ANALYSIS OF MULTIPLEX GENE EXPRESSION MAPS OBTAINED BY VOXELATION

Li An[1], Hongbo Xie[1], Mark Chin[2], Zoran Obradovic[1], Desmond Smith[2], and Vasileios Megalooikonomou[1]*
[1]*Department of Computer and Information Sciences, Temple University, PA, USA*
[2]*Department of Human Genetics, David Geffen School of Medicine, UCLA, CA, USA*
** Corresponding author: Li An E-mail: anli@temple.edu*

Abstract

In this paper we present an approach for identifying the relationships between gene expression maps and gene functions based on the multiplex gene expression maps of mouse brain obtained by voxelation. To analyze the dataset, we choose typical genes as queries and aim at discovering similar gene groups. We use the wavelet transform for extracting features from the left and right hemispheres averaged gene expression maps, and the Euclidean distance between each pair of feature vectors to determine gene similarity. We also perform a multiple clustering approach on the gene expression maps, combined with hierarchical clustering. Among each group of similar genes and clusters, the gene function similarity is measured by calculating the average gene function distances in the gene ontology structure. The experimental results confirm the hypothesis that genes with similar gene expression maps might have similar gene functions. The voxelation data takes into account the location information of gene expression level in mouse brain, which is novel in related research. The proposed approach can potentially be used to predict gene functions and provide helpful suggestions to biologists.

Keywords

Voxelation; gene expression maps; gene function; clustering

1. Introduction

Gene expression signatures in the mammalian brain hold the key to understanding neural development and neurological disease. A new approach is developed by combining voxelation with microarrays for acquisition of genome-wide atlases of expression patterns in the brain [1-2]. Voxelation involves dicing the brain into spatially registered voxels (cubes). Each voxel is then assayed for gene expression levels and images are reconstructed by compiling the expression data back into their original locations [3-4]. It produces multiple volumetric maps of gene expression analogous to the images reconstructed in biomedical imaging systems [5-7]. Related research work suggests that voxelation is a useful approach for understanding how genome constructs the brain. Gene expression patterns obtained by voxelation show good agreement with known expression patterns [1].

Researchers at David Geffen School of Medicine at UCLA used voxelation in combination with microarrays to analyze whole mouse brains at low resolution [1]. They acquired 2-dimensional images of gene expression for 20,847 genes, obtained by using microarrays in combination with voxelation for a 1mm slice of the mouse brain at the level of striatum (Fig.1). The coronal slice from a mouse brain is put on a stage and is cut with a matrix of blades that are spaced 1 mm apart thus resulting in cubes (voxels) which are $1mm^3$. There are voxels like A3, B9..., as Fig.2 shows. A1, A2... are in red signifying that voxels were not retrieved from these spots, but empty voxels were assigned to maintain a rectangular. So, each gene is represented by the 68 gene expression values composing a gene expression map of mice brain (Fig.2). In other words, the dataset is a 20847 by 68 matrix, in which each row represents a particular gene, and each column is the log2 ratio expression value for the particular probe in a given voxel. The data was found to be of good quality based on multiple independent criteria and insights provided by others into the molecular architecture of the mammalian brain. Known and novel genes were identified with expression patterns localized to defined substructures within the brain.

Previous work [8-10] has been done to detect gene functions, without though taking into account the

978-0-7695-3452-7/08 $25.00 © 2008 IEEE
DOI 10.1109/BIBM.2008.67

IEEE computer society

location information of a gene's expression in a mouse brain to study gene functions. Based on the multiple volumetric maps of gene expression of mice brain, in this study we identify the relations between gene expression maps and gene functions. Our analysis consists of similarity queries and clustering analysis of

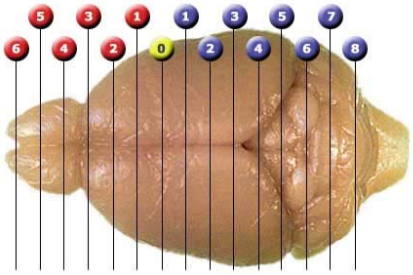

Fig. 1 The mouse brain at bregma = 0

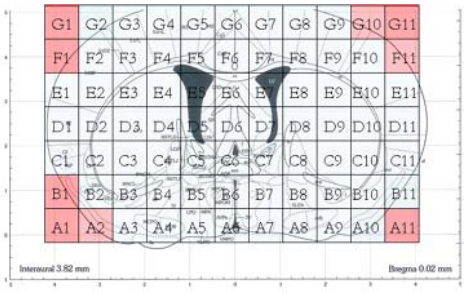

Fig.2 Voxels of the coronal slice

the gene expression maps. The proposed approach is based on the features extracted by the wavelet transform from the original gene expression maps. Among each group of similar genes, we calculate the average gene function distance in the gene ontology structure to indicate the gene function similarity. K-means is used for clustering gene expression maps. The significant clusters that have both similar gene expression maps and similar gene functions are obtained by a proposed technique, which we call multiple clustering.

The experimental results from the similarity analysis confirm the hypothesis that genes with similar gene expression map might have similar gene functions. The clustering analysis also detects certain clusters of genes that have similar functions. The proposed approach and analysis can potentially be used to predict gene functions and provide suggestions to biologists.

2. Methods

The proposed approach includes two parts. The first part consists of similarity queries based on gene expression maps. For this part we choose typical genes as queries and search for similar genes based on their expression maps and features. The second part consists of clustering analysis of the gene expression maps and computation of the average function distance for each cluster. In addition to these two parts, we attempted to identify the relations between each gene's expression map and its participatory functions. The hypothesis is that genes with similar gene expression map have similar gene functions. The results in Section 4 show that this hypothesis holds for certain groups of genes.

2.1 Finding similar genes

In this part of the analysis we choose typical genes as queries and attempt to discover groups of genes similar (w.r.t. the gene expression maps) to the query gene.

2.1.1 Reducing Noise

The original dataset we analyzed consists of data for 20847 genes. Data with no significant gene expression value can be viewed as noise. We eliminate this kind of data to improve the results. If none of the expression values of a gene is bigger than 1 or smaller than -1, we consider the gene insignificant. After normalizing (making sure the mean is 0 and standard deviation is 1) the rest of the data, we obtain a new dataset which has 13576 significant genes. We observe that only half of the genes in the dataset are known genes whose annotation information can be found from an online database, including the function information. The genes with unknown function might confuse our results. So we only consider 7783 genes (from the 13576 significant genes) whose functions are known as the basic dataset for our analysis.

We also take advantage of the inherent bilateral symmetry of the mouse brain by averaging the left and right hemispheres, which proves (as our experimental results demonstrate) very useful in decreasing noise. Mice do not have "handedness" or speech-centers of the brain which are known to be localized to one hemisphere in humans. Therefore, a voxel or two that stands out is probably more believable if it has corresponding voxel(s) located in the same general location in the other hemisphere.

2.1.2 Wavelet Features Extraction

In order to take into account spatial information about the 68 voxels we consider in the brain map, we employ wavelets in feature extraction. Working directly with the original 68-element vectors of gene expression values ignores the spatial information.

Intuitively, we expect to have correlation among the values of voxels in the same spatial neighborhood. The wavelet transform is a tool that cuts up data, functions or operators into different frequency components and study each component with a resolution matched to its scale [11]. Here, we use the discrete wavelet transform (DWT) with single-level two-dimensional wavelet decomposition employing the Daubechies wavelet function to extract features based on the gene expression matrix (Fig. 2). The outputs of the wavelet transformation involve approximation coefficients, which are the average of gene expression values in neighborhood voxels, and detail coefficients, which indicate the difference of each voxel from the average. By using multilevel 2-D wavelet decomposition on the 7 by 11 matrix (Fig. 2) at level 4, we obtain 75 coefficients including approximation and detail coefficients to achieve the best results.

2.1.3 Gene Maps Similarity

To determine the gene maps (gene expression matrix) similarity, the Euclidean distance between each pair of vectors (each with 75 wavelet features) is used. Let *S* be a set of Euclidean distances between the query and all the other genes in the dataset, and *Dis* be a special distance between the query and a general gene. Then *Num* is the number of distances *Si*, where $Si < Dis, Si \in S$. We define the p-value of *Dis* as $\frac{Num}{n}$, where *n* is the number of elements in set *S*. So for each query, we can find a number of genes which are similar to the query with the corresponding small p-value.

2.1.4 Gene Functions Similarity

To identify the functions similarity, we use the average function distance in the gene ontology structure among each group of similar genes. For example, Fig.3 shows a part of the gene ontology structure. Each node corresponds to a gene function, so the function distance between functions B and E is 3. The smaller the function distance the more similar the two functions are.

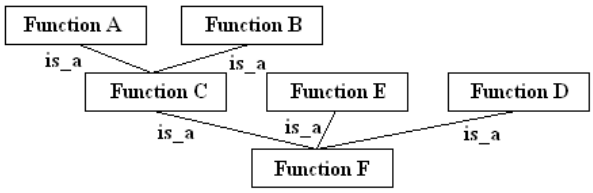

Fig.3 A part of the gene ontology structure

Because each gene holds more than one gene function, we take all the functions of all the genes in the group to build a set of functions. The average gene function distance is obtained by averaging the distances between each pair of functions in the set; thus, it can be used to determine the function similarity in the group. To rank the function distance values, we randomly choose 1000 gene groups, each consisting of 1000 genes. The average function distance in each group is calculated, resulting in a set *U* of 1000 values, called set *rand_func_dis*. For a given average function distance *G_Dis*, the p-value is defined as $\frac{Num\_func}{1000}$, where *Num_func* is the number of *Ui* with $Ui < G\_Dis$, $Ui \in U$. So the gene function similarity in a group of genes can be identified by how smaller the p-value of the average function distance of the group is.

2.2 Clustering analysis

In addition to similarity analysis we propose clustering analysis of the gene expression maps and computation of the average function distance in each cluster. Here, we attempt to find the significant clusters that have both similar gene expression maps and similar gene functions. After comparing different clustering methods [12-14], we chose the K-means algorithm [15] as the clustering tool. We also propose a clustering method which is a combination of multiple clustering and hierarchical clustering.

2.2.1 Multiple clustering

We propose a multiple clustering method to perform the clustering. This method consists of multiple steps. In each step, K-means is used on the current dataset producing n clusters. Among the n clusters, suppose there are m significant clusters ($m<n$) whose p-value of average function distance is smaller then 0.05. The new dataset for the next step is obtained by removing the m clusters, previously determined as significant, from the current dataset. Then K-means is repeated again on the newly formed dataset. The process is repeated many times until there are no significant clusters (i.e., with p-value<0.05) that can be found, or the size of clusters obtained is too small to be meaningful.

2.2.2 Hierarchical clustering

For the K-means clustering algorithm, the number of clusters is predefined. Without prior knowledge, the estimation of the appropriate number of clusters becomes a challenge in clustering analysis to

accurately get the most significant clusters. In this paper divisive hierarchical clustering is used to determine the number of clusters for K-means. In each step of multiple clustering, the number of clusters *n* starts at a minimum value and is incremented. At the first step, *n* starts at 2 and is incremented by 1 until the significant clusters are found. At that time, we assume *n*=K. Then the significant clusters are removed from the dataset and the clustering repeats on the remaining genes. The clustering proceeds to the next step with the number of clusters *n* in this step starting at K-1.

2.2.3 Cluster Validation

In this paper, we use the following strategy to judge the performance of clustering. The point-to-centroid distance is used to determine whether the clusters are compact. The intra-cluster distance is defined as

$$Intra\_Cluster\_Dist = \frac{1}{N}\sum_{i=1}^{k}\sum_{x_j \in S_i}\left|x_j - \mu_i\right|^2$$

where N is the total number of data points, S_i, $i=1,2,..,k,$ are the k clusters and μ_i is the centroid or mean point of all the points $x_j \in S_i$.

Another measure of cluster performance is the inter-cluster distance, i.e., the distance between clusters. This is calculated by taking the minimum of the distances between each pair of cluster centroids as follows:

$$Inter\_Cluster\_Dist = \min\left(\left|\mu_i - \mu_j\right|^2\right), i = 1,2,...,k-1$$
$$j = i+1,...,k$$

We take the minimum of the distance between clusters because it is the upper limit of cluster performance and is expected to be maximized. The ratio of intra-cluster distance to inter-cluster distance can serve as an evaluation function for cluster performance. The validity of a k-clustering result is defined as

$$Validity = \frac{Inter\_Cluster\_Dist}{Intra\_Cluster\_Dist}$$

Since we want to maximize the inter-cluster distance and minimize the intra-cluster distance, we want the validity value to be maximized.

3. Results

3.1 Finding similar genes

In these experiments we chose eight genes as queries (similarly to [1]). Fig.4 shows the gene expression maps of the eight queries. The eight genes are selected [1] as having restricted expression patterns based on the micro-array voxelation data. For example, cortically expressed genes include Clstn1, Ppp1r1b is with spatially restricted expression in the striatum, and Ndn shows expression restricted to the hypothalamic region. Different colors represent different levels of gene expression. Here, we try to find similar genes to a query gene based on the reduced dataset (7783 genes) and the wavelet features.

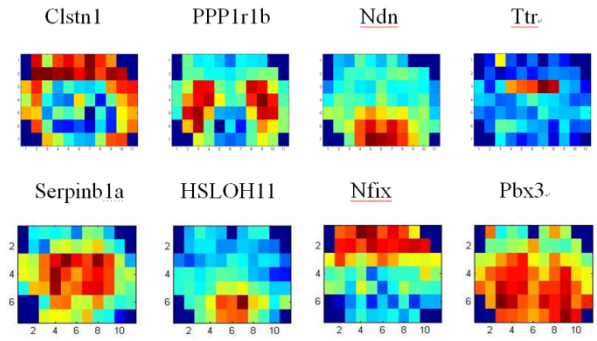

Fig.4 Typical genes used as queries

We consider increasing thresholds of the p-value (from 0.0005 to 0.009) and find a number of similar genes whose distance to the target gene is smaller than the threshold. Then we calculate the average function distance in the group of the selected similar genes. Tables 1 - 2 show the results of genes Clstn1, and PPPlrlb. We highlight p-values of function distance that are smaller than 0.05. We consider the function distance with respect to three categories: cellular component, molecular function and biological process.

Examining the group of similar genes of target1 (Clstn1), Table 1 shows that there are very small p-values of function distance in the category of biological process, meaning that these similar genes have functions that are very close with respect to position in the gene ontology structure (i.e., these similar genes have similar functions in the category of biological process). The experimental results of the other target (Table 2) also show that genes with similar gene expression maps have very close function position in gene ontology structure, at least in one of the three biological categories (these results have not been reported in detail here due to paper size restrictions).

3.2 Finding significant clusters

In these experiments, we apply clustering iteratively to get the significant clusters with both low p-value (<0.05) of Euclidean Distance of gene expression and low p-value of Function Distance. The experiments are applied on the data set of 7883 genes that consists of both significant and known genes.

Table 1. Results for Gene Clstn1

Euclidean Distance (P-value)	Number of similar genes	Average Function Distance (p-value)		
		Cellular Component	Molecular Function	Biological Process
0.0005	10	0.379	0.191	**0.001**
0.001	21	0.081	**0.041**	**0.016**
0.002	42	0.557	**0.041**	**0.001**
0.003	63	0.720	0.728	**0.001**
0.004	83	0.775	0.913	0.071
0.005	104	0.610	0.705	0.130
0.006	125	0.729	0.899	0.111

Table 2. Results for Gene PPP1r1b

Euclidean Distance (P-value)	Number of similar genes	Average Function Distance (p-value)		
		Cellular Component	Molecular Function	Biological Process
0. 0005	10	**0. 010**	1. 000	**0. 001**
0. 001	21	**0. 020**	0. 989	**0. 047**
0. 002	42	0. 574	0. 400	0. 166
0. 003	63	0. 172	0. 834	0. 064
0. 004	83	**0. 035**	0. 998	0. 231
0. 005	104	0. 082	0. 998	0. 441
0. 006	125	0. 162	0. 998	0. 449

Each gene is represented by the full 75 wavelet features extracted from the hemi-averaged gene expression map. The multiple clustering combined with hierarchical clustering is repeatedly applied until there are no significant clusters found, or the size of clusters obtained is too small. Fig.5 - Fig.7 show the average of gene expression maps of significant clusters obtained by k-means for different datasets. Each gene expression map corresponds to one cluster.

Since there are three categories of gene functions in gene ontology, we attempted to identify significant clusters for each one of the three different categories (separately) and then with respect to all of the three categories together. For example, when considering the category "Cellular Component", we only searched for significant clusters with low p-value of Functions Distance in the category "Cellular Component". In the case where we considered all three categories together, we searched for significant clusters with low p-value of Functions Distance in any one of the three categories.

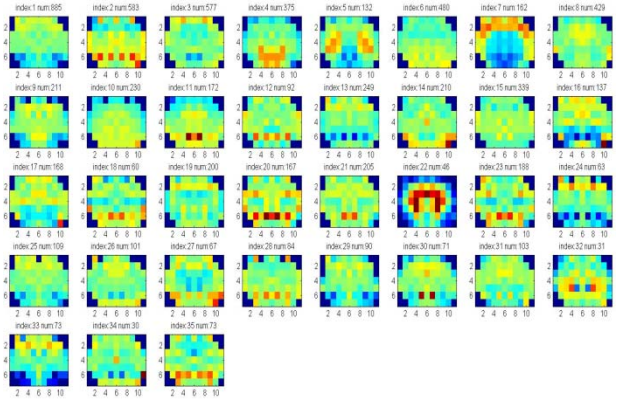

Fig.5 Cellular Component, 35 significant clusters found

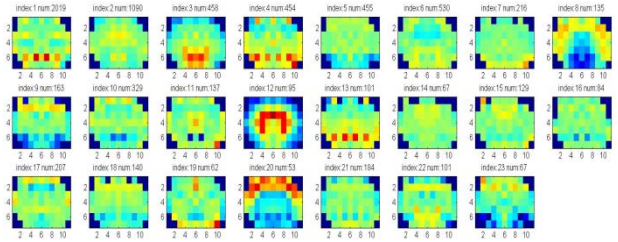

Fig.6 Molecular Function, 23 significant clusters found

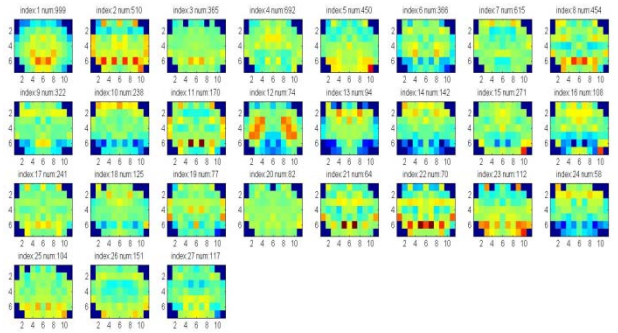

Fig.7 Biological Process, 27 significant clusters found

3.2 Cluster validation

In order to evaluate the proposed hierarchical clustering approaches, we used two different clustering algorithms in each step of the multiple clustering to find out the significant clusters. One is k-means with a selected k number, where k is the square root of the size of the data set. The other algorithm is using hierarchical clustering to decide the most suitable k. We evaluated the significant clusters we obtained by calculating cluster distance and compared the results of the two kinds of clustering methods.

Table 3 shows that the validity value of the hierarchical clustering (used in our experiments) is larger than the validity value of the selected k clustering in each category.

Function Category	Method	Intra Cluster Distance	Inter Cluster Distance	Validity
Cellular Component	Selected k	4.0212	0.6355	0.1580
	Hierarchical	4.6096	0.8928	0.1937
Molecular Function	Selected k	4.0469	0.5148	0.1272
	Hierarchical	5.0396	1.1211	0.2225
Biological Process	Selected k	3.8917	0.6472	0.1663
	Hierarchical	4.7262	0.7971	0.1687
All the three categories	Selected k	4.0110	0.5543	0.1382
	Hierarchical	4.8385	0.9813	0.2028

Table 3. Comparing two clustering methods: Intra_Cluster_Dist measures the intra distance inside a cluster, Inter_Cluster_Dist measures the distance between clusters, and Validity indicates the overall performance of the clustering.

4. Discussion

Although research work has been done to detect gene functions, not much effort has focused on identifying the relation between gene expression maps in mice brain and related gene functions. By using wavelet features to determine the similarity of gene expression maps, and the function distance in ontology structure to determine the similarity of gene functions, our 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, clustering analysis detected certain clusters of genes that have both similar gene expression maps and gene functions. So, the obtained results confirm the hypothesis that genes with similar gene expression map might have similar gene functions. This paper tries to quantify this hypothesis presenting a way to evaluate it as well as a set of genes for which the hypothesis holds.

To obtain the significant clusters, we only analyze the genes which are both significant and have known functions, i.e., genes whose annotation information can be found at online databases, including the function information. The results based on the dataset we considered support the following claim. By examining the known and unknown genes together to find groups of similar genes (which are obtained either by similarity finding or clustering), one might provide helpful suggestions to biologists about unknown genes having similar gene functions to the known genes in the same group. Therefore the proposed approach has the potential to be used in predicting gene functions.

Acknowledgements

This work was supported in part by the National Science Foundation under grants IIS-0705215, IIS-0237921 and the National Institutes of Health under grant R01 MH68066-04.

References

[1] Mark H. Chin, Alex B. Geng, Arshad H. Khan, Wei-Jun Qiang, Vladislav A. Petyuk, Jyl Boline, Shawn Levy, A genome-scale map of expression for a mouse brain section obtained using Voxelation, Physiological Genomics, 2007, p. 313-321.

[2] Brown, P.O. and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. Nat. Genet. 21, p.33–37.

[3] Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. 1999. High density synthetic oligonucleotide arrays. Nat. Genet. 21, p. 20–24.

[4] Ram P. Singh, Vanessa M. Brown, Abhijit Chaudhari, Arshad H. Khan, Alex Ossadtchi, Daniel M. Sforza, A. Ken Meadors, Simon R. Cherry, Richard M. Leahy, Desmond J. Smith, High-resolution voxelation mapping of human and rodent brain gene expression. J Neurosci Methods, 2003. 125(1-2): p. 93-101.

[5] Dahai Liu and Desmond J. Smith, Voxelation and gene expression tomography for the acquisition of 3-D gene expression maps in the brain, Methods, Volume 31, Issue 4, 2003, p. 317-325.

[6] Brown VM, Ossadtchi A, Khan AH, Gambhir SS, Cherry SR, Leahy RM, Smith DJ. Gene expression tomography, Physiol Genomics 8, 2002, p. 159-167.

[7] Brown VM, Ossadtchi A, Khan AH, Cherry SR, Leahy RM, Smith DJ., High-throughput imaging of brain gene expression. Genome Res, 2002. 12(2): p. 244-54.

[8] Bruce Alberts, Adil E. Shamoo, Alexander Johnson, Felix A. Khin-Maung-Gyi, Julian Lewis, Martin Raff, Keith Roberts, Molecular Biology of the Cell, 2002, ISBN-10: 0815332181.

[9] Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. Science 270, p.484–487.

[10] Boguski, M.S. and A.R. Jones, Neurogenomics: at the intersection of neurobiology and genome sciences. Nat Neurosci, 2004. 7(5): p. 429-33.

[11] Gerald Kaiser, A Friendly Guide to Wavelets, Birkhauser, 1994, ISBN 0-8176-3711-7 .

[12] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95, 1998, 14863-8.

[13] Hartigan, J. A. Clustering algorithms, New York,: Wiley, 1975.

[14] Jain, A. K., and Dubes, R. C. Algorithms for clustering data, Englewood Cliffs, N.J.: Prentice Hall, 1988.

[15] J. B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967, p.281-297