Gene expression profiling for subtyping of glial tumors

I.S. Knyazeva

Central Astronomical Observatory of the RAS at Pulkovo, Saint-Petersburg, Russia, iknyazeva@gmail.com

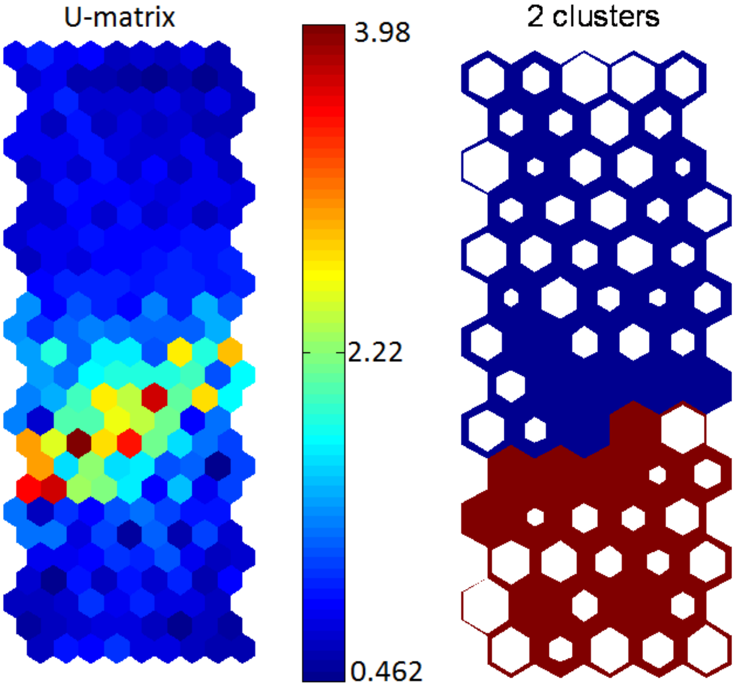
A.A.Mekler

The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russia, mekler@yandex.ru

V.V. Dmitrenko, A.V. Iershov, V.M. Kavsan

Institute of Molecular Biology and Genetics of NASU, Kiev, Ukraine, kavsan@imbg.org.ua

Entropy analysis and Kohonen self-organizing maps (SOMs) were used for the selection of the most informative genes, whose expression levels are applicable to distinguish the molecular variants of glioblastoma, the most malignant brain tumor. For this purpose, data concerning the expression of 12480 genes in 224 glioblastoma samples were recovered from Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). Clusterization of so big number of features (gene expression values) is not correct. Thus, the main task of the study was to select a subset of features that may give an obvious clusterization of the analyzed samples. Entropy analysis was chosen for the primary feature selection because it is very helpful for determination of the most varying features and increases in this way chances to find subgroups. The entropy of a parameter *x,* partitioned into *K* equal intervals $-\Delta x_1, \Delta x_2, ..., \Delta x_K$, could be calculated using a box-counting method [1] as $H(x) \approx -\sum_{k=1}^{K} P_k \ln P_k$. According to this calculation, 474 genes with highest entropy of their expression levels in glioblastomas were selected. Genes with bimodal expression distribution were selected (92 genes in total) and each of all possible pairs of these genes (4186) was mapped onto SOM and the clusterization quality on the map was evaluated by the quantization error criteria which represent average distance between the data vectors and the corresponding best matching units vectors [2]. 30 genes were selected according to the quantization error. SOM was trained by these 30-dimensional vectors and a very well pronounced division of glioblastoma samples was revealed. Obtained U-matrix (left part of the figure) was used for the clusterization of glioblastomas by the *k*-means method. In order to select proper number of clusters, clusterization was performed with different *k* and the clusterization quality was evaluated for each *k* using the Davies-Bouldin clustering evaluation index [3]. The best clusterization was at *k*=2.



U-matrix for 30 genes and SOM with clusters border and the grade of the nods filling

Obtained results showed that glioblastomas may be divided at least into two molecular subtypes.

Study was supported by the RFBR grant #12-04-90434-Укр_a and NAS Ukraine grant #07-0412.

1. C. Beisbart et al. Probabilities in Physics, Oxford University Press, USA (2011), 432 p.
2. T. Kohonen. Self-Organizing Map, 2nd ed., Springer-Verlag, Berlin, (1995), pp. 113.
3. D. L. Davies et al. (1979) A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227.