A New Sparse Simplex Model for Brain Anatomical and Genetic Network Analysis

Heng Huang[1], Jingwen Yan[2], Feiping Nie[1], Jin Huang[1], Weidong Cai[3], Andrew J. Saykin[2], and Li Shen[2⋆]

[1] Computer Science and Engineering, University of Texas at Arlington, TX, USA
[2] Radiology and Imaging Sciences, Indiana University School of Medicine, IN, USA
[3] BMIT Research Group, School of IT, University of Sydney, Australia

Abstract. The Allen Brain Atlas (ABA) database provides comprehensive 3D atlas of gene expression in the adult mouse brain for studying the spatial expression patterns in the mammalian central nervous system. It is computationally challenging to construct the accurate anatomical and genetic networks using the ABA 4D data. In this paper, we propose a novel sparse simplex model to accurately construct the brain anatomical and genetic networks, which are important to reveal the brain spatial expression patterns. Our new approach addresses the shift-invariant and parameter tuning problems, which are notorious in the existing network analysis methods, such that the proposed model is more suitable for solving practical biomedical problems. We validate our new model using the 4D ABA data, and the network construction results show the superior performance of the proposed sparse simplex model.

1 Introduction

In recent research, the large-scale screenings for gene expression profiles across all different brain regions have been done by the Allen Institute for Brain Science, known as Allen Brain Atlas (ABA) project [1]. ABA provides spatially mapped large-scale gene expression database and enables quantitative comparison of data measurements across genes, anatomy, and phenotype. Detection of gene-anatomy association in brain structure is crucial for understanding brain function based on the molecular and genetic/genomic information. Particularly in the mouse or human brain where there are over thousands of genes expressed, systematic and comprehensive quantification of the expression densities in the whole three-dimensional (3D) anatomical context is critical.

The ABA database provides cellular resolution 3D expression patterns for both mouse and human (ongoing project). The image data are generated by *in situ* hybridization using gene-specific probes, followed by slide scanning and

⋆ HH, FN, and JH were supported by CCF-0917274, DMS-0915228, IIS-1117965. JY, AS, and LS were supported in part by NSF IIS-1117335, NIH UL1 RR025761, U01 AG024904, NIA RC2 AG036535, NIA R01 AG19771, NIH R01 LM011360, and NIA P30 AG1013318S1.

3D image registration to the Allen Reference Atlas (ARA) [2] and expression segmentation [3]. The resulted mouse brain 4D expression data are in a set of spatially aligned 3D volumes of size $67 \times 41 \times 58$. The genes' values expressed in each voxel of the mouse brain are recorded.

The ABA contains information about the spatial distribution of genes within the human and mouse brain. Efficient and effective analysis of these high throughput data can shed light on the global function of mammalian central nervous system [4] and provide important information for understanding the connections of human brain anatomy, genome, and transcriptome. However, most previous research works are limited to retrieve correlation values between the spatial patterns of genes [5], or cluster the brain regions into co-expressed groups [6].

Network analysis provides a productive approach to analyze the high throughput biomedical and biological data. Transforming the data into a network framework offers distinct advantages for directly relating specific biomedical and biological interactions or outcome states with the network properties and dynamics. Thus, it is desired to model and analyze the spatial gene expression data of human brain in ABA in network format. Existing approaches to construct biomedical and biological networks usually have three deficiencies: (1) shift variant, *i.e.* when the data are shifted with a value, the network construction result will be totally different; (2) tedious parameter tuning is needed and not suitable for the practical applications; (3) the network edge weight has no probability interpretation to help the analysis. In this paper, to tackle these problems, we propose a novel sparse simplex learning model and applied it to ABA mouse brain data to create both anatomical and transcriptomic networks, which provide important insights into the global structure of the anatomy and transcriptome.

2 Related Work

The ABA brain microarray data provide the great opportunity to model the neuroanatomical and transcriptomic networks, where each vertex represents a spatial location or a gene and the edges between vertices encode the correlations between locations and genes. In recent related studies, the weighted gene co-expression network analysis (WGCNA) [7] based computational tools were mainly used to construct the co-expression network. More recently, Ji [8] used an approximate formulation for Gaussian graphical modeling [9] to model the mouse brain networks and showed more efficient and stable construction results. Given the input data $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$, this approximation model calculates the edge weight matrix $W \in \Re^{n \times n}$ (all values on the diagonal are "0"s) by solving a series of sparsity regularized regression problems. In this paper, we write matrices as capital letters and vectors as boldface lowercase letters. Given a matrix $W = [w_{ij}]$, its i-th row and j-th column are denoted as $\mathbf{w}^i$ and $\mathbf{w}_j$, respectively.

In [8], the weights of edges from vertex $\mathbf{x}_i$'s neighboring vertices to $\mathbf{x}_i$ are learned by solving the standard sparse representation problem:

$$\min_{\boldsymbol{\alpha}_i} ||\mathbf{x}_i - X_{-i}\boldsymbol{\alpha}_i||^2 + \lambda||\boldsymbol{\alpha}_i||_1, \tag{1}$$

where $X_{-i} = [\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_n]$ is the data matrix obtained from X by removing the i-th data point, $\boldsymbol{\alpha}_i \in \Re^{(n-1)\times 1}$ is a weight vector from $\mathbf{w}_i$ by removing the i-th weight, which is zero. The network links are constructed by applying the thresholding value 0.5 to the edge weights. The above model constructs the network/graph using Lasso for variables selection [9]. However, this approach has three key deficiencies: 1) This method is not shift invariant, *i.e.*, if data are shifted with an arbitrary value, such as $\mathbf{x}_i = \mathbf{x}_i + t\mathbf{1}$, the network construction result will be totally different. 2) The parameter λ has to be tuned to get good results. Although [8] provided a strategy to learn this parameter, the strategy also depends on the link thresholding value. Thus, the network construction results are not robust as expected. 3) The edge weights cannot be interpreted as probabilities. To solve these deficiencies, we propose a new sparse simplex learning model to construct brain networks with non-parameter tuning, shift invariant, and probability interpretation advantages.

3 Methodology

3.1 Sparse Simplex Learning Model

Sparse learning models have been actively applied to solve problems in computational neuroscience [10–14]. To effectively construct the brain networks, the sparse representation model can be utilized as in [8]. When we build the neuroanatomical and transcriptomic networks, we hope the edge weight has the probability meaning, which can directly tell us the link strength between two nodes. Thus, we add two constraints on the sparse representation model: $\boldsymbol{\alpha}_i \geq 0$ and $\boldsymbol{\alpha}_i^T \mathbf{1} = 1$, where $\mathbf{1} \in \Re^{(n-1)\times 1}$ is a vector with all "1" as elements. The new objective will solve:

$$\min_{\boldsymbol{\alpha}_i} \|\mathbf{X}_{-i}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 , \quad s.t. \;\; \boldsymbol{\alpha}_i \geq 0, \; \boldsymbol{\alpha}_i^T \mathbf{1} = 1 . \tag{2}$$

After imposing these two constraints, the solutions $\boldsymbol{\alpha}_i$ will have the probability interpretations. The $\boldsymbol{\alpha}_i(j)$ is the edge weight between nodes i and j. Because $\sum_j \boldsymbol{\alpha}_i(j) = \boldsymbol{\alpha}_i^T \mathbf{1} = 1$, $\boldsymbol{\alpha}_i(j)$ can be interpreted as the probability to have an edge between nodes i and j.

In the network construction, we hope the learning model is shift-invariant, such that the network constructions have small changes when the data have an arbitrary shift value. The shift-invariant property is important for practical biomedical applications, because the data collection processes are often effected by instruments and environment factors and the collected data may include a shifted value caused by these factors. Fortunately, after imposing the above two constraints, the new sparse learning model becomes shift-invariant.

When the data are shifted by a constant t, *i.e.*, $\mathbf{x}_k = \mathbf{x}_k + t\mathbf{1}$ for all $k = 1, \cdots, n$, the computed similarities between the pairs of nodes will be changed. The objective function becomes:

$$\left\|(\mathbf{X}_{-i} + t\mathbf{1}\mathbf{1}^T)\boldsymbol{\alpha}_i - (\mathbf{x}_i + t\mathbf{1})\right\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 . \tag{3}$$

Because $\boldsymbol{\alpha}_i^T \mathbf{1} = 1$, the above objective can be written as:

$$\|\mathbf{X}_{-i}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 \,, \tag{4}$$

which is the original one. Thus, the new objective in (2) is shift-invariant.

More important, the constraints in problem (2) make the second regularization term as a constant. Thus, the problem (2) becomes

$$\min_{\boldsymbol{\alpha}_i} \|\mathbf{X}_{-i}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 \,, \quad s.t.\ \ \boldsymbol{\alpha}_i \geq 0,\ \boldsymbol{\alpha}_i^T \mathbf{1} = 1 \,. \tag{5}$$

Thus, the new model has no parameter, such that it is suitable for biomedical and biological applications, in which we usually lack information/data to tune the parameter.

Because the constraints in problem (5) are the simplex formulation, we call the new method as the sparse simplex learning model. Note that the above constraints (ℓ_1 ball constraints) indeed introduce sparse solution α_i.

The ABA 4D data are large-scale with high-dimensionality. Thus, we need to derive the efficient optimization algorithm to solve the new objective in Eq. (5). It is more appropriate to apply the first-order methods, *i.e.*, use function values and their (sub)gradient at each iteration. There are many first-order methods, including gradient descent, subgradient descent, and Nesterov's optimal method [16]. In this paper, we use the accelerated projected gradient method to optimize Eq. (5).

3.2 Optimization Algorithm

When we use the accelerated projected gradient method to solve this problem, the critical step of the projected gradient method is to solve the following proximal problem:

$$\min_{\boldsymbol{\alpha}_\mathbf{i}} \frac{1}{2} \|\boldsymbol{\alpha}_i - \mathbf{v}\|_2^2 \,, \quad s.t\ \ \boldsymbol{\alpha}_\mathbf{i} \geq 0, \boldsymbol{\alpha}^T \mathbf{1} = 1 \,. \tag{6}$$

We write the Lagrangian function of problem (6) as:

$$\frac{1}{2}\|\boldsymbol{\alpha}_i - \mathbf{v}\|_2^2 - \gamma(\boldsymbol{\alpha}_i^T \mathbf{1} - 1) - \boldsymbol{\lambda}^T \boldsymbol{\alpha}_i, \tag{7}$$

where γ is a Lagrangian multiplier and $\boldsymbol{\lambda}$ is a Lagrangian multiplier vector, both of which are to be determined. Suppose the optimal solution to the proximal problem (6) is $\boldsymbol{\alpha}^*$, the associate Lagrangian coefficients are γ^* and λ^*. Then according to the KKT condition [17], we have the following equations:

$$\begin{cases} \forall j, \quad \alpha_{ij}^* - v_j - \gamma^* - \lambda_j^* = 0 & (8) \\ \forall j, \quad \alpha_{ij}^* \geq 0 & (9) \\ \forall j, \quad \lambda_j^* \geq 0 & (10) \\ \forall j, \quad \alpha_{ij}^* \lambda_j^* = 0 & (11) \end{cases}$$

where α_{ij}^* is the j-th scalar element of vector $\boldsymbol{\alpha}_i^*$. Eq. (8) can be written as $\alpha_{ij}^* - v_j - \gamma^* 1 - \lambda_j^* = 0$. According to the constraint $\mathbf{1}^T \boldsymbol{\alpha}_i^* = 1$, we have $\gamma^* = \frac{1 - \mathbf{1}^T \mathbf{v} - \mathbf{1}^T \boldsymbol{\lambda}^*}{n}$. Thus, $\boldsymbol{\alpha}^* = (\mathbf{v} - \frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{v} + \frac{1}{n}\mathbf{1} - \frac{\mathbf{1}^T \lambda^*}{n}\mathbf{1}) + \lambda^*$.

Denoting $\bar{\lambda}^* = \frac{\mathbf{1}^T \boldsymbol{\lambda}^*}{n}$ and $\mathbf{u} = \mathbf{v} - \frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{v} + \frac{1}{n}\mathbf{1}$, we can write $\boldsymbol{\alpha}^* = \mathbf{u} + \boldsymbol{\lambda}^* - \bar{\lambda}^* \mathbf{1}$. Thus, $\forall j$ we have:

$$\alpha_{ij}^* = u_j + \lambda_j^* - \bar{\lambda}^*. \tag{12}$$

According to Eqs. (9)-(12) we know $u_j + \lambda_j^* - \bar{\lambda}^* = (u_j - \bar{\lambda}^*)_+$, here $x_+ = max(x, 0)$. Then we have

$$\alpha_j^* = (u_j - \bar{\lambda}^*)_+. \tag{13}$$

Therefore, we can obtain the optimal solution $\boldsymbol{\alpha}^*$ if we know $\bar{\boldsymbol{\lambda}}^*$.

We write Eq. (12) as $\lambda_j^* = \alpha_{ij}^* + \bar{\lambda}^* - u_j$. Similarly, according to Eqs.(9)-(11), we know $\lambda_j^* = (\bar{\lambda}^* - u_j)_+$. Since $\mathbf{v}$ is a $n-1$-dimensional vector, we have $\bar{\boldsymbol{\lambda}}^* = \frac{1}{n-1} \sum_{j=1}^{n-1} (\bar{\lambda}^* - u_j)_+$. Defining a function as

$$f(\bar{\lambda}) = \frac{1}{n-1} \sum_{i=1}^{n-1} (\bar{\lambda} - u_j)_+ - \bar{\lambda}, \tag{14}$$

such that $f(\bar{\lambda}^*) = 0$ and we can solve the root finding problem with Newton method to obtain $\bar{\lambda}^*$.

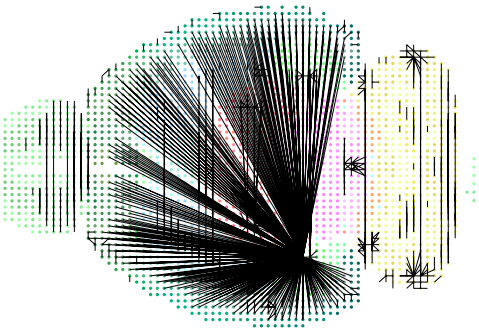
The convergence rate of our algorithm is $O(\frac{1}{t^2})$, where t is the number of iterations. The detailed proof can be found at [15, 16].

4 Experiments and Discussions

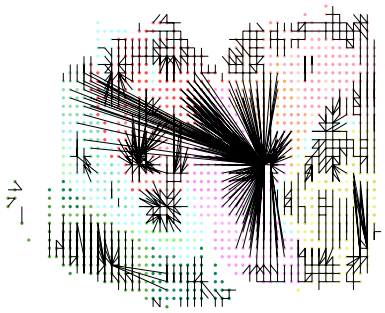
4.1 Experimental Results on ABA Data

In our experiment, we use the ABA cellular resolution 3D expression patterns in the male, 56-day-old C57BL mouse brain. The 4D spatial gene data are a 4D tensor $2980 \times 67 \times 41 \times 58$, in which the first index corresponds to genes, and the other three indices represent the rostral-caudal, dorsal-ventral and left-right spatial directions, respectively. The newest database provides 2980 genes which are slightly different to the data used in [8].

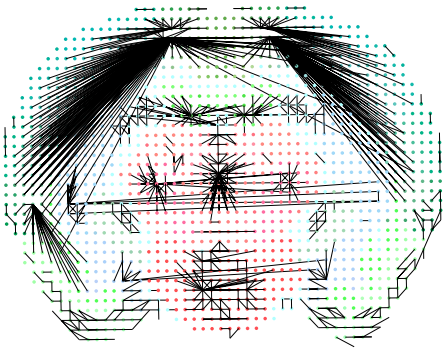
When we create the genetic network, each node is one gene of 2980 genes. In [8], the tensor factorization method was used to reduce the dimensionality. However, this is an improper process. Although the voxels on the boundary of brain have no gene values, the tensor factorization includes the values of these voxels into calculation. We used the PCA method to reduce the dimensionality. Although the number of voxels is very large, the number of genes is not large. The PCA calculation is still affordable. For each gene, we reduce its 3D tensor data to $25 \times 15 \times 20$ and then concatenate its all values into a feature vector. The resulted 7500×2980 data matrix is used as input of sparse simplex learning model to construct the genetic network.



(a) Slice 20 on rostral-caudal direction



(b) Slice 29 on left-right direction



(c) Slice 33 on rostral-caudal direction

Fig. 1. We select and visualize the center slices of three directions of the 3D neuroanatomical network. (a) The 20-th slice on the dorsal-ventral direction. (b) The 29-th slice on the left-right direction. (c) The 33-rd slice on the rostral-caudal direction. The region with the largest number of connections corresponds to the brain structure dentate gyrus.

When we build the anatomical network, we directly use the $2980 \times 67 \times 41 \times 58$ tensor data. We have total $67 \times 41 \times 58$ nodes in brain spatial structure and the gene values are features. The sparse simplex model is performed to construct the neuroanatomical network. Because the network is 3D, we cannot visualize the whole 3D network. Thus, in Figure 1, we select three slices of brain data (center slices on three directions) and plot the networks on them. Figure 1(a) shows the 20-th slice on the dorsal-ventral direction. Figure 1(b) plots the 29-th slice on the left-right direction. Figure 1(c) visualizes the 33-rd slice on the rostral-caudal direction. The region with the largest number of connections corresponds to the brain structure dentate gyrus. We don't plot the genetic network here, because there is no spatial structure in genes. The genetic network cannot show meaningful visualization.

4.2 Model Evaluation Using Clustering Tasks

In above experimental results, we showed that the proposed sparse simplex model can efficiently construct both genetic and neuroanatomical networks. Because there is no ground-truth results for network constructions, we cannot directly compare the performance of our sparse simplex method. Thus, we use the clustering task results to compare our sparse simplex model to other graph construction methods. We use the sparse representation method [8] to construct graph and then perform Normalized Cut (NCut) and Self-Tuning Spectral Clustering (STSC) methods. After that, we build the graph using the proposed model and then perform the Normalized Cut (SSM+NCut). The clustering accuracy on six public computer vision benchmark image datasets are shown in Table 1. We also show the clustering results of K-means and NMF as baseline results. Although these data are not biomedical image data, we only use them for validation purpose because they have ground truth labels. In all results, our new sparse simplex model shows the promising graph/network construction results.

Datasets	K-means	NMF	NCut	STSC	SSM+NCut
AR	0.133	0.143	0.158	0.130	**0.324**
AT&T	0.664	0.678	0.698	0.685	**0.763**
JAFFE	0.789	0.774	0.795	0.813	**0.902**
MNIST	0.641	0.636	0.647	0.693	**0.796**
PIE	0.229	0.241	0.234	0.186	**0.325**
UMIST	0.475	0.457	0.443	0.394	**0.514**

Table 1. Clustering accuracy using different graph construction methods.

5 Conclusion

In this paper, we propose a novel sparse simplex learning model to construct the genetic and neuroanatomical networks using ABA 4D spatial gene patterns. Compared to the existing methods, the new model has three advantages: (1) it is shift-invariant such that the noise in data collection won't dramatically effect the network construction; (2) it doesn't require the parameter tuning, thus it is suitable for practical biomedical and biological applications; (3) it has probability interpretations on the resulted network weights, which can help the further data analysis. We validate the proposed model using the ABA mouse brain data and construct both genetic and anatomical networks. Our new model can also be applied to other biomedical network construction and analysis problems.

References

1. Lein, E.S.: Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168–176 (2007)

2. Dong, H.W.: The Allen Reference Atlas: A Digital Color Brain Atlas of the C57BL/6J Male Mouse (2009)
3. Ng, L., Pathak, S.D., Kuan, C., Lau, C., Dong, H., Sodt, A., Dang, C., Avants, B., Yushkevich, P., Gee, J.C., Haynor, D., Lein, E., Jones, A., Hawrylycz, M.: Neuroinformatics for genome-wide 3-D gene expression mapping in the mouse brain. IEEE/ACM Trans. Comput. Biol. Bioinformatics 4, 382–393 (2007)
4. Jones, A.R., Overly, C.C., Sunkin, S.M.: The Allen Brain Atlas: 5 years and beyond. Nat. Rev. Neurosci. 10, 821–828 (2009)
5. Ng, L., et al.: An anatomic gene expression atlas of the adult mouse brain. Nat Neurosci 12, 356–362 (2009)
6. Bohland, J.W., Bokil, H., Pathak, S.D., Lee, C.K., Ng, L., Lau, C., Kuan, C., Hawrylycz, M., Mitra, P.P.: Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods 50, 105–112 (2010)
7. Zhang, B., Horvath, S.: A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology 4(1), 17 (2005)
8. Ji, S.W.: Computational network analysis of the anatomical and genetic organizations in the mouse brain. Bioinformatics 27(23), 3293–3299 (2011)
9. Meinshausen, N., Buhlmann, P.: High-dimensional graphs and variable selection with the Lasso. The Annals of Statistics 34(3), 1436–1462 (2006)
10. Wang, H., Nie, F.P., Huang, H., Risacher, S., Saykin, A.J., and Shen, L.: Identifying adsensitive and cognition-relevant imaging biomarkers via joint classification and regression. In: MICCAI, pp. 115–123 (2011)
11. Wang, H., Nie, F.P., Huang, H., Kim, S., Nho, K., Risacher, S., Saykin, A.J., Shen, L.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. Bioinformatics 28(2), 229–237 (2012)
12. Wang, H., Nie, F.P., Huang, H., Risacher, S., Saykin, A.J., and Shen, L.: Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. Bioinformatics 28(12), i127–i136 (2012)
13. Wang, H., Nie, F.P., Huang, H., Yan, J., Kim, S., Nho, K., Risacher, S., Saykin, A.J., Shen, L.: From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant SNPs. Bioinformatics 28(18), i619–i625 (2012)
14. Wang, H., Nie, F.P., Huang, H., Yan, J., Kim, S., Risacher, S., Saykin, A.J., Shen, L.: High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In: NIPS, pp. 1286–1294 (2012)
15. Nesterov, Y.: Method for solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math Dokl 1983(2), 372–376 (1983)
16. Nesterov, Y.: Gradient methods for minimizing composite objective function (2007)
17. Boyd, S., Vandenberghe, L.: Convex Optimization, Cambridge University Press (2004)