

Marker Gene Selection Evaluation in Brain Tumor Patients Using Parallel Coordinates

Atiq U. Islam, Khan M. Iftekharuddin, David J. Russomanno
Department of Electrical and Computer Engineering
The University of Memphis
Memphis, TN 38152

Abstract—Based on microarray gene expression datasets, many statistical methods have been proposed to locate the significant differentially expressed genes (marker genes) among different sample groups. Although robust models for identifying marker genes more accurately is an area of intense research, effective tools for the evaluation of results is often ignored in the literature. In this paper, we propose a novel visualization method to evaluate the marker gene selection process in brain tumors. We use parallel coordinates to visualize the expression patterns of the marker genes in a way that facilitates the qualitative measure of the overall accuracy of the selection process. We exploit our proposed method to evaluate the robustness of 2 statistical tests as examples of gene selection methods. To measure the reliability of our evaluation process we exploit a brain tumor prediction mechanism based on the selected marker genes. It is anticipated that if the marker genes are precisely located, tumor prediction accuracy is improved. We check if our parallel coordinate based visual understanding about the exactness of the selected marker genes is in agreement with tumor prediction method. We surmise that the prediction results support our conclusion based on visualization.

Index Terms— DNA microarray, marker gene, validation.

1.0 INTRODUCTION

Since the advent of microarray technology, numerous statistical methods have been proposed to locate the significant differentially expressed genes (marker genes) among different sample groups. Although robust models for identifying marker genes more accurately is an area of intense research, effective tools for the evaluation of results is important to assess the supremacy of one method over the others. Furthermore, the statistical tests are sensitive to parameter tuning. For example, a tight-fisted cutoff may miss some of the important marker genes, whereas a generous threshold increases the number of false positives. Therefore, an evaluation process is important to fine tune the parameters. In this paper, we propose a novel visualization approach to qualitatively measure the performance of marker gene selection process in brain tumors.

2.0 DATASET AND PREPROCESSING

The data set is DNA microarray gene expression derived from 34 patient samples with 2 types of medulloblastomas [1]: 9 samples belong to desmoplastic and the remaining samples belong to classical medulloblastoma. HuGeneFL arrays containing 5,920 known genes and 897 sequence tags were used for hybridization. Data preprocessing is done per the procedure identified in [1] and subsequent steps use log scaled values of the gene expressions.

3.0 METHODS AND IMPLEMENTATION

In our proposed visualization method, average gene expression levels of any specific brain tumor are drawn as polylines in a parallel coordinate plot. Parallel Coordinates is a multi-dimensional data visualization scheme that exploits 2D pattern recognition capabilities of humans [2]. Here, the parallel and equally spaced axes represent individual genes. For each group, separate polylines are drawn. The more the gene expression levels differ between groups, the more space appears among the polylines in the plot. The purpose of the plot is not to locate marker genes, but rather to qualitatively measure the performance of the marker gene selection process. The steps are formalized as follows:

1. Locate a set G of genes that are differentially expressed among different treatment groups.
2. Set the average expression of each gene within any specific group as: $\overline{X}_{gk} = \sum_{i=1}^N X_{igk}$, where X_{igk} ($i=1..N$, $g=1..G$, $k=1..2$) is the gene expression value of g -th gene, i -th sample, and k -th treatment group. N is the number of replication. $[\overline{X}_{gk}]$ can be termed as meta-gene expression level of g -th gene for group k
3. Partition the genes into two clusters: (i) C_1 where genes are up-regulated in treatment group 1 compared to group 2; and (ii) C_2 where genes are up-regulated in treatment group 2 compared to group 1.
4. For each cluster C_j ($j=1,2$), again cluster the gene expression values \overline{X}_{gk} into C_{cp} ($p=1, \dots, P$) clusters exploiting self-organizing maps (SOM) [3]. SOM consists of a regular grid of units and the units learn to represent statistical data, described by model vectors $x \in \mathcal{R}^n$.
5. For each cluster C_{cp} ($p=1, \dots, P$):
 - a. Sort the genes according to $\|\overline{X}_{g1} - \overline{X}_{g2}\|$
 - b. Normalize each meta-gene expression values \overline{X}_{gk} between 0 and 1
 - c. Plot meta-gene expression values \overline{X}_{gk} using parallel coordinates, where each parallel axes corresponds to a specific gene and each polyline corresponds to a specific treatment group.

4.0 BRAIN TUMOR PREDICTION

We exploit a brain tumor prediction method to evaluate the performance of our visualization method. We derive prototypes for each category of tumor samples such as desmoplastic and classical medulloblastoma. Such a prototype consists of the mean expression levels of the marker genes of any specific category. To predict the tumor category of a new sample, we calculate the Euclidian distance between the new sample and each of the prototypes. The category of the new sample is predicted as that of the nearest one (desmoplastic or classical). Leave-one-out method is used to measure the performance of the prediction method. It is intuitive that as more “true” marker genes and/or less “false” marker genes are used in building tumor prototypes, the between-prototype distance in the Euclidean space increases. Hence, it is reasonably anticipated that a better marker gene selection method helps to build better tumor prototypes which yield better prediction accuracy.

5.0 RESULTS AND DISCUSSION

In this section, we explore 2 marker gene selection methods such as t-test [4] and empirical Bayesian analysis (ebam) [5]. The purpose is to illustrate the usefulness of our visualization method in comparing the effectiveness of the methods.

Figure 1(a) shows genes selected using individual student t-test (with $p < 0.0005$). Genes are partitioned into 4 clusters. The first two clusters (numbered from left to right and top to bottom) consist of genes that are up-regulated in classical medulloblastoma while the next two clusters consist of genes up-regulated in desmoplastic. We observe that on *some* of the parallel coordinates the lines are not far apart. This observation can be interpreted as the corresponding genes are not significantly differentially expressed between the treatment groups. These marker genes result in 85% prediction accuracy based on the prediction method discussed in Section 4.0. In Figure 1 (b) the average expression levels of 52 genes are plotted that are selected using ebam test (with $\delta = 0.80$). With these selected marker genes we obtain prediction accuracy of 82%. Note that although ebam locates more genes as marker genes, many of the axes have lines that are not far apart and hence do not represent marker genes. From our visual comparison, we conclude that for the current parameter setting, the t-test offers less false positives. This visual observation is clearly supported by the tumor prediction accuracy (85% for t-test vs. 82% for ebam).

6.0 CONCLUSION

In this paper we proposed a novel approach to qualitatively evaluate the performance of marker gene selection methods using visualization. We provided examples to illustrate how this visualization helps to assess the supremacy of one marker gene selection method over others. It is also intuitive that this method is very useful in parameter tuning of the gene selection methods. Because of space limitations we can not provide examples to illustrate this case. Thus, the proposed visualization helps to obtain a set of “true” marker genes with less impurity. Such sets of “pure” marker genes are very

crucial for accurate brain tumor diagnosis, prognosis and therapy.

7.0 ACKNOWLEDGMENTS

The research in this paper is supported in-part through research grants [RG-01-0125, TG-04-0026] provided by the Whitaker Foundation with Khan M. Iftikharuddin as the principal investigator.

8.0 REFERENCES

- [1] Pomeroy S.L., Tamayo P., Gaasenbeek M., Sturla L.M., Angelo M., McLaughlin M.E., Kim J.Y., Goumnerova L.C., Black P.M., Lau C., Allen J.C., Zagzag D., Olson J.M., Curran T., Wetmore C., Biegel J.A., Poggio T., Mukherjee S., Rifkin R., Califano A., Stolovitzky G., Louis D.N., Mesirov J.P., Lander E.S., and Golub T.R., “Prediction of central nervous system embryonal tumor outcome based on gene expression,” *Nature*, vol. 415, p. 436-442, 2002.
- [2] Inselberg, A. and Dimsdale B., “Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry,” In *Proceedings of IEEE Conference on Visualization*, pp. 361-378, 1990.
- [3] Kohonen T., *Self-Organization and Associative Memory*, 2nd Edition, Springer-Verlag, Berlin, 1987.
- [4] Crawley, M., “Statistics: An Introduction using R,” *John Wiley & Sons Ltd.*, 2005.
- [5] Efron B., Tibshirani R., Storey J.D., and Tusher V., “Empirical Bayes analysis of a microarray experiment,” *Journal of American Statistics Assoc.*, vol. 96, no. 115, pp. 1-60, 2001.

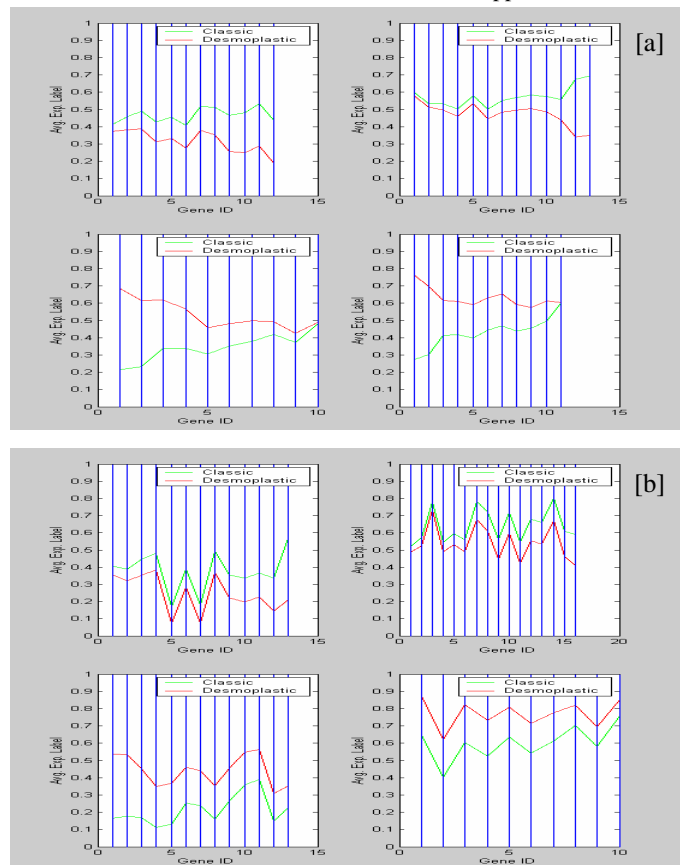


Figure 1: Expression patterns of the marker genes associated with desmoplastic (red line) and classical (green line) medulloblastoma. Genes are selected using (a) t-test [4] (46 genes, $p < 0.0005$, 85%), (b) ebam [5] (52 genes, $\delta = 0.80$, 82%).