Gene expression analysis for tumor classification using vector quantization

Edna Márquez[1] Jesús Savage[1], Ana María Espinosa[2], Jaime Berumen[2], Christian Lemaitre[3]

[1] IIMAS, Universidad Nacional Autónoma de México
Cd. Universitaria, Coyoacán , México D.F.
[2]Unidad de Medicina Genómica, Hospital General de México,
[3] Universidad Autónoma Metropolitana
mcmm@servidor.unam.mx

Abstract. Gene expression analysis is one of the most important tasks for genomic medicine, using these it is possible to classify tumors, which are directly related with the development of cancer. This paper presents a clustering method for tumor classification, vector quantization, using gene expression profiles from microarrays of mRNA with samples of cervical cancer and normal cervix. Vector quantization is used to divide the space into regions, and the centroids of the regions represent patients with tumors or healthy ones. Also the regions found by the vector quantizer are used as the base for classifying other tumors, that could help in the prognostics of the illness or for finding new groups of tumors.

Keywords: Gene expression analysis, clustering, vector quantization, tumor classification.

1 Introduction

A few years ago, the traditional way to diagnose and identify a tumor was by direct observation and with histopathology studies and immunohistochemistry. The pathologist identified the cancer cells in a tumor sample and by its shape and staining determines cell lineage. Actually, through molecular biology with the use of gene expression surges another way to identify the tumor type. Unlike that traditional way this is quantifiable and can be automated through the implementation of processing algorithms more accurately and reproducible. In response to these new needs of biology, we propose a clustering method for tumor classification, the vector quantization.

The classification of tumors is very important, since the prognostic and treatment response of the cancer is based on its type, so it is very important to know the exactly class of tumor for the patients. The aim of this work is to contribute for tumor classification with a better precision according the gene expression profile in the

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008,
Supplementary Conf. Proc., pp.95-103, 2008.

tumor tissue, also do a class prediction of new tumors through the assignment of them to one defined class, which could be used in prognostics of clinical parameters relevant to therapy response.

Cervical cancer represents the second most common malignancy in women around the world [6], in Mexico the cervical cancer (CaCu) is the second morbidity cause for the women, and the principal agent for CaCu is the Human Papillomavirus type 16 (HPV16). According its histology the cervical cancer could be of two types: escamous or adenocarcinoma, the first is found in the epithelial cells and the second in the glandular cells, the most aggressive and with poor prognostics is the adenocarcinoma.

For this work, we used 39 tumor samples and 12 control cases, the first 39 samples came from women with detected cervical cancer in invasor state caused by HPV16, and the 12 controls correspond of women who have a normal cervix.

The samples are fixed on oligonucleotide microarrays, with this technology the information of thousands of genes is represented at same time that needs computational methods to processes this huge data biology. In this paper it was applied an unsupervised learning algorithm, the vector quantization (VQ), for classifying the type of cervical cancer from gene expression of thousands genes included in the microarrays of mRNA, and the principal component analysis to reduce the dimension that makes possible the visualization of the clusters found by VQ.

In the next sections, it is presented the general issues about the microarray's data and their preprocessing, it is explained the vector quantization algorithm for analyzing the gene expression profiles, and it is discussed the results derived from the experiments and finally the conclusions.

2 Materials and methods

The data of gene expression were generated using the Affymetrix HGFocus GenChip of mRNA that contains the expression of 8793 genes. The samples came from 39 Mexican women with diagnostic of cervical cancer and 12 control cases which came from women without cervical cancer.

All sample tumors correspond to Human Papillomavirus 16 (HPV16) infection that represents the most important risk factor for the development of cervical cancer and is linked to a high incidence of cervical cancer in Mexico and HPV16 is the most frequently detected (50%) all over the world [7]. The tumor samples have one of the two types of cervical cancer: adenocarcinoma and epidermoid (escamous), 13 samples correspond to adenocarcinoma and 26 samples correspond to epidermoid and 12 samples are controls of normal cervix.

2.1 Normalization

The data of the matrix that contains the intensity values of thousands genes generated by the mRNA microarrays were passed through a process of normalization called

Robust Multi-chip Analysis (RMA) [2]. The normalization is used as a factor of adjustment to the data in compensation of the experimental variability and it is necessary for the comparition of samples. In the GeneChips of mRNA typically each gene is represented by 16-20 pairs of oligonucleotides called probe sets. The first component of these pairs is referred to as a perfect match (PM) probe that is paired with a mismatch (MM) probe, the PM and MM are referred to as a probe pair, in this algorithm the MM is ignored and the PM is normalized with the background. The RMA consists on $\log_2$ of the correction of PM with the background given to the probes. This RMA normalization is carried out using the statistical language R[3]. All the samples used for the experiments were normalized with RMA and its appearance is viewed in figure 1.

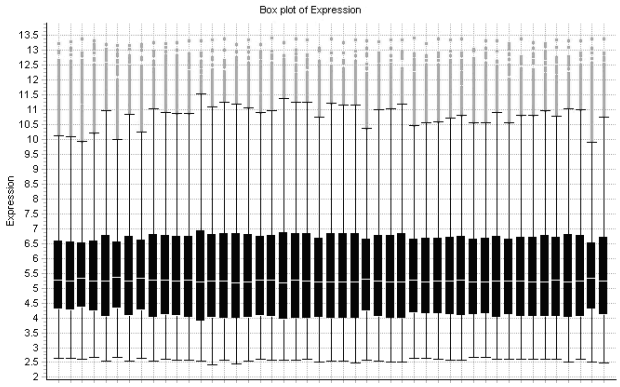


Fig. 1. With RMA normalization the samples have the $\log_2$ intensity, and that box plot shows the dispersion of the normalized microarrays.

2.2 Vector Quantization

Vector Quantization techniques [4] have been extensively used for data compression in digital signal processing and telecommunications. And for gene expression, VQ is used as a good approximation for finding classes of tumors and also of genes, as is presented in [8].

Given a set of N vectors, $p_j = \{x_{j1}, x_{j2}, \ldots, x_{jm}\}$; j = 1...N, and m=1...8793 that represents a gene expression profile, a set of centroids is found. The centroids represent the vectors in each cluster, and the collection of centroids is called a codebook. Then the codebook is designed from a long training sequence that is representative of all p_j vectors to be encoded by the system. The codebook is created with the Linde-Buzo-Gray (LBG) algorithm [4], which is based on the generalized Lloyd Algorithm [5].

Following is the VQ algorithm:

1. Find an initial codebook D_1, with one centroid C_1, by averaging all the vectors p_j, with $L_m=1$;
2. Modify each of the centroids C_i; i = 1...L_m in D_m by adding them a vector φ of small magnitude to generate new centroids from each of them, generating a new codebook D_m+1, L_m+1;
3. Given a codebook $D_m = C_i$; i = 1...L_m assign each vector p_j into the clusters R_k whose centroid C_k is the closest to p_j according to some distortion measure $d(p_j,C_k)$. In this case, the distortion measure is the Euclidean distance between two vectors (1);
4. Recompute the centroids C_k for each of the clusters, R_k, by averaging all the vectors p_j that belong to R_k;
5. If the average distortion of D_m is bigger than a certain ε, go to 3;
6. If L_m < codebook size go to 2, where codebook size is the number of clusters of the environment.

To calculate the distortion between the two vectors, the centroid of the cluster and the gene expression profile, is used the Euclidean distance:

$$d(p_j, C_k) = \sqrt{\sum_{i=1}^{m} (x_{ji} - y_{ki})^2} \qquad (1)$$

In the gene expression analysis for tumor classification one vector corresponds to a sample, tumor or control, with the dimension equals to number of expressed gene. In this case there are 51 vectors, and their dimension is 8793, $p_j = \{x_{j1}, x_{j2}, ..., x_{j8793}\}$; j=1...51, this is the set of vectors in gene expression analysis for tumor classification:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \ldots\ldots x_{1m} \\ x_{21} & x_{22} & x_{23} \ldots\ldots x_{2m} \\ x_{31} & x_{32} & x_{33} \ldots\ldots x_{3m} \\ . & . & . \quad . \\ . & . & . \quad . \\ . & . & . \quad . \\ x_{N1} & x_{N2} & x_{N3} \ldots\ldots x_{Nm} \end{pmatrix}$$

N=1…51 samples and

m=1…8793 genes

For the visualization of the clusters in the space is necessary to reduce the dimension of the vectors that represent the centroids, from N-dimension to 2-dimension or 3-dimension, this task is through the Principal Components Analysis (PCA) [1], where a component is equivalent to one dimension. With PCA it can see graphically the distribution in the space of the found clusters represented by the vectors of centroids.

3 Experiments and results

The VQ algorithm was applied to the data of gene expression of 39 tumors and 12 control cases of cervical cancer. Previously, the data of microarrays were normalized

with RMA algorithm [2], so it was found their classification in few experiments changing the dimension of the codebook. And finally, we made some experiments of classification of new samples, when a new sample is presented to the clusters, it is possible to know what cluster represents it, with the similarity of gene expression profile, the pattern expression of all samples which into one cluster is expressed with the centroid of the cluster.

The results of the experiments were compared with the classification done with the histological studies of the same samples.

3.1 Clustering

The first experiment is the creation of 2 groups or classes: tumor and control, it was used the 51 examples (39 tumors and 12 control cases). The results in this experiment were accurate, one group was created with the 39 tumors and the other group has the 12 controls, in table 1 it can see the 2 clusters, one with the tumors and another with the control cases.

Table 1. Tumor and control classification with a codebook size of 2. The results in this experiment were 100% accurate.

	Cluster1	Cluster2	Total
Tumor	39	0	39
Controls	0	12	12
Total	39	12	51

In the second experiment were created three groups with the 51 samples, one group with the control samples, another group with the adenocarcinoma samples and the third group with the epidermoid samples. The difference with their histological classification was 7.8%.

Table 2. In this case the control samples are well classified. In the other two clusters are 4 samples misclassified, 3 epidermoids and 1 adenocarcinoma.

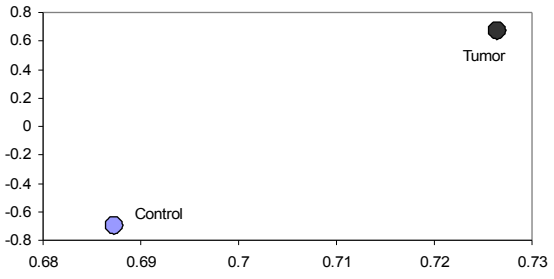
	Cluster1	Cluster2	Cluster3	Total
Epidermoid	23	3	0	26
Adenocarcinoma	1	12	0	13
Control	0	0	12	12
Total	24	15	12	51

To improve the classification in the third experiment with the 51 samples we define 1 cluster for control samples, 2 clusters for epidermoids and 1 for adenocarcinoma, however again 4 tumors were misclassified according the histological type of tumors.

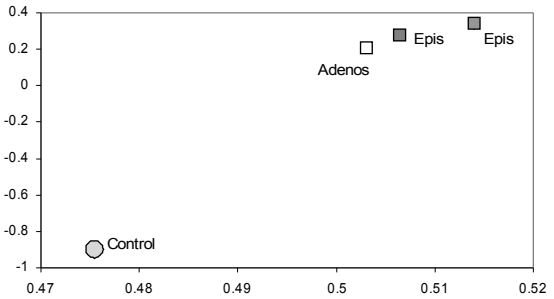
Table 3. Tumor and control classification with a codebook size of 4. In this separation between controls and tumors with 4 clusters, 3 epidermoids and 1 adenocarcinoma were misclassified. The difference of classification was 7.8%

	Cluster1	Cluster2	Cluster3	Cluster4	Total
Epidermoid	14	0	3	9	26
Adenocarcinoma	0	0	12	1	13
Control	0	12	0	0	12
Total	14	12	15	10	51

For visualization of the clusters' graphics was used the PCA algorithm[1], the dimension of centroids was reduced from 8793 to only 2 dimensions, with the representation of the 2 main components. Figure 2 presents the 2-dimenssion graphic of the centroids. The distance between controls and tumors are very long, it expresses clearly the difference of gene expression between a normal cell and a tumoral cell.



a)



b)

Fig. 2. Graphics present a 2-dimenssional separation of tumor and control centroides with PCA. a) representation of the 2 clusters of table 1 and, b) representation of 4 clusters of table 3.

In the follwing experiments we used only the tumors samples for classifying the cervical cancer in two histological types: epidermoids and adenocarcinoma, the results are similar to previous experiments, that is, the same 4 tumors were classified in different clusters again when we compare with the results of histopathology.

In the table 4, with 39 tumor samples the classification in two clusters: 3 epidermoid tumors were put in the adenocarcinoma's cluster and 1 adenocarcinoma tumor in the epidermoid's cluster. In table 5, with 3 clusters we got the same result with 4 classified samples. Results for 4 clusters are presented in table 6. In this case, the number of misclassified samples incressed.

Table 4. Tumor classification with a codebook size of 2. In this separation there are 4 tumors classified not in the expected cluster, 3 epidermoids and 1 adenocarcinoma.

	Cluster1	Cluster2	Total
Epidermoid	3	23	26
Adenocarcinoma	12	1	13
Total	15	24	39

Table 5. Tumor classification with a codebook size of 3. In this separation there are 2 clusters for epidermoid tumors and 1 for adenocarcinoma, 4 tumors were classified not in the expected cluster.

	Cluster1	Cluster2	Cluster3	Total
Epidermoid	3	13	10	26
Adenocarcinoma	12	0	1	13
Total	15	13	11	39

Table 6. Tumor classification with a codebook size of 4. In this separation there are 2 clusters for epidermoid tumors and 2 for adenocarcinoma, but the number of tumors misclassified increased to 5 samples.

	Cluster1	Cluster2	Cluster3	Cluster4	Total
Epidermoid	3	13	9	1	26
Adenocarcinoma	9	0	1	3	13
Total	12	13	10	4	39

In all the previous experiments the samples that were misclassified are the same, then the classification is not exactly the same when it is used the gene expression profile, that is the information from molecular biology through the numerical gene expression and with the histopathological observation have some variants.

3.2 Classification of new samples

With the clusters it is possible to find the classification for a new gene expression profile and to know its representative class. Each cluster is represented by its centroid and the new sample is compared with the similarity measure, the centroid with the biggest similarity belongs to the representative class. In the genomic medicine this task could help to know about the development of the tumor, the response to the treatment or the expectations for the tumor evolution, according with the tumors knowing yet.

In this case the classification was performed with 22 new samples, 16 tumors and 6 control cases, and we did their prediction in two basic classes: tumor and control, and finally using the type of tumor. For table 7 were presented the 22 new samples to two clusters: one of tumors and another of controls, and 1 new control sample was classified into the tumor cluster and 2 tumors into the control cluster.

Table 7. The results of prediction with new samples presented to the clusters, the prediction of 19 was accurate and 3 was wrong (2 tumors and one control).

Samples	Control	Tumor	Wrong
22	5	14	3

When the classification is done with the histological tumor types: adenocarcinoma and epidermoid, with a cluster for each one and other for the control cases, the result is showed in table 8.

Table 8. The prediction of 22 new samples presented to 3 clusters shows 3 errors: 1 epidermoid and 1 adenocarcinoma tumors and 1 control.

Cluster	Right	Wrong	Total
Epidermoid	12	1	13
Adenocarcinoma	2	1	3
Control	5	1	6
Total	19	3	22

The clusters previously found with VQ method represent the patterns of gene expression of the tumors or control cases which were used in the classification of new samples. The qualification of right or wrong classification is according with the comparison against the histological analysis. The difference type of information make necessary to apply another taste with the samples to identify the type of tumor in the cases which the results were not equals between VQ method and histopathological observation.

4 Conclusion

In this paper, VQ was used for gene expression analysis in the tumor classification, with the clusters found it is possible the identification of tumor classification using thousands of genes in the microarrays. The tumor classification with VQ, using numerical information of thousands expressed genes, gave very similar results to the classification with histopathological observation where the information is not quantifiable. VQ represents another way for classifying diseases like cancer, where

the identification of the type is very important for the treatment and the prediction of its evolution.

The four tumor samples misclassified in the first part of the experiments were classified very similar with other clustering method, self-organizing map, then the classification of diseases is an open problem where the clustering methods could help to solve them.

References

1. Everitt B., Dunn G.: Applied Multivariate Data Analysis, Oxford, University Press, New York (1992)
2. Irizarry R. et al.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat., 4, (2003) 249-264
3. http://www.r-project.org/index.html, The R Project for Statistical Computing
4. Linde Y., Buzo A., Gray R.: An Algorithm forVector Quantizer Design. IEEE Transactions on Communications, (1980) 84-95
5. Lloyd. S.: Least Squares Quantization in PCM. IEEE Transactions on information Theory, IT-28: (1982) 127-135
6. Sanjosé S. et al.: Worldwide prevalence and genotype distribution of cervical human papillomavirus DNA in women with normal cytology:a meta-analysis, Review of http://infection.thelancet.com, Vol 7 (2007)
7. Snijders J, Steenbergen R, Heiderman D, Meijer C.: HPV-mediated cervical carcinogenesis: concepts and clinical implications. J Pathol. (2006) 208:152-64
8. Pham T., Dominik B., and Hong Y.:Spectral Pattern Comparison Methods for Cancer Classification Based on Microarray Gene Expression Data, IEEE transactions on circuits and systems, vol. 53, no. 11, november, (2006) 2425-2430