Combining gene expression and function in a spatially localized approach

Evangelia I. Zacharaki, Angeliki Skoura
Department of Computer Engineering and Informatics
University of Patras
Patras, Greece
ezachar@upatras.gr, skoura@ceid.upatras.gr

Li An
Data Engineering Laboratory, Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, USA
anli@temple.edu

Desmond J. Smith
Department of Molecular and Medical Pharmacology
David Geffen School of Medicine, UCLA
Los Angeles, USA
dsmith@mednet.ucla.edu

Vasileios Megalooikonomou
Department of Computer Engineering and Informatics
University of Patras
Patras, Greece
Data Engineering Laboratory, Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, USA
vasilis@ceid.upatras.gr

Scott H. Faro
Department of Radiology,
Temple University School of Medicine,
Philadelphia, USA
scott.faro@tuhs.temple.edu

***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. Gene function enrichment analysis of clusters enabled exploration of the relationships among brain regions, gene expression and gene function. In addition, the analysis revealed new function categories, for example biological process for the gene Pbx3, cellular component and biological process for the gene Ppp1r1b. Our work confirms the hypothesis that genes of similar spatial expression patterns display similar functions indicating that our methodology could assist in the function identification of unannotated genes.**

Keywords-gene expression maps; brain mapping; voxelation; gene function; gene ontology

I. 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 on a genomic scale permit the localization of gene products onto maps of the brain allowing the acquisition of Gene Expression Maps (GEMs). The techniques for spatial mapping of transcripts in the brain are various and offer compelling information. A less expensive and fast method, which allows acquisition of both transcript and protein mapping data in parallel and simplifies co-registration of multiple genes, is voxelation [1, 2], 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. Since about 40% of the proteins encoded in eukaryotic genomes are proteins of unknown function, a challenging issue in the field is to associate GEMs with gene function information in order to reveal function characterization of unannotated genes [3]. The methodologies for the analysis of GEMs involve the application of feature extraction techniques combined with data mining methods such as clustering, classification and similarity search. Furthermore, gene information from other sources, such as Gene Ontology, is usually employed to validate biological hypothesis or to strengthen the fidelity of research outcomes. For example, aiming to identify 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 genes of similar expression maps display function similarity, where the identification of function similarity was based on Gene Ontology. Focusing on the study of spatial expression patterns in GEMs, Chin et al. [5] identified clusters of genes with expression patterns localized to defined substructures of the brain. However, the extracted information regarding anatomical expression was not associated with gene function annotation.

In order to explore gene function and gene expressions differences with regard to brain regions, in this paper we propose an anatomy-oriented framework for the analysis of GEMs obtained by voxelation. Firstly, we examine if the down-weighting of inconsistent measurements, such as in voxels with high partial volume effects helps generate more informative clusters relevant to function categories. Afterwards, we identify clusters containing genes whose expressions display similar anatomical distribution in respect

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

to specific brain regions such as white matter, gray matter and the hippocampal region. We then investigate the hypothesis that gene clusters with similar expression patterns also have similar gene function. Our investigation concludes that clusters of genes with similar localized expression patterns display function similarity. These results indicate that our work has the potential to create comprehensive atlases of gene expression in the mammalian brain and to provide insight into the identification of unannotated genes based on the analysis of their GEMs.

II. METHODS

Associating gene expression anchored to brain anatomy with functional activity can provide a better understanding of the role of the gene's products. In this study we investigate the hypothesis that genes with similar expression maps have similar gene functions. For this purpose GEM's similarity is calculated based on the expression patterns acquired with the voxelation technique. The voxelation technique allows acquisition of expression images in parallel, simplifying cross-analysis of multiple genes and also is less expensive and faster than traditional approaches. Voxelation data however have much lower resolution (e.g. 1 mm^3) than single cell resolution data, and thus suffer from partial volume effect in which 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, while in homogenous regions where similar expression patterns are expected, averaging does not alter significantly the gene's expression profile.

Hence, 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: the three maps represent spatial distribution of different tissue types and one map reflects the confidence on the measurements over the whole brain. We furthermore explain how these spatial maps are used in the calculation of similarity between GEMs. Finally, we provide the definition of GEM similarity and function similarity and also describe the two sets of experiments that have been performed. A summarized illustration of the analysis steps described next is shown in Fig.1.

A. Brain partitioning and partial volume correction

The data included coronal slices at the level of the striatum of adult C57BL/6J male mice of same age (8 weeks old) [5]. The voxel map is shown in Fig. 2a and has a resolution of 1 mm^3. Since there is a large amount of noise in microarray experiments, we take advantage of the inherent bilateral symmetry of the mouse brain and the lack of "handedness" or speech-centers in mice by averaging left and right hemispheres to decrease noise. We choose the voxels A6, B6, C6, D6, E6, F6 and G6 as midline of the two hemispheres. The averaged gene expression is obtained by considering for each row of the GEM pairs of symmetric cells, for example C1 and C11 (Fig. 2a). Thus the analysis is performed for all GEMs with $n = 42$ voxels (out of 68). For clearness of illustration however, the obtained results are reflected about the midline to create a full coronal image of the brain. The brain's anatomical morphology is explored by mapping a mouse brain atlas [6] on the space of the GEM as illustrated in Fig. 2b.

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. 2b and are used to construct

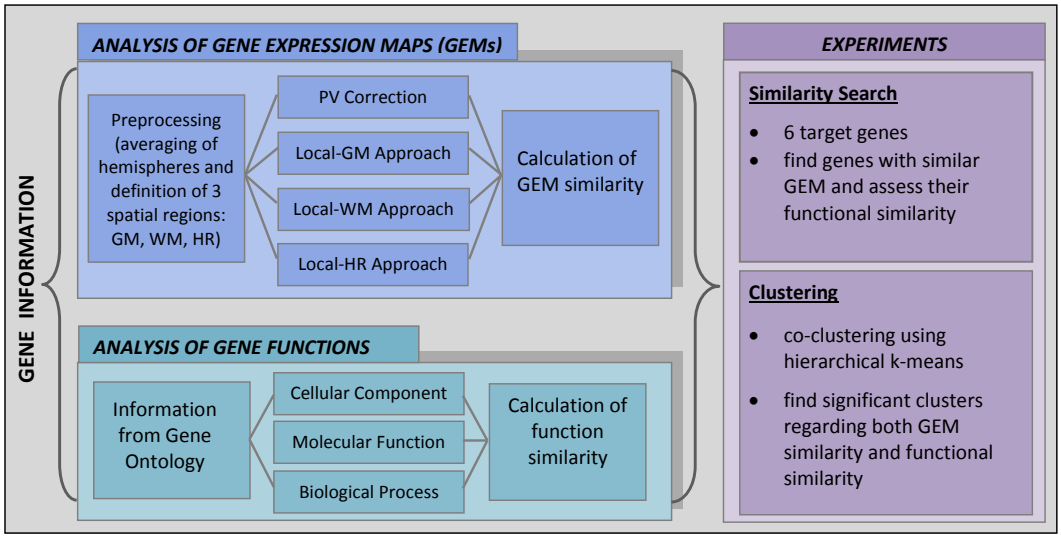


Figure 1. Illustration of the performed analysis

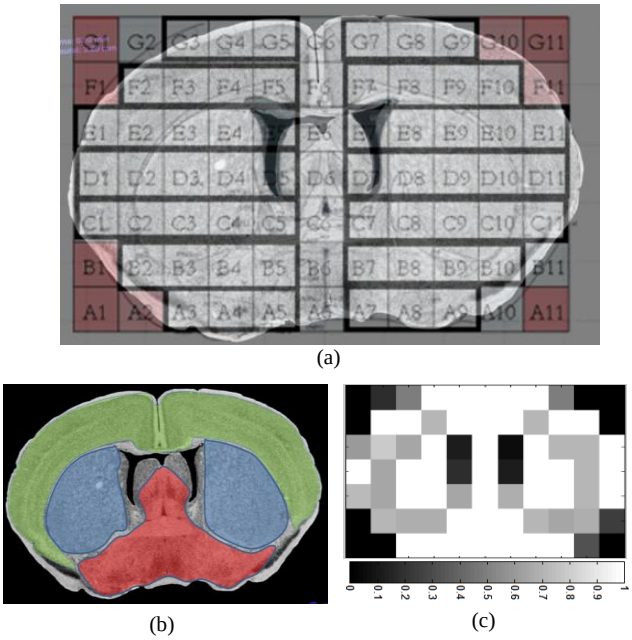
spatial maps by assigning the value of 1 to a voxel if it belongs to the corresponding brain segment and 0 otherwise. Voxels on region boundaries are assigned a value equal to the partial volume in each tissue compartment. The acquired spatial maps are denoted as $w_{GM}, w_{WM}, w_{HR} \in R^n$ for GM, WM and HR, respectively, where n is the number of voxels. The amount of partial volume (*PV*) for each brain voxel j is then calculated by the following measure of fuzziness:

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

It is easy to see that the more equally distributed is the tissue to the three compartments, the higher is $w_{PV}(j)$. 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, $w_C \in R^n$, indicating the certainty of each voxel value, is then defined as the negative of the uncertainty map as shown next:

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

and is illustrated in Fig. 2c. Similarly to the voxelation data, the two hemispheres of the spatial maps are averaged across the midline.



(a) (b) (c)

Figure 2. Mouse brain partitioning. (a) Brain atlas with superimposed voxelation grid (red voxels indicate background), (b) Coronal brain tissue maps based on atlas [6] at bregma=0 (green: GM, blue: WM, red: HR), (c) Confidence map $\mathrm{w_C}$ (the darker, the less certainty).

B. Definition of gene expression similarity

Let $\mathrm{x}_i \in R^n$ be the expression profile of gene i, where n is the number of voxels of the particular slice of mouse brain we consider. The gene expression maps similarity, Sim_G, between two genes, x_1 and x_2, is defined as the squared weighted Euclidean distance function formalized below:

$$Sim_G(\mathrm{x}_1, \mathrm{x}_2) = \frac{\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{w(j)}\left(\mathrm{x}_1(\mathrm{j}) - \mathrm{x}_2(\mathrm{j})\right)^2}{\frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{w(j)}} \qquad (3)$$

The weight vector w is used to emphasize dissimilarity on selective spatial locations. If the weight is 1 for all voxels then this corresponds to the standard definition of similarity that weights all voxels equally. If non-equal weights are used then two case scenarios can be investigated: the weight vector emphasizing voxels over the whole brain or emphasizing specific local anatomic regions. The first case scenario allows investigating whether by down-weighting the measurements on locations with high uncertainty a more informative similarity measure is formed that is not affected by partial volume and artifacts due to ventricles. This is tested by using the confidence map as weight vector and the method is then denoted as *global approach with PV correction*. The second case scenario allows investigating whether gene function correlates with gene expression in specific anatomic locations. Here we investigate whether genes with similar expression in some anatomic locations have similar gene functions. For this purpose we use the three spatial maps as weight vectors in Eq. 3 and investigate each local region independently. The method is then denoted as *local-GM*, *local-WM* and *local-HR approach*, respectively.

Furthermore, similarity of expression in a group S of genes is defined as the average similarity of all genes in the group to the group center ($\bar{\mathrm{x}}$), as shown below:

$$GSim_G(\mathrm{S}) = \left(\frac{1}{M}\right)\sum_{j=1}^{M} Sim_G(\mathrm{x_j}, \bar{\mathrm{x}}),$$

where $\bar{\mathrm{x}}$ is the average GEM and M the number of genes in the group. In order to assess the significance level of this similarity, we define a p-value. We randomly select N (=1000) groups, each one consisting of M genes and calculate the gene expression similarity of each group, $GSim_F(\mathrm{S_t}),\ t = \{1, .., N\}$. The p-value indicates how many of the N groups have group similarity larger than the similarity in the studied group and is calculated as shown below:

$$p\text{-}value_G(\mathrm{S}) = \left(\frac{1}{N}\right)\sum_{t=1}^{N}\left(GSim_G(\mathrm{S}) < GSim_G(\mathrm{S_t})\right).$$

C. Definition of gene function similarity

The gene function similarity $Sim_F(f_1, f_2)$ between two functions f_1 and f_2 is calculated using Lin's method [7] to evaluate function distance in Gene Ontology structure. This method applies an information theoretic definition of

similarity as long as there is a probabilistic model. The similarity values are publicly available and obtained within each of the three categories of Gene Ontology (GO version: January 2009) that refer to "Cellular Component", "Molecular Function" and "Biological Process". The values are based on frequencies from the Mouse Genome Informatics (MGI) annotation dataset (MGI version: 01/31/2009).

The function similarity in a group of genes is calculated as average pairwise similarity, as explained next. Let S be a group of genes and $F_i = \{f_{ik}\}$ be the set of functions of gene i, where $i \in$ S, $k = \{1,2,..,|F_i|\}$ and $|\cdot|$ denoting the cardinality of a set. The function similarity of the group is then defined as

$$GSim_F(\mathrm{S}) = \frac{\sum_{i,j \in S} \sum_{k=1}^{|F_i|} \sum_{\substack{l=1 \\ j \neq i}}^{|F_j|} Sim_F(f_{ik}, f_{jl})}{|\mathrm{S}|(|\mathrm{S}|-1)/2} \quad (4)$$

The denominator of $GSim_F(\mathrm{S})$ equals the number of all possible pairs of genes in S.

In order to estimate the range of values of $GSim_F$ and determine the significance (p-value) of a calculated value, all possible group similarities should be calculated. Due to the huge number of possible gene combinations, we did not generate all of them but randomly selected N (=10000) groups, each one consisting of M (=1000) genes and the corresponding average similarities $GSim_F(\mathrm{S_t})$, $t = \{1, .., N\}$,were calculated from Eq. 4. The p-value of function similarity indicates how small is the function similarity of the respective group (S) in respect to the N groups:

$$p\text{-}value_F(\mathrm{S}) = \left(\frac{1}{N}\right) \sum_{t=1}^{N} \left(GSim_F(\mathrm{S}) < GSim_F(\mathrm{S_t})\right)$$

By presetting M, we were able to assess the range of function similarity values by a single offline calculation and avoid costly repetitive calculations based on the number of genes per studied group.

D. Similarity search and clustering experiments

In order to investigate the relation between expression and function, two sets of experiments are performed. The first set of experiments helps us investigate whether the application of an anatomy mask reveals a clearer association between gene expression and gene function focusing on six prototype genes. These genes are selected by geneticists [5] as they display diverse expression patterns throughout the GEM. Considering each prototype gene as a query, we retrieve genes of similar expression maps based on expression similarity of Eq. 3. The groups are formed independently for each query, thus the same genes can appear in each group. The results are assessed for increasing group sizes containing the 0.1%, 0.2%, ..., 1%, closest to the query genes. For each group size (= 7, 15, ..., 78 genes) we then calculate the average function similarity. Function similarity is assessed by $p\text{-}value_F$; smaller p-value indicates a functionally more coherent group of genes.

In the second set of experiments we employ clustering analysis aiming to classify the available genes into clusters that have both similar GEMs and similar gene functions. First clusters of GEMs are determined by the k-means algorithm [8] using the weighted Euclidean distance function (Eq. 3). According to this criterion, the clusters consist of genes with similar expression in the studied region of interest (all 4 spatial maps are tested independently). Then only the clusters with significant expression maps similarity and function similarity are retained, whereas the rest of the clusters are further split into an increasing number of smaller clusters until they reach 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) is not pre-defined, but calculated in a hierarchical fashion [4]. The results of this set of experiments are 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, $i = 1,2, ...,\ k$, 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, $D_{inter} = \min\left(\left|\mu_i - \mu_j\right|^2, i = 1, 2, \dots, k-1,\ j = i+1, \dots, k\right)$.

III. RESULTS

Our dataset is composed of 2-dimensional reconstructed images of expression of 7783 genes acquired by voxelation at a resolution of 1 mm^3 [5]. All 7783 genes are annotated genes i.e. their function information can be found from online databases.

A. Similarity search based on prototype genes

Fig. 3 shows the gene expression maps for 6 genes selected as queries. Protein phosphatase 1, regulatory (inhibitor) subunit 1B (PPP1r1b) is strongly expressed in striatum, necdin (Ndn) and Homo sapiens loss of heterozygosity (HSLOH11) are expressed in hypothalamus, serine (or cysteine) peptidase inhibitor, clade B, member 1a (Serpinb1a) is weakly expressed in striatum, nuclear factor I/X (Nfix) is expressed in a gradient pattern in cortex and pre-B-cell leukemia transcription factor 3 (Pbx3) is expressed in striatum and adjacent ventral structures. For each prototype gene, we detected an increasing number of genes most similar to the query by using different spatial maps and we calculated the significance of function similarity ($p\text{-}value_F$) in the group. The function similarity was considered with respect to the three function categories,

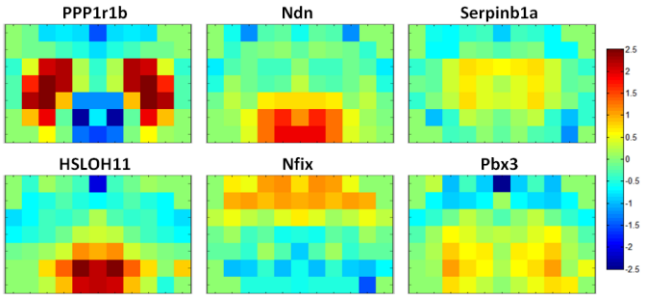


Figure 3. Gene expression maps of the selected prototype genes.

Cellular Component, Molecular Function and Biological Process. Some indicative results are shown for the prototype genes *Pbx3* and *PPP1r1b* in Tables I and II, correspondingly. We highlight p-values that are smaller than 0.05.

An example demonstrating the importance of the use of the confidence map is presented in Table I. It can be seen that the p-values regarding Biological Process are much smaller when applying the confidence map (global approach with PV correction) in the computation of Sim_G 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 similar gene expression in specific anatomic regions might indicate similar gene function, we calculated the GEM's similarity by using each one of the three spatial maps w_{GM}, w_{WM}, w_{HR} and by using no mask (global approach). In most cases the use of a spatial map provided better function similarity (i.e. smaller p-value) compared to the absence of any map. Among the three spatial maps, each gene was associated mainly with one of them; for example *PPP1r1b* displayed higher function similarity when focusing on WM region, whereas *Nfix* displayed higher results when focusing on GM region. Furthermore, the assessment of function similarity by applying the global approach revealed that the proposed spatially-oriented approach of GEMs achieved not only higher function similarity (e.g. regarding the Biological Process for *PPP1r1b*) but also revealed new function categories that are related with the particular gene (e.g. Cellular Component and Biological Process for *PPP1r1b*). These comparative results regarding *PPP1r1b* gene are illustrated in Table II.

In order to summarize the results for all 6 prototype genes and all spatial maps, we performed paired t-tests to determine if the use of any spatial map was effective. We compared the 10 function similarity values (p-values) for each function category under the null hypothesis of no average difference between the results produced with and without a spatial map. Table III shows all cases with function similarity difference significantly greater than zero (two-tail $p < 0.05$). We can see that in 15 out of 72 cases (6 genes × 3 function categories × 4 spatial maps) the function similarity changed significantly. In 11 cases it increased (indicated by an up-ward arrow ↑) and in 4 cases it decreased (shown by a down-ward arrow ↓) when one of the spatial maps ($w_C, w_{GM}, w_{WM}, w_{HR}$) was used. This means that more groups of genes with similar function were constructed than destructed by the localized approach.

If we focus on the 4 cases with function similarity reduction, we notice the following; *PPP1r1b* and *HSLOH11* are mainly expressed in WM or HR. Thus the inspection of the expression pattern in GM brings no significant information gain which might explain the function similarity reduction by the local-GM approach. Similarly, *Nfix* is mainly expressed in GM and shows significant function similarity related to Molecular Function when all anatomic locations are considered (with or without PV correction). By focusing on WM, which shows more subtle patterns, the relation to Molecular Function similarity is not revealed. On the other hand, *Ndn* shows under-expression in primary motor area or somatosensory area. Focusing on GM seems to increase similarity related to Cellular Component and Biological Process and decrease similarity related to Molecular Function. It might be possible that Molecular Function is not related to primary motor area or somatosensory area.

Regarding the cases with function similarity increase, it is interesting to notice that there are genes with low expression in a particular location which have very similar function. For example *Ndn* is not expressed in WM. The analysis showed that similar genes (also not expressed in WM) have similar function related to Cellular Component and Biological Process. This indicates that the lack of expression might also be meaningful.

Moreover, the PV correction (use of w_c) seems to significantly improve the results for the Biological Process and leave the results unaltered for the other two functions. Also there are no cases in which the PV correction causes a reduction in function similarity.

Summarizing, the gene expression in specific anatomic regions of the brain seems to play an important role in the identification of gene function. However comparison with other studies is required to validate the previous interpretations of the results for the selected genes. The application of the method to a significantly larger number of target genes will reveal the repeatability of the drawn conclusions.

B. Clustering based on all genes

The GEMs of each significant cluster obtained by the hierarchical k-means algorithm for different ontologies are averaged and illustrated in Fig. 4. The average maps, each of them representing one cluster, are shown only for the first 6 significant clusters for each gene function and each spatial map used in the clustering process. The significant clusters with low $p\text{-}value_F$ in any one of the three function categories are shown in the right column of Fig. 4. Although red pixels of GEMS indicate strong positive expression

while blue-like pixels indicate strong negative expression, both of them encode important information for the analysis of gene expression in brain regions. As can be seen in Fig. 4, both function categories of Cellular Component and Molecular Function are expressed prominently in regions of GM and HR, while Biological Process is expressed throughout the brain (all three regions).

Also, it can be noticed that for each function category different patterns are detected according to the incorporated spatial map. Hence, the application of such a localized approach allows the detection of spatially local patterns that would not be discovered otherwise. For example, the incorporation of w_{GM} showed that 619 genes (cluster 1) with strong expression in motor area or somatosensory area have high similarity related to Cellular Component. This pattern was not detected when the whole GEMs were considered (with or without correction for partial volume).

As a note, a few clusters are formed showing strong negative expression in the anterior cingulate area (corresponding to voxel G6 in Fig. 2a) and almost no expression elsewhere. In the current approach we did not consider the partial volume and background effect in the anterior cingulate area thus this strong expression value influences the clustering when w_{GM} is used. We will change this in the future by down-weighting this voxel.

Table III shows the clustering validity score (D_{inter} / D_{intra}) with the highest 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 global approach with PV correction 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.

IV. Discussion and Conclusion

Comparison of expression patterns obtained from microarray voxelation can be performed using cluster analysis, which is an unbiased discovery-driven method. Clusters of many genes with high similarity of both expression and function give greater clarity of consensus expression images and highlight distinct expression patterns in various brain regions such as cortex, white matter, striatum and hypothalamus. Here, new function categories were revealed by introducing spatial information from brain anatomy into the analysis of GEMs. For example, we obtained improved fit for Biological Processes with respect to the gene Pbx3, and uncovered Cellular Component and Biological Processes for the gene Ppp1r1b. Such clearer identification of genes' function categories enriches our understanding of the genetic circuitry contributing to brain development. Moreover, the use of a confidence map in the calculation of GEMs similarity resulted in higher average function similarity, improving the robustness of function categorization.

By clustering GEMs of genes with known and unknown function together, our approach has the potential for use in predicting unknown gene functions in the brain. Our future research plans include the application of the proposed methodology to voxelation data of mouse models of neurological diseases to provide insights into the genomic mechanisms of these disorders. Also, the collection of data and assessment of the method in different time points will allow to study the dynamic change in gene correlation and spatial localization.

The classification methods described here would also be applicable to the analysis of human brain voxelation data. Such data is available [2]. However, compared to the mouse, the main difficulties in using human data are the greater variances in inter-individual anatomy and post-mortem conditions of the human samples. The resulting additional noise may render the analyses less reliable. Furthermore, obtaining multiple brain samples is more difficult for humans than for mice.

Acknowledgment

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.

References

[1] R. P. Singh, et al., "High-resolution voxelation mapping of human and rodent brain gene expression", Journal of Neuroscience Methods, vol. 125, May 2003, pp. 93-101.

[2] V. M. Brown, A. Ossadtchi, A. H. Khan, S. R. Cherry, R. M. Leahy and D. J. Smith, "High-throughput imaging of brain gene expression", Genome Research, vol. 12(2), Oct. 2002, pp. 244-254.

[3] K. Horan, et al., "Annotating genes of known and unknown function by large-scale coexpression analysis", Plant Physiology, vol. 147(1), May 2008, pp. 41-57.

[4] L. An, H. Xie, M. H. Chin, Z. Obradovic, D. J. Smith and V. Megalooikonomou, "Analysis of multiplex gene expression maps obtained by voxelation", BMC Bioinformatics, vol. 10(Suppl. 4), Apr. 2009.

[5] M. H. Chin, et al., "A genome-scale map of expression for a mouse brain section obtained using voxelation", Physiological Genomic, vol. 30 (3), 2007, pp. 313-321.

[6] Mouse Brain Atlas: C57BL/6J Coronal, http://www.mbl.org/atlas170/atlas170_frame.html

[7] D. Lin, "An Information-Theoretic Definition of Similarity", Proc. International Conference on Machine Learning (ICML 98), Morgan Kaufmann, 1998, pp. 296-304.

[8] J. B. MacQueen, "Some Methods for Classification and Analysis of MultiVariate Observations", Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.

TABLE I. AVERAGE FUNCTION SIMILARITY ($p\text{-}value_F$) OF GENES SIMILAR TO THE GENE PBX3

Group size	Global approach with PV correction			Global approach		
	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 II. AVERAGE FUNCTION SIMILARITY ($p\text{-}value_F$) OF GENES SIMILAR TO THE GENE PPP1R1B

Group size	Local-WM approach			Global approach		
	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

TABLE III. CHANGE IN FUNCTION SIMILARITY BY THE LOCALIZED APPROACH. THE ARROWS INDICATE SIMILARITY INCREASE (↑) OR DECREASE (↓) WHEN ANY OF THE SPATIAL MAPS ($w_C, w_{GM}, w_{WM}, w_{HR}$) WAS USED

	Over-expressed	Under-expressed	Cellular Component	Molecular Function	Biological Process
PPP1r1b	WM	HR	w_{GM} ↓		w_C ↑
Ndn	HR	GM	w_{GM} ↑, w_{WM} ↑	w_{GM} ↓	w_{GM} ↑, w_{WM} ↑, w_{HR} ↑
Serpinb1a	WM			w_{WM} ↑	
HSLOH11	HR		w_{GM} ↑	w_{WM} ↑	w_{GM} ↓
Nfix	GM	HR		w_{WM} ↓	
Pbx3	WM, HR	GM			w_C ↑, w_{WM} ↑

TABLE IV. VALIDITY SCORE (D_{INTER} / D_{INTRA}) FOR ALL SIGNIFICANT CLUSTERS. THE NUMBER OF SIGNIFICANT CLUSTERS IS SHOWN IN PARENTHESES FOR EACH CASE

	Global approach	Global approach with PV correction	Local-GM approach	Local-WM approach	Local-HR approach
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)

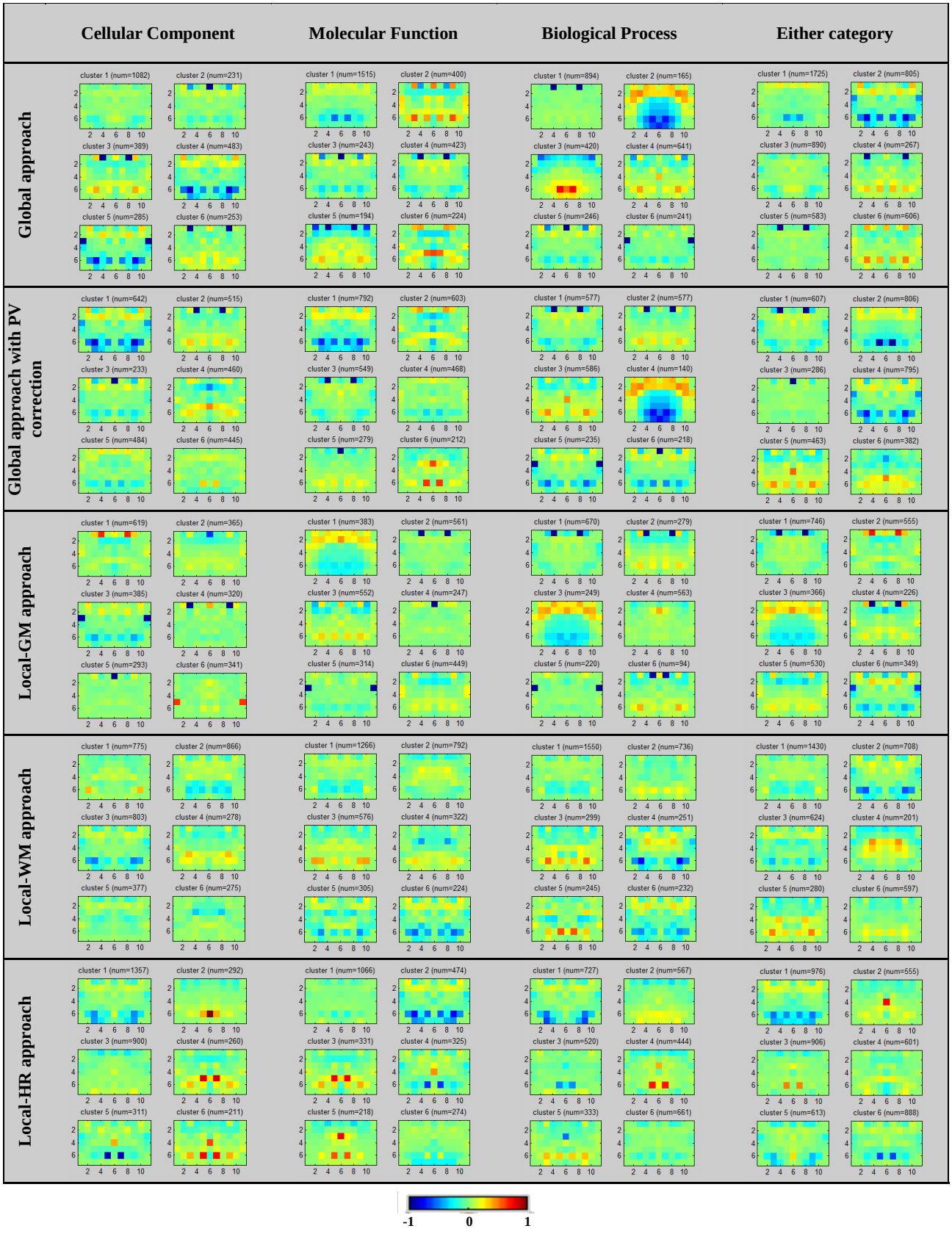


Figure 4. Average gene expression maps for the first 6 significant clusters for each function