Tumor Classification by Gene Expression Profiling: Comparison and Validation of Five Clustering Methods

M. Granzow
Intelligent Bioinformatics Systems
German Cancer Research Center
69120 Heidelberg, Germany
Tel.: +49-(0)6221-423603
M.Granzow@DKFZ.de

D. Berrar
Intelligent Bioinformatics Systems
German Cancer Research Center
69120 Heidelberg, Germany
Tel.: +49-(0)6221-423604
D.Berrar@DKFZ.de

W. Dubitzky
Intelligent Bioinformatics Systems
German Cancer Research Center
69120 Heidelberg, Germany
Tel.: +49-(0)6221-423603
W.Dubitzky@DKFZ.de

A. Schuster
Faculty of Informatics
University of Ulster, Jordanstown
Co. Antrim BT37 0QB, N. Ireland
Tel.: +44 (0)28-90-366093
A.Schuster@ulst.ac.uk

F.J. Azuaje
Department of Computer Science
Trinity College
Dublin, 2, Republic of Ireland
Tel.: +353-1-608-2459
Francisco.Azuaje@cs.tcd.ie

R. Eils
Intelligent Bioinformatics Systems
German Cancer Research Center
69120 Heidelberg, Germany
Tel.: +49-(0)6221-423600
R.Eils@DKFZ.de

ABSTRACT

The considerable "algorithmic complexity" of biological systems requires a huge amount of detailed information for their complete description. Current high-throughput technology such as microarrays is generating an overwhelming amount of data of biological systems at the molecular and cellular level. To adequately organize, maintain, analyze and interpret this deluge of information the adaptation of existing and the development of new computational methodologies and tools is required. The principal approach to analyzing and interpreting biological data is to abstract them into logical structures that support and incrementally promote the development of a more general conceptual framework for characterizing, explaining, and predicting processes in living systems. Cluster analysis refers to a computing methodology that discovers and describes meaningful patterns or structures in data. Generally, cluster algorithms are governed by a learning-by-observation process. A plethora of specific algorithms has been suggested in the literature. In the context of microarray gene expression profiling of tumors, this work describes a comparative study of five clustering methods.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**] I.5.3 [**Clustering**]: Algorithms – *comparison of classical, fuzzy, neural and mixed methods; verification.* I.5.2 [**Design Methodology**] Classifier design and evaluation – *decision trees*; Pattern analysis – *feature selection*

General Terms

Algorithms, verification

Keywords

Microarray, gene expression, clustering analysis, discriminant analysis, data mining, acute leukemia, cancer classification

1. INTRODUCTION

Tumors are generally classified by means of nonmolecular parameters such as clinical course, morphology and pathohistological characteristics. Nevertheless, the classification criteria obtained with these methods are not sufficient in every case. For example, it creates classes of cancer with significantly different clinical course or treatment response. As advanced molecular techniques are being established, more information about tumors is captured and stored electronically. One of these techniques, cDNA microarrays, is used for profiling the expression of many thousand genes in one single experiment of a tissue sample, e.g. a tumor. The generated data may contribute to a more precise tumor classification, identification or discovery of new tumor subgroups, and to the prediction of clinical parameters relevant to prognosis or therapy response.

In hematology, distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) has been established on the basis of clinical phenotype, morphology, enzyme reactions, and cell surface antigens. Recently, subgroups of the two classes have been described exhibiting specific chromosomal translocations [1,2,3]. Yet, misclassification occurs in single cases. Additionally, different behavior, e.g. in clinical course or treatment response, can be observed indicating a further discrimination within these classes. Furthermore, the differentiation between ALL and AML requires a number of

methods available in specialized laboratories. Thus, a classification based upon one single experiment is lacking to date.

DNA microarrays may contribute to this. Yet, to gain a deeper understanding of the vast amount of data acquired by this technology more than sorting in spreadsheets and plotting few graphs is needed. Advanced data analysis methods are required to discover and describe hidden patterns within such data. This includes classical statistical methods but also computing methodologies such as data mining, machine learning, artificial intelligence, knowledge based techniques, and database technology. In the context of gene expression monitoring, this paper focuses on a comparison of classical and more advanced approaches to cluster analysis. Furthermore, we present a new framework for validation of clusters by a symbolic learning algorithm.

Both discriminant analysis and cluster analysis are often referred to as classification. However, cluster analysis is quite different from discriminant analysis in that it actually establishes groups (i.e., classes) of objects or entities, whereas discriminant analysis assigns objects to groups that were defined in advance. Thus, cluster analysis is often used when a set of cases or observations is to be divided into natural groups. In addition to establishing such groups or clusters, cluster analysis is also concerned with describing the resulting groups in a concise and intelligible manner. More specifically, clustering is a process or task that is concerned with establishing classes or groups from a set of observations, and with the definition or description of the classes that are identified. Because of this added requirement and complexity, clustering is considered a higher-level process than discriminant analysis. A general definition of clustering is provided in Definition 1 (based on [4]).

Definition 1: Given a set X with n objects $X = \{x_1, x_2, ..., x_n\}$ with each object, x_i, $i = 1 \ldots n$, described by m attributes, $x_i = (x_{i1}, x_{i2}, ..., x_{im})$, determine a classification (grouping, clustering) that is most likely to have generated the observed objects.

A frequently applied approach to clustering attempts to produce classes or groups that maximize similarity within classes but minimize similarity between classes. In the context of microarray data analysis, clustering methods may be useful for automatically detecting new subgroups of clinical or biological entities (e.g., tumors, genetic risk groups) in the data.

A variety of cluster algorithms have been applied to gene expression data generated from either yeast or human tissue; for a review see [5]. Golub et al. [6] describe the analysis of gene expression profiles of 72 patients with either AML or ALL. They apply a *Kohonen neural network* (also called *self-organizing map*) cluster analysis method to identify and define a small number of prominent classes within the data [7]. This approach is widely used in the microarray community.

An alternative to classical clustering methods are the so-called *fuzzy clustering* approaches [7] and the *growing cell structures* neural networks [8]. To asses how these advanced methods fare in comparison to the classical approaches to cluster analysis, this study investigates five clustering methods: *k*-means, Kohonen networks, growing cell structure networks, fuzzy c-means, and fuzzy Kohonen networks. Going beyond cluster detection, the presented analytical framework also proposes an automatic cluster validation method based on a post-clustering discriminant analysis step. The particular algorithm used in this phase is the C5.0 decision tree algorithm [9,10]. Through this well-known technique it is possible to generate rules of expression stages for a number of genes corresponding to certain clusters. The biological importance of these genes in the context of acute leukemias and/or tumor biology is discussed in section 4.

2. ANALYZING MICROARRAY DATA

The problem of analyzing microarray data lies in its characteristic complexity. Generally microarrays describe a set of n entities (e.g., patients) by means of m characterizing features such as gene expression levels. Typically, m is larger than n by a factor of 10 to 100 and characterizing features are real number values. The following subsections describe the nature of the data used in this work and the analysis methods.

2.1 Data Description

The original data sets provided at http://bioinformatics.duke.edu/camda/dataset.htm included expression profiles of 7130 human DNA probes spotted on Affimetrix Hu6800 micro-arrays of 72 patients with either AML (25 cases) or ALL (47 cases). Tissue samples were collected at time of diagnosis before treatment, taken either from bone marrow (62 cases) or peripheral blood (10 cases) and reflected both childhood and adult leukemias. Furthermore, a description of subtypes (B-cell or T-cell ALL), treatment response (15 cases), French-American–British (FAB) classification (20 cases), gender and source was given. RNA preparation, however, was performed using different protocols.

The cluster analyses are directly applied to the normalized (linearly re-scaled) integral number values of the expression profiles of all 72 patients. After clustering and before discriminant analysis, the expression profiles are subject to a discretization step that produces three different symbolic values representing underexpressed, balanced, and overexpressed states (Section 2.5.1). Furthermore, genes showing the same expresion value in all 72 cases were excluded from further analysis as they do not carry discriminatory information.

2.2 Basic Methodology

The basic analysis framework of this study is characterized by five distinct phases or steps:

1. *data preparation*: Remove control genes and discretize data for decision tree analysis in step 4.
2. *clustering*: Apply five clustering techniques with different settings for controlling expected cluster number.
3. *inspection*: Use visualization techniques to inspect initial cluster results and select "promising" clusters.
4. *validation*: Apply discriminant analysis (decision tree) to identify and validate interesting clusters.
5. *parameter identification*: Analyze discriminant models to isolate the main parameters (genes, clinical factors) describing the resulting clusters.

Beyond the insight provided by applying diverse clustering methods, the validation procedure outlined in step 4 is the most interesting part of the entire process. By applying a symbolic discriminant method to those clusters found most "promising" after visual inspection, it is not only possible to evaluate the validity of the clusters but also to identify and describe which parameters are characteristic for the clusters. Furthermore,

discriminant models are a useful "by-product" of this method. These models can be used to classify new observations in the future.

The following subsections provide a brief overview of the five clustering techniques and the discriminant method used in this study.

2.3 Five Clustering Techniques

In general the aim of any clustering technique is to discover and describe structures contained within data. Typically, these structures are classes, categories, or groups of objects. Most classical clustering techniques, such as the k-means algorithm and Kohonen nets, assign to each object exactly one class [11]. In some situations this can be an oversimplification, because objects can be partially assigned to two or more classes. Fuzzy clustering algorithms try to address this problem by allowing objects to be gradual members of two or more groups or classes. In the following subsections we briefly describe a classical clustering method (k-means), a fuzzy clustering method (c-means), two neural methods (Kohonen and growing cell structures networks), and a fuzzy-neural hybrid method (fuzzy Kohonen network).

2.3.1 K-Means Clustering

The k-means algorithm is one of the most widely studied "automatic classification" algorithm [12]. This simple algorithm is initialized with the number of sought clusters (the parameter k). Then, in its basic standard implementation: (1) k points are chosen at random as cluster centroids; (2) the cases are assigned to the clusters by finding the nearest centroid; (3) Next new centroids of the clusters are calculated by averaging the positions of each point in the cluster along each dimension moving the position of each centroid; (4) this process is repeated from step (2) until the boundaries of the clusters stop changing. The performance of k-means clustering is highly dependent on the initial seed centroids. Therefore the result of this method is often suboptimal.

2.3.2 Fuzzy C-Means Clustering

Imagine a set X with n objects with each object described by m attributes (see Definition 1). Given a set C of k classes, $C = \{c_1, c_2, ..., c_k\}$, one of the key features of the fuzzy c-means algorithm is that for every single object x in X the algorithm assigns a membership degree $\mu(x,c)$ to every single class c in C such that $\mu(x,c)$: $X \times C \rightarrow [0,1]$ [13]. The algorithm stops if the change of membership degree of all objects falls below a predefined threshold. Using an Euclidean distance measure, this method does not work well if the underlying classes or clusters deviate strongly from hyperspherical structures. Fuzzy Kohonen nets try to address this problem.

2.3.3 Kohonen Networks

Kohonen networks or *self-organizing feature maps* [14] can be described as two layer networks that consist of an input layer comprised of sensory receptors and an output layer comprised of processing units. Each processing unit is connected with a set of sensory receptors. It is assumed that the lateral interaction between the processing units follows a center-excitatory and surrounding-inhibitory scheme. This scheme implements the local cooperation and global competition between processing units. To reduce the potential complexity of this computation strategy, a "winner takes all" logic is usually adopted. That is, for a particular input pattern, the unit that has the highest activation is picked as the winner. Synaptic weights of the winning unit and those within a defined neighborhood are then updated. A crucial problem with Kohonen networks is that the number of classes that a network can determine is directly related to the number of neurons or units in the network. Since this number is pre-defined and fixed the formation of new classes is not possible. So-called growing cell structures neural networks (see below) address this problem. Other problems of the Kohonen net approach include its dependence on the sequence in which the learning data are presented and its lack of robust convergence criteria. These problems are addressed by fuzzy Kohonen networks.

2.3.4 Growing Cell Structures Networks

Growing cell structures (GCS) neural networks are a variation of the Kohonen networks. The GCS method offers several advantages over Kohonen nets [9]: (1) self-adaptive topology which is highly independent of the user (no pre-definition necessary); (2) in contrast to Kohonen nets, a GCS net requires only a small number of constant parameters and there is no need to define time-dependent or decay schedule parameters; and (3) its ability to interrupt a learning process or to continue a previously interrupted one permits the construction of incremental and dynamic learning systems.

2.3.5 Fuzzy Kohonen Neural Networks

A Fuzzy Kohonen network is a model in which concepts of fuzzy sets and Kohonen networks are combined. The two major parts of the model are a Kohonen network and a fuzzy c-means algorithm (see above). The use of both techniques in one model aims at benefiting from the advantages of different techniques and at overcoming some of the shortcomings of each individual technique such as the Kohonen learning parameters discussed before [15].

2.4 The C5.0 Decision Tree Algorithm

Decision trees are a supervised machine learning technique used for classification and prediction. Decision tree learning follows a kind of top-down, divide-and-conquer learning process. The basic algorithm for decision tree learning can be described as follows:

1. Based on an information gain measure, select an attribute to place at the root of the tree and branch for each possible value of the tree. Thereby, the underlying case set is split up into subsets, one for each value of the considered attribute.
2. Recursively repeat this process for each branch, using only those cases that actually reach that branch.
3. If at any time all instances at a node have the same classification, stop developing that part of the tree.

Once trained, a decision tree can predict a new data set by starting at the top of the tree and following a path down the branches until a leaf node is encountered. The path is determined by imposing the split rules on the values of the independent variables in the new data set.

The decision tree approach in this study is based on the C5.0 implementation of SPSS' Clementine [16]. The C5.0 decision tree learning algorithm is a commercial decision tree and rule induction engine developed by Ross Quinlan [10,11]. It is the state-of-the-art successor of the widely used C4.5 decision tree algorithm [10]. In contrast to other decision tree algorithms such as CART [13], C5.0 is able to generate trees with a varying number of branches per node. Decision trees based on C5.0

algorithm provide a clear indication of which attributes are important for the classification task at hand.

2.5 Implementation

2.5.1 Preprocessing

The gene expression profiles of the original data set are represented as absolute integral-numbered expression intensities. Although the C5.0 algorithm is in principle able to handle continuous inputs, a mapping to dicrete values is performed to represent the data in a way that directly reflects the underlying biological interpretation in the representation scheme. Here, we distinguish between balanced expression, underexpression, and overexpression of genes. Since no further information about the cut-off levels are available, the gene expression profiles are discretized according to the following rules: (1) values smaller than the median of all negative values are considered as *underexpressed*, (2) values larger than the median of all positive values are considered as *overexpressed*, and (3) all other values are considered as *balanced*.

A comparison with other statistical measures (e.g. mean plus standard deviation) shows that the negative and positive medians seem to be biologically plausible cut-off levels for these data with respect to their distribution. The expression profile data have been discretized as depicted in Fig. 1.

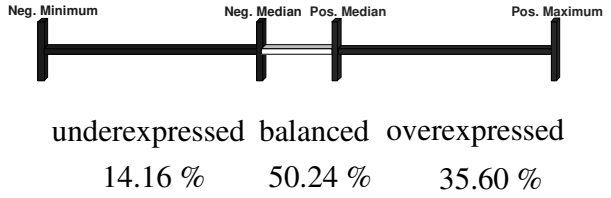

Figure 1: Discretization of the absolute gene expression profile data into underexpressed, balanced, and overexpressed genes. The percentages given reflecct the distribution of expression levels within the data set.

The discretization in three different expression classes resulted in some genes to carry little or no discriminative information for C5.0, for example, a gene that is balanced (or over- or underexpressed) for *all* of the 72 patients. Therefore, the discrete data set is filtered and non-discriminative genes are removed.

2.5.2 Clustering

Here, the five different clustering algorithms are applied to gene expression profiles of 72 patients with AML or ALL. The algorithms are implemented in various software packages: SPSS' Clementine (*k*-means, Kohonen networks) [16], MIT's DataEngine (fuzzy c-means networks, fuzzy Kohonen networks) [14], and a non-proprietary implementation of the GCS algorithm [19].

2.5.3 Defuzzification

For this study it was necessary to defuzzify the results of fuzzy cluster analyses, where each case is assigned to all clusters to a degree drawn from the unit interval. In this case, defuzzification is governed by a rule that assigns a single cluster to each case. One way to achieve this is to choose the cluster which is assigned the highest membership degree. This rule was adopted in this study. This process was necessary only in the case of fuzzy c-means algorithm, as the fuzzy Kohonen algorithm encompasses the defuzzification step.

2.5.4 Inspection

In the context of discriminating AML and ALL cases, the generated clusters are inspected with respect to within-cluster homogeneity. This evaluation is based on subject visual inspection using visualization techniques. Clusters with high degrees of homogeneity are deemed “promising” and selected for further analysis.

2.5.5 Validation

Based on selected "promising" clusters, a post-clustering discriminant analysis is performed to automatically validate and characterize the clusters. This is done by using the discrete data representation (balanced, underexpressed, overexpressed) and the C5.0 decision tree algorithm.

2.5.6 Parameter identification

The output of the decision tree analysis provides a description of the logical rules that make up the classification model and the relative importance of the parameters. Thus, it is possible to determine the genes that are characteristic for the underlying clusters.

3. RESULTS

To compare and evaluate the five clustering methods, each method is applied to the entire data set of 72 patients. Each algorithm produced a different number of clusters in each run (between 2 and 16). Table 1 shows an example of the cluster sizes based on a "expected" number of 4 clusters. Note, that depending on the algorithm it may or may not be possible to explicitly pre-determine the number of clusters that will actually be formed.

Table 1: The number of cases per cluster for performing 4 clusters.

	Cluster			
Method	**# 1**	**# 2**	**# 3**	**# 4**
k-means	1	2	23	46
fuzzy c-means	13	13	19	27
GCS	14	17	19	22
Kohonen	72	0	0	0
fuzzy Kohonen	7	13	20	32

The k-means method generated 2 clusters containing only 1 and 2 cases (cluster #1 and #2), whereas 1/3 and 2/3 of the remaining 69 cases are placed into cluster #3 and #4, respectively. Setting the map topology to 2x2, the Kohonen network was not able to partition the data into 4 disjoint sets, but produced only one cluster with 72 cases and 3 empty clusters. Thus, for a predefined grid of 2x2, the Kohonen net failed to discover any structure in the data. The two fuzzy approaches and the GCS created clusters of similar sizes.

All clustering results (number of clusters between 2 and 16) are evaluated on the basis of the distribution of the biological classes (AML and ALL) within the clusters. By this method, the highest degree of intra-cluster homogeneity is obtained by partitioning the data set into 9 clusters. the Fuzzy Kohonen network yielded the most homogeneous clusters with respect to the two tumor classes (cf. Figure 2). Five clusters contain only ALL cases (cluster #2, #3, #4, #5, #9), 1 cluster only AML cases (cluster #1), and within clusters #6 and #7, there is only one single mismatch. Cluster #8

comprises 2 AML cases together with 10 ALL cases. Additionally, with respect to the subclasses of ALL (i.e. B-cell or T-cell ALL), this algorithm also performed well: Five clusters contain either B-Cell or T-cell only (cluster #3, #6, #7 #8, #9); in 2 clusters, only 1 case does not match (cluster #2, #5); in one cluster, 2 cases fail to produce a match (cluster #4).

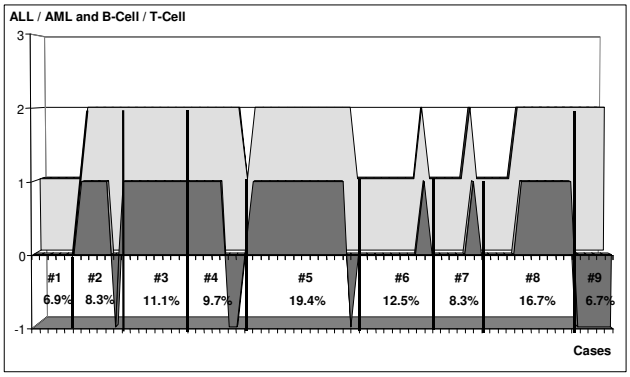

Figure 2: Distribution of AML and ALL cases over 9 clusters generated by the fuzzy Kohonen network. The 72 cases are indexed on the X-axis. On the Y-axis, AML cases are encoded as 1 and ALL as 2. Both AML and ALL cases are portrayed in light grey. The B-cell subclass of the ALL cases are encoded as 1 whereas the T-cell subclass is encoded as -1. Both subclasses are dark grey colored. The numbers #1 to #9 indicate to which cluster these cases belong. The percentages beneath the cluster numbers indicate how many cases are grouped into the different clusters.

Using a different number of clusters did not seem to improve the output of the fuzzy Kohonen network or any of the other algorithms (data not shown).

The C5.0 decision tree algorithm is used to understand and to characterize the results of the fuzzy Kohonen network. Basically, the "promising" clusters are marked with unique class labels and the decision tree is trained to differentiate these "hypothetical" classes. Through this procedure it is possible to establish the objective measures of cohesiveness or validity of the clusters and to describe their logical structure by symbolic rules intelligible to humans. These rules may be represented as tree-like structures, i.e. as decision trees, or as sets of *if-then* rules (see Fig. 3).

Applying this post-clustering validation procedure, the average classification accuracy over 10 cross-validation runs was 41.6%. C5.0 generates only one source node for all clusters. Starting at the top node the tree could take three directions: first, if this gene is downregulated, take the –1-branch, if it is balanced take the way straight down, but if it is upregulated, take the +1-branch (cf. Fig. 3). Each following node has the same degree of freedom or decision choice (-1, 0, +1). In C5.0, the most important node of a tree is the root node at the top. Thus, for all clusters, SMAD4 was the most important gene node.

Cluster #1 is described by two rules: the genes *SMAD4*, *PRPS2*, *nk mRNA*, *TSNAX*, *RB1*, *TTK PK* are balanced compared to normal tissue expression, and *YAP65* is overexpressed. The second rule for this cluster comes into force if *YAP65* additionally is balanced and SPN is upregulated. With these two rules, all cases in cluster #1 are covered.

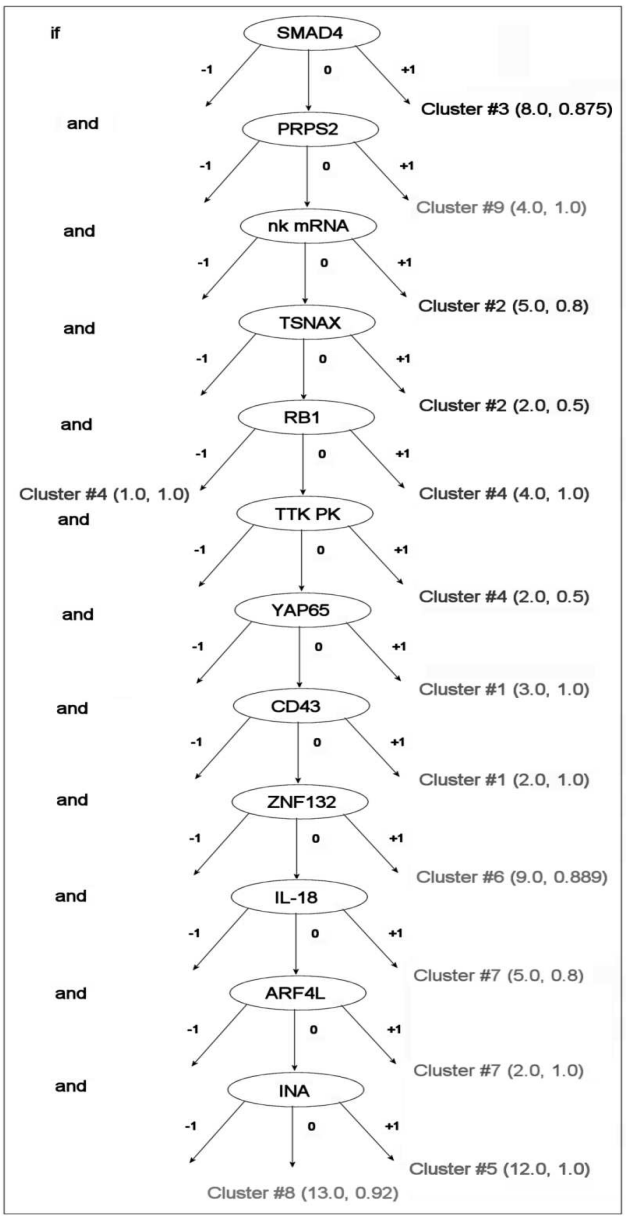

Figure 3: The generated rule sets for the 9 clusters by Fuzzy Kohonen network. The expression states downregulated, balanced, and upregulated are indicated by –1, 0, and +1, respectively. The underlying rule set has to be read as follows. For example, the rule for cluster 9: if *SMAD4*-expression is balanced and *PRPS2* is overexpressed then sort into cluster 9. This pattern was found 4 times in the whole data set and applied only for cluster 9. Another example is: if *SMAD4* is overexpressed then sort into cluster 3. This pattern is found in 8 cases in the data set, 7 times in cluster 3, i.e. 87.5% of all cases with this particular expression profile and once in a different cluster. Abbreviations: *SMAD4*: Sma- and mad-homologues in man 4; *PRPS2*: phosphoribosyl pyrophasphate synthetase 2; *nk mRNA*: normal keratinocyte messenger RNA; *TSNAX*: translin associated factor X; RB1: Retinoblasoma susceptibility gene 1; *TTK PK*: T-cell threonin kinase protein kinase; *YAP65*: yes-associated protein 65kD; *CD 43*: cluster of differentiation 43; *ZNF132*: zinc finger protein; *IL 18*: Interleukin 18; *ARF4L*: Adenosindiphostphate-ribosylation factor 4-like; *INA*: internexin neuronal intermediate filament protein alpha.

Similarly, there are two rules for cluster #2, including upregulated profiles of *nk mRNA* and *TSNAX*. For cluster #3, only one rule was obtained, namely *SMAD4* is upregulated. Cluster #4 is described by three rules. *SMAD4*, *PRPS2*, *nk mRNA*, and *TSNAX* are balanced, and *RB1* is either up- or downregulated. Furthermore, *SMAD4*, *PRPS2*, *nk mRNA*, *TSNAX*, and *RB1* are balanced, and *TTK PK* is upregulated. In cluster #5 which includes 1 AML and 11 ALL cases, only one rule could be found: *SMAD4*, *PRPS2*, *nk mRNA*, *TSNAX*, *RB1*, *TTK PK*, *YAP65*, *CD 43*, *ZNF132*, *IL 18*, and *ARF4L* remain balanced, and *INA* is overexpressed. All 12 cases are characterized by this rule and it occurs only in cluster #5. The tree of rules extends to *ZNF132* for cluster #6, which is upregulated for all 9 cases. Cluster #7 on the other hand is depicted by two rules, spanning from *SMAD4* balanced to *IL 18* and *ARF4L*, respectively, both genes overexpressed. The most heterogeneous cluster #8 is described by one rule, leaving all 12 genes in a balanced state. This is true for 13 cases, 92% of which are grouped in cluster #8. The last cluster, containing only ALL cases, is referred to by one single rule. *SMAD4* is balanced, whereas *PRPS2* is overexpressed in 4 patients in the whole data set, all of them belonging to this cluster.

Overall, between 1 and 3 rules per group are needed to describe each cluster. Notably, the inherent rules C5.0 extracts from the data required only a total of 12 genes for differentiating the 9 clusters.

4. INTERPRETATION AND DISCUSSION

This study presents a novel analysis framework for validating and characterizing clusters obtained by different clustering methods. Furthermore, we systematically compared five different clustering methods in the context of microarray analysis, namely: *k*-means, fuzzy c-means, Kohonen networks, growing cell structures, and fuzzy Kohonen networks. The fuzzy Kohonen approach proves to be the most performant in partitioning the data set into 9 mostly homogeneous clusters. Furthermore, through a post-clustering discriminant analysis algorithm, it is shown that it is possible to objectively validate and characterize cluster results. Overall, only 12 genes and 14 rules (ranging from 1 to 3 rules per cluster) are necessary to characterize the 9 clusters.

Judging and validating the performance of the various clustering methods focuses on the ability to generate homogeneous clusters that are predominantly populated by either AML or ALL patients (see Fig. 2). The experience gained from this study demonstrates that a two-stage validation process is useful. This process relies on visualization techniques and manual visual evaluation to screen and select "promising" clusters, and an automatic post-clustering validation step involving machine learning approaches such as decision trees.

The 12 genes needed to characterize the clusters obtained by the fuzzy Kohonen network include genes encoding proteins involved in cell surface activities (*CD 43*, *IL 18*), and a DNA binding protein (*TSNAX*), that was shown to bind specifically to consensus sequences at breakpoint junctions of chromosomal translocations in many cases of lymphoid malignancies [20]. Furthermore, a gene encoding *SMAD4*, a protein playing a pivotal role in the transforming growth factor (TGF)-ß pathway is playing the most important role in all clusters (cf. Fig. 3). TGF-ß mediates cell growth inhibition and has a pleiotropic and profound effect on the immune system and hematologic malignancies [21]. Additionally, a known tumor suppressor gene, *RB1*, a regulator of the nucleotide production pathway, *PRPS2* [22], a potential transcription regulator (*ZNF132*), and some genes encoding proteins with unknown relationship to tumor biology are found important. Of one of the latter, *nk mRNA*, the gene identifier on the microarray chip (U43374) cannot be found in any tumor genetic/protein related public database. Thus, no information, as to what kind of gene/clone this mRNA reflects is available.

Golub et al. [6] provided a list of 50 genes necessary for distinction between ALL and AML. The C5.0 algorithm needed only 12 genes for this task, indicating that the number of genes necessary for the discrimination between the two leukemias in the respective data set is indeed low.

In conclusion, it is helpful not to rely on one single clustering method but to apply a variety of approaches. The proposed framework in this study, i.e. to identify and validate interesting clusters by a discriminant analysis, provided further insight in the results obtained by the most promising clustering approaches.

5. REFERENCES

[1] Golub T.R., Barker G.F., Bohlander S.K., Hiebert S.W., Ward D.C., Bray-Ward P., Morgan E., Raimondi S.C., Rowley J.D., Gilliland D.G.. Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia. Proc Natl Acad Sci, 92: 4917-4921, 1995.

[2] McLean T.W., Ringold S., Neuberg D., Stegmaier K., Tantravahi R., Ritz J., Koeffler H.P., Takeuchi S., Janssen J.W., Seriu T., Bartram C.R., Sallan S.E., Gilliland D.G., and Golub T.R.. TEL/AML-1 dimerizes and is associated with a favorable outcome in childhood acute lymphoblastic leukemia. Blood 88: 4252-4258, 1996.

[3] Shurtleff S.A., Buijs A, Behm FG, Rubnitz JE, Raimondi SC, Hancock ML, Chan GC, Pui CH, Grosveld G, Downing JR. TEL/AML1 fusion resulting from a cryptic t(12;21) is the most common genetic lesion in pediatric ALL and defines a subgroup of patients with an excellent prognosis. Leukemia. 9: 1985-1989, 1995.

[4] Upal M.A., and Neufeld, E.. Comparison of unsupervised classifiers. in Proceedings of the First International Conference on Information, Statistics and Induction in Science, 342-353, World Scientific, Singapore, 1996.

[5] Sherlock G. Analysis of large-scale gene expression data. Curr Opin Immunol, 12: 201-205, 2000.

[6] 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. 286: 531-537, 1999.

[7] Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub T.R.. Interpreting patterns of gene expression with self-organising maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci, 96: 2907-2912, 1999.

[8] Hoeppner, F., Klawonn, F., Kruse. R. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. John Wiley & Sons, Inc., 1999.

[9] Fritzke, B. Growing Cell Structures: A Self-Organizing Network for Unsupervised an Supervised Learning. Neural Networks, vol. 7, 1441-1460, 1994.

[10] Quinlan J.R.. C4.5 : Programs for machine learning. Morgan Kaufmann, San Francisco, 1993.

[11] Witten, I.H., and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Pub., San Francisco, 1999.

[12] Anderberg, M.R. Cluster analysis for applications. Academic Press, New York, San Francisco, London, 1973.

[13] Berry, M.J. and Linoff, G. Data Mining Techniques For Marketing, Sales and Customer Support, John Wiley & Sons, Inc., New York, 1997.

[14] DataEngine. Manuals of the DataEngine V2.1. MIT Management Intelligenter Technologien GmbH. Aachen, Germany, 1998.

[15] Kohonen, T. Self-organized formation of topologically correct feature maps. Biol Cybern, 43:59-69, 1982.

[16] Huntsberger, T.L. and Aijimarangsee P. Parallel self-organising feature maps for unsupervised pattern recognition. In: Bezdek J.C. and Pal N.R, Editors: Fuzzy models for pattern recognition, 483-495. IEEE Press, New York. 1992

[17] SPSS Clementine. http://www.spss.com/clementine

[18] RuleQuest Research Data Mining Tools. http://www.rulequest.com.

[19] Azuaje F., Dubitzky W., Black N., and Adamson K.. Discovering Relevance Knowledge in Data: A Growing Cell Structure Approach IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, Vol. 30, No 3, 448-460, 2000.

[20] Aoki K., Ishida R., Kasai M.. Isolation and characterization of a cDNA encoding a Translin-like protein, TRAX. FEBS Lett Jan 20;401(2-3):109-12, 1997

[21] De Visser K., and Kast W.M.. Effects of TGF-beta on the immune system: implications for cancer immunotherapy. Leukemia, Aug;13(8):1188-99, 1999

[22] Ahmed M., Taylor W., Smith P.R., Becker M.A.. Accelerated transcription of PRPS1 in X-linked overactivity of normal human phosphoribosylpyrophosphate synthetase. J Biol Chem, Mar 12;274(11):7482-8, 1999