Identifying Gene Functions using Functional Expression Profiles obtained by Voxelation

Li An
Data Engineering Laboratory, Dept. of Computer and Information Sciences, Temple University
415 Wachman Hall,1805 N. Broad St., Philadelphia, PA 19122
215-204-9260
anli@temple.edu

Hongbo Xie
Center for Information Science and Technology, Temple University
Wachman Hall, 1805 N. Broad St.
Philadelphia, PA 19122
267-475-5147
michaelxie@ist.temple.edu

Zoran Obradovic
Center for Information Science and Technology, Temple University
303 Wachman Hall, 1805 N. Broad St.
Philadelphia, PA 19122
215-204-6265
zoran@ist.temple.edu

Desmond J. Smith
Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, UCLA
CHS, Los Angeles, CA 90095
2nd line of address
310-206-0086
dsmith@mednet.ucla.edu

Vasileios Megalooikonomou
Data Engineering Laboratory, Center for Information Science and Technology, Temple University
314 Wachman Hall,1805 N. Broad St.,
Philadelphia, PA 19122
215-204-5774
vasilis@temple.edu

ABSTRACT
Gene expression profiles have been widely used in functional genomic studies. However, not much work in traditional gene expression profiling takes into account the location information of a gene's expressions in the brain. Gene expression maps, which contain spatial information regarding the expression of genes in mice˝s brain, are obtained by combining voxelation and microarrays. Based on the idea that genes with similar gene expression maps may have similar gene functions, we propose an approach to identify gene functions. A gene function can potentially be associated with a specific gene expression profile. We name this specific gene expression profile, Functional Expression Profile (FEP). A functional expression profile can be obtained either by directly finding genes with a certain function, or by analyzing clusters of genes that have similar expression maps and similar functions. By taking advantage of the identified FEPs, we can annotate gene functions with high accuracy. Compared to the traditional K-nearest neighbor method, our approach shows higher accuracy in predicting functions. The images of FEPs are in good agreement with anatomical components of mice˝s brain, and provide valuable insight in terms of function prediction to biological scientists.

ACM-BCB 2010, Niagara Falls, NY, USA

Categories and Subject Descriptors
G.4 [**Programming Languages**]: H.2.8 Database Applications, *Data mining, Image databases;* I.5 PATTERN RECOGNITION; I.5.3 Clustering; I.5.4 Applications; J.3 LIFE AND MEDICAL SCIENCES, *Biology and genetics.*

General Terms
Algorithms, Management, Experimentation.

Keywords
Gene function annotation, voxelation, gene expression maps, and functional expression profile.

1. INTRODUCTION
The use of microarrays for gene expression profiling has been widely used in recent functional genomic studies. Gene expression signatures in the mammalian brain hold the key to understanding neural development and neurological disease. While research [1-4] has been done to detect gene functions, most of the time it has not taken into account the locations of a gene's expressions in the brain to identify gene functions. Voxelation is a new approach that involves dicing the brain into spatially registered voxels (cubes). It produces multiple volumetric maps of gene expression analogous to the images reconstructed in biomedical imaging systems [5-7]. Related research suggests that voxelation is a useful approach for understanding how the genome constructs the brain. Gene expression patterns obtained by voxelation show good agreement with known expression patterns [8-9]. Based on the genome-wide

atlases of expression patterns in the brain [10-11], gene function identification can be greatly improved in terms of accuracy.

Researchers at the David Geffen School of Medicine at UCLA used voxelation in combination with microarrays for acquisition of genome-wide atlases of expression patterns in the brain [10]. They acquired 2-dimensional images of gene expression for 20,847 genes. The procedure of obtaining the raw data is described here briefly. A fresh brain is removed from a sacrificed mouse, and then a 1mm slice of the brain at the level of striatum is obtained (Figure 1). The coronal slice is put on a stage and is cut by a matrix of blades that are spaced 1 mm apart, thus resulting in cubes (voxels) that are 1mm3. There are voxels like A3, B9..., as Figure 2 shows. A1, A2, B1... are in red, signifying that voxels were not retrieved from these spots, and are empty cubes that were assigned to maintain a rectangular. So, each gene is presented by the 68 gene expression values in 68 voxels to compose a gene expression map of a mouse‟s brain. This data has been found to be of good quality based on multiple independent criteria and insights provided by others [8-10] into the molecular architecture of the mammalian brain.

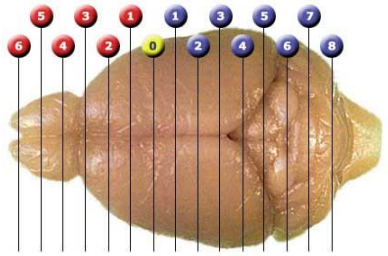

Figure 1 The mouse brain at bregma = 0

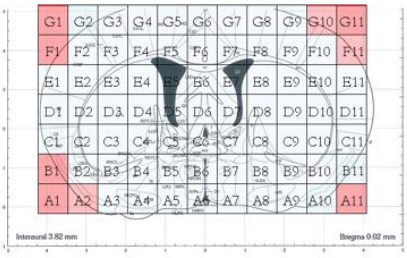

Figure 2 Voxels of the coronal slice

Our previous analysis of the gene expression maps [11] focused on the identification of the relationship between the gene functions and gene expression maps. We used wavelet features to determine the similarity of gene expression maps and a function distance in the gene ontology structure to determine the similarity of gene functions. In certain cases, the group of genes that was identified as similar to a target gene shared very similar gene functions in at least one gene function category. Moreover, clustering analysis detected a number of clusters of genes that have both similar gene expression maps and similar gene functions. These clusters were denoted as significant clusters. That work confirmed that the hypothesis that genes with similar gene expression maps have similar gene functions holds for a certain set of genes. Therefore, genes with currently unknown gene functions may have functions similar to those of known genes with which they have similar expression maps.

In this study, our goal is to identify gene functions based on the multiple volumetric maps of gene expression in mice brains. We take advantage of the relationship between gene expression maps and gene functions to predict gene functions. For a given gene function, there might be a specific gene expression map (profile) that is associated with the given function. The genes that have similar gene expression maps to the specific profile are supposed to hold similar gene functions. We name this specific gene expression profile, Functional Expression Profile (FEP). An FEP can be obtained directly by studying each gene function related to the dataset and identifying if the function has a specific gene expression profile, or it can be obtained through the average profiles of significant clusters of gene expression maps obtained by our previous analysis. We propose a gene function annotation method that takes advantage of the results of identified FEP. We compare the method with the traditional K-nearest neighbor (KNN) method that has been used in identifying gene functions [15, 16], which simply annotates a given gene with the functions of the top k genes in the training set with the highest correlation coefficient to that gene. The experimental results show that the accuracy of the identifying gene functions is high, in some cases reaching 99 percent, and the proposed approach compares favorably to the K-nearest neighbor method. Moreover, the FEPs obtained directly from gene functions have better performance in function prediction than the FEPs obtained by significant clusters.

2. DATA AND METHODS

2.1 Gene expression maps

The data set we consider in this study is a 20,847 by 68 matrix, in which each row represents the 68 expression values of a particular gene, and each column represents the log2 ratio expression values for all the probes (genes) in a given voxel. The 68 voxels are located in mice‟s brain, as Figure 2 shows. By using different colors to show different values of gene expression, the expression map for a certain gene can be visualized as in Figure 3.

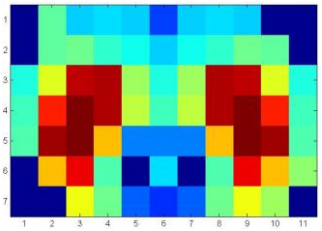

Figure 3 A visualized gene expression map

From the analysis we discard genes whose 68 expression values are all between -1 and 1. The remaining 13,576 genes‟ IDs are imported into the SOURCE [12] database [13] to retrieve their Gene Ontology (GO) annotation information. Out of the 13576 genes, 7883 genes are known genes and are annotated with at least one GO term. There are 2416 unique GO terms in total. Among those 2416 GO terms, 1065 are biological processes

where 693 of them are associated with at least two genes. 1103 GO terms belong to molecular functions where 707 of them are associated with at least two genes. 248 GO terms belong to cellular components where 207 of them are associated with at least two genes.

2.2 Averaging hemispheres

Since there is a large amount of noise in microarray experiments, we average the data over both hemispheres to improve the signal. Additionally, the averaging of hemispheres takes advantage of the inherent bilateral symmetry of mice"s brain. Mice do not have "handedness" or speech-centers in the brain, which are known to be localized to one hemisphere in humans. Therefore, a voxel or two that stands out is probably more reliable if it has a corresponding voxel located in the same general location in the other hemisphere. In the process of averaging, for each row of the map, we average the framed cells, as shown in Figure 4. Then, we replace B1 with B11, A2 with A10, and the averaged gene expression map is obtained as in Figure 5.

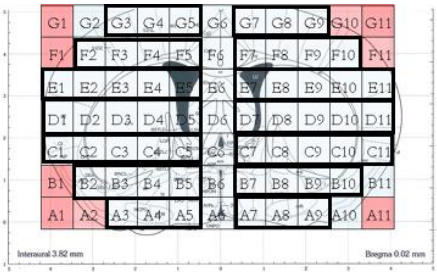

Figure 4 Averaging left and right hemispheres

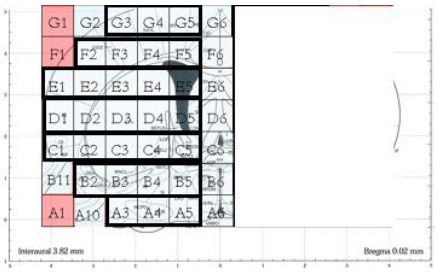

Figure 5 Averaged gene expression map

2.3 Wavelet transformation

Here we describe how we use the wavelet transform to extract new features from the original 68 expression values of each gene expression map. In the data set that we study, each row represents the gene expression values corresponding to the 68 voxels in the selected slice of mice"s brain. Intuitively, if an expression value is similar to other values in its spatial neighborhood, it is more reliable. However the original vectors of gene expression values ignore the spatial information. In order to measure the spatial consistency of expression values with others in their spatial neighborhood and to take into account the spatial factors of voxels in the brain map, we employ the wavelet transform to extract new features.

The wavelet transform is a tool that cuts up data, functions or operators into different frequency components and studies each component with a resolution matched to its scale [14]. 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 (see Figure 2). The output of the wavelet transformation consists of 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. For the averaged map of 6 by 7 cells, by employing multilevel 2-D wavelet decomposition at level 3, we obtain 42 coefficients (combining approximation and detail coefficients to approach the best results).

2.4 Functional expression profiles

2.4.1 Identifying FEPs by non-cluster-based method

One method to obtain FEPs is to explore each GO term (gene function) and identify all the genes that contain this GO term. Since not all genes with similar gene expression map have similar gene functions, we need to rank the group of gene expression patterns to determine if the genes with identical function have similar expression profiles. We study GO terms associated with at least two genes and use a statistical procedure to identify GO terms with average pair-wise gene profile correlation significantly higher than the correlation expected to be present at random. The random model assumes that genes corresponding to a given GO term are selected at random from the available pool of genes. The algorithm we use to test the null hypothesis assuming the random model is shown below.

Algorithm for identifying FEPs

1. Calculate the average pairwise correlation coefficient between n gene expression profiles associated with a given GO term;
2. Select n genes randomly from the dataset. Compute the average pairwise correlation coefficient in the random set of genes;
3. Repeat Step 2 M times, and report as p-value the proportion of the random sets with average pairwise correlation larger than that of the original gene set.
4. If the p-value obtained from Step 3 is less than a given threshold r, average the gene expression profiles, where genes are associated with the given function to create the Functional Expression Profile (FEP).

The remaining GO terms with p-values larger than the threshold are discarded since there is no sufficient evidence to demonstrate that the corresponding genes are correlated. We call this method non-cluster-based FEP method. The number of iterations M has been set to 10,000 and the threshold r has been set to 0.05 in our experiments, as discussed in Section 3.

2.4.2 Identifying FEPs by cluster-based method

During our previous analysis of this gene expression dataset [11], a number of significant gene clusters were identified with both similar gene expression maps and similar gene functions. Based on the 7883 known genes, the significant clusters were detected for the three categories of gene ontology (Cellular Component, Molecular Function, and Biological Process) separately.

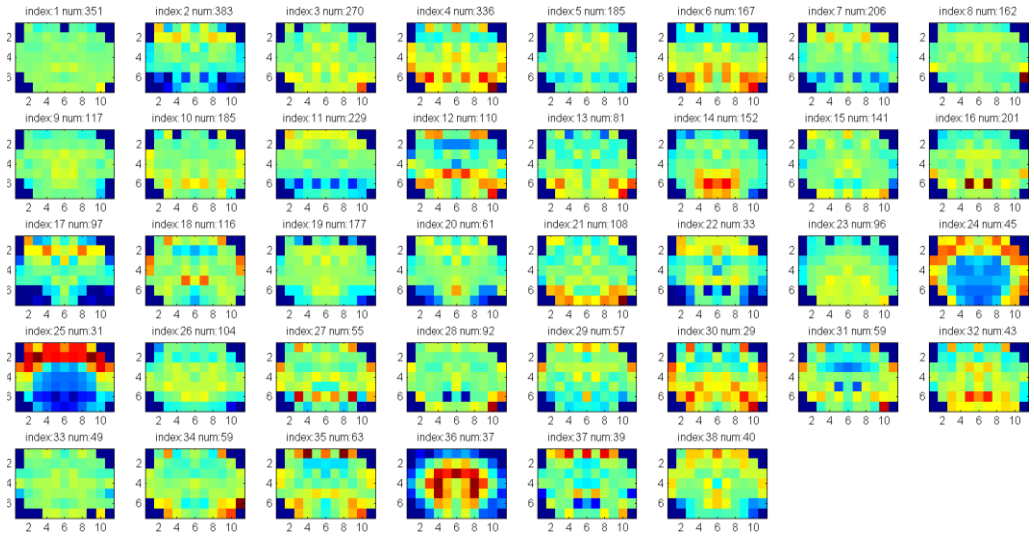

Figure 6 The 38 significant clusters found with respect to Cellular Component.

Table 1 shows the number of significant clusters we detected and the total number of genes in these clusters. Figure 6 shows the average of gene expression maps of significant clusters with respect to the category "Cellular Component". Each small image in this figure shows an average map of the 68 gene expression values for all genes in a cluster. This image can be viewed as FEP. We call this method cluster-based FEP method.

Table 1. Number of significant clusters

GO Category	Number of Significant Clusters	Number of genes in all significant clusters
Cellular Component	38	5651
Molecular Function	50	6112
Biological Process	43	5520

For each significant cluster, we examined all unique GO terms shown in the genes of the cluster. Suppose that there are N genes that include a certain GO term in the cluster, and that the size of the cluster is S. The GO terms with ratio N/S larger than a given threshold are reported. Those GO terms are associated with the average gene expression maps of the corresponding cluster, which are viewed as FEPs. Because the same GO term can appear frequently in different significant clusters, a GO term can be associated with several FEPs. Moreover, there can be several frequent GO terms within a significant cluster, or there can be no frequent GO terms within a cluster. The strategies to deal with the above cases are presented in Section 2.5.

2.5 Annotating unknown gene functions using identified FEP

In this study, our objective is to identify gene function by using gene expression maps. Traditional approaches for identifying unknown gene functions have numerous difficulties, e.g., the naive KNN method in which the neighborhood of a given point becomes very sparse in a high dimensional space [15]. Here we propose a gene function annotation method which takes advantage of the identified FEPs. We show the benefit of our approach by comparing it to the traditional K-nearest neighbor method.

The voxelation dataset is randomly split into two disjoint subsets: a training set and a test set. The training set contains 3/4 of the data. The remaining data forms the test set. For those 3126 gene ontology terms associated with the known and significant 7883 genes, we evaluate our approach for each function separately. That is, for each given function GO_j {1<j<3126}, our target label set is a 7883 by 1 vector. The value of the entry i {1<i<7883} of this vector is a binary variable where „1" indicates that gene G_i is annotated with function GO_j, and „0" otherwise. We first build the prediction model using the training set. Then we use the obtained model to label the test set and compare the assigned binary labels to the real labels. The accuracy is measured as the average of specificity and sensitivity. We repeat our approach for all functions and for each function we report the accuracy of the prediction.

For the functional expression profile approach, the training set is used to obtain the biological process and molecular function FEPs using the algorithm described earlier. For a given test set of genes, if its gene expression profile (map) is significantly correlated with a given FEP of GO term GO_j, the gene will be identified as annotated with function GO_j. We consider a gene expression map as significantly correlated with a FEP if the correlation coefficient of the gene expression map and FEP is higher than 95% of the 10,000 randomly selected pairs of gene expression maps. For the K-NN approach, we set K to 1. For a given function GO_j and a given gene in the test set, we compute the correlation coefficients between the given gene and all genes in the training set and rank the correlations. The function label (0 or 1) of the gene with the highest correlation coefficient in the training set is used as the predicted label for the given gene in the test set.

For the FEPs obtained by the cluster-based method, a GO term might be connected to several FEPs. In this case, a given gene will be annotated with this GO term if its gene expression profile is significantly correlated with any one of these FEPs associated with the GO term. In the case that a number of GO terms are shown in one cluster, i.e., several GO terms are associated with one FEP, the same FEP (average gene expression profile of the cluster) is assigned to these GO terms.

3. RESULTS

3.1 Identifying functional expression profiles

The training set is used to identify functional expression profiles using the algorithm described earlier in Section 2.4.1. A sample set of gene expression profiles and its corresponding functional expression profile ("thyroid hormone generation; GO: 0006590") is shown in Figure 7. The two gene expression maps are much correlated to each other. The FEP preserves the characters of the expression maps fairly well.

For these experiments, we set the threshold r of the p-value of step 3 of the algorithm to 0.05, and set M to 10,000. The method identifies 48, 12 and 52 FEPs for biological processes, molecular functions, and cellular components respectively. These FEPs are visualized in Figures 8-10. Each small image in the figures denotes an FEP for a certain GO term. The GO ID of the GO term is given above each small image. The FEPs are sorted in descending order of the prediction accuracies of gene functions. The top 10 GO terms and their accuracies are presented later in Tables 2-4 for biological process, molecular function, and cellular component respectively.

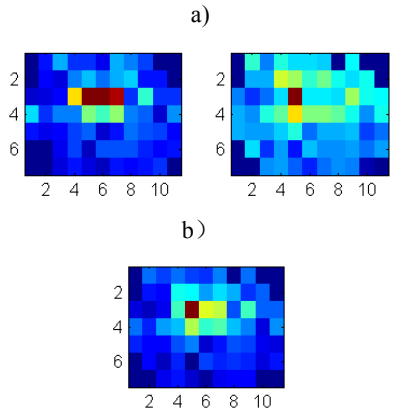

Figure 7 a) Gene Expression Profiles of genes with the function "thyroid hormone generation; GO: 0006590". b) Functional Expression Profile (FEP) of the function.

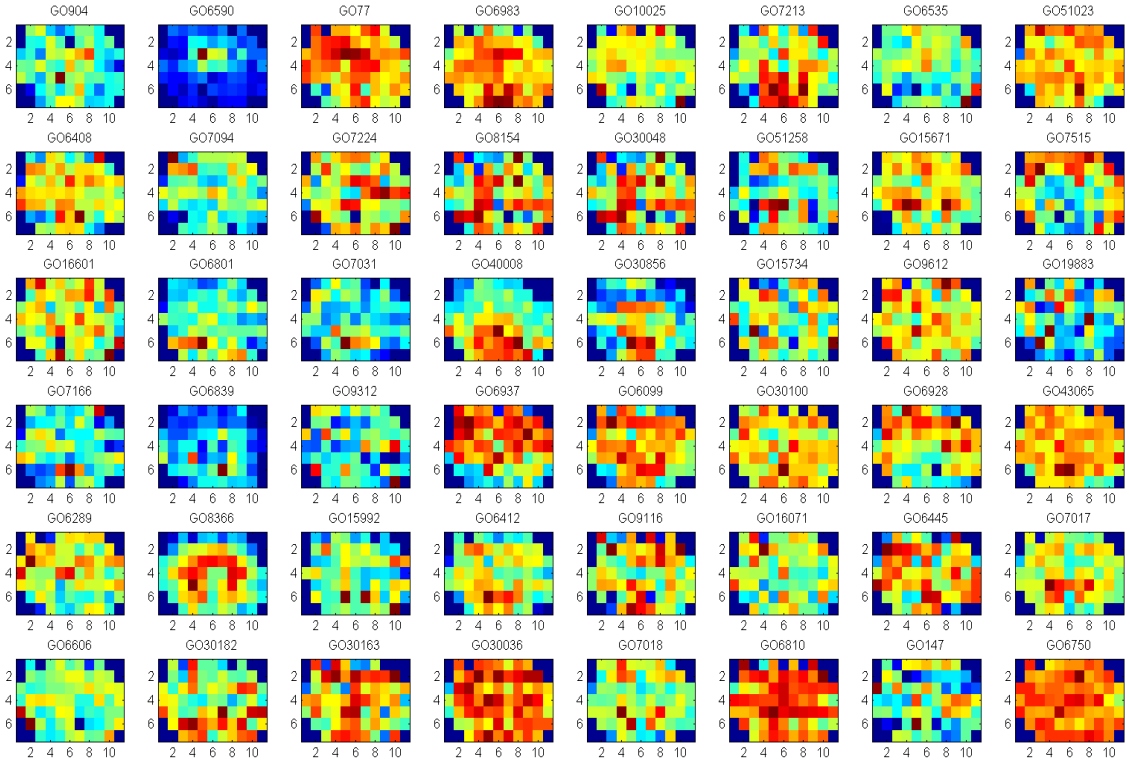

Figure 8 The 48 FEPs for biological processes

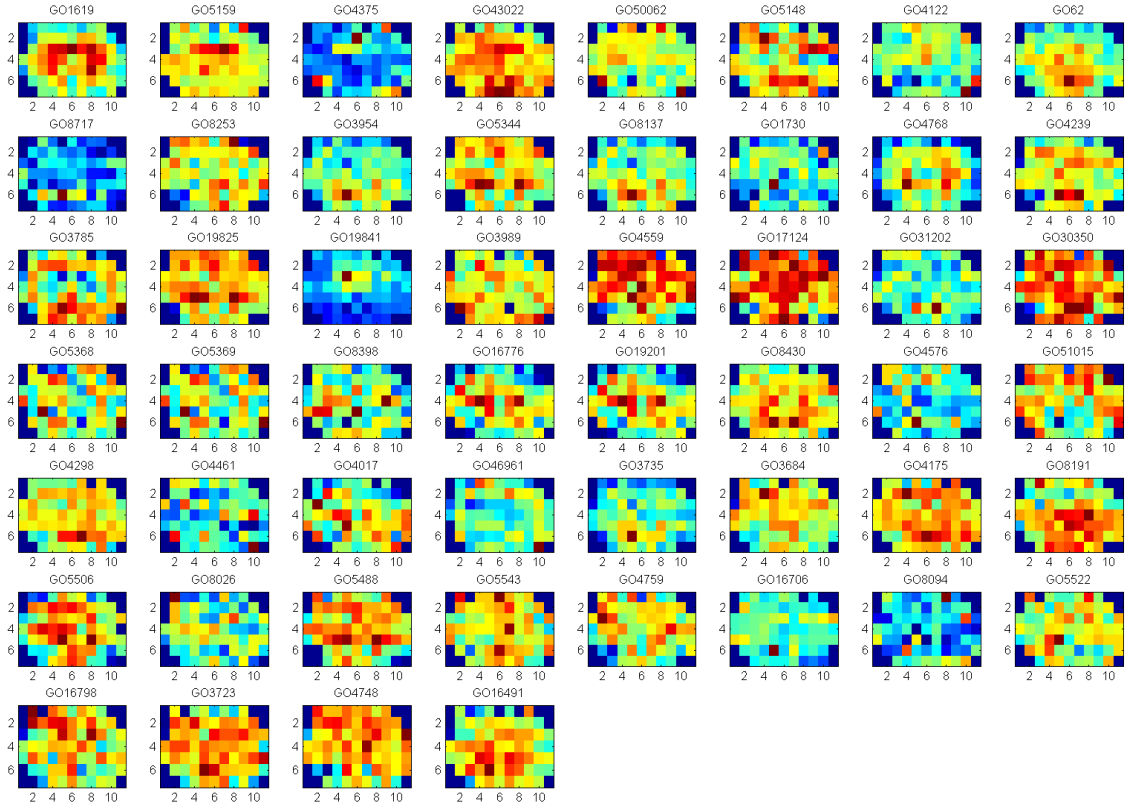

Figure 9 The 52 FEPs for molecular functions

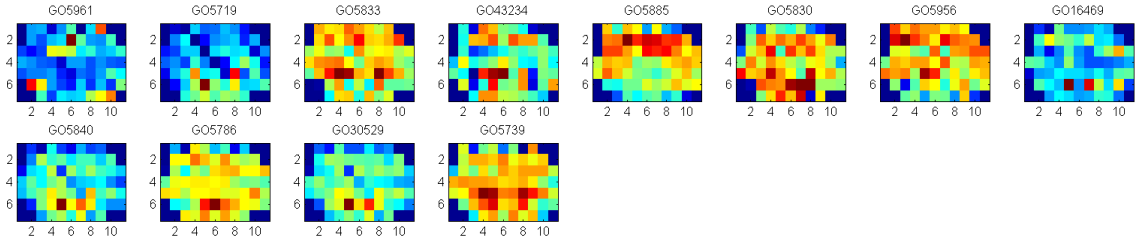

Figure 10 The 12 FEPs for cellular components

3.2 Function prediction using the KNN approach

The KNN approach results are summarized in Tables 2-4. We see that the KNN method fails to perform better than an arbitrary classification model (accuracy ~ 50%) for almost all functions. This is due to the extremely unbalanced data distribution for the given function annotation. Only a very small fraction of genes are annotated with the given functions. Although the specificities are very high, the sensitivities are almost close to 0 (Tables 2-4). As shown in Figure 11, over 90 percent of the functions are annotated with less than 10 genes out of 7883 genes. This makes it extremely hard for the KNN method to correctly identify gene functions based on its nearest neighbor.

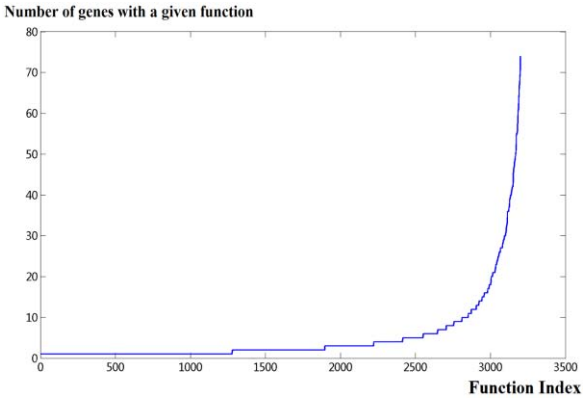

Figure 11 Number of genes annotated with a given function. The function index is sorted by the number of genes with the function

3.3 Function prediction using non-cluster-based FEPs

Using Functional Expression Profiles improves the accuracy for a large fraction of functions. We compare the results of the non-cluster-based FEP method to those obtained by the KNN method. Here the FEPs are obtained directly by finding genes with a given GO term. We show the top 10 results for biological processes, molecular functions and cellular components in Tables 2, 3 and 4 respectively. From these tables we see that the prediction accuracy is up to 99%, and a large number of functions show significant improvement compared to the traditional KNN method. Some of the functions such as "*snRNA export from nucleus*" have more than 40% improvement as shown in Table 2.

Table 2 Top 10 GO terms of Biological Process

Gene Ontology Term	Accuracy		Specificity		Sensitivity	
	Non-cluster-based FEP	KNN	Non-cluster-based FEP	KNN	Non-cluster-based FEP	KNN
cellular morphogenesis during differentiation	0.99	0.75	0.97	1	1	0.5
thyroid hormone generation	0.98	0.71	0.96	1	1	0.43
DNA damage checkpoint	0.97	0.69	0.94	1	1	0.39
ER overload response	0.97	0.63	0.94	1	1	0.25
wax biosynthetic process	0.96	0.58	0.93	1	1	0.16
acetylcholine receptor signaling, muscarinic pathway	0.95	0.57	0.91	1	1	0.13
cysteine biosynthetic process from serine	0.94	0.55	0.88	0.98	1	0.12
regulation of immunoglobulin secretion	0.93	0.5	0.85	1	1	0
snRNA export from nucleus	0.92	0.5	0.85	1	1	0
mitotic cell cycle spindle assembly checkpoint	0.87	0.5	0.94	1	0.8	0

Table 3 Top 10 GO terms of Molecular Function

Gene Ontology Term	Accuracy		Specificity		Sensitivity	
	Non-cluster-based FEP	KNN	Non-cluster-based FEP	KNN	Non-cluster-based FEP	KNN
lysosphingolipid and lysophosphatidic acid receptor activity	0.99	0.71	0.98	1	1	0.43
insulin-like growth factor receptor binding	0.98	0.69	0.97	1	1	0.39
glycine dehydrogenase (decarboxylating) activity	0.97	0.67	0.95	1	1	0.34
ribosome binding	0.97	0.67	0.94	1	1	0.34
long-chain-fatty-acyl-CoA reductase activity	0.96	0.63	0.93	1	1	0.25
prolactin receptor binding	0.96	0.55	0.91	1	1	0.12
cystathionine beta-synthase activity	0.94	0.5	0.88	1	1	0
acyl-CoA binding	0.93	0.5	0.86	1	1	0
inhibition of cell differentiation	0.93	0.5	0.86	1	1	0
5'-nucleotidase activity	0.92	0.5	0.85	0.98	1	0

Table 4 Top 10 GO terms of Cellular Component

Gene Ontology Term	Accuracy		Specificity		Sensitivity	
	Non-cluster-based FEP	KNN	Non-cluster-based FEP	KNN	Non-cluster-based FEP	KNN
glycine cleavage complex	0.97	0.75	0.95	1	1	0.5
nuclear euchromatin	0.97	0.69	0.94	1	1	0.38
hemoglobin complex	0.84	0.58	0.88	1	0.81	0.16
protein complex	0.83	0.58	0.99	0.99	0.67	0.16
Arp2/3 protein complex	0.79	0.56	0.86	0.95	0.71	0.16
A ribosome that is found in the cytosol of the cell	0.75	0.57	0.87	1	0.63	0.13
protein kinase CK2 complex	0.73	0.54	0.87	0.98	0.6	0.1
proton-transporting two-sector ATPase complex	0.71	0.5	0.93	1	0.5	0
ribosome	0.71	0.5	0.87	1	0.55	0
signal recognition particle, endoplasmic reticulum targeting	0.68	0.5	0.85	1	0.5	0

3.4 Function prediction using cluster-based FEPs

Here the FEPs are obtained from the average gene expression maps of the significant clusters with respect to biological process, molecular function and cellular component. For each significant cluster, we find out the most frequent GO terms, i.e., those terms for which the ratio of genes with the GO term to the size of the cluster is bigger than 0.2. So there might be several frequent GO terms for some clusters, but also there could be no frequent GO terms for a cluster. We find 17 frequent GO terms of the clusters with respect to biological process, 21 GO terms with respect to molecular function, and 24 GO terms with respect to cellular component. Table 5 shows the top 10 GO terms and their accuracies of gene function prediction based on the corresponding FEPs obtained from the clusters with respect to biological process. Tables 6 and 7 show the top 10 results with respect to the other two ontologies.

Table 5 Top 10 GO terms with respect to significant clusters of Biological Process by cluster-based FEPs

Gene Ontology Term	Accuracy	Specificity	Sensitivity
hemoglobin complex	0.83	0.81	0.85
oxygen transporter activity	0.83	0.81	0.85
oxygen binding	0.8	0.81	0.8
oxygen transport	0.8	0.81	0.79
Binding	0.62	0.81	0.44
Mitochondrion	0.53	0.81	0.26
Membrane	0.52	0.93	0.12
hydrolase activity	0.52	0.79	0.25
integral to membrane	0.52	0.47	0.57
Transport	0.52	0.63	0.41

Table 6 Top 10 GO terms with respect to significant clusters of Molecular Function by cluster-based FEPs

Gene Ontology Term	Accuracy	Specificity	Sensitivity
hemoglobin complex	0.8	0.84	0.77
oxygen transporter activity	0.8	0.84	0.77
oxygen binding	0.79	0.84	0.73
oxygen transport	0.78	0.84	0.72
binding	0.62	0.84	0.39
mitochondrion	0.55	0.82	0.28
transport	0.52	0.77	0.27
membrane	0.52	0.65	0.38
integral to membrane	0.51	0.3	0.72
nucleic acid binding	0.51	0.96	0.06

Table 7 Top 10 GO terms with respect to significant clusters of Cellular Component by cluster-based FEPs

Gene Ontology Term	Accuracy	Specificity	Sensitivity
hemoglobin complex	0.8	0.84	0.77
oxygen transporter activity	0.8	0.84	0.77
oxygen binding	0.79	0.84	0.73
oxygen transport	0.78	0.84	0.72
binding	0.61	0.84	0.38
aminoacyl-tRNA ligase activity	0.54	0.98	0.1
tRNA aminoacylation for protein translation	0.54	0.98	0.09
mitochondrion	0.53	0.84	0.23
integral to membrane	0.52	0.34	0.7
ligase activity	0.51	0.98	0.04

Although significant clusters are obtained by considering different categories of gene ontology [11], the frequent GO terms are not restricted to the categories. So, there are GO terms that are common in the three tables. The prediction accuracy reaches up to 83%, which is lower than the accuracy of the prediction obtained by non-cluster-based FEP. Also observe that although some functions have very low accuracy, they have very high specificity, such as „*membrane'* in Table 5, „*nucleic acid binding'* in Table 6, and „*ligase activity'* in Table 7.

3.5 Comparing cluster-based FEP method and non-cluster-based FEP method

Here we compare the cluster-based FEP method with the non-cluster-based FEP method besides the prediction accuracy. Table 8 shows the differences of the two methods. Although the cluster-based FEP method has better prediction accuracy than non-cluster-based FEP method, the former one makes use of a larger number of genes to build FEPs than the later one, and more genes are correctly annotated with at least a gene function than those annotated by non-cluster-based FEP.

Table 8 Comparing cluster-based FEP method and non-cluster-based FEP method

Gene Ontology		Biological Process	Molecular Function	Cellular Component
Number of genes used	*Cluster-based FEPs*	5414	5891	5586
	Non-cluster-based FEPs	1182	872	571
Number of genes correctly annotated	*Cluster-based FEP*	3234	4009	3560
	Non-cluster-based FEP	271	273	162

Non-cluster-based method detects 48, 52, and 12 FEPs respectively to the categories of gene ontology, whereas cluster-based method detects 17, 21 and 24 respectively to the same categories. The cluster-based method finds out fewer FEPs because over 90 percent of the functions are annotated with less than 10 genes out of 7883 genes, so that a very small number of functions are found to be frequent in a significant cluster.

There are few intersections between the FEPs (or gene functions) obtained by the two methods. Table 9 shows the intersections and their prediction accuracies. For the same gene functions, KNN has the worst prediction accuracy and non-cluster-based

FEP method is slightly better than the cluster-based FEP method. The small number of intersections indicates that the two methods can detect different kinds of gene functions.

Table 9 Prediction accuracies of FEP Intersections between non-cluster-based and cluster-based methods

Intersection		KNN	Non-cluster-based FEP	Cluster-based FEP
Biological Process	*Transport*	0.5	0.51	0.52
	oxygen transport	0.5	0.82	0.8
Molecular Function	*oxygen transporter activity*	0.5	0.84	0.8
	Binding	0.5	0.61	0.6
	oxygen binding	0.5	0.81	0.79
Cellular Component	*mitochondrion*	0.5	0.55	0.53
	hemoglobin complex	0.58	0.84	0.8

4. DISCUSSION

In this study, we focus on the analysis of gene expression maps obtained by voxelation and processed with microarrays. Information about the genes that are being expressed in each one of the voxels which are spatially registered in the mice brain is being collected and analyzed. Gene expression profiles are extracted by wavelet transformation on the averaged hemisphere of the mice brain, taking into account the correlation between neighboring voxels. This work tries to improve upon the results obtained by related research work based on gene expression profiles that do not consider spatial information.

Our study is based on the hypothesis that genes with similar gene expression maps may have similar gene functions. This hypothesis was confirmed for a number of genes by our previous analysis [11]. The hypothesis might not hold for all genes, but we showed that it holds for at least a set of genes. Therefore, a gene function might be related to a certain gene expression map. In this paper, we examine the gene functions associated with at least two genes to see if these genes have similar gene expression profiles. By ranking the pair-wise correlation coefficient for all the expressions profiles of genes associated to a given function, a significant expression profile is reported. We call this specific expression profile which is associated with the given gene function, Functional Expression Profile (FEP). FEPs are used here to annotate genes with functions by comparing the unknown gene"s profile with all the identified FEPs. The function whose FEP is significant similar to the gene"s profile is assigned to the unknown gene. Another method we propose for getting FEPs is identifying significant clusters of genes which have both similar gene functions and gene expression profiles. We denote the average profile of a cluster as FEP and we assign frequent functions observed in the cluster to this FEP.

Experimental results show that our proposed function annotation approach which uses non-cluster-based FEPs can reach accuracy of up to 99 percent. The cluster-based FEP approach can reach up to 84 percent accuracy and has high specificity and low sensitivity. The K-NN method fails to achieve better than 50 percent accuracy which may be due to the extremely unbalanced data distribution for a given function annotation. The reason for the reduced performance of the cluster-based FEP compared to the non-cluster-based FEP might be that the significant clusters do not include all the genes of the dataset (see Table 1). This is related to the fact that cluster-based FEP has high specificity and low sensitivity. Also it can be attributed to the fact that FEPs obtained by the cluster-based method have much more blurred images (see Figure 6) than those obtained by the non-cluster-based approach (see Figures 8-10). So, the non-cluster-based FEPs have better quality.

We further compare the cluster-based FEP and the non-cluster-based FEP methods by examining the number of genes included in FEPs, the number of genes correctly annotated with at least a gene function, and the number of FEPs detected. The cluster-based method identifies less FEPs, but it has many more genes involved in FEPs and more genes correctly annotated by at least one function. This means that the functions associated with FEPs obtained by the two methods are quite different. The functions identified by the non-cluster-based method are specific and infrequent in our dataset, whereas the functions identified by the cluster-based method are common and frequently appearing in genes. Therefore, although the non-cluster-based method has better prediction accuracy than the cluster-based method, the cluster-based one remains a useful approach and it is more advanced in some cases, such as the case where we need to study common gene functions.

Another way, we can examine the results is by connecting gene expression images of FEPs with their prediction accuracies. For example, let us consider the "thyroid hormone generation" which has a prediction accuracy of 98 percent (Table 2). The FEP of this function is visualized in the second small image on the top row of Figure 8. This image is very interesting because it shows very high expression in a single voxel. Some of the images of FEPs are in good agreement with anatomical components of mice"s brain. We believe that these findings can provide meaningful information to biologists.

So far, the FEPs we found are based on a single gene function. This means that one FEP is associated with one function. A question to explore is what will happen if we take into account two or three functions together, such as the frequent itemsets of functions that have been identified in [17]. In future work, we will consider the problem of detecting FEPs for a set of certain gene functions.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under grants IIS-0705215, IIS-0237921 and by the National Institutes of Health under grant R01 MH68066-05 funded by NIMH, NINDS and NIA, and NINDS grant 1 R01 NS050148. The funding agencies specifically disclaim responsibility for any analyses, interpretations and conclusions.

6. REFERENCES

[1] Yingyao Zhou, Jason A. Young, Andrey Santrosyan, Kaisheng Chen, S. Frank Yan and Elizabeth A. Winzeler. *In silico gene function prediction using ontology-based pattern identification*. Bioinformatics 2005; 21(7):1237-1245.

[2] Jason Li, Saman K Halgamuge, Christopher I Kells, and Sen-Lin Tang. *Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages*. BMC Bioinformatics 2007 May 22;8 Suppl 4:S6.

[3] Xianghong Jasmine Zhou, Ming-Chih J Kao, Haiyan Huang, Angela Wong, Juan Nunez-Iglesias, Michael Primig, Oscar M Aparicio, Caleb E Finch, Todd E Morgan, and Wing Hung Wong. *Functional annotation and network reconstruction through cross-platform integration of microarray data*. Nature Biotechnology 23, 238 - 243, 2005.

[4] Mohammed Kashani-Sabet, Yong Liu, Sylvia Fong, Pierre-Yves Desprez, Shuqing Liu, Guanghuan Tu, Mehdi Nosrati, Chakkrapong Handumrongkul, Denny Liggitt, Ann D. Thor, and Robert J. Debs. *Identification of gene function and functional pathways by systemic plasmid-based ribozyme targeting in adult mice*. PNAS March 19, 2002 vol. 99 no. 6, p. 3878-3883

[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] Vanessa M. Brown, Alex Ossadtchi, Arshad H. KHAN, Sanjiv S. Gambhir, Simon R. Cherry, Richard M. Leahy and and Desmond J. Smith. *Gene expression tomography*, Physiol Genomics 8, 2002, p. 159-167.

[7] Vanessa M. Brown, Alex Ossadtchi, Arshad H. Khan, Simon R. Cherry, Richard M. Leahy, and Desmond J. Smith. *High-throughput imaging of brain gene expression*. Genome Res, 2002. 12(2): p. 244-54.

[8] 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.

[9] 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.

[10] Mark H. Chin, Alex B. Geng, Arshad H. Khan, Wei-Jun Qian, Vladislav A. Petyuk, Jyl Boline, Shawn Levy, Arthur W. Toga, Richard. Smith, Richard M. Leahy, and Desmond J. Smith. *A genome-scale map of expression for a mouse brain section obtained using Voxelation*. Physiol. Genomics 30: p. 313-321. 2007.

[11] Li An, Hongbo Xie, Mark H. Chin, Zoran Obradovic, Desmond J. Smith, and Vasileios Megalooikonomou. *Analysis of multiplex gene expression maps obtained by voxelation*, BMC Bioinformatics. 2009; 10(Suppl 4): S10.

[12] Maximilian Diehn, Gavin Sherlock, Gail Binkley, Heng Jin, John C. Matese, Tina Hernandez-Boussard, Christian A. Rees, J. Michael Cherry, David Botstein, Patrick O. Brown and Ash A. Alizadeh. *SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data*. Nucleic Acids Res. 2003 Jan 1;31(1):219-23.

[13] *Stanford Genomic Resources*, http://genome-www.stanford.edu/.

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

[15] Zizhen Yaocorresponding and Walter L Ruzzo. *A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data*, BMC, Bioinformatics, 2006, 7 (Suppl 1), S11.

[16] Saket Kharsikar, Dale Mugler, Daniel Sheffer, Francisco Moore and Zhong-Hui Duan. *A Weighted k-Nearest Neighbor Method for Gene Ontology Based Protein Function Prediction*. Proceedings of the Second International Multi-Symposiums on Computer and Computational Sciences table of contents, p. 25-31, 2007.

[17] Li An, Zoran Obradovic, Desmond Smith, Olivier Bodenreider, and Vasileios Megalooikonomou. *Mining Association Rules among Gene Functions in Clusters of Similar Gene Expression Maps*. Workshop Proceedings of 2009 IEEE International Conference on Bioinformatics and Biomedicine, p. 254 – 259.