of 2 and 3 clusters. These observations indicate that the clustering structures obtained by our approach are less dependant on the used clustering algorithm. This means that the discretization and feature selection steps of our approach assist significantly the discovery of the most prominent genes.

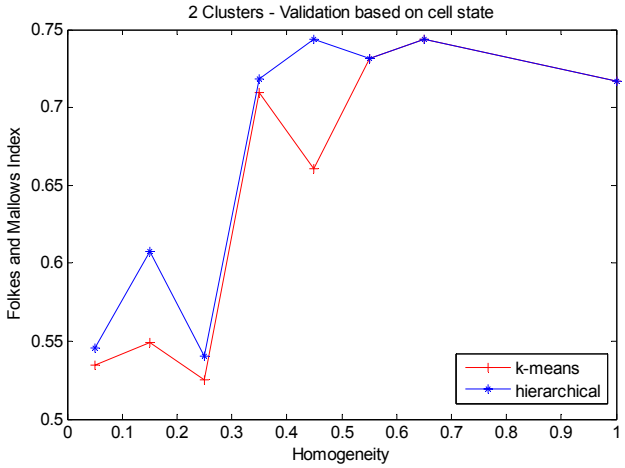

Figure 14. Comparison of k-means and hierarchical algorithm for 2 clusters and various homogeneity intervals on the 74x822 dataset (comparison is based on cell state).

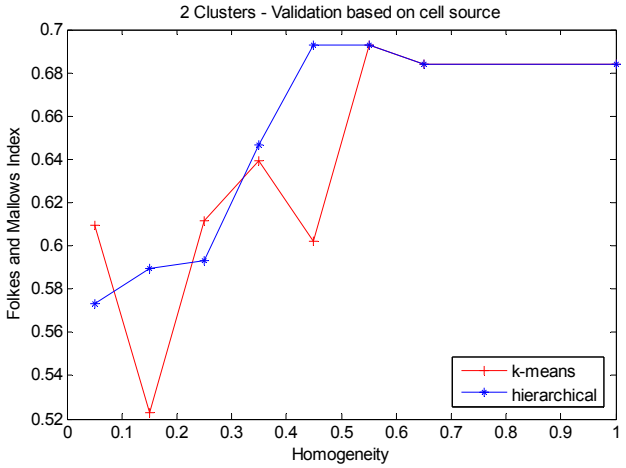

Figure 15. Comparison of k-means and hierarchical algorithm for 2 clusters and various homogeneity intervals on the 74x822 dataset (comparison is based on cell source).

6. CONCLUSIONS

In this paper we proposed an approach for clustering gene expression data that are collected with the SAGE method. In particular we focused on clustering SAGE libraries according to their cell state (cancerous or normal) and compared with the clustering of the libraries according to their cell source (bulk or cell line). SAGE libraries are usually clustered almost equally well according to either their cell state, or their cell source. However, a clustering structure that discriminates cancerous samples from normal ones is more interesting than one that discriminates bulk from cell line samples. Moreover, the presence of two cell sources in combination to the high dimensionality of the data and the existence of many cancer types makes more difficult the discrimination of the samples. In addition, if we think of the considerable contamination that is present in the bulk tissue samples, the problem becomes even harder. By saying contamination we mean that a cancerous bulk tissue sample may also contain surrounding normal cells. The phenomenon of bulk tissue contamination was also observed in previous works [11, 18].

In order to deal with these problems and uncover those features (gene tags) that are more relevant to the cell state grouping of SAGE libraries, we have utilized the under-expressions and over-expressions of genes for studying their variation across the SAGE libraries. Our approach successfully discovers the number of clusters that are intrinsic in the two datasets that were used in this paper. Also, the clustering structures obtained using our approach, conform quite better to the cell state partition of the data. These clustering structures also conform better to the cell source partition in comparison to the clustering structures obtained using all the features. Moreover, the experiments shown that a number of genes (sometimes very small number of genes is adequate) can encapsulate the most valuable information for clustering. So, the benefit is twofold. First, we manage to obtain considerably better clustering structures, and second we drastically reduce the data dimensionality and consequently the computational cost.

Our future plans include the improvement of the gene selection step of our approach. In particular, we plan to use a better method for partitioning the homogeneity space in a number of intervals. Moreover, we are thinking about the modification and application of our approach on gene expression data that were collected with other techniques, like DNA arrays. Finally, we intend to study more in-depth the impact of sequencing errors and other possible sources of noise on the effectiveness of gene expression clustering.

7. REFERENCES

[1] Adams, M.D., Kelley, J.M., Gocayne, J.D. Dubnick, M., Polymeropoulos, M.H., Xiao, H. Merril, C.R. Wu, A., Olde, B., Moreno, R.E. Kerlavage, A.R., McCombie, W.R, and Venter, J.C. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252, 1991, pp. 1651-1656.

[2] Alves, A. Zagoruiko, N., Okun, O., Kutnenko, O., and Borisova, I.. Predictive Analysis of Gene Expression Data from Human SAGE Libraries. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 60–71.

[3] Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.F., and Gandrillon, O.. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biol*ogy, 3(12).

[4] Ben-Dor, A., Shamir, R., and Yakhini, Z. Clustering Gene Expression Patterns. Journal of Computational Biology, 6(3/4), 1999, 281-297.

[5] Cai, L., Huang, H. Blackshaw, S., Liu, J.S., Cepko, C., and Wong, W.H. Clustering analysis of SAGE data using a Poisson approach. Genome Biology, 2004, 5:R51.

[6] Eisen, M.B., Spellman, P., Brown, P.O., and Botstein, D., Cluster Analysis and Display of Genome-Wide Expression Patterns, In Proceedings of the National Academy of Sciences, 95(25): 14863--14868, 1998.

[7] Esseghir, M.A., Ben Yahia, S., and Abdelhak, S. Localizing compact set of genes involved in cancer diseases using an evolutionary connectionist approach. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 78-83.

[8] Gamberoni, G., and Storari, S. Supervised and unsupervised learning techniques for profiling SAGE results. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 121–126.

[9] Gandrillon, O. Guide to the gene expression data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 116–120

[10] Gasmi, G., Hamrouni, T., Abdelhak, S., Ben Yahia, S., and Mephu Nguifo, E.. Extracting Generic Basis of Association Rules from SAGE Data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 84–89.

[11] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, Vol. 286, 1999, 531-537.

[12] Halkidi, M., Batistakis, Y, and Vazirgiannis, M. Cluster Validity Methods: Part I. *SIGMOD Record* 31(2): 40-45 (2002).

[13] Hebert, C., Blachon, S., and Cremilleux, B.. "Mining δ-strong Characterization Rules in Large SAGE Data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 90-101.

[14] Lin, H.-T. and Li, L. Analysis of SAGE Results with Combined Learning Techniques. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 102–113.

[15] MacQueen, J.B Some Methods for Classification and Analysis of Multivariate Observations, In *Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics*, 1967, 281-297.

[16] Martinez, R., Christen, R., Pasquier, C., and Pasquier, N. Exploratory Analysis of Cancer SAGE Data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 72–77.

[17] Mitra, P. and Majumder, D.D.. Feature Selection and Gene Clustering from Gene Expression Data. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), 2004, pp. 343 - 346.

[18] Ng, R.T., Sander, J., and Sleumer, M.C. Hierarchical cluster analysis of SAGE data for cancer profiling. In *Proceedings of Workshop on Data Mining in Bioinformatics*, 2001, pp. 65-72.

[19] Rioult, F. "Mining strong emerging patterns in wide SAGE data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 484–487.

[20] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 1995, pp. 467-470.

[21] Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. Serial analysis of gene expression, *Science*, 270 (5235), 1995, pp. 484-487.

[22] Wang, H., Zheng, H., Azuaje, F. Poisson-Based Self-Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(2), 2007, pp. 163-175.