Using an atlas-based approach in the analysis of gene expression maps obtained by voxelation

E. I. Zacharaki[1*], A. Skoura[1] , L. An[2] , D. Smith[3] , V. Megalooikonomou[1,2]

[1]Department of Computer Engineering and Informatics, University of Patras, Patras, Greece
[2]Data Engineering Laboratory, Center for Data Analytics and Biomedical Informatics, Temple University, PA, USA
[3]Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, UCLA, CA, USA

*(ezachar@upatras.gr)

Abstract. The integration of gene expression datasets with gene function information provides valuable insights in unraveling the molecular mechanisms of the brain. In this paper, gene expression maps, acquired by the technique of voxelation, are analyzed using an atlas-based framework and the extracted spatial information is employed to organize genes in significant clusters. Moreover, gene function enrichment analysis of clusters enables exploring the relationships among brain regions, gene expressions and gene functions. Our work confirms the hypothesis that genes of similar spatial expression patterns display similar functions indicating that our methodology could assist in the functional identification of unannotated genes.

Keywords: gene expression maps, brain mapping, voxelation, gene function, gene ontology

1 Introduction

The mammalian brain is a complex organ exhibiting a rich variety of gene expression patterns across a broad range of cell types. Expression of genes is manifested by the production of RNA transcripts within cells and recent advances in the quantitative detection of mRNA and proteins on a genomic scale permit localization of gene products onto maps of the brain. Concerning that about 40% of the proteins encoded in eukaryotic genomes are proteins of unknown function, a challenging issue in the field of analysis of gene expression maps (GEMs) is their association with gene function information in order to reveal functional characterization of unannotated genes [1]. Preliminary results concerning the amounts of transcripts detected within different tissues or the same tissues under different states (e.g., physiological or disease), which are visualized using GEMs, shed light on the etiology and pathology of neurological diseases and may lead to the discovery of biomarkers and of proteins responsible for diseases[2].

The methodologies for spatial mapping of proteins and transcripts in the brain are various, offering compelling information. A cheap and fast method is voxelation [3]. Voxelation allows acquisition of both transcript and protein mapping data in parallel simplifying co-registration of multiple genes, however it offers gene expression maps of intermediate resolution. According to this approach, the brain is divided into spatially registered voxels (cubes) and using microarrays or mass spectroscopy spatial images with quantitative information on transcripts or proteins are reconstructed.

The analysis of GEMs involves the application of feature extraction techniques combined with data mining methodologies such as clustering, classification and similarity search. Gene information from other aspects, such as Gene Ontology, is usually employed to validate biological hypothesis or to strengthen the fidelity of research outcomes. For example, aiming at the identification of unannotated genes An et al. [4] analyzed GEMs by extracting wavelet features and by using a multiple clustering technique. The authors confirmed the hypothesis that a subset of genes with similar expression maps display function similarity, where the identification of function similarity was based on Gene Ontology. Regarding the analysis of GEMs for the discrimination between normal and disease, Brown et al. [5] investigated the expression differences between normal and Parkinson's disease (PD) mice brains using voxel expression maps of 9000 genes acquired by voxelation. The analysis was based on two genexgene matrices (normal and PD matrix) whose elements represented the cross correlation of corresponding GEMs. Gene expression in normal and PD brains revealed significant global expression differences when averaged across voxels for known genes. Moreover, the singular Value Decomposition method was applied to the two matrices and global shifts of gene expression between normal and PD brains were indicated. Concerning spatial differentiation through a GEM produced by voxelation, a clustering analysis of gene expression patterns from mouse voxelation data revealed four distinct groups of genes corresponding to different mapping patterns [6].

In this paper, we extend previous work [4] exploring gene expressions differences with regard to brain anatomy. Our aim is the identification of genes whose expressions display similar anatomical distribution regarding specific brain regions such as white matter, gray matter and the hippocampal region. We also want to investigate if the gene clusters with similar expression patterns have also similar gene function. Finally, we examine if we can extract more informative clusters by down-weighting inconsistent measurements, such as in voxels with high partial volume effect. Our investigation concludes that clusters of genes with similar localized expression patterns display functional similarity indicating that our work has the potential of creating comprehensive atlases of gene and protein expression in the mammalian brain.

2 Methods

The association of gene expression in the brain anatomy with functional activity can provide a better understanding of the role of the gene's products. In this study we are investigating the hypothesis that genes with similar expression maps have similar

gene functions. For this purpose GEM's similarity is calculated as in previous work [4] based on the expression patterns acquired with the voxelation technique. Voxelation data however have much lower resolution than single cell resolution data, thus suffer from partial volume effect, meaning that the acquired expression values represent an average over the gene expression of all cells in each voxel. This limitation becomes especially prominent in regions where different tissue types mix, whereas in homogenous regions where similar expression patterns are expected, averaging does not alter significantly the gene's expression profile.

Thus, in this study we examine whether partial volume effect and unreliable measurements can affect GEMs similarity and therefore alter the relationship between gene expression and function. The idea is that measurements on untrustworthy locations should have less effect on the calculation of the overall similarity between genes. Such regions include the background and ventricles and also voxels with high partial volume effect. Next we describe the construction of four spatial maps: three of them represent spatial distribution of different tissues and one map reflects the confidence on measurements. We furthermore explain how these spatial maps are used in the calculation of similarity between GEMs.

2.1 Brain partitioning and confidence map

Let Ω be the set of genes and $\mathbf{x}_i \in R^n,\ i \in \Omega$, be the expression profile of gene i, where n is the number of voxels of the particular slice (at the level of the striatum) of mouse brain we consider.

We explore the brain's anatomical morphology by mapping a mouse brain atlas [7] on the space of the GEM as illustrated in Fig. 1 (left). Then the registered atlas image is partitioned into three regions: (i) gray matter (GM) in the cerebral cortex and anterior cingulated area, (ii) white matter (WM) including the striatum and caudoputamen and (iii) hippocampal region (HR) including the nucleus accumbens, substantia innominata, diagonal band nucleus and medial septal nucleus and excluding the lateral septal nucleus. The three brain segments are visualized in Fig. 1 (in the midde) and are used to construct spatial maps by assigning the value of 1 to a voxel if it belongs to the corresponding brain segment, or 0 if it doesn't belong. Voxels on region boundaries are assigned a value equal to the partial volume in each tissue compartment. The acquired spatial maps are denoted as $\mathrm{w_{GM}}, \mathrm{w_{WM}}, \mathrm{w_{HR}} \in R^n$ for GM, WM and HR, respectively. The amount of partial volume for each brain voxel j is then calculated by the following measure of fuzziness:

$$w_{PV}(j) = \sqrt{1 - \left(w_{GM}^2(j) + w_{WM}^2(j) + w_{HR}^2(j)\right)} \tag{1}$$

It is easy to see that the more equally distributed is the tissue to the three compartments, the higher is $\mathrm{w_{PV}}(j)$. The uncertainty map is calculated by averaging the amount of partial volume and the volume outside brain tissue (background or ventricular regions). A confidence map, $\mathbf{w_C} \in R^n$, indicating the certainty of each voxel value, is then defined as the negative of the uncertainty map as illustrated in Fig. 1 on

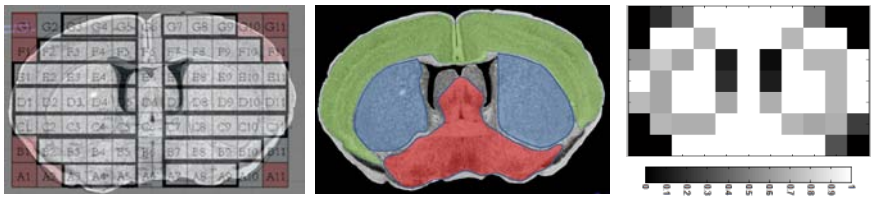

Fig. 1. Mouse brain partitioning. Left: Brain atlas with superimposed voxelation grid (red voxels indicate background). Middle: Brain tissue maps based on atlas [7] at bregma=0 (green: GM, blue: WM, red: HR). Right: Confidence map w_C (the darker, the less certainty).

$$w_C\ (j) = 1 - \frac{w_{PV}(j) + 1 - \left(w_{GM}(j) + w_{WM}(j) + w_{HR}(j)\right)}{2} \tag{2}$$

the right and defined in equation 2.
Next, due to the inherent bilateral symmetry of the mouse brain and lack of "handedness" or speech-centers in mice, we decrease the amount of data by retaining the voxels of only one brain hemisphere. Similarly for the GEMs, left and right hemispheres are averaged and only one hemisphere is retained, thus decreasing noise. For all GEMs and spatial maps the number of voxels is therefore reduced to $n = 42$.

2.2 Definition of gene similarity of expression and function

The gene expression maps similarity between two genes, $\mathbf{x}_1$ and $\mathbf{x}_2$, is defined as the squared weighted Euclidean distance function, D, formalized below:

$$\mathrm{D}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{n} w(j)\left(x_1(j) - x_2(j)\right)^2}{\frac{\sum_{j=1}^{n} w(j)}{n}} \tag{3}$$

The weight vector w is used to emphasize dissimilarity on selective spatial locations. By incorporating such a weight vector, we can test two hypotheses. First, we can investigate whether by down-weighting the measurements on locations with high uncertainty we can form a more informative similarity measure that is not affected by partial volume and artifacts due to ventricles. This hypothesis is tested by using the confidence map $\mathrm{w_C}$ as weight vector. The second hypothesis is that the gene function might correlate with gene expression in specific anatomic locations. Thus we investigate whether genes with similar expression in some anatomic locations have similar gene functions. For this purpose we use the three spatial maps $\mathrm{w_{GM}}, \mathrm{w_{WM}}, \mathrm{w_{HR}}$ as weight vectors in equation (3) and investigate each region independently.

The gene function similarity is calculated using Lin's method [8] to evaluate function distance in Gene Ontology structure. The similarity values are obtained within each of the three categories of Gene Ontology (GO version: January 2009), and are based on frequencies from the Mouse Genome Informatics (MGI) annotation dataset (MGI version: 01/31/2009). The three categories of gene function refer to "Cellular Component", "Molecular Function" and "Biological Process". The function similarity

between two clusters of genes is calculated as average pairwise function distance, as explained in [4]. The significance of the function distance (p-value) is calculated as a percentile in respect to average pairwise function distances. Due to the huge number of possible gene combinations, groups of genes are randomly selected and the corresponding average distances are calculated. The p-value then indicates how small is the respective function similarity with respect to the average function distances. More details are provided in [4].

2.3 Clustering analysis

For each spatial map, clustering analysis was performed by two sets of experiments. In the first set we selected prototype genes with diverse expression patterns as queries and detected genes with similar expression maps. Then we calculated the average function distance in each cluster of similar genes. Similarity was assessed by the p-value; smaller p-values result in smaller clusters. In the second set of experiments we attempted to find clusters that have both similar GEMs and similar gene functions. First clusters of GEMs were determined by the k-means algorithm using the weighted Euclidean distance function (Eq. 3). According to this criterion, the clusters consisted of genes with similar expression in the investigated region of interests (all 4 spatial maps were tested). Then only the clusters with significant expression maps similarity and average function similarity were retained, whereas the rest of the clusters were further split into an increasing number of smaller clusters until they reached the significance threshold (p-value = 0.05) for both gene expression and function. Thus the parameter K in the k-means algorithm (representing number of clusters) was not predefined, but calculated in a hierarchical fashion [4].

The first set of experiments will help us investigate whether the average function distance for each group of genes is reduced when specific spatial maps are used. The results of the second set of experiments will be used to extract the common gene expression patterns for each significant cluster and examine whether these patterns are related with specific anatomical locations. Moreover, connectivity relations might be revealed, if distinctive expression patterns will be identified in locations different from the applied spatial maps.

The validation of clustering is performed by the commonly used ratio of inter-cluster distance (D_{inter}) to intra-cluster distance (D_{intra}). The intra-cluster distance is defined as the average distance of each point to its cluster centroid, $D_{intra} = \frac{1}{N}\sum_{i=1}^{k}\sum_{x_j \in S_i}\left|x_j - \mu_i\right|^2$, where N is the total number of data points, S_i, $i = 1,2, \ldots, k$, k is the number of clusters and μ_i is the centroid of the cluster S_i). The inter-cluster distance is the minimum of the distances between each pair of cluster centroids, $D_{inter} = \min\left(\left|\mu_i - \mu_j\right|^2, i = 1,2, \ldots, k-1,\ j = i+1, \ldots, k\right)$.

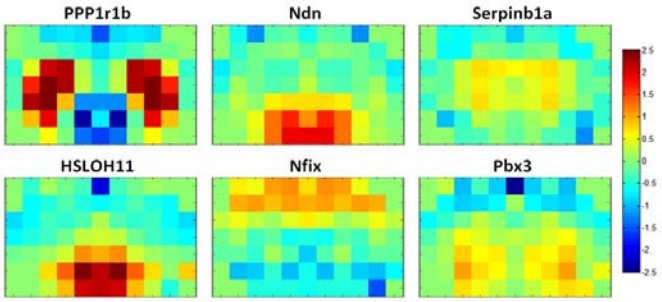

Fig. 2. Gene expression maps of the selected prototype genes

3 Results

3.1 Clustering based on prototype genes

Fig. 2 shows the gene expression maps for 6 genes selected as queries. PPP1r1b is strongly expressed in striatum, Ndn and HSLOH11 are expressed in hypothalamus, Serpinb1a is weakly expressed in striatum, Nfix is expressed in a gradient pattern in cortex and Pbx3 is expressed in striatum and adjacent ventral structures. For each prototype gene, we detected increasing number of genes (7, 15, 23,..., 78) with similar expression maps based on the use of different spatial maps and we calculated the average function distance in the group. The function similarity was considered with respect to the three function categories, Cellular Component, Molecular Function and Biological Process. Some indicative results are shown for the prototype genes PPP1r1b and Nfix in Tables 1 and 2, correspondingly. We highlight p-values that are smaller than 0.05.

A gene example demonstrating the importance of the use of the confidence map w_C is presented in Table 1. It can be seen the p-values regarding Biological Process are much smaller when applying the confidence map in the computation of the similarity of GEMs leading to the conclusion that the expression map of the gene Pbx3 is affected by partial volume effects and artifacts due to background and ventricles. The use of the confidence map which reflects the down-weighting of unreliable regions improves the similarity rate between the query gene Pbx3 and retrieved similar genes, confirming the hypothesis that a more informative similarity measure is obtained when incorporating weights of significance in the calculation of GEM's similarity.

Regarding the hypothesis that genes expression in specific anatomic regions might be correlated with similar gene functions, we calculated the gene expression map similarity using each one of the three spatial maps w_{GM}, w_{WM}, w_{HR} and using no mask. In most cases the use of a spatial mask provides better function similarity (i.e. smaller p-value) results compared to the absence of any mask. Among the three spatial maps, each gene is associated mainly with one of them; for example PPP1r1b displays higher function similarity when focusing on WM region, whereas Nfix displays higher results when focusing on GM region. Furthermore, the comparison with the function similarity results when no spatial map is used, revealed that the

Table 1. Comparison of similarity results of the gene Pbx3 with respect to the use of the confidence map w_C .

Number of similar genes	**Using the confidence map w_C**			**Without using a confidence map**		
	Cellular Component	**Molecular Function**	**Biological Process**	**Cellular Component**	**Molecular Function**	**Biological Process**
7	1.00	**0.00**	**0.00**	1.00	**0.00**	0.60
15	1.00	**0.00**	**0.00**	1.00	0.08	0.38
23	1.00	**0.00**	**0.00**	1.00	**0.00**	**0.00**
31	0.88	**0.00**	**0.00**	0.98	**0.00**	**0.04**
39	0.24	**0.00**	**0.00**	0.86	**0.00**	**0.00**
47	0.16	**0.00**	**0.02**	0.59	**0.00**	0.06
55	**0.02**	**0.00**	**0.00**	0.30	**0.00**	0.82
63	0.76	**0.00**	0.32	0.11	**0.00**	0.73
70	0.59	**0.00**	0.92	**0.02**	**0.00**	0.94
78	**0.04**	**0.00**	0.97	0.07	**0.00**	0.98

Table 2. Average Function Distance (p-values) using Lin's method [8] for the gene PPP1r1b. The clusters of similar GEMs are created using the proposed spatially-oriented methodology focusing on WM region (left) and without using a spatial map for comparison (right).

Number of similar genes	**Using the spatial map w_{WM}**			**Without using a spatial map**		
	Cellular Component	**Molecular Function**	**Biological Process**	**Cellular Component**	**Molecular Function**	**Biological Process**
7	**0.00**	**0.00**	**0.00**	1.00	1.00	0.18
15	**0.00**	**0.00**	**0.00**	0.82	0.14	0.23
23	**0.00**	**0.01**	**0.00**	0.71	0.06	**0.01**
31	**0.00**	**0.00**	**0.00**	1.00	**0.03**	0.39
39	**0.00**	**0.00**	**0.00**	0.63	**0.02**	0.91
47	**0.00**	**0.00**	**0.00**	0.50	**0.01**	1.00
55	**0.00**	0.06	**0.00**	0.48	**0.00**	0.97
63	0.89	0.54	**0.04**	0.75	**0.00**	0.99
70	1.00	0.41	0.36	0.78	**0.01**	0.93
78	1.00	0.13	0.22	0.98	0.09	0.82

proposed spatially-oriented approach of GEMs achieves not only higher function similarity (e.g. regarding the biological process for PPP1r1b) but also revealed new function categories that are related with this gene (e.g. cellular component and biological process for PPP1r1b). These comparative results regarding PPP1r1b gene are illustrated in Table 2. Obviously, focusing on the anatomic region of WM the retrieved genes display much higher function similarity indicating that specific anatomic regions of the brain play important role in identification of gene function.

3.2 Clustering based on all genes

The GEMs of each significant cluster obtained by the hierarchical k-means algorithm for different function categories are averaged and some of them are illustrated in Fig. 3. The average maps, each of them representing one cluster, are shown only for the first 6 significant clusters for each one of the three function categories when the proposed confidence map was used (right column of Fig. 3) and when no spatial map was used in the clustering process (left column of Fig.3). We also searched for significant clusters with low p-value of functions distance in any one of the three function categories; the corresponding results are shown in the last row of Fig. 3. The cardinality of each cluster is presented above the corresponding average GEM. Comparing the average GEMs in Fig. 3, we remark that many significant clusters are present regardless the use of the confidence map; a characteristic example is the cluster 2 in biological process without any spatial map and the cluster 4 in biological process with the use of confidence map. Although red-like pixels of GEMS indicate strong positive expression while blue-like pixels indicate strong negative expression, both of them are useful as encode important information for the analysis of gene expression. Furthermore, the location of such a pixel plays an important role; for example a red pixel in the region of ventricles is not informative. As it can be seen in Fig. 3, the clusters when using the confidence map contain more informative pixels, i.e. pixels which both display strong gene expression and are localized in meaningful regions within the brain.

Table 3 shows the clustering validity score (D_{inter} / D_{intra}), with the highest validity score for each function category highlighted. The results indicate that when function similarity for each of the three function categories is sought independently, the best clustering is achieved when the confidence map is used. When function similarity for any of the three function categories is aimed, the best clustering is achieved when the gene expression in GM is considered.

4 Discussion and Conclusions

Clustering analysis on voxelation data showed that the group of genes that was identified as similar to a target gene shares very similar gene functions in at least one gene function category. Moreover for some genes when a spatial map was used in the calculation of GEMs similarity, the average function similarity was increased or a new function category was revealed. By clustering GEMs of genes with known and unknown function together, the proposed approach has the potential to be used in predicting unknown gene functions.

Acknowledgments
This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

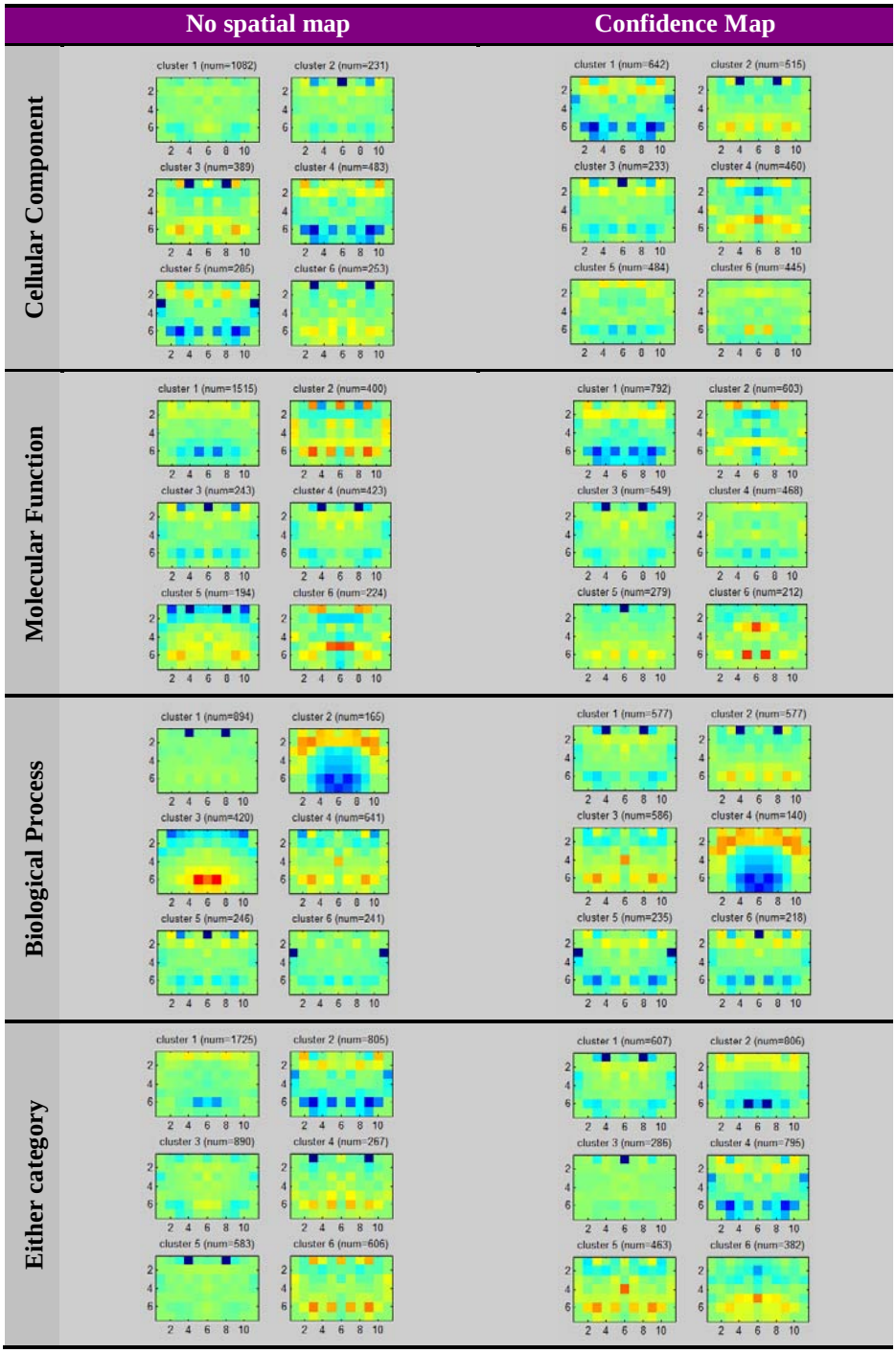

Fig. 3. Average gene expression maps for the first 6 significant clusters for each function.

Table 3. Validity score for all significant clusters. The number of significant clusters is shown in parentheses for each case.

	No spatial map	Confidence Map	GM	WM	HR
Cellular Component	0.099 (78)	**0.128** (72)	0.078 (86)	0.056 (65)	0.098 (60)
Molecular Function	0.116 (81)	**0.154** (75)	0.062 (91)	0.079 (67)	0.109 (41)
Biological Process	0.123 (82)	**0.162** (70)	0.095 (83)	0.067 (60)	0.125 (46)
Either category	0.107 (32)	0.089 (34)	**0.110** (45)	0.055 (39)	0.088 (21)

References

1. Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J.F., Zhu, J.K., Cushman, J.C., Gollery, M., Girke, T.: Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol, 147(1), 41-57, (2008)
2. Zhang, X., Zhou, J.Y., Chin, M.H., Schepmoes, A.A., Petyuk, V.A., Weitz, K.K., Petritis, B.O., Monroe, M.E., Camp, D.G., Wood, S.A., Melega, W.P., Bigelow, D.J., Smith, D.J., Qian, W.J., Smith, R.D.: Region-specific protein abundance changes in the brain of MPTP-induced Parkinson's disease mouse model. J Proteome Res, 9(3), 1496-1509, (2010)
3. Chin, M.H., Geng, A.B., Khan, A.H., Qiang, W.J., Petyuk, V.A, Boline, J., Levy, S.: A genome-scale map of expression for a mouse brain section obtained using voxelation. Physiological Genomic, 313-321, 2007
4. An, L., Xie, H., Chin, M.H., Obradovic, Z., Smith, D.J., Megalooikonomou, V.: Analysis of multiplex gene expression maps obtained by voxelation. BMC Bioinformatics, 10(Suppl. 4), (2009)
5. Brown, V.M., Ossadtchi, A., Khan, A.H., Yee, S., Lacan, G., Melega, W.P., Cherry, S.R., Leahy, R.M., Smith, D.J.: Multiplex three-dimensional brain gene expression mapping in a mouse model of Parkinson's disease. Genome Res, 12(6), 868-884, (2002)
6. Park, C.C., Petyuk, V.A., Qian, W.J., Smith, R.D., Smith, D.J.: Dual spatial maps of transcript and protein abundance in the mouse brain. Expert Rev Proteomics, 6(3), 243-249, (2009)
7. Mouse Brain Atlas: C57BL/6J Coronal, http://www.mbl.org/atlas170/atlas170_frame.html
8. Lin, D.: An Information-Theoretic Definition of Similarity. Proc. of the Fifteenth International Conference on Machine Learning, Madison, Wisconsin, 296-304, (1998)