ISSN 1060-992X, Optical Memory and Neural Networks (Information Optics), 2010, Vol. 19, No. 2, pp. 181–186. © Allerton Press, Inc., 2010.

47glioblastoma Gene Expression Profile Diagnostics by the Artificial Neural Networks[1]

Y. A. Kuperin[a], A. A. Mekler[b], I. Kniazeva[c], D. R. Schwartz[d], V. V. Dmitrenko[e], V. I. Rimar[e], and V. M. Kavsan[e]

[a]*Saint Petersburg State University, Saint Petersburg, Russia*
[b]*Institute of Human Brain, Russian Acad. Sci., Saint Petersburg, Russia*
[c]*Central Astronomical Observatory at Pulkovo, Russian Acad. Sci., Saint Petersburg, Russia*
[d]*Saint Petersburg State Polytechnical University, Saint-Petersburg, Russia*
[e]*The Institute of Molecular Biology and Genetics of NASU, Kiev, Ukraine*
e-mail: mekler@narod.ru

Received April 5, 2010; in final form, April 16, 2010

Abstract—Two artificial neural networks of different types were applied to profiles of gene expression in most aggressive human brain tumor – glioblastoma – and in normal brain tissue. The results of gene expression profiles classification are presented. First method – self organizing maps – gave good discrimination of profiles on the trained map. Another ANN – perceptron – showed a good result of classification – more then 95% of the test data set were successfully classified. Due to high correlations between some gene expression values one can suppose, that number of genes necessary for successful classification may be reduced.

Key words: Glioblastoma, gene, artificial intelligence, perceptron, self-organizing maps.

DOI: 10.3103/S1060992X10020098

INTRODUCTION

Development of molecular biomarkers and molecular therapeutics forms the basis for the personalized cancer medicine in the 21st century. Now the cancer therapy is changing from a “one size fits all” approach to more personalized one, in which each patient is treated according to the genetic defects in his tumor. Identification and characterization of specific gene expression profiles (so-called gene-expression signatures) in tumors is a significant contribution to understanding of the molecular features of malignancies, mechanisms of their arising and development.

Individual molecular markers have limited significance for the diagnostic evaluation of tumors due to the high heterogeneity of their biological properties, and only simultaneous analysis of changes in many marker genes can characterize individual tumor reliably. Despite some differences in the multigene signatures, gene profiles for the same tumor types have similar distribution on functional groups that can give similar prognostic information. In an effort to identify genes, which might be used as molecular markers for glial tumors, we compared gene expression in glioblastoma and normal adult human brain. Obtained results demonstrated 129 genes with more than 5-fold change of expression in tumors compared to the normal brain cells [1]. Further characterization of these genes will result in the development of so-called cDNA-panels, which can be used for molecular typing of human brain tumors, i.e. determination of certain molecular variants for the tumors of identical histological type may serve for diagnostics and as targets for molecular therapy.

There are several examples of signatures developed for prognostic evaluation of the definite cancer types. 70-gene signature was tested successfully on the big group of patients for the prognostic evaluation of breast cancer [2]. Another research group independently identified a 76-gene signature by similar approach [3]. 16-gene signature was identified also using 250 genes tested on 400 tumors [4]. Amongst all gene-expression signatures that have been identified up to now, only three are commercially available: the 70-gene signature for breast cancer prognosis under the name MammaPrint (Agendia); the 16-gene sig-

[1] The article is published in the original.

Table 1. Input data format

N	Name	CAMK2B (#1)	STMN2 (#2)	…	PSMB8 (#19)	KIF20 A (#20)
1	GSM97800	1.83	1.47	…	0	0
2	GSM97803	1.69	1.14	…	0	0
…	…	…	…	…	…	…
295	GSM96975	0.44	0.03	…	0.11	0.03
296	GSM96986	0.1	0.79	…	0.25	0.08

nature as Oncotype DX (Genomic Health); and a 2-gene signature, which has been recently released, under the name the H/I test (AviaraDx) [5].

Artificial neural networks were used in some cases for the classification of genes according to their expression. For example, such approach was used for prostate cancer [6]. Changes of gene expression profile during the diauxic shift in Saccharomyces cerevisiae were characterized by Kohonen self-organizing map [7]. This method was considered as one of the alternatives for the differential diagnostics of leukemia [8].

To-day, there is no similar commercial signature for glial tumors, although several publications described the gene expression profiles allowing to characterize differences mostly between glioblastoma and normal brain [9, 10], including the implementing of artificial neural map [11] were published. However, the identification of gene-expression signature for glioblastoma, which can be used for molecular typing and prognostic evaluation of glial tumors, is an actual problem still. It has theoretical as well as practical importance.

PROBLEM DEFINITION AND DATA DESCRIPTION

The main task of the study is revelation of group of genes, which expression profiles are most easy to be classified by means of ANN with the purpose of diagnostics of some oncological diseases. One of the most important subtasks of this task is the clusterization of the group of subjects into two subgroups — healthy and ill — judging the levels of expression of the genes set. In the present work this task is performed implementing Kohonen self-organized maps and multilayer perceptron. Selection of the genes for classification is performed using database of two modern methods of the expressive genetics — series analysis of the gene expression (SAGE) and microchip analysis. Database included two sets of vectors corresponding to healthy subjects and patients. Each subject was represented by the vector, each component of which had a value of the respective gene expression. In total training set value was 296 vectors. 71 of them were from the healthy people, all the rest — from glioblastoma patients. Vectors dimension was 20 — according to number of genes under study.

The input data format is shown in the Table 1. In this table column “Name” contains encoded names of subjects while codes CAMK2B (#1), STMN2 (#2) and so on — codes of genes and their indexes in profile — 20 genes in total.

In the beginning of our study we performed correlation analysis of data. It was found out, that expression values of some genes are highly correlated. Correlations in the group of healthy tissue differed from correlations in the glioblastoma (Fig. 1). This says about possibility of reliable clusterization of gene profiles by SOM or perceptrons.

DATA CLUSTERIZATION BY SOM

A self-organizing map (SOM) [12] is a neural network algorithm that reduces the dimension of a data set while being topology-preserving, i.e., proximity of cases in high-dimensional space is preserved in the space of reduced dimensionality. Typically, data dimension is reduced to two dimensions for easy visualization. The SOM algorithm works by arranging artificial neurons on a two-dimensional grid, with each neuron being connected to its neighbors. This grid is then reshaped according to the similarity of cases in order to reflect the high-dimensional data distribution. This is very important for medicine diagnostics because it gives opportunity observe patients with similar gene expression and compare their medical reports.

SOM TRAINING PARAMETERS SELECTION

The practice of SOM implementation shows some difficulties. Different program realizations [13–17] lead to different results on the same training data sets. Different training parameters also can lead to different results. This can be related with the program realization of method, the initialization of SOM, the number of neurons, the choice of neighborhood function, the initial and finite radius of training, etc. In order to identify the right way of realization in the aspect of solving the FCP (Fundamental Clustering Problem) tasks we compared the most of available software-based implementations of SOM algorithm [18]. As a result we used in our research a batch realization of the algorithm in Matlab SOM Toolbox. Also it should be noted that the choice of network parameters and parameters of its training is a crucial stage of the training. For example if the number of neurons is too large, then overtraining is possible, when the nodes of the network "stick" to each input vector. The effect of overtraining makes the network lose its smoothness – the generalizing ability. Elastic resilient maps [14] are in certain way the analog of SOM where the parameters of the network flexibility and elasticity regulating are applied.

It is possible to control the SOM overtraining without test dataset by means of training cessation when the neighborhood radius becomes smaller than radius R_ρ (1) (2-D rectangle topology).

$$R_\rho \approx \sqrt{\frac{SOMSize}{DataLen}}, \qquad (1)$$

where *SOMSize* – size of network; *DataLen* – length of training data.

Finally, at the last stage of training there is a competition between the winner neurons for the data which fall between them. This leads to the smoothing of the space between the winner neurons. The detailed method description is in [17].

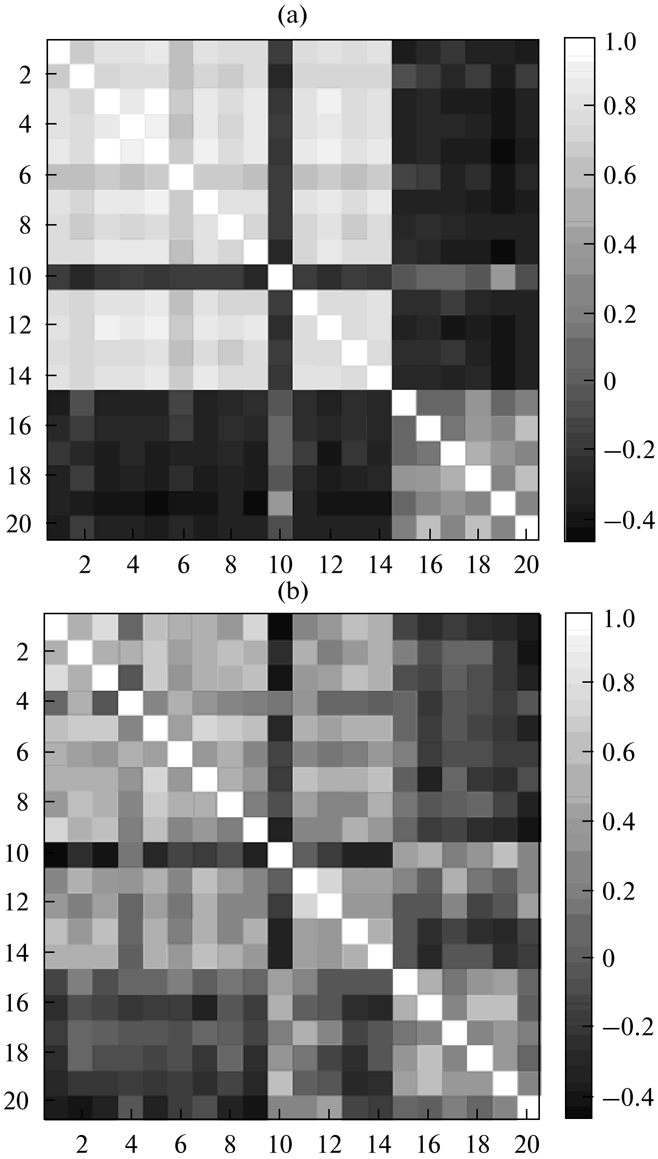

Fig. 1. Correlations between gene expression values. a) glioblastoma; b) normal tissue.

The SOM initialization is performed in the space of two main components. The initial radius of training is selected in the following way:

$$R_{init} = \frac{\max(XSize,\ YSize)}{2}. \qquad (2)$$

Here *XSize* and *Ysize* are the sizes of the sides of rectangular network, *SOMSize* = *XSize* × *Ysize*.

SOM TRAINING RESULTS

After the Kohonen map training was completed, the vectors corresponding to normal gene expression profiles were pictured on the map as white marks, and to the pathological – as black marks. The parameters of SOM were the following: topology – 2-D sheet, bond type – rectangular, number of neurons 50 × 50, initial training radius – 1500, final training radius – 3, neighborhood function – Gaussian. Fig. 2 shows the

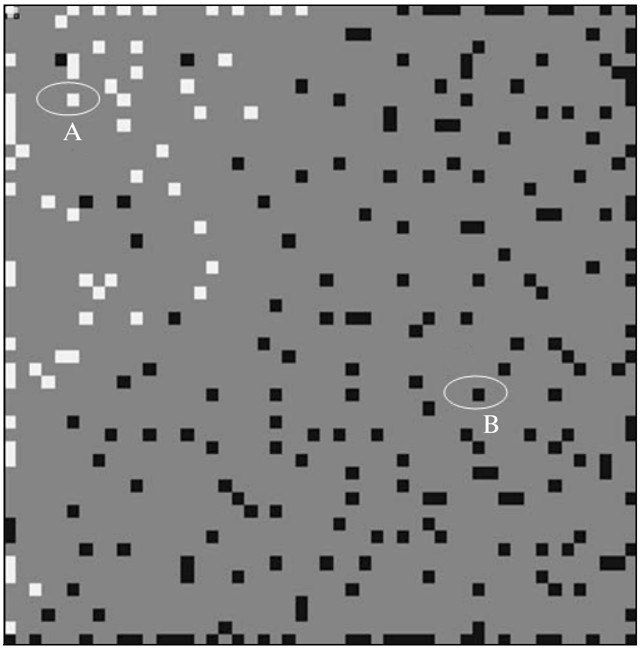

Fig. 2. Marks distribution throughout the map.

marks distribution throughout the map after its training. One can see that the data set is divided into two big groups (normal and glioblastoma).

The winner neurons with frequencies 2 or more (2 or more profiles mapped onto them) may be the most of interest for further study, because they contain data with very similar patterns. Neurons placed in the opposite corners of the map correspond to the most different profiles of gene expression. Some profiles mapped onto the "wrong" area – normal onto the "pathological" area and vice versa. This can be explained by insufficient informativeness of the inputs or wrong diagnosis, as well as by individual peculiarities of people under study.

Figure 3 shows typical gene expression profiles in normal tissue and glioblastoma that were mapped respectively onto neurons A and B in two clusters on Fig. 3.

Classification by the Two-Layer Feedforward Neural Network

We used two-layer feedforward artificial neural network for this classification task. We choose neural network with 20 neurons in hidden layer and hyperbolic tangent as an activation function, and two neurons in output layer with linear activation function. Input data set was divided into the training, validation and test subsets in proportion 60% : 20% : 20% respectively. Feature vector of dimension 20 was used as an input vector. Output vector consists of two elements. Their values are [0 1] when input vector represents healthy tissue or [1 0] when it is glioblastoma. The results of classification are shown in the Table 2.

Table 2. The results of gene expression profiles classification by feed forward neural network

	Sample size		
	Training set	Validation set	Test set
Normal tissue (total)	160	32	33
Normal tissue (misclassified as glioblastoma)	1 (0.6%)	0	1 (3%)
Glioblastoma (total)	48	12	11
Glioblastoma (misclassified as normal)	1 (2.1%)	0	0
Overall classification error	2 (0.96%)	0	1 (2.3%)

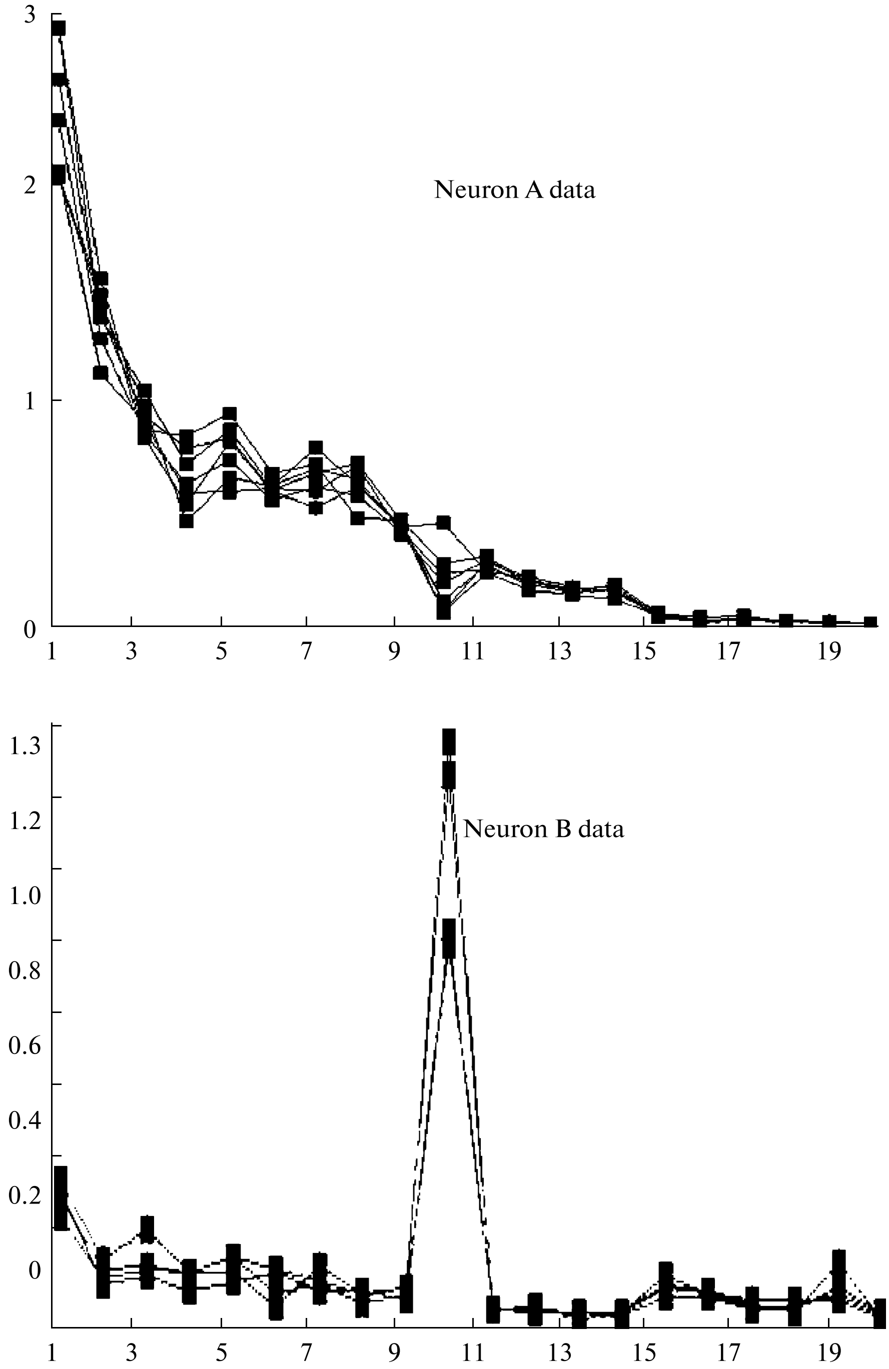

Fig. 3. Gene expression profiles in the healthy tissue (neuron A) and glioblastoma (neuron B). Several profiles were mapped onto each of these neurons. X-axis – index of genes in profile. Y-axis – gene expression value.

CONCLUSIONS

Performed study shows that applied methods allow discriminating data set into two parts – normal and pathology – with quite high reliability. This means, that datasets, we used, have been well prepared. High correlations between gene expression values (Fig. 1) show, that number of genes may be reduced.

In the future we are going to reduce training vector dimensionality. This may be done by different means: mutual entropy reduction, principal components extraction etc. We are planning to make such a data preprocessing on the next step of our study. Also, we are planning to include in data analysis information about expression of more genes. We intent to get in this way more detailed information from SOM aiming the possibility of typologization of tumors and their prognostic assessment.

ACKNOWLEDGEMENTS

Study was partly supported by Ukrainian Center of Science and Technology, project #4688.

REFERENCES

1. Kavsan, V.M., Shostak, K.O., Dmitrenko, V.V., Zozulya, Y.A., Rozumenko, V.D., and Demotes-Mainard, J., Characterization of Genes with Increased Expression in Human Glioblastomas, *Tsitol Genet.*, 2005, vol. 39, pp. 37–49.

1 2. van De Veer, L.J., Dai, H., van De Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, J.L., van Der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., and Friend, S.H., Gene Expression Profiling Predicts Plinical Outcome of Breast Cancer, *Nature,* 2002, vol. 415, pp. 530–536.

3. Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M., Atkins, D., and Foekens, J.A., Gene-Expression Profiles To Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer, *Lancet*, 2005, vol. 365, pp. 671–679,.

4. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., Hiller, W., Fisher, E.R., Wickerham, D.L., Bryant, J., and Wolmark, N., A Multigene Assay To Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer, *N. Engl. J. Med.*, 2004, vol. 351, pp. 2817–2826.

5. Ma, X.J., Hilsenbeck, S.G., Wang, W., Ding, L., Sgroi, D.C., Bender, R.A., Osborne, C.K., Allred, D.C., Erlander, M.G., The HOXB13: IL17BR Expression Index is a Prognostic Factor in Early-Stage Breast Cancer, *J. Clin. Oncol.*, 2006, vol. 24, pp. 4611–4619.

6. Venu Gopala Rao, K., Prem Chand, P., and Ramana Murthy, M.V., A Neural Network Approach in Medical Decision Systems, *Journal of Theoretical and Applied Information Technology*, 2007, vol. 3, no. 4, pp. 97–101.

7. Törönen, P., Kolehmainen, M., Wong, G., Castrén, E., Analysis of Gene Expression Data using Self-Organizing Maps, *FEBS Letters,* 1999, vol. 451, pp. 142–146.

8. Granzow, M., Berrar, D., Dubitzky, W., Schuster, A., Azuaje, F.J., and Eils, R., Tumor Classification By Gene Expression Profiling: Comparison and Validation of Five Clustering Methods, *ACM SIGBIO Newsletter*, 2001, vol. 21, no. 1, pp. 16–22,.

9. Demuth, T., Rennert, J.L., Hoelzinger, D.B., Reavie, L.B., Nakada, M., Beaudry, C., Nakada, S., Anderson, E.M., Henrichs, A.N., McDonough, W.S., Holz, D., Joy, A., Lin, R., Pan, K.H., Lih, C.J., Cohen, S.N., and Berens, M.E., Glioma Cells on The Run - The Migratory Transcriptome of 10 Human Glioma Cell Lines, *BMC Genomics*, 2008, vol. 9, p. 54

10. Li, A., Walling, J., Ahn, S., Kotliarov, Y., Su, Q., Quezado, M., Oberholtzer, J.C., Park, J., Zenklusen, J.C., and Fine, H.A., Unsupervised Analysis of Transcriptomic Profiles Reveals Six Glioma Subtypes, *Cancer Res.*, 2009, vol. 69, no. 5, pp. 2091-9.

11. Petalidis, L.P., Oulas, A., Backlund, M., Wayland, M.T., Liu, L., Plant, K., Happerfield, L., Freeman, T.C., Poirazi, P., and Collins, V.P., Improved grading and Survival Prediction of human Astrocytic Brain Tumors By Artificial Neural Network Analysis of Gene Expression Microarray Data, Mol Cancer Ther May 2008, vol. 7, pp. 1013–1024.

12. Deboeck, G., Kohonen, T., Eds., *Visual Explorations in Finance with Self-Organizing Maps*, London: Springer-Verlag, 1998.

13. Laboratory of Computer and information Science Adaptive Informatics Research Centre, Projects: SOM_PAK, WEBSOM, Toolbox MatLab, www.cis.fi

14. ESOM – DataBionics. Marburg, http://www.mathematik.uni-marburg.de

15. Gorban, B., Kegl, D., Wunsch, A., and Zinovyev, Eds., *Principal Manifolds For Data Visualization and Dimension Reduction*, Berlin – Heidelberg – New York: Springer, 2007.

16. Viscovery SOMMine – Eudaptics Software Viscovery SOMine, www.eudaptics.com

17. Ellipse. Ellipse Self Organizing Maps. www.ellipse.fi

18. Schwartz, D.R., *Algorithmic Peculiarities of Multidimensional Data Clusterization Method, Based on Kohonen Networks*, Science and Innovations in The Technical Universities, St.Petersburg, 2007, pp. 90–93 (in Russian).

SPELL: 1. ok