Candidate gene discriminating gliomas identification via a supervised iteration of bipartitive k-means initialised via partititve division according to principal components

ROSSELLA MELCHIOTTI and DIEGO LIBERATI
National Research Council of Italy
Department of Electronics,
Information and Biomedical Engineering,
Politecnico di Milano
ITALY

Abstract: - In this paper, the candidate gene discriminating gliomas identification via a supervised iteration of bipartitive k-mean is presented. Gliomas are supervisedly discriminated by identifying, via iterative bipartitive division according to principal directions initializing k-means, salient genes able to cluster representative patients, thus also giving an insight about degrees of epigenetic similarity among different kinds of gliomas.

Key Words: -K-means, PCA, clustering, salient genes identification

1 Introduction

The purpose of this paper is to describe the results obtained by analyzing a dataset containing genetic information about patients affected by different kinds of gliomas (a type of brain tumor). The main goal is to discover a set of genes for classifying gliomas with respect to gene expression. With this work we are trying to establish if the histological classification, which refers to the microscopic structure of tissue, has a genetic foundation. We would like to find some genes whose expression allows discriminating different categories of gliomas (in our example we are focusing on astrocytomas (AD), oligodendrogliomas and oligoastrocytomas (O) and anaplastic astrocytomas (AA)). Our dataset also contains a certain number of control subjects, so that we can analyze the genetic differences between people affected by this kind of tumor and healthy ones. Even if glioma is a rare type of tumor, the fact that life expectancy could still be quite low makes this kind of study pretty relevant. Life expectancy for gliomas depends on the typology of the tumor, ranging from 1 year for the resistant multiform Glioblastoma to 10 years for the Oligodendroglioma.
The principal treatments for this kind of cancer are a mix of chemotherapy, radiation therapy and surgery, depending on the typology of the glioma and on its position in the nervous system [1]. That is why it is so important to be able to make an early classification in order to decide the proper treatment for the kind of tumor affecting the very patient.

Case	Gender	Age	Tumor location	Histology
1 - O1	M	27	right frontal	Oligodendroglioma grade II
2- O20	F	69	left temporal	Oligodendroglioma grade II
3- O4	F	50	right frontal	Oligodendroglioma grade II
4- O23	F	59	left frontal	Oligodendroglioma grade II
5- O19	M	24	right frontal	Mixed oligoastrocytoma grade II
6- O15	M	56	left frontal	Mixed oligoastrocytoma grade II
7- O10	F	30	right frontal	Mixed oligoastrocytoma grade II
8 - O7	M	32	left frontal	Mixed oligoastrocytoma grade II
9- O8	M	38	left temporal	Mixed oligoastrocytoma grade II
10- O6	F	47	left temporal	Mixed oligoastrocytoma grade II
11 - O16	F	31	right parietal	Astrocytoma grade II

12 - AD6	M	40	left temporo-occipital	Astrocytoma grade II
13 - AD9	M	38	left frontal	Astrocytoma grade II
14 - AD10	M	70	left temporal	Astrocytoma grade II
15 - AD11	M	25	left temporal	Astrocytoma grade II
16 - AD12	M	34	left temporal	Astrocytoma grade II
17 - O2	F	26	right frontal	Anaplastic Oligodendroglioma
18 - O3	M	21	right frontal	Anaplastic Oligodendroglioma
19 - O17	F	24	left temporal	Anaplastic Oligodendroglioma
20 - O18	F	45	right parietal	Anaplastic Oligodendroglioma
21 - O9	F	46	left parietal	Anaplastic oligodendroglioma
22 - O24	M	57	left fronto-temporal	Anaplastic Oligoastrocytoma
23 - O12	F	27	left temporal	Anaplastic Astrocytoma
24 - AA3	F	25	right frontal	Anaplastic Astrocytoma
25 - AA6	F	36	left fronto-temporal	Anaplastic Astrocytoma
26 - AA5	M	44	right temporal	Anaplastic Astrocytoma

Table 1 Description of the dataset

This kind of study could also help understanding which genes are involved in a particular kind of glioma so that geneticists could concentrate their efforts on studying them in order to get a better understanding of the tumor and to elaborate an aimed strategy for fighting it. The paper is organized as follows. We will start by describing the dataset we have worked on. Afterwards we will focus on the approach adopted for classifying data, lingering briefly over each step of the procedure. We will finish by describing the obtained results and the limits of the methods we have used.

2 The dataset description

Data have been provided by the Hammersmith Hospital of West London. The dataset, obtained using the microarray technology, was constituted by the expressions of 54675 genes in 36 patients. Each patient has been associated to a vector containing 54675 variables, each measuring the activation level of the corresponding gene. RNA was hybridized to Affymetrix U133_Plus_2 arrays (Affymetrix, Santa Clara,CA) according to the manufacturer protocol. The arrays were scanned in an Affymetrix/Hewlett-PackardGeneChip Scanner 3000[2]. Data analysis was performed using Matlab and its Statistics Toolbox.

For the dataset we have worked on, an a priori classification of patients was available. The dataset was composed by 13 control subjects, 5 patients affected by astrocytomas, 3 patients affected by anaplastic astrocytomas and 15 patients affected by oligodendrogliomas and mixed gliomas. The analysis was carried out using this information. We have in fact adopted a supervised approach to solve this classifying problem. A concise description of clinical data of patients we analyzed is in table 1.

3 Proposed approach

Our analysis have been performed on a matrix 36x54675 where rows represented patients and columns represented gene expression from each of the DNA spots on the microarray. The first step in our work was to get a better knowledge of the dataset. Since it couldn't be plotted because of its dimension, we applied a dimension reduction using a Principal Component Analysis (PCA). PCA is a technique for reducing multidimensional datasets to lower dimensions[3]. After mean centering the data for each attribute, eigenvalue decomposition was performed in order to retain those features of the dataset that contributed the most to its variance, by keeping lower-order principal components and ignoring higher-order ones. A very few low order principal components contain, in a variety cases, all the most important features of the data. To characterize the trends exhibited by a dataset, PCA extracts directions where the cloud of samples in the original multidimensional space is more extended. The best low-dimensional space is determined by the eigenvectors (called

principal components) associated to the "largest" eigenvalues of the covariance matrix of the original dataset X [4]. If we apply this method to our original set we obtain the representation of Figure 1 showing the projection of our samples on the plan defined by the two first principal components.

At a first glance we can observe four main clusters. The first, on the upper right side of the image, is composed uniquely by N elements; the second on the lower right side contains heterogeneous data; the third on the upper left side contains only O elements, and the fourth, in the lower left part of the graph, is heterogeneous.

In order to easy analysis on our set of data it could be useful to reduce the dimension of the original matrix (patients x variables) to decrease the computational burden of working with such a large (and dimensionally unbalanced) data structure.

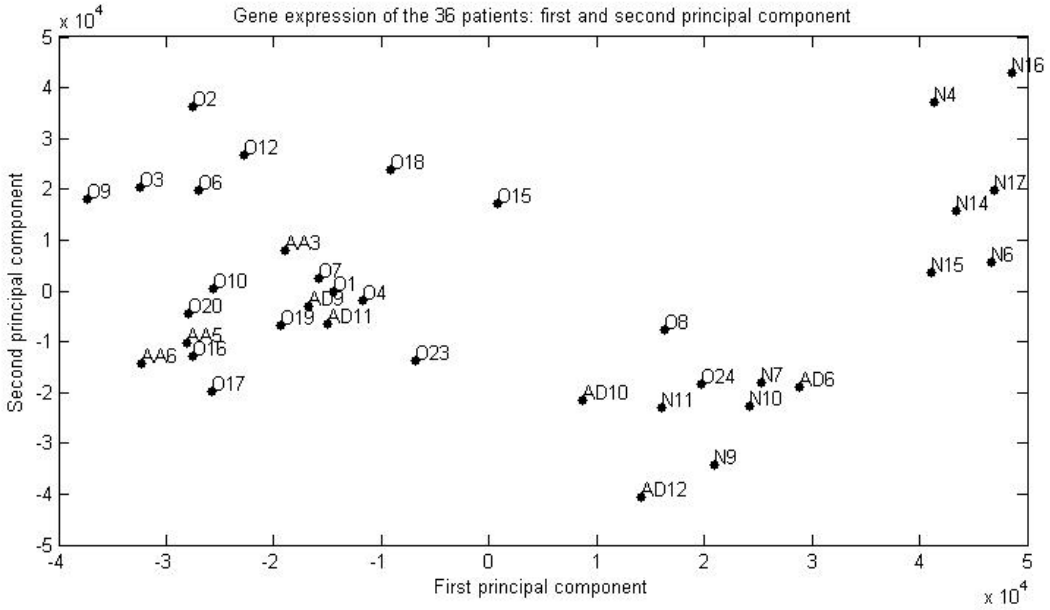

Figure 1 Dataset plotted on the two first principal components

A first pruning of genes not likely to be significant for the final classification has been performed on the dataset. For doing this we have eliminated all those genes whose variability was below a certain threshold. The critical point of adopting this approach is the choice of the threshold which has to be determined in such a way not to eliminate genes that could be relevant for our classification process. The threshold tuned for this application was 50 000: all those genes with a variance less than the threshold were eliminated from the clustering analysis (the maximal variance of a single gene in samples we have considered was 2.10^7 , the minimal variance was 0.019 and the average variance was 4.10^4). The threshold we have chosen is just above the average variance meaning that we mainly eliminated from our dataset the variables less varying across subjects (thus, probably, are the less salient in classifying different patients, considering that they have less variant values across all samples).

This process led to a more manageable matrix, composed by 36x3405 elements (36 patients, 3405 gene expressions). The clustering approach adopted for the classification consisted in the use of a K-means algorithm initialized using a Principal Direction Divisive Partitioning (PDDP) as successfully done for discriminating leukemias[5]".

K-means is an iterative procedure used for partitioning a dataset into k sub-clusters. The procedure begins with the choice of k points considered as the initial centroids of the k clusters we are trying to build. At each step every element of the dataset is associated to the closest centroid and the centroids of the obtained clusters are computed. The procedure is re-iterated till the algorithm stabilizes, i.e. till when two following iterations result in the very same clustering[6].

The main problem using this technique is the initialization: different initial choices of centroids can lead to completely different results. To solve this problem we have used a one-shot algorithm, called Principal Direction Divisive Partitioning, to choose these initialization points. In our particular case we have adopted a bisecting approach: at each step a cluster was divided into two sub clusters (k=2).

PDDP is a non iterative clustering technique. In this context PDDP is used to obtain the two initial centroids for initializing K-means. The idea behind this algorithm is that data are often aggregated in two clusters (possibly spherically shaped) so that the first

principal component (or at least the second one) is oriented from one cluster to the other.

When the configuration of data shows such characteristics, the partition can be obtained by means of a cut along a hyper-plane orthogonal to the first principal component (or to the second one) and passing through the centroid of the data set. We have to put some efforts in choosing along which component is more useful to perform the cut since choosing the wrong direction could lead to a completely wrong classification [5].

In order to detect the genes relevant for the classification we have adopted a technique known as "gene shrinking". Once we have partitioned in the two clusters of interest, we start to eliminate from the original set one of the variables and we see if, by applying the clustering procedure to this modified set, we obtain the same clusters as before. We continue to eliminate variables till the obtained clusters change.

Suppose u is the vector containing all the variables we are working with $u = u_1 u_2 \ldots u_{3405}$ where the variables are sorted by decreasing values of u_i (which measure the projection, thus the importance, of the i^{th} variable for the classification). Let's consider the vector $u' = u_1 u_2 \ldots u_{3404}$. If by applying the partitioning algorithm using u' instead of u the obtained clusters do not change, we continue our procedure by eliminating the next variable. The procedure stops when it is not anymore possible to eliminate other variables without changing the results [5]. In this step we also have applied some empirical considerations. Once most of the genes had been eliminated we have tried to use different combinations with the remaining genes to see if we could reduce the subset of relevant genes further. In this last phase we could also use the analysis of the graph (like in Figure 3) showing the variation of gene expression over the set of patients to empirically find genes that could play an important role in obtaining a particular cluster (supervised technique). The purpose was hopefully to find some genes which allowed obtaining homogeneous clusters: one composed uniquely of control subjects and three others composed only by patients affected by each particular category of glioma under examination.

4 Results

Thanks to our analysis we have been able to identify two genes allowing to separate control subjects from patients affected by gliomas. The genes identified are 236024_at and 236024_at. Applying PDDP + K-means to the dataset containing all patients but only these two variables yields the results shown in Figure 2. The samples have been projected on the first two principal components for a better visualization.

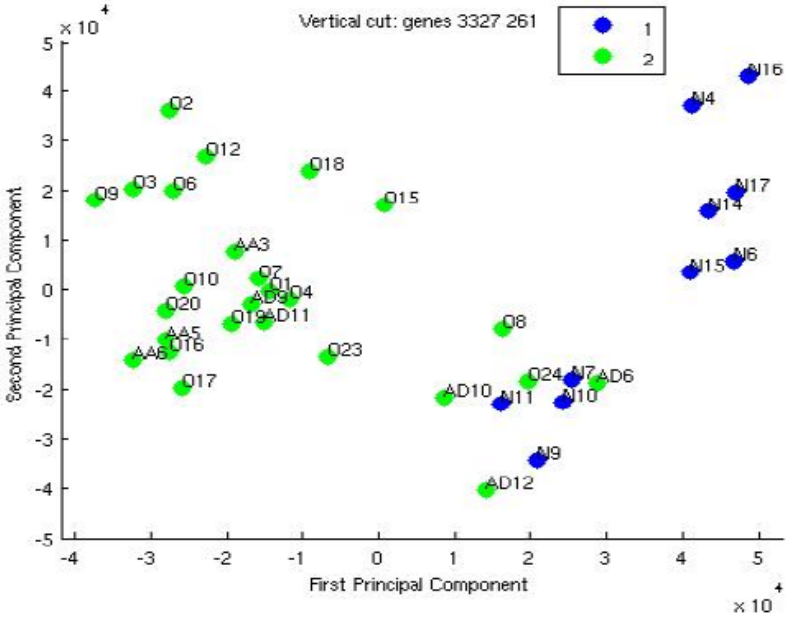

Figure 2 Clusters obtained with the two genes 236024_at and 200774_at

The points labeled by letter N refer to control subjects while the points labeled by letters O, AA, AD refer to patients affected respectively by oligodendrogliomas or mixed gliomas, anaplastic astrocytomas and astrocytomas. The gene expression, over the 36 samples, of the two genes identified is shown in Figure 3.

Gene 236024_at is the most important in separating Ns patients from the others as we can clearly see from the graph representing the variation of gene expression over the entire set of patients. Gene 200774_at(261) improves the quality of the separation. Without this gene two AD patients would probably have been misclassified. Gene expression of gene 236024_at is

higher in normal patients with respect to sick patients. On the other hand gene 200774_at is over expressed in sick patients.

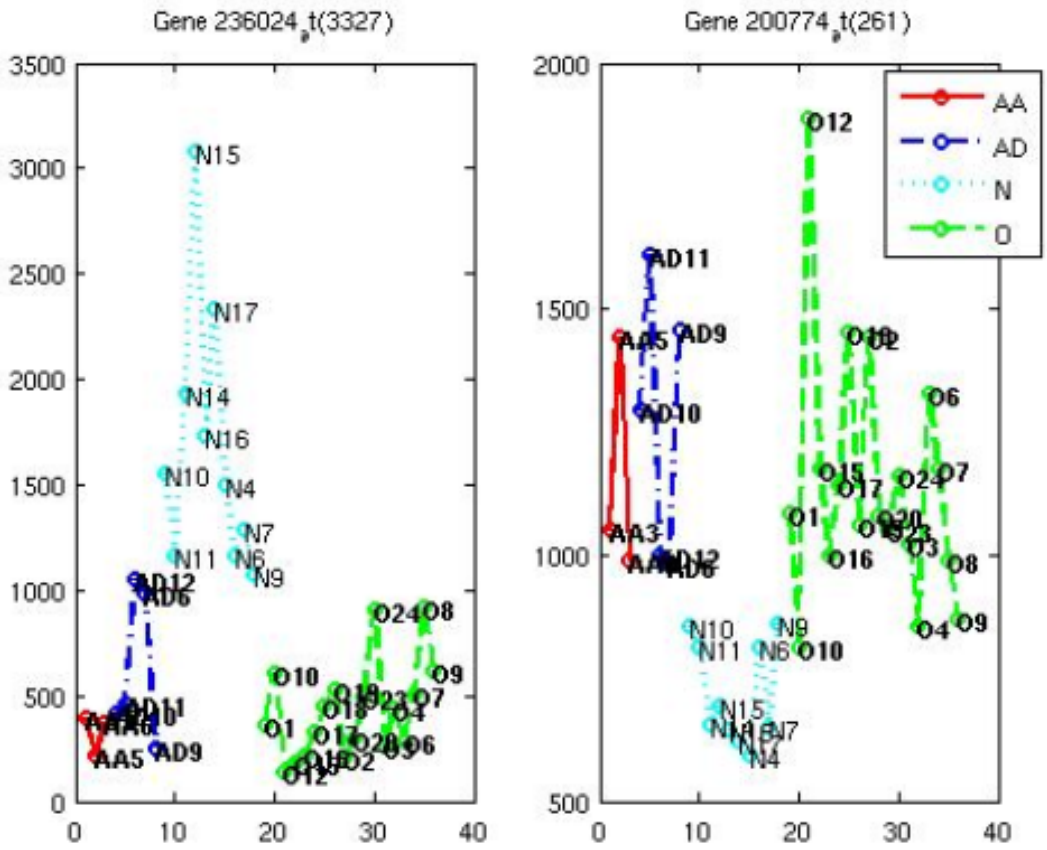

Figure 3 Gene expressions of genes 236024_at and 200774_at

To separate healthy patients from sick ones we can also use some thresholds on gene expression instead of applying a PDDP+K-means clustering procedure which is costly. We can clearly see this from Figure 4 showing the projection of patients on the plan defined by such two identified genes. The black line represents the partitioning line used by the PDDP + K-means algorithm while the vertical and horizontal lines indicate thresholds – as it could also have been automatically identified via Hamming Clustering[7] - on the values of these two genes for establishing if a patient must be considered as affected by a glioma or not. From this plot we can see that if gene 236024_at has an expression greater than 1000 and gene 200774_at has an expression smaller than 900 then the patient is not affected by a glioma at least within our data base: a generalization to bigger and/or different data set could be worth experimenting on larger data bases.

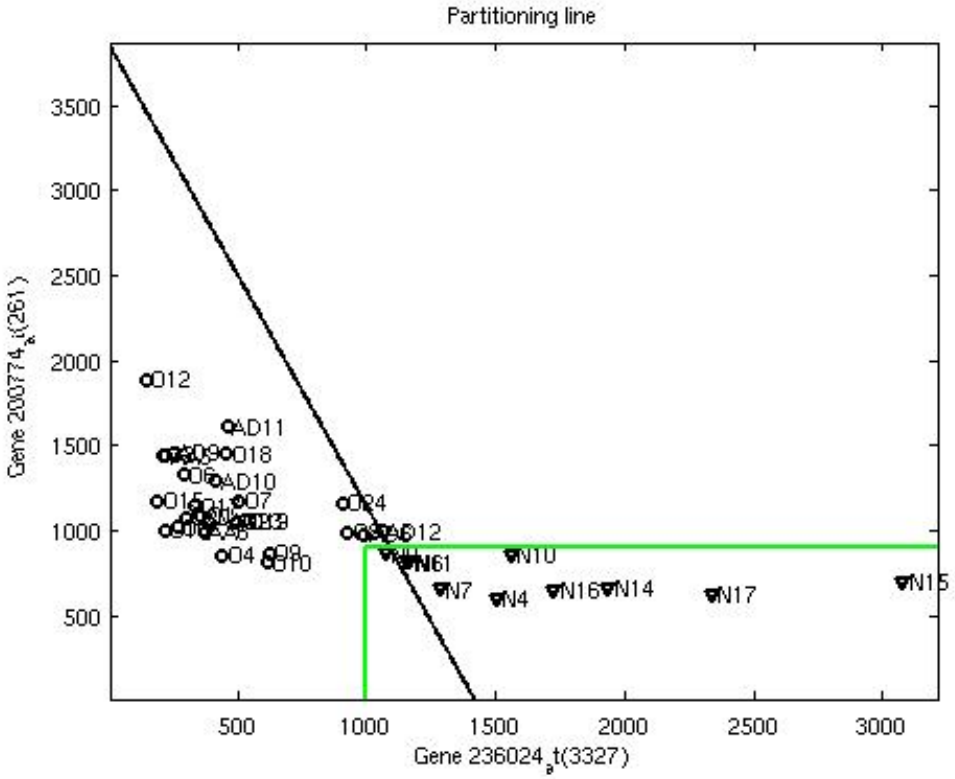

Figure 4 Partitioning line on the plan defined by genes 236024_at and 200774_at

If we analyze the biological function of these two genes we can see that gene 236024_at, whose name is GPM64, is a gene which plays significant roles in neural cell adhesion and some aspects of neurite growth. It synthesizes the neuronal membrane glycoprotein M6A[8]. On the other hand gene 200774_at, called C9orf10, plays a role in nuclear mRNA splicing. It looks to be of interest to further investigate a direct link between the high expression of this gene and the presence of gliomas[9]. If we want to evaluate the degree of correctness of our approach we can apply a method known as cross one out validation. It consists in eliminating one of the elements from the training set and see how the obtained results change [10].

For example for our first step (separating sick patients from healthy ones) we try our algorithm eliminating at every step one of the elements (working then on a data set of 35 patients instead of the complete set of 36 patients). The results obtained are presented in Table 2. We can see that the percentage of errors using these two genes is pretty small. Few elements are misclassified if we consider a dataset smaller than the original one. If we remove the control patients from the dataset and continue with our analysis we can see that using two more genes it is possible to separate patients affected by anaplastic astrocytomas (labeled by the letter AA) from all the other sick patients.

Eliminated Patient	# Errors	Elements placed in the wrong cluster
AA3	3	AD9 N9
AA5	5	AD9 N10 N4 N7 N9
AA6	2	AD9 N9
AD10	2	AD9 N9
AD11	2	AD9 N9
AD12	2	AD9 N9
AD6	3	AD9 N9 AD12
AD9	3	AD9 N9 AD12
N10	2	N7 N9
N11	2	N7 N9
N14	2	N9 AD12
N15	3	N9 AD12 AD6

N16	1	N9
N17	2	AD12 AD6
N4	2	N9 N7
N6	2	N9 N7
N7	2	N9 N7
N9	1	N9
O1	0	
O10	0	
O12	1	N9
O15	0	
O16	0	
O17	0	
O18	0	
O19	0	
O2	0	
O20	0	
O23	0	
O24	1	AD12
O3	0	
O4	0	
O6	0	
O7	0	
O8	1	AD12
O9	0	

Table 2 - Cross one out validation: results

The two genes involved are 234985_at and 224848_at. Results are shown in figure 5.

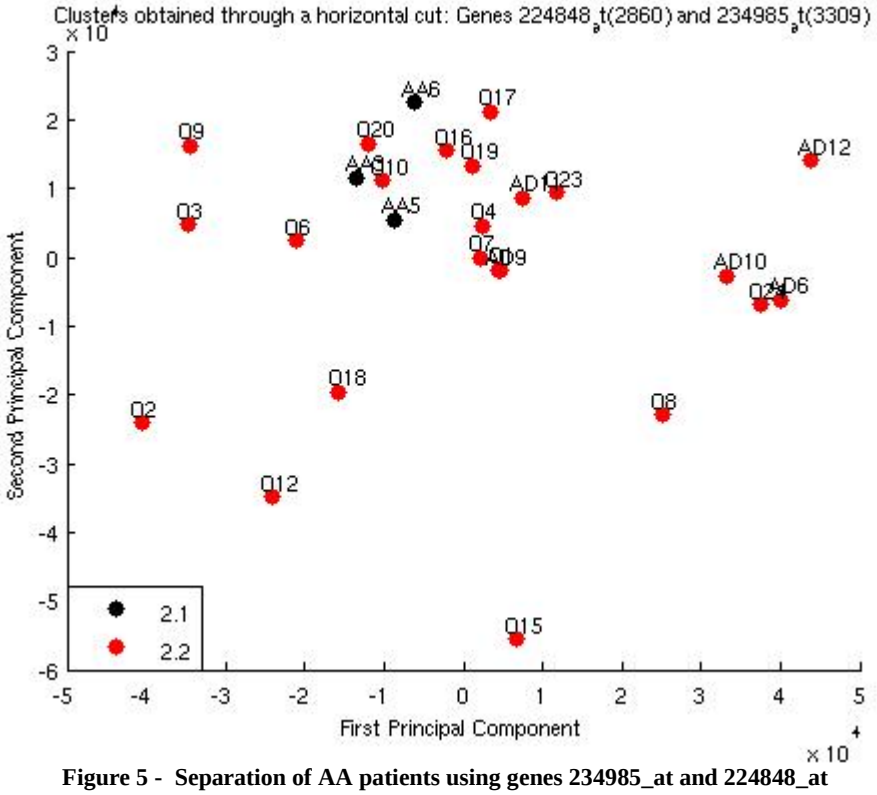

Figure 5 - Separation of AA patients using genes 234985_at and 224848_at

The expression of the two genes discriminating anaplastic astrocytomas are shown in figures 6 and 7

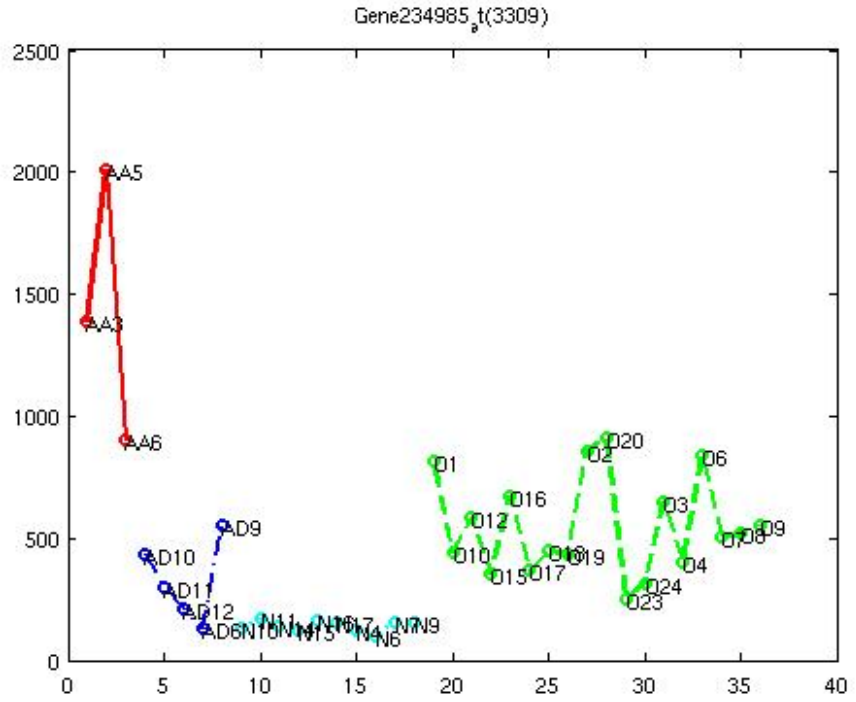

Figure 6 - Gene expression of gene 234985_at

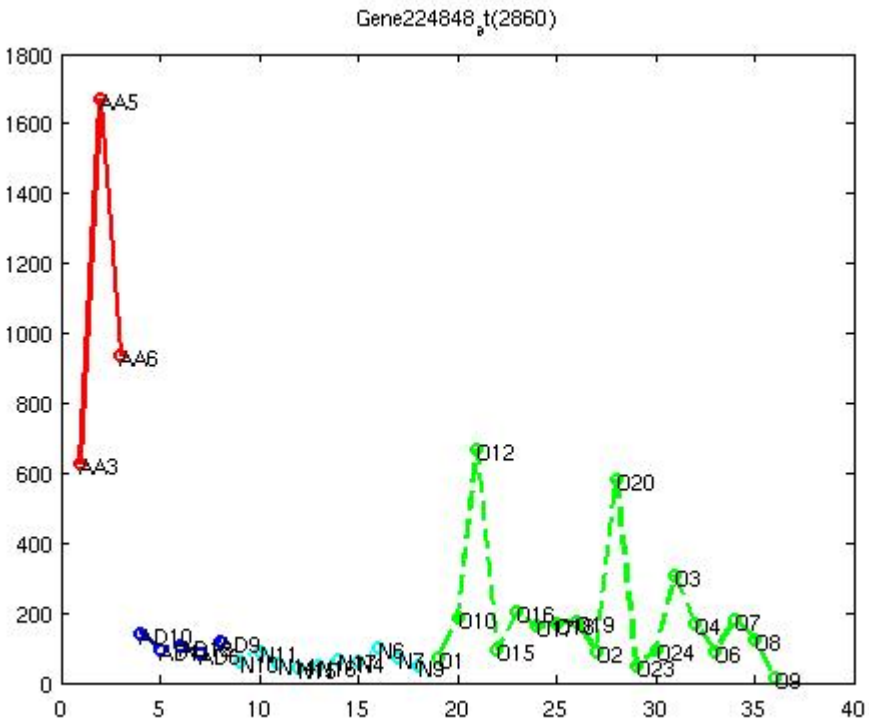

Figure 7 Gene expression of gene 224848_at

We can see that both genes are over expressed in AA patient with respect to other types of gliomas and even to control subjects.

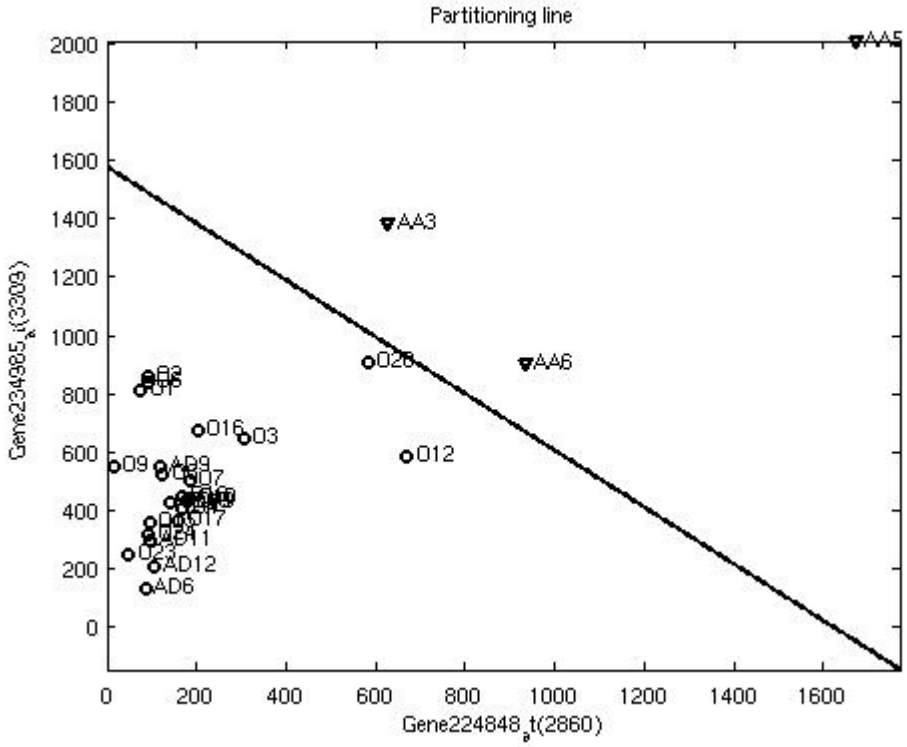

Figure 8 - Partitioning line on the plan defined by gene 234985_at and 224848_at

If we project patients on the plan defined by these two genes like in figure 8, we can see that in this case, in order to identify a threshold on the value of these two genes for characterizing AA patients without using the PDDP + K-means algorithm, one shoud search for instance for a threshold on the sum of their expression, probably around 1600.

If we look for the biological function of these two genes we have that gene 224848_at, called cyclin-dependent protein kinases (CDK6) regulates major cell cycle transitions in eukaryotic cells. It has been suggested that CDK6, and the homologous CDK4, link growth factor stimulation with the onset of cell cycle progression [11]. Gene 234985_at, called Hypothetical protein LOC143458 hasn't been clearly associated with a biological process yet [11]. If we apply the cross-one out validation for evaluating the second step of our hierarchical tree (separating AA patients from all the other sick ones) we obtain the results synthesized in table 3

Eliminated Element	# Errors	Elements placed in the wrong cluster
AA3	1	AA4
AA5	7	O1 O12 O16 O2 O20 O3 O6
AA6	0	
AD10	0	
AD11	0	
AD12	0	
AD6	0	
AD9	0	
O1	0	
O10	0	
O12	0	
O15	0	
O16	0	
O17	0	
O18	0	
O19	0	
O2	0	
O20	0	

O23	0
O24	0
O3	0
O4	0
O6	0
O7	0
O8	0
O9	0

Table 3 - Cross one out validation: results on second hierarchical step

At this step we have even fewer errors. We only obtain misclassified elements if we eliminate from our training set one of the AA patients. This is probably due to the fact that we have such few patients affected by this pathology in our dataset that each of them plays a key role in the classification process.

Eliminating AA patients from our analysis and applying the usual approach, we have been able to find three genes capable of separating most patients affected by astrocytomas from those affected by oligodendrogliomas or mixed gliomas. Results are shown in Figure 9.

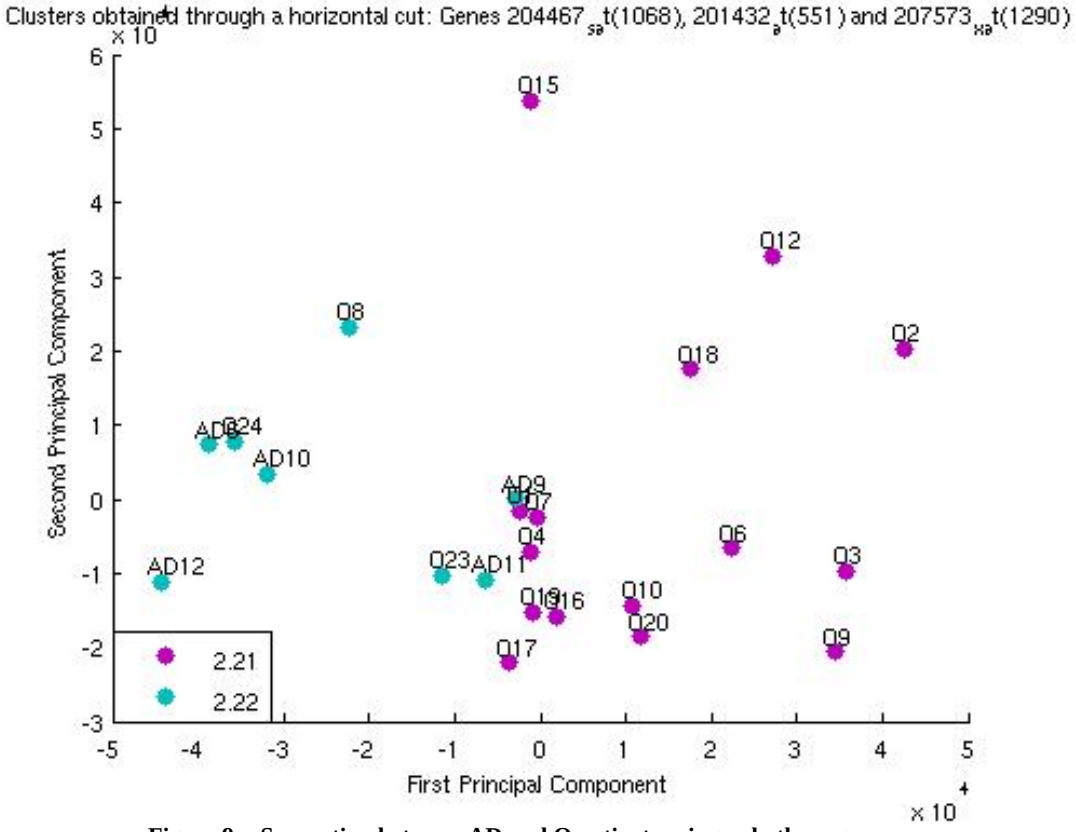

Figure 9 - Separation between AD and O patients using only three genes

All astrocytomas are classified in the same cluster (points labeled by letters AD) but three patients affected by oligodendrogliomas and mixed gliomas are still considered in the wrong cluster.

The three genes involved in this classification step are 204467_s_at, 201432_at, 207573_x_at. Their expression can be seen in figure 10-12.

Gene 204467_s_at, called SNCA, provides instructions for making a small protein called alpha-synuclein, quite abundant in the brain, mainly localizing at the tips of nerve cells (neurons) in specialized structures called presynaptic terminals [11].

Gene 201432_at, called CAT is responsible for the production of catalase. Catalase is an antioxidant protecting cells from hydrogen peroxide. Some experiments in invertebrates suggest a role for CAT in ageing [10]. Gene 207573_x_at, called ATP5L plays a role in ATP synthesis [11].

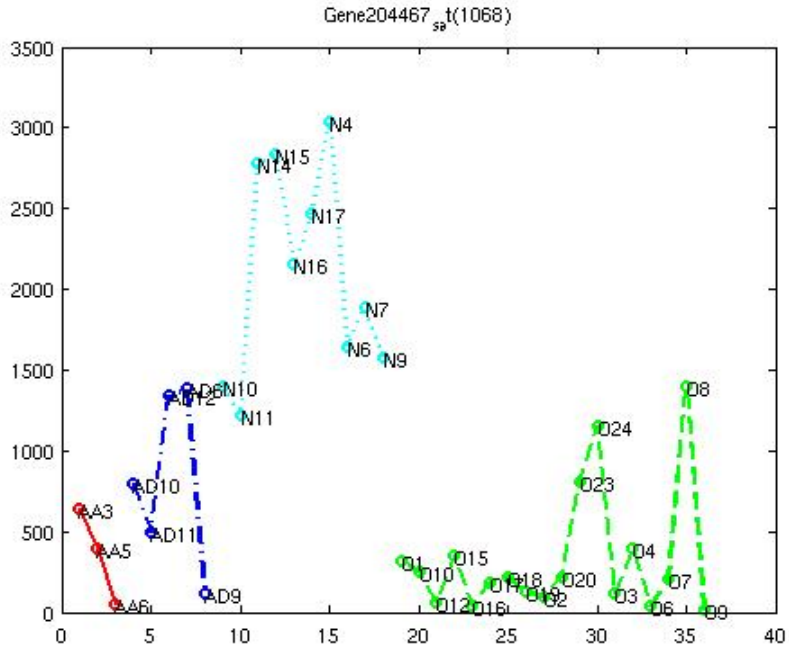

Figure 10 - Gene expression for gene 204467_s_at

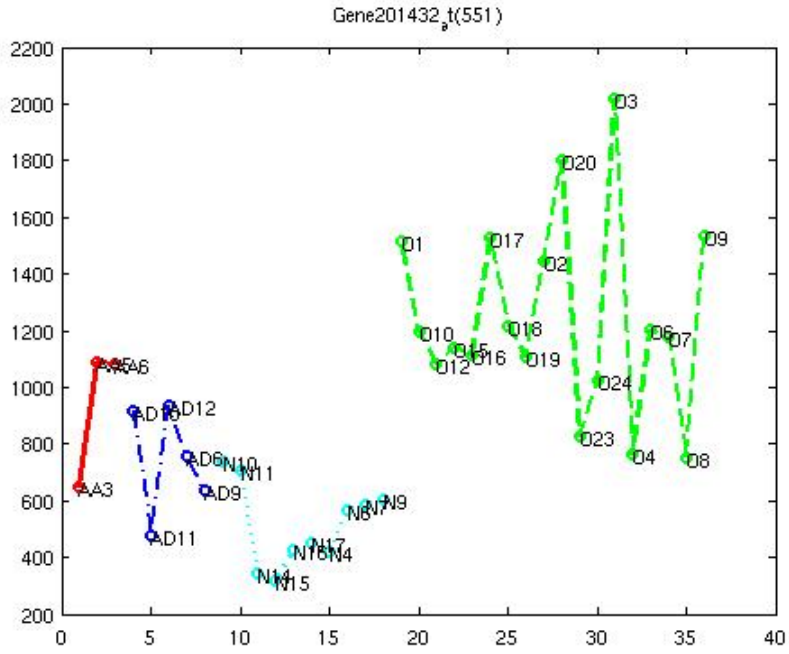

Figure 11 - Gene expression for gene 201432_at

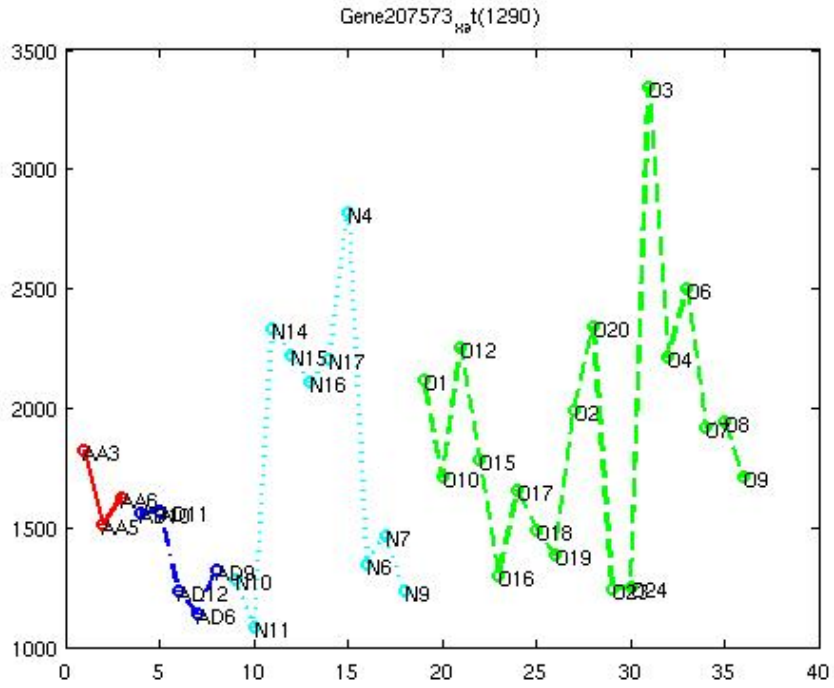

Figure 12 - Gene expression for gene 207573_x_at

Such three genes do not allow alone to obtain a perfect classification. If we apply the cross one out approach to the set containing only patients affected by gliomas of type O and AD we obtain Table 4

We can see that the three elements O23 O24 O8 which had been classified in the same cluster with all the AD patients in the original data set still continue to be considered in the same cluster when we eliminate one patient from the data set. Here the error percentages are much higher. The explanation for the three misclassified O patients could be that three genes are not powerful enough to distinguish these two typologies of gliomas. Other combinations of genes might be better indicated for doing this kind of discrimination. The best approach would be to use powerful parallel systems for evaluating all the possible combinations and see if there are combinations which lead to better results. The same reasoning could be applied to the two previous steps. It could also be possible that a linear classification is not powerful enough for distinguish between AD patients and O ones. In that case we could envisage the possibility to use heavily non-linear methods like the already recalled Hamming Clustering whose nonlinearities being logical does allow fast computing besides keeping inferring understandable rules.

5 Conclusions

Through our work we have been able to discover a small set of genes seeming to play an important role in discriminating healthy patients from sick ones. One gene in particular, called GPM64, seems to be strongly related to this kind of tumor: its expression levels in control patients differ greatly from those of patients affected by gliomas. It is a gene which plays a significant role in neural cell adhesion and some aspects of neurite growth. It synthesizes the neuronal membrane glycoprotein M6A: its pathways are thus probably worth to be deeply investigated with respect to gliomas.

Eliminated Element	# Errors	Elements placed in the wrong cluster
AD10	4	AD9 O23 O24 O8
AD11	4	AD9 O23 O24 O8
AD12	3	O23 O24 O8
AD6	8	O10 O15 O16 O18 O19 O23 O24 O8
AD9	3	O23 O24 O8
O1	3	O23 O24 O8
O10	3	O23 O24 O8
O12	3	O23 O24 O8
O15	3	O23 O24 O8
O16	3	O23 O24 O8

O17	3	O23 O24 O8
O18	3	O23 O24 O8
O19	3	O23 O24 O8
O2	6	O16 O18 O19 O23 O24 O8
O20	4	AD9 O23 O24 O8
O23	3	AD9 O24 O8
O24	2	O23 O8
O3	4	AD9 O23 O24 O8
O4	3	O23 O24 O8
O6	3	O23 O24 O8
O7	3	O23 O24 O8
O8	2	O23 O24
O9	3	O23 O24 O8

Table 4 - Cross one out validation: results with the three salint genes

Regarding the classification of sick patients we have been able to find two genes, CDK6 and LOC143458, helping in discriminating tumors of type AA from all the other categories: their pathways are thus also probably worth to be investigate especially if also involving the previously mentioned gene

CDK6 regulates major cell cycle transitions in eukaryotic cells. It has been suggested that CDK6, and the homologous CDK4, link growth factor stimulation with the onset of cell cycle progression. The function of gene LOC143458 is still not perfectly known: investigating its pathways also in relation to gliomas could even offer further insights even beyond the very scope of our present research

The most difficult step in our classification was the separation of AD patients from O ones. These two kinds of gliomas show similar behavior in gene expression. Nevertheless we have been able to found three genes, SNCA, CAT, ATP5L that seems to be relevant in differentiating the tumors: their expression levels are pretty different in the two typologies. Unluckily the classification we have obtained is only partial since a small set of patients is still misclassified. Further investigation, also with more extended casuistic is surely still needed in this respect

SNCA provides instructions for making a small protein called alpha-synuclein which is abundant in the brain. It localizes mainly at the tips of nerve cells (neurons) in specialized structures called presynaptic terminals.

CAT is responsible for the production of catalase which is an antioxidant protecting cells from hydrogen peroxide. Some experiments in invertebrates suggest a role for CAT in ageing.

ATP5L plays a role in ATP synthesis.

In conclusion we have been able to prove that there's a correlation between the histological classification of gliomas and the expression of certain few genes. The genes we have identified probably play an important role in the set-in and development of the tumor and could be the object of specific studies in order to find aimed strategies for fighting this kind of cancer.

As we have seen such partially linear approach gives good results but it is not perfect especially for separating AD patients from O ones. It could probably be useful to use other techniques in order to see if the impossibility to separate some O patients from AD ones come from the limits of a linear approach or if it is intrinsically related to a genetic similarity between these two typologies of gliomas.

Acknowledgement

Professor Federico Turkheimer, a former great pupil, is warmly acknowledged for having made data available and for key suggestions

References

[1] P. Kleihues, L.H. Sobin, "World Health Organization classification of tumors", *Cancer*, 2000, pp. 88-2887.

[2] F. E. Turkheimer, F. Roncaroli, B. Hennuy, C. Herens, M. Nguyen, D. Martin, A. Evrard, V. Bours, J. Boniver, M. Deprez, "Chromosomal Patterns of Gene Expression from Microarray Data: Methodology, Validation and Clinical Relevance", *BMC Bioinformatics*, 2006, vol. 7, pp. 526-543.

[3] *K. Y. Yeung, W. L. Ruzzo, "Principal component analysis for clustering gene expression data", Bioinformatics, 2001, vol. 17, pp. 763-774.*

[4] I.T. *Jolliffe, Principal Component Analysis, 2nd ed., NY: Springer, 2002.*

[5] S. Garatti, S. Bittanti, D. Liberati, P. Maffezzoli, " An unsupervised clustering approach for leukemia classification based on DNA micro-arrays data", *Intelligent Data Analysis*, 2007, vol. 11, pp. 175-188.

[6] C. Ja-Shen, K.H.C. Russell, L. Yi-Shen, "An extended study of the K-means algorithm for data clustering and its applications" *Journal of the Operational Research Society*, 2004, vol. 55, pp. 976-987.

[7] M Muselli and D Liberati "Binary rule generation via Hamming Clustering", IEEE Transactions on Knowledge and Data Engineering, 14 (6), 1258-1268, 2002
[8] J. Alfonso, M.E. Fernández, B. Cooper, G. Flugge, A.C. Frasch, "The stress-regulated protein M6a is a key modulator for neurite outgrowth and filopodium/spine formation", *Proc. Natl. Acad. Sci.*, 2005, vol. 102, pp. 17196-17201.
[9] Nuclear Protein Database. [Online]. Available: http://npd.hgu.mrc.ac.uk . [Accessed: Feb. 25, 2008].
[10] J. Shao, "Linear Model Selection by Cross-Validation", *Journal of the American Statistical Association*, 1993, vol. 88, pp. 486-494.
[11] National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov .