



Bioimage informatics

Identify connectome between genotypes and brain network phenotypes via deep self-reconstruction sparse canonical correlation analysis

Meiling Wang ^{1,2}, Wei Shao ^{1,2}, Xiaoke Hao³, Shuo Huang^{1,2} and Daoqiang Zhang^{1,2,*}

¹College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, ²MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China and ³School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China

*To whom correspondence should be addressed.

Associate Editor: Olga Vitek

Received on November 10, 2021; revised on January 21, 2022; editorial decision on January 30, 2022; accepted on February 2, 2022

Abstract

Motivation: As a rising research topic, brain imaging genetics aims to investigate the potential genetic architecture of both brain structure and function. It should be noted that in the brain, not all variations are deservedly caused by genetic effect, and it is generally unknown which imaging phenotypes are promising for genetic analysis.

Results: In this work, genetic variants (i.e. the single nucleotide polymorphism, SNP) can be correlated with brain networks (i.e. quantitative trait, QT), so that the connectome (including the brain regions and connectivity features) of functional brain networks from the functional magnetic resonance imaging data is identified. Specifically, a connection matrix is firstly constructed, whose upper triangle elements are selected to be connectivity features. Then, the PageRank algorithm is exploited for estimating the importance of different brain regions as the brain region features. Finally, a deep self-reconstruction sparse canonical correlation analysis (DS-SCCA) method is developed for the identification of genetic associations with functional connectivity phenotypic markers. This approach is a regularized, deep extension, scalable multi-SNP-multi-QT method, which is well-suited for applying imaging genetic association analysis to the Alzheimer's Disease Neuroimaging Initiative datasets. It is further optimized by adopting a parametric approach, augmented Lagrange and stochastic gradient descent. Extensive experiments are provided to validate that the DS-SCCA approach realizes strong associations and discovers functional connectivity and brain region phenotypic biomarkers to guide disease interpretation.

Availability and implementation: The Matlab code is available at <https://github.com/meimeiling/DS-SCCA/tree/main>.

Contact: dqzhang@nuaa.edu.cn

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Among brain diseases appearing in the world, Alzheimer's disease (AD) lies in a high incidence level. Such disease seriously damages human behaviors, particularly in memory decline, which further injures the abilities of language, visuospatial perception, arithmetic abilities and executive functioning. It is very valuable to early diagnose and prevent AD (Brookmeyer *et al.*, 2007; Lambert *et al.*, 2013; Winkler *et al.*, 2010).

Brain imaging genetics is a rising research topic. In such research, genotyping and neuroimaging data are integrated for revealing the associations from genotypes to phenotypes (Du *et al.*, 2018; Fu *et al.*, 2015; Hao *et al.*, 2017; Lin *et al.*, 2014; Shen and Thompson, 2020; Wang *et al.*, 2021c,b; Yan *et al.*, 2014). With genetic analysis

of imaging measures, not only risk variants closely related to diseases can be detected, but the potential biological mechanism of pre-clinical brain variations is deeply investigated. Most existing works are focused on the correlation analysis between structural neuroimaging and genetic variants. For example, Yan *et al.* (2014) have exploited sparse canonical correlation analysis (SCCA) with prior knowledge to associate the APOE gene with structural information of 78 regions of interests (ROIs). In Hao *et al.* (2017), a three-way SCCA (T-SCCA) framework has been designed to associate genetic variations in the APOE gene with brain ROIs measured via voxel-based morphometry (VBM). In Du *et al.* (2018), a truncated l_1 -norm penalized SCCA has been developed, which can accurately associate APOE single nucleotide polymorphism (SNP) rs429358 with the

gray matter density of hippocampus for the brain region related to AD. Wang et al. (2021c) have incorporated useful discriminant similarity information into SCCA approach, and successfully discovered common ROIs associated with genetic variations in the APOE gene.

At present, these studies have only attached importance for feature extraction from brain ROIs, and few successfully depict relations among brain areas. Many works have insisted that internal functional connections are included in resting networks (RSNs). The brain networks are able to be collected from resting state functional magnetic resonance imaging (fMRI) for quantifying brain region connections. Within a brain network (a.k.a., connectome), nodes are represented as regions-of-interest (ROIs) and edges can be denoted to be the correlation or connectivity strength between brain regions. Therefore, such brain network is viewed as a set of nodes and edges. We note that the network connections of an individual are considered as the reflection of the comprehensive properties of the corresponding brain system (Sporns, 2014). For instance, in literature (Thompson et al., 2013), relevant works on the relationship between genes and brain structural connections and functional networks (fMRI) are reviewed. Jie et al. (2014) developed a connection network-based classification method for classifying patients suffering from mild cognitive impairment (MCI) and normal people, and the classification accuracy is significantly improved. In Fu et al. (2015), Fu et al. (2015) concluded that the functional connectivity is genetically limited by heritability analysis. Wang et al. (2019) proposed a general framework so that the risk genetics are associated with brain network by exploiting both network voxel and connectivity. These voxel information and connectivity information are obtained from the structural magnetic resonance imaging (sMRI) and fMRI, which are deemed to be intermediate traits bridging genetic risk factors and disease state. As a result, the function connection network is tightly associated with both genes and diseases.

In our work, firstly, we parcellate the brain fMRI for each subject (as the phenotype input) into 90 ROIs by using the Automated Anatomical Labeling (AAL) atlas, where the cerebellum is removed. Then, for describing the connection strength of the brain region, we exploit the average time series of each brain region and the Pearson correlation coefficient of the brain region. The brain functional connection network of a subject is a 90×90 weighted brain network. Studies have shown that correlation-based methods are more effective in carrying the interconnection between brain regions (Smith et al., 2011). At last, for brain network phenotype data, we construct a functional connection matrix. In addition, some related research works have indicated that PageRank can be not only applied to directed graph (i.e. social network analysis), but also used for undirected graph (i.e. protein structure analysis) (Sarma et al., 2015). For example, these works in both Jiang et al. (2017) and Liu et al. (2020) apply PageRank to undirected protein network analysis. The study in Gleich (2015) has proposed that PageRank can also be utilized in human brain network analysis, although it is an undirected graph. Inspired by those works, we tried to use PageRank to evaluate node importance information of our brain network. After that, its upper triangle elements are extracted to be connectivity features and the PageRank algorithm (Florescu and Caragea, 2017) is adopted for estimating the importance of different brain regions as the brain region features. Based on the above consideration, our study focuses on identifying the connectome (including the brain region features and connectivity feature) by realizing the analysis of correlation between functional brain network and genetic variants. As a bi-multivariate technique, SCCA is strongly able to discover complex multi-SNP-multi-QT correlations in imaging genetics. Nevertheless, for previous SCCA methods, there exist three challenges in the calculation of precise bi-multivariate correlations and selection of relevant features, including high-dimensionality (across all 4005 network edges between 90 brain regions), non-linearity and fewer subjects. In order to remedy the challenges mentioned-above, a novel kernel approach for brain imaging genetics, which is called deep self-reconstruction sparse canonical correlation analysis (DS-SCCA). More specifically, the deep network, i.e. multiple stacked layers of non-linear transformation, is adopted to be the kernel, and the self-representation matrix is learned for reconstructing the

source data at the top layer. With the parametric algorithm, augmented Lagrange and stochastic gradient descent for optimization, the parameters in the designed model will be iteratively learned. The motivations are two-fold. First, a multi-layer feed-forward neural network is utilized to map each sample into a non-linear feature space, which aims to well discover the non-linear relationship of samples. Second, when the feature dimension is high and the sample size is small, it is difficult to calculate accurate bi-multivariate relationships and selecting relevant features. The self-representation idea (Elhamifar and Vidal, 2013) is facilitated to conduct the subspace clustering iteratively at the top layer of network. Subspace clustering at the top layer of the network is conducted, aiming to iteratively learn the mapping function. In the mapping function, the local structure is manipulated by the grouping effect which reflects the similarity structure of data to help identify these relevant markers.

In our preliminary work, only genotype data and connectivity edge data are combined for imaging genetics, which has been reported on ISBI 2021 (Wang et al., 2021a). In this journal article, the new contributions are given as following: *First*, extending the framework is a connectome genetics work, which is conducted for exploring the relationship between SNPs and brain network. This study can better discover how genetic factors affect brain connectivity than our preliminary work. *Second*, the proposed DS-SCCA approach is assessed on three additional phenotype datasets (i.e. brain network, the brain ROI node features of the non-network structure, the brain ROI node features of the network structures), with results validating its effectiveness. *Third*, more comprehensive biomarkers have been provided to guide disease interpretation than our preliminary work. *Fourth*, our method is compared to several state-of-the-art algorithms. *Fifth*, the analysis of the parameter and grouping effect in the proposed DS-SCCA model is studied.

2 Materials and methods

For most existing works, only the correlations between structure imaging and genetic variants are concerned, whereas the connectome is ignored, which are represented by a set of nodes and edges, where nodes denote ROIs in the brain and edges denote the connectivity strength or correlation between brain regions. Our work focuses on identifying the connectome (including the brain region features and connectivity features) of functional brain networks derived by realizing the relationships between genotypes and brain network phenotypes. Figure 1 is adopted to show a visual framework of our approach. Specifically, a connection matrix is firstly built, whose upper triangle elements are extracted to be connectivity features. Then, the PageRank algorithm is exploited so that the importance of different brain regions is estimated as the brain region features. At last, a novel DS-SCCA method is proposed to identify genetic associations with functional connectivity phenotypic markers.

In this section, we will describe the connectivity features based on Pearson correlation coefficient in the connectome, the brain region features based on the PageRank algorithm in the connectome, SCCA, the proposed DS-SCCA and optimization algorithm.

2.1 Brain connectivity features based on Pearson correlation coefficient

Using the AAL atlas, the brain fMRI for each subject (as the phenotype input) is firstly decomposed into 90 ROIs, where the cerebellum is removed. Many related works have been validated that correlation-based methods can effectively realize the interconnection between brain regions (Smith et al., 2011). The functional connectivity between the ROI pairs is calculated by exploiting Pearson correlation coefficients. A functional connectivity network is built for each subject, where the network vertices correspond to the ROIs and the edge weights are the correlation coefficients. We here apply Fisher's r -to- z transformation to the functional connectivity network (matrix) for improving the normality of the coefficients below:

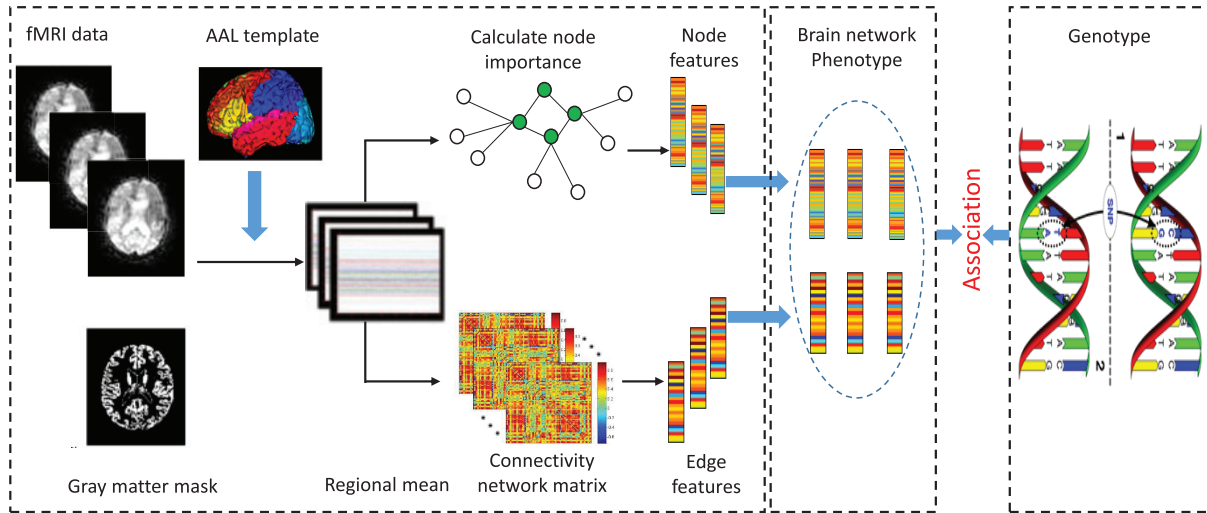


Fig. 1. Visual framework of the developed method

$$z = 0.5[\ln(1+r) - \ln(1-r)], \quad (1)$$

where z represents approximately normal distribution with respect to standard deviation $\sigma_z = \frac{1}{\sqrt{n-1}}$, in which n represents the amount of ROIs. r denotes the Pearson correlation coefficient. Moreover, for extracting valuable network measures, all negative correlations are dislodged. Thus, the brain functional connection network of a subject is a weighted brain network matrix with 90×90 . At last, the upper triangle elements of the functional connection matrix are exploited to be connectivity feature of the connectome, and which yields a set of brain connectivity features for each subject.

2.2 Brain region features based on PageRank algorithm

As shown in Florescu and Caragea (2017), the PageRank algorithm is the extensively utilized in Web page analysis. One argument is pointed out that if a node has significant links, its links to other nodes are also of significance. Gleich (2015) insisted that due to similar properties included in brain networks, the PageRank algorithm can be exploited for the analysis of brain networks. Inspired by this, brain regions are compared with web pages, as well as connections between regions are compared with ones between web pages. The PageRank algorithm is further utilized for estimating the significance of brain regions. Suggested by Yan and Ding (2011), a simplified PageRank form can be described below:

$$s_A = \sum_{a \in P(A)} \frac{s_a}{N_a}, \quad (2)$$

where A denotes a node, and $P(A)$ represents the set of nodes connecting to node A . N_a is the amount of links of node a , and s_A represents score of node A . The PageRank has the following property, i.e. the bigger value of s_A , the higher significance of A th node. Figure 2 is adopted to illustrate a visual example of the algorithm and corresponding calculation steps. Here, $P(A) = \{a, b\}$ and $s_A = \frac{s_a}{2} + \frac{s_b}{2}$.

Finally, for the PageRank algorithm in Equation (2), the node importance vectors of brain network are computed to be the brain region features of the connectome, and a set of brain region features for each subject is produced.

2.3 Sparse canonical correlation analysis

In a brain network (i.e. connectome), nodes are represented as ROIs, and edges can be denoted to be the connectivity strength or correlation between brain regions. Therefore, such brain network is able to be viewed as a set of nodes and edges. According to this argument, above-discussed connectivity edge features in Section 2.1 and node features in Section 2.2 are directly fused into brain network phenotype data. Suppose that the SNP genotype data is denoted by

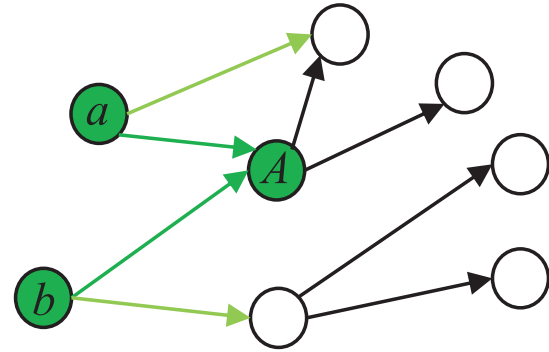


Fig. 2. An example of the PageRank algorithm and the calculation steps

$X = [x_1, \dots, x_n, \dots, x_N]^T \in R^{N \times p}$, and the brain network phenotype data is represented as $Y = [y_1, \dots, y_n, \dots, y_N]^T \in R^{N \times r}$, in which N means the subject amount, as well as p and r denote respectively the feature amounts of both SNPs and brain network.

As we know, the SCCA (Chi et al., 2013; Lin et al., 2014; Parkhomenko et al., 2009; Wan et al., 2011; Witten et al., 2009) is a common multivariate approach and a powerful association method, which aims to linearly transform X and Y so that Xu and Yv are maximally correlated. This theory is modeled to be the following form:

$$\max_{u,v} u^T X^T Y v \quad (3)$$

$$s.t. \quad u^T X^T X u \leq 1, v^T Y^T Y v \leq 1, \|u\|_1 \leq c_1, \|v\|_1 \leq c_2,$$

where u and v denote canonical loadings or weights, which reflect the contribution of each feature in the identified canonical correlation. It should be noted that the first two constraints are able to describe the covariance of data. The last two constraints can control the sparsity, which aim to select few relevant features from the SNP and imaging data.

2.4 Deep self-reconstruction SCCA

With the SCCA bi-multivariate technique, complex multi-SNP-multi-QT correlations can be identified in imaging genetics. It should be noted that three challenges of SCCA are existed in the calculation of precious bi-multivariate correlations and selection of relevant features, including high-dimensionality, non-linearity and fewer subjects. In order to address such challenges, a multi-layer

feed-forward neural network is utilized. By this way, each sample is mapped into a non-linear space, and the non-linear correlation of samples could be effectively explored. Further, at the top $M+1$ layer, subspaces are clustered, aiming to iteratively learn the mapping functions. Thus, we define the following DS-SCCA:

$$\begin{aligned} \max_{u,v,\theta_f,\theta_g} & u^T f^T(X; \theta_f) g(Y; \theta_g) v \\ \text{s.t. } & u^T f^T(X; \theta_f) u \leq 1, \quad v^T g^T(Y; \theta_g) v \leq 1, \quad (4) \\ & \|u\|_1 \leq c_1, \|v\|_1 \leq c_2 \end{aligned}$$

where $\theta_g = \{W_g^{(m)}, b_g^{(m)}, C_g, m = 1 : M\}$ and $\theta_f = \{W_f^{(m)}, b_f^{(m)}, C_f, m = 1 : M\}$ represent the parameters of the deep network, in which $m = 1, 2, \dots, M$ is the amount of the layers. Here, in order to clearly present our method, $\theta = \{W^{(m)}, b^{(m)}, C, m = 1 : M\}$ is defined to be θ_f (or θ_g), input sample x_i (or y_i) is denoted by $h_i^{(0)} = x_i \in \mathbb{R}^p$ (or $h_i^{(0)} = y_i \in \mathbb{R}^r$), $l = 1, 2, \dots, N$, the output of the m th layer is defined to be:

$$h_l^{(m)} = G(W^{(m)} h_l^{(m-1)} + b^{(m)}) \in \mathbb{R}^{d_m}, \quad (5)$$

where $W^{(m)} \in \mathbb{R}^{d_m \times d_{m-1}}$ and $b^{(m)} \in \mathbb{R}^{d_m}$ represent the weight and bias matrices in the m th layer, in which d_m indicates the dimension. $G(\cdot)$ denotes the activation function.

For data X (or Y), we define the output $H^{(M)}$ of the top layer as following:

$$H^{(M)} = [h_1^{(M)}, h_2^{(M)}, \dots, h_N^{(M)}] \quad (6)$$

In order to obtain $H^{(M)}$, the data matrix is firstly transformed into a non-linear space via using a multi-layer feed-forward neural network. The self-representation idea (Elhamifar and Vidal, 2013) is then facilitated to iteratively cluster subspaces at the top layer, so that the reconstruction ability of the self-representation is ensured and the local data structure can be captured. The corresponding model is formulated into:

$$\min_{\{W^{(m)}, b^{(m)}\}_{m=1}^M, C} \frac{1}{2} \sum_{l=1}^N \|h_l^{(M)} - c_l H^{(M)}\|_F^2 + \frac{\lambda}{2} \text{tr}(CLC^T), \quad (7)$$

where the first term indicates the loss function, which ensures the reconstruction ability. The second term exploits the grouping effect so that the local data structure is captured. Its effectiveness has been validated in literature (Han et al., 2014), and which has the strong capacity of describing the similarity structure for the reconstruction of the source data. $C \in \mathbb{R}^{N \times N}$ and $c_l \in \mathbb{R}^{l \times l}$ represent the self-reconstruction matrix. $H \in \mathbb{R}^{N \times p}$ or $(\mathbb{R}^{N \times r})$ and $h_l \in \mathbb{R}^{l \times p}$ or $(\mathbb{R}^{l \times r})$ represent the output of the top layer for the given data matrix X (or Y). $L = D - S$ represents the Laplacian matrix, where S measures the data similarity. Following this study in Han et al. (2014), the similarity matrix S can be computed via exploiting the k -nearest neighbor graph, in which neighbor size k is fixed to four in our

experiments. D denotes the diagonal matrix containing element $D_{ii} = \sum_{j=1}^n S_{ij}$. $\text{tr}(\cdot)$ is used to compute the trace of matrix.

Some remarks are here provided. For our designed model, the self-representation (Elhamifar and Vidal, 2013) is facilitated for learning the transformation function. By neural network, the non-linearity of data can be utilized. With the grouping effect, the local structure is manipulated, which can well reflect the similarity structure for the reconstruction of the source data. Figure 3 is used to visually present embedding deep self-reconstruction mapping from source data to the mapped data space.

2.5 Optimization algorithm

For obtaining optimal solution of Equation (4), we design a proximal alternating optimization scheme. The iterative update of $W^{(m)}$, $b^{(m)}$, C , u and v is as follows:

Update $W^{(m)}$ and $b^{(m)}$: Keeping with C and $H^{(M)}$ fixed, variables $W^{(m)}$ and $b^{(m)}$ can be updated via:

$$\min_{\{W^{(m)}, b^{(m)}\}_{m=1}^M} \frac{1}{2} \sum_{l=1}^N \|h_l^{(M)} - c_l H^{(M)}\|_F^2 \quad (8)$$

The Equation (8) can be solved by the sub-gradient descent algorithm. We take the derivatives of the objective function in Equation (8) respectively with respect to parameters $W^{(m)}$, $b^{(m)}$ to zero, and the chain rule (Peng et al., 2016) is utilized for obtaining the forms below:

$$\frac{\partial J}{\partial W^{(m)}} = \Delta^{(m)} (h_l^{(m-1)})^T, \quad (9)$$

$$\frac{\partial J}{\partial b^{(m)}} = \Delta^{(m)}, \quad (10)$$

where $\Delta^{(m)}$ satisfies the form below:

$$\Delta^{(m)} = \begin{cases} (W^{(m+1)})^T \Delta^{(m+1)} \odot G'(t_l^{(m)}), & m = 1, \dots, M-1 \\ (h_l^{(m)} - c_l H^{(m)}) \odot G'(t_l^{(m)}), & m = M \end{cases}, \quad (11)$$

where $t_l^{(m)} = W^{(m+1)} h_l^{(m-1)} + b^{(m)}$, $G(\cdot)$ represents the activation function and $G'(\cdot)$ denotes the corresponding derivative. The operator \odot is used to calculate the element-wise multiplication.

Thus, the neural network can be updated by the following paradigm:

$$\begin{cases} W^{(m)} = W^{(m)} - \tau \frac{\partial J}{\partial W^{(m)}} \\ b^{(m)} = b^{(m)} - \tau \frac{\partial J}{\partial b^{(m)}} \end{cases}, \quad (12)$$

where τ indicates a small positive constant, which denotes the step size and is fixed to $\tau = 10^{-4}$ in the experimental evaluation.

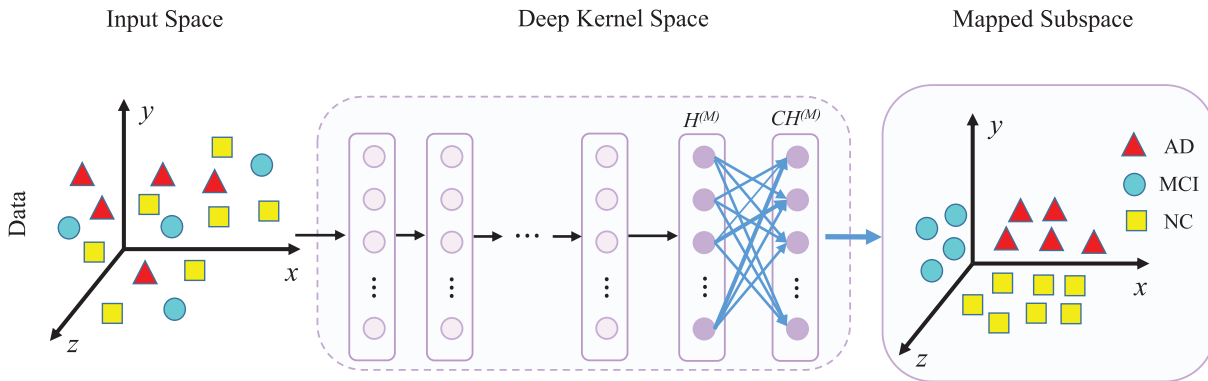


Fig. 3. Visual presentation of embedding deep self-reconstruction mapping from source data to the mapped data subspace

Update C: Once fixing $W^{(m)}$ and $b^{(m)}$, C can be updated via:

$$\min_C \|H^{(M)} - CH^{(M)}\|_F^2 + \lambda \text{tr}(CLC^T) \quad (13)$$

The derivative of Equation (13) with C is set to zero, and we can obtain the following equation:

$$(H^{(M)})^T (H^{(M)})C + \lambda CL = (H^{(M)})^T (H^{(M)}) \quad (14)$$

This equation meets the form of the continuous Lyapunov equation, whose solution could be acquired by utilizing 'lyap' function in MATLAB. The $W^{(m)}$, $b^{(m)}$ and C will be iteratively updated until the objective function Equation (8) satisfies the convergence condition.

In Equation (4), for convenient and clear description, $\theta = \{W^{(m)}, b^{(m)}, C, m = 1 : M\}$ is defined as $\theta_f = \{W_f^{(m)}, b_f^{(m)}, C_f, m = 1 : M\}$ (or $\theta_g = \{W_g^{(m)}, b_g^{(m)}, C_g, m = 1 : M\}$), and we define the output $H^{(M)}$ of the top layer in the neural network as $H_f^{(M)}$ (or $H_g^{(M)}$). According to the above step, we iteratively update $W_f^{(m)}$, $b_f^{(m)}$, C_f for given data X and $W_g^{(m)}$, $b_g^{(m)}$, C_g for given data Y . Once the objective function converges, we can obtain the C_f , $H_f^{(M)}$, C_g and $H_g^{(M)}$.

Update u and v : To update u and v , we consider network parameters $W_f^{(m)}$, $b_f^{(m)}$, C_f for the data matrix X and $W_g^{(m)}$, $b_g^{(m)}$, C_g for the data matrix Y are in an optimum state. Then, the mapped neural activities are able to be represented into:

$$f(X; \theta_f) = C_f H_f^{(M)} \triangleq \Lambda, \quad (15)$$

$$g(Y; \theta_g) = C_g H_g^{(M)} \triangleq \Omega \quad (16)$$

Therefore, for solving the DS-SCCA problem—Equation (4), the Lagrange multiplier is used, and the objective function with the penalties can be expressed into:

$$L(u, v) = \min_{u, v} \|\Lambda u - \Omega v\|_2^2 + \frac{\alpha_1}{2} \|\Lambda u\|_2^2 + \frac{\alpha_2}{2} \|\Omega v\|_2^2 + \gamma_1 \|u\| + \gamma_2 \|v\| \quad (17)$$

At last, the corresponding solutions in each iteration step are described below:

$$u = (\Lambda^T \Lambda + \gamma_1 D_1)^{-1} \Lambda^T \Omega v, \quad (18)$$

$$v = (\Omega^T \Omega + \gamma_2 D_2)^{-1} \Omega^T \Lambda u, \quad (19)$$

where D_1 and D_2 represent both diagonal matrices, whose k_1 th element is denoted by $\frac{1}{2\|u^{k_1}\|_1}$ ($k_1 \in [1, r]$) and k_2 th element is $\frac{1}{2\|v^{k_2}\|_1}$ ($k_2 \in [1, r]$). As can be found, the k_1 th element $\frac{1}{2\|u^{k_1}\|_1}$ and k_2 th element $\frac{1}{2\|v^{k_2}\|_1}$ cannot be computed when $|u| = 0$, $|v| = 0$. Therefore, $\frac{1}{2\|u^{k_1}\|_1}$ and $\frac{1}{2\|v^{k_2}\|_1}$ are here respectively rewritten as $\frac{1}{2\sqrt{u^{k_1} + \xi}}$ and $\frac{1}{2\sqrt{v^{k_2} + \xi}}$, where ξ is a real number with a small value.

Because D_1 , D_2 rely on u , v , we present an optimization procedure with iterative update to solve this objective. In each iteration, u is updated via fixing v , and v is then updated via fixing u . The procedure stops once the predefined stopping criterion is satisfied. The general optimization procedure of the proposed algorithm is provided in Algorithm 1.

3 Experiment

We will provide extensive experiments to assess the performance of the proposed DS-SCCA approach.

3.1 Materials and experimental settings

(1) Data Preparation

The imaging data and genotyping data are prepared from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database in this article. The related new information is updated in www.adniinfo.org. The subject data consist of fMRI and SNP data. We compare imaging and gene subjects and eliminate ones with missing values; thus, 157 valid subjects are used, which contain 26 AD, 40 early mild cognitive impairment (EMCI), 34 late mild cognitive impairment (LMCI), 19 significant memory concern (SMC) and 38 Normal Control (NC).

(2) Experimental Settings

In order to avoid any bias caused by random data division, 10 times independent non-repetitive 5-fold cross-validation are implemented

Algorithm 1 DS-SCCA method

Input: SNP genotype $X = [x_1, \dots, x_n, \dots, x_N]^T \in R^{N \times p}$; Brain network phenotype $Y = [y_1, \dots, y_n, \dots, y_N]^T \in R^{N \times r}$; Subjects with label information (i.e. NC, SMC, EMCI, LMCI or AD)

Output: u , v , θ_f , θ_g

Optimization:

- 1: Initialize $W^{(m)}$, $b^{(m)}$, $H^{(M)}$ and C ;
- 2: **while** not converge **do**
- 3: **for** $i = 1, 2, \dots, N$ **do**
- 4: Randomly select a sample x_i and let $b_i^{(0)} = x_i$;
- 5: Update $W^{(m)}$ and $b^{(m)}$ by Equation (8);
- 6: Compute $H^{(M)}$ by Equation (5);
- 7: Update C by Equation (14);
- 8: **end**
- 9: **end while**
- 10: Initialize u and v ;
- 11: **while** not converge **do**
- 12: Calculate the diagonal matrix D_1 , where the k_1 th element is $\frac{1}{2\|u^{k_1}\|_1}$;
- 13: Update $u = (\Lambda^T \Lambda + \gamma_1 D_1)^{-1} \Lambda^T \Omega v$;
- 14: Scale u so that $\|\Lambda u\|_2 = 1$;
- 15: Calculate the diagonal matrix D_2 , where the k_2 th element is $\frac{1}{2\|v^{k_2}\|_1}$;
- 16: Update $v = (\Omega^T \Omega + \gamma_2 D_2)^{-1} \Omega^T \Lambda u$;
- 17: Scale v so that $\|\Omega v\|_2 = 1$;
- 18: **end while**

to evaluate the average regression performance. A grid search is employed to optimally adjust the regularization parameters. The correlation coefficient between actual and predicted response values is computed, which is a common index for evaluating performances of association analysis. The actual value represents the feature value of the genotype data or phenotype data obtained by the association analysis method in this article.

In order to discover the association between SNPs and brain network, four competing QTs properties are included, i.e. (i) the brain ROI node features of the non-network structure: VBM, fluorodeoxyglucose positron emission tomography (FDG), and F-18 florbetapir PET scans amyloid imaging (AV45) in ADNI datasets, (ii) the brain ROI node features of the network structures: clustering coefficients, which is denoted as CC, (iii) the brain ROI node features of the network structures: node importance by using the pagerank algorithm, which is denoted as NIP and (iv) the brain connectivity edge features of the network structures: construct a functional connection matrix whose upper triangle elements are extracted to be connectivity feature, which is denoted as edge.

In our experiments, the benchmark algorithm SCCA (CCA with lasso) (Chi et al., 2013; Lin et al., 2014; Wan et al., 2011; Yan et al., 2014) is selected for comparison in this study. In addition, since the kernel method (Melzer et al., 2003) is a common tool that can well reveal and describe non-linear structural information in source input space, the SCCA is extended to the corresponding kernel version (KSCCA) so that the non-linear problems are considered for the comparison in this study. We exploit the Gaussian kernel (i.e.

$K_x(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma_x^2}}$ and $K_y(y_i, y_j) = e^{-\frac{\|y_i - y_j\|^2}{2\sigma_y^2}}$). The optimal kernel widths σ_x^2 and σ_y^2 are sought via searching parameter spaces $[2^{(-2)}, 2^{(-1)}, 2^0, 2^1, 2^2] \times \sigma_{x_0}^2$ and $[2^{(-2)}, 2^{(-1)}, 2^0, 2^1, 2^2] \times \sigma_{y_0}^2$. $\sigma_{x_0}^2$ and $\sigma_{y_0}^2$ are the mean square distances of samples. In our developed DS-SCCA method, we select dimensions of SNPs and OTs to be the L value for DNN-based models.

3.2 Improved association between genotype and brain network phenotype

Figure 4 is adopted to show the correlation coefficients estimated by different methods for different competing QTs properties via 5-fold cross-validation, where its mean is shown. From this figure, it is observed that all approaches demonstrate stable results across five experiments for different QTs properties. Although acceptable correlation coefficients are yielded via all approaches, DS-SCCA is obviously and consistently advantageous over SCCA method. Hence, it can be validated that the good ability to identify strong imaging genetic associations. In addition, for different QTs properties, obviously, DS-SCCA yields the best mean on brain network QTs. Specifically, the brain ROI node features (including CC and NIP) and connectivity edge features of the network structures reveal significant correlation, indicating the relation between the functional features and SNPs. Furthermore, the brain ROI node features (NIP)

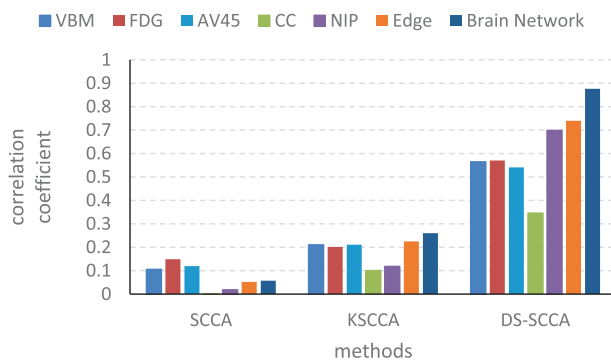


Fig. 4. The listed experimental average results by using fivefold cross validation, which is obtained on different QTs data

of the network structures reveal higher correlation than the brain ROI node features (including VBM, FDG and AV45) of the non-network structure, which also proves that the technology of brain network analysis and the hidden biological information of the brain network have great potential for mining the association between SNPs and imaging.

Figure 5 also presents the canonical loadings tested from 5-fold cross-validation in one experiment, where heat maps are used for the brain ROI node features and each row corresponds to a brain ROI node feature. We list the estimated canonical loading u on the top panel, which indicates weights for SNPs. We present the estimated v on the bottom, which contains weights for the imaging markers. It could be observed that our proposed DS-SCCA method in the brain ROI node features (NIP) yields a much clear picture for u and v . Only the APOE e4 SNP rs429358 is emphasized on the genetic side, which is tightly correlated with AD (Lambert et al., 2013). Our method also discovered the APOE e4 SNP in the brain ROI node features (CC, VBM, FDG and AV45), but presented much additional SNPs compared with the brain ROI node features (NIP). Thus, their results are not as sparse as NIP QTs properties. As a result, we could see that with DS-SCCA for NIP QTs properties, a much clean pattern is exhibited as well as a small amount of relevant imaging signals are reported, including amygdala-left, frontal-inferior-left, hippocampus-left and hippocampus-right that are closely associated with AD (de Leon et al., 1995; Du et al., 2016; Horinek et al., 2007). Briefly, the proposed DS-SCCA in brain network QTs properties successfully discovered a biologically valuable correlations between APOE SNP rs429358 and the metabolic alteration, pathological amyloid depositions and structure atrophy at the brain regions closely associated with AD. This further demonstrates that the proposed DS-SCCA in NIP QTs properties is able to achieve both the revealing of strong imaging genetic associations and the identification of valuable and relevant genetic and imaging markers.

3.3 Analysis of the most related ROI marker

In order to detect brain imaging ROIs, we average the acquired sparse coefficients through 5-fold cross-validation using our proposed method in brain network QTs data. Then, we choose the top 10 maximum weight ROIs to be the significant ROI markers. Table 1 is adopted to present the top 10 selected imaging features, whose average regression coefficients across five cross-validation trials are visually shown in Figure 6 via mapping onto the brain (Xia et al., 2013).

From Table 1 and Figure 6, it is seen that most of the selected ROIs identified via the proposed approach agree with those reported in previous works. Relevant structural imaging studies (de Leon et al., 1995; Horinek et al., 2007) reported that several diagnostic AD markers have been identified. Literature (Barnes et al., 2007) has proved that cingulate atrophy and hippocampus atrophy are important features for judging whether AD, and cingulate gyrus atrophy is more discriminative for AD. Meanwhile, Anterior cingulate gyrus and Orbitofrontal cortex (superior) are both related to cognitive and emotional control (Bechara et al., 2000; Bush et al., 2000), and patients with AD usually have greater emotional changes, which further confirms the effectiveness of the DS-SCCA approach. Additionally, many works have concluded that the abnormalities of the default model network (DMN) are included in destroyed brains suffering from mental illness and behavioral disorders. DMN is an important functional system in the brain. In DMN, misconnection of nodes may lead to AD, depression and even schizophrenia. In AD, DMN is first harmed by amyloid deposits (Zhang and Raichle, 2010), which mainly consists of precuneus, posterior cingulate cortex, inferior parietal lobule, bilateral temporal cortex and medial prefrontal cortex. Previous studies have found that the default network function of patients or high-risk groups has local activity abnormalities (Wang et al., 2013). According to our results, it is further proved that the gray matter density of DMN is strongly associated with genotype. Besides prior findings are confirmed, APOE rs429358 is also associated with other eminent AD markers like middle frontal gyrus and inferior occipital gyrus. Particular

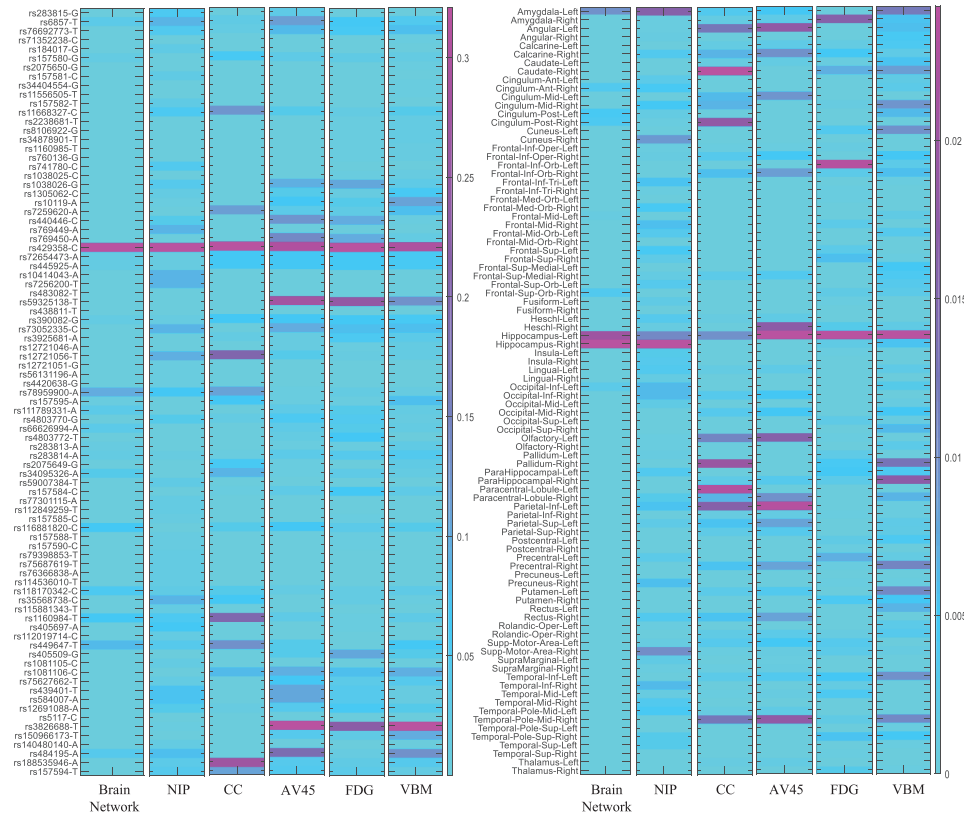


Fig. 5. The experimental results with five-fold cross validation on different QTs properties data

Table 1. The top 10 ROIs selected for brain network QTs by the DS-SCCA method

ID	ROI	Related studies
38	R. Hippocampus	de Leon <i>et al.</i> (1995)
37	L. Hippocampus	de Leon <i>et al.</i> (1995)
41	L. Amygdala	Horinek <i>et al.</i> (2007)
35	L. Posterior cingulate gyrus	Barnes <i>et al.</i> (2007)
36	R. Posterior cingulate gyrus	Barnes <i>et al.</i> (2007)
32	R. Anterior cingulate gyrus	Barnes <i>et al.</i> (2007)
6	R. Orbitofrontal cortex (superior)	Bechara <i>et al.</i> (2000)
53	L. Inferior occipital gyrus	—
7	L. Middle frontal gyrus	—
5	L. Orbitofrontal cortex (superior)	Bechara <i>et al.</i> (2000)

Note: L, left; R, right.

correlations among genotypes, phenotypes and neuropsychiatric symptoms seem to be existed, which are worthy to be further investigated.

3.4 Connectivity analysis

Brain network is viewed as a simplified representation of brain system, where nodes indicate brain regions as well as edges are represented as the connections between regions. The thickness represents the weight of the edge after feature selection, which can be understood as the importance of the edge in correlation problems.

For the edge features in the brain network, we use the Brainnet Viewer (Xia *et al.*, 2013) to plot top 10 maximum weight edges chosen via the designed DS-SCCA approach in Table 2 and Figure 7.

From Table 2 and Figure 7, we observe that the internal structure of DMN has obvious influence on the identification of genotype.

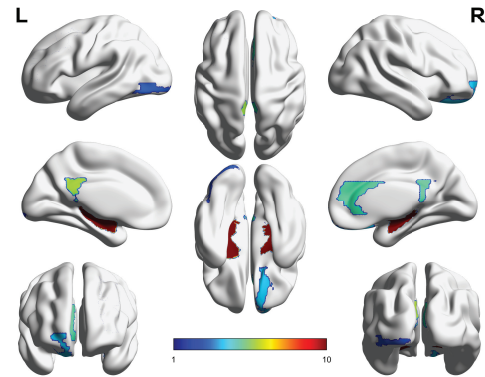


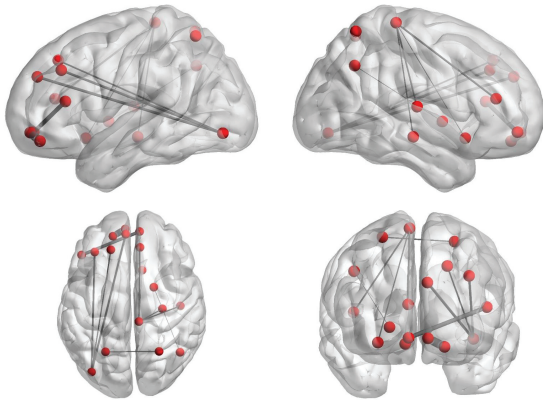
Fig. 6. The mapping top 10 ROIs for brain network QTs by our DS-SCCA approach are visually presented. The color denotes the regression coefficients of imaging markers

Additionally, the connection between DMN and other brain regions (angular gyrus, middle frontal gyrus, caudate) are closely related to genotype. Many existing works have validated that the local abnormal activity of the DMN node area can cause the variation of topological network attributes of other network areas (Ward *et al.*, 2015). The results of this article also further indicate that DMN is the important purpose of AD. At present, literature (de Jong *et al.*, 2008) illustrates that the volume of the pallidum and putamen is obviously related to that of the neocortical gray matter in subjects. But, different degrees of brain gray matter atrophy are included in AD patients (Karas *et al.*, 2004). The visual cortex in the brain, which is mainly responsible for processing visual information, is located around calcarine. The literature (Yip *et al.*, 2005) has found that 90% of the subjects had amyloid deposition around calcarine,

Table 2. The edges are selected via the designed DS-SCCA approach

Significance	Edge
1	L. Pallidum L. Angular gyrus
2	L. Middle frontal gyrus L. Calcarine cortex
3	L. Pallidum L. Caudate
4	R. Middle frontal gyrus L. Posterior cingulate gyrus
5	R. Putamen R. Inferior frontal gyrus (triangular)
6	L. Pallidum L. Orbitofrontal cortex (middle)
7	L. Calcarine cortex L. Middle frontal gyrus
8	L. Posterior cingulate gyrus R. Middle frontal gyrus
9	R. Precentral gyrus R. Supplementary motor area
10	R. Paracentral lobule L. Amygdala

Note: L, left; R, right.

**Fig. 7.** Visualization of the edges selected via the designed DS-SCCA approach

which is one of the manifestations of AD. In addition, the structural characteristics of precentral gyrus, paracentral lobule and supplementary motor area are closely associated with AD (Iwai et al., 1995; Jenkins et al., 1992). According to numerous clinical examples, it is proved that in this area certain obstacles in the movement and perception of advanced AD patients are included like stiff hands and feet, curl and incontinence. Our results further confirm that patients with AD have movement and perception obstacles in clinical manifestations.

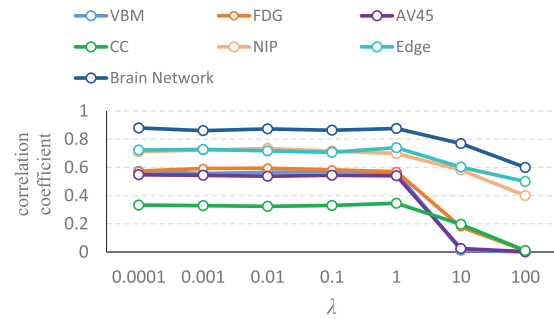
4 Discussion

In this section, the effect of the parameter λ is analyzed, grouping effect of our proposed deep self-reconstruction network is further investigated, state-of-the-art algorithms are compared, and we finally discuss the limitations of the proposed approach and the directions of further study.

4.1 Effect of parameter λ

In our proposed DS-SCCA method, the parameter λ for the identification between genotypes and different QTs properties is an important parameter by DS-SCCA. The optimal value of parameter λ is selected from the range of [0.0001, 0.001, 0.01, 0.1, 1, 10, 100] in our experiments. For observing how the λ values affect the final performance, the curves of DS-SCCA methods are plotted with the selected parameter range. In addition, we also use correlation coefficient to measure the performance of DS-SCCA. Figure 8 shows the experimental results with different λ on different QTs properties data.

From Figure 8, it is found that the λ values affect the final performance. More specifically, when λ lies in the range of [0.0001, 1], DS-SCCA achieves better correlation coefficient. This means that the weight matrices got by Equation (7) have the ability to hold the

**Fig. 8.** The experimental results with different λ on different QTs properties imaging data

intrinsic local data structure and the great performances could be obtained. Conversely, when the λ value is bigger than 1, the correlation coefficient decreases down obviously. The reason for this is that with λ increasing, the role of the regularization term in Equation (7) tends to be weakened and it will appear to many non-zero values in the reconstructive weight vector, which implies that the intrinsic geometric structure of data will be neglected.

4.2 Grouping effect of our proposed deep self-reconstruction network

In Supplementary Figure S1, the t-distributed stochastic neighbor embedding (t-SNE) plot (Laurens and Hinton, 2008) is used for visually presenting the original genetic data and reconstructed data. We can find that our proposed deep self-reconstruction network produces grouping effect for different classes of patients (i.e. LMCI, EMCI, SMC, NC and AD).

In addition, in order to validate the effectiveness of such grouping effect at the top $M+1$ layer of the network by self-representation, we compare DS-SCCA with the DS-SCCA without grouping effect by self-representation (denoted as D-SCCA) and DS-SCCA only with grouping effect by self-representation (denoted as S-SCCA), with results shown in Supplementary Figure S2. As can be seen in Supplementary Figure S2, S-SCCA can yield much better results. This implies that self-representation promotes the correlation coefficient of DS-SCCA. Moreover, the correlation coefficient of D-SCCA is obviously smaller than that of DS-SCCA. In other words, such grouping effect at the top $M+1$ layer is effective in DS-SCCA to assist to improve the association performance.

4.3 Comparison with state-of-the-art algorithms

In this part, the proposed approach is compared with some state-of-the-art algorithms, which is tested on ADNI data for brain imaging genetics. The details of these algorithms and their results are presented in Supplementary Table S1, in which five competing methods are used for comparison and each method exploits different subsets and different QTs.

As found in Supplementary Table S1, our method (DS-SCCA) acquires the best mean \pm SD values compared with four competing methods for the brain ROI features of the non-network structure on the ADNI databases. Thus, further proves that the functional connectivity for brain regions can assist to discover the important brain ROIs closely related to SNPs.

4.4 Limitations and future work

Experimental results presented above have validated the effectiveness and superiority of our method in brain imaging genetics association analysis. However, three limitations are included in our work. First, we only study the connectome genetics, which can be used for exploring the relationship between SNPs and brain network and has no ability to fully analyze the topological structure of brain network. Such topological structure has been proved to provide effective structural information and further promote the imaging genetic study. The future work is further focused on the construction of the

brain network model, combination of the topological structure, and the mining of more brain network features to discover biologically meaningful results. Second, in this initial work, our DS-SCCA model could be well exploited to calculate accurate bi-multivariate correlations and select relevant features. In order to obtain optimal solution of this model, we update $W_f^{(m)}$, $b_f^{(m)}$, C_f , $W_g^{(m)}$, $b_g^{(m)}$, C_g , u and v iteratively. There are eight learnable parameters and the training time is about 2.5 h. Nevertheless, once more features are contained in datasets, it will be more challenging to identify truly relevant ones. Thence, how to design an improved DS-DSCCA model and reduce the training time should be focused on the future study. Third, in our study, our used data (brain network phenotype data and SNP genotype data) here are not the biological sequence, therefore these related tools for biological sequence analysis such as BioSeq-BLM (Li et al., 2021), BioSeq-Analysis2.0 (Liu et al., 2019), ilearn (Chen et al., 2020), etc. cannot be directly used in our present work. In the future study, we will also focus on the biological sequence data and use the related tools for biological sequence analysis to explore the brain imaging genetics.

5 Conclusion

In this article, a comprehensive association analysis for both edge-level and network-level brain connectomic features is conducted. Different from previous works focusing on the tracts (i.e. ROIs), we develop a DS-SCCA method for investigating the genetic variants underlying these connectomic features. In addition, the proposed DS-SCCA method is a regularized, deep extension, scalable multi-SNP-multi-QT method, which is well-suited for applying imaging genetic association analysis to the ADNI datasets with non-linearity, high-dimensionality and a small amount of subjects. Extensive experimental results have validated the effectiveness of our method. Furthermore, this study initially attempts to consider how genetic factors affect brain connectivity, which is the main contribution of this article.

Funding

This work was supported in part by the National Natural Science Foundation of China [62106104, 62136004, 61902183, 61876082, 61861130366, 61732006]; and the National Key Research and Development Program of China [2018YFC2001600, 2018YFC2001602].

Conflict of Interest: none declared.

References

Barnes, J. et al. (2007) Atrophy rates of the cingulate gyrus and hippocampus in AD and FTL. *Neurobiol. Aging*, **28**, 20–28.

Bechara, A. et al. (2000) Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, **10**, 295–307.

Brookmeyer, R. et al. (2007) Forecasting the global burden of Alzheimer's disease. *Alzheimers Dementia*, **3**, 186–191.

Bush, G. et al. (2000) Cognitive and emotional influences in anterior cingulate cortex. *Trends Cognit. Sci.*, **4**, 215–222.

Chen, Z. et al. (2020) iLearn: an integrated platform and meta-learner for feature engineering, machine learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinf.*, **21**, 1047–1057.

Chi, E.C. et al. (2013) Imaging genetics via sparse canonical correlation analysis. In: *IEEE International Symposium on Biomedical Imaging*, San Francisco, USA, pp. 740–743.

de Jong, L.W. et al. (2008) Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: a MRI study. *Brain*, **131**, 3277–3285.

de Leon, M. et al. (1995) The hippocampus in aging and Alzheimer's disease. *Neuroimaging Clin. N. Am.*, **5**, 1–17.

Du, L. et al.; Alzheimer's Disease Neuroimaging Initiative. (2016) Structured sparse canonical correlation analysis for brain imaging genetics: an improved graphnet method. *Bioinformatics*, **32**, 1544–1551.

Du, L. et al.; Alzheimer's Disease Neuroimaging Initiative. (2018) A novel SCCA approach via truncated l1-norm and truncated group lasso for brain imaging genetics. *Bioinformatics*, **34**, 278–285.

Elhamifar, E. and Vidal, R. (2013) Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, 2765–2781.

Florescu, C. and Caragea, C. (2017) A position-biased pagerank algorithm for keyphrase extraction. In: *American Association for Artificial Intelligence, San Francisco, USA*.

Fu, Y. et al. (2015) Genetic influences on resting-state functional networks: a twin study. *Hum. Brain Mapping*, **36**, 3959–3972.

Gleich, D.F. (2015) Pagerank beyond the web. *SIAM Rev.*, **57**, 321–363.

Han, H. et al. (2014) Smooth representation clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA*, pp. 3834–3841.

Hao, X. et al.; Alzheimer's Disease Neuroimaging Initiative. (2017) Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimers disease. *Sci. Rep.*, **7**, 44272.

Horinek, D. et al. (2007) Magnetic resonance analysis of amygdalar volume in Alzheimer's disease. *Curr. Opin. Psychiatry*, **20**, 273–277.

Iwai, A. et al. (1995) The precursor protein of non-a beta component of Alzheimer's disease amyloid is a presynaptic protein of the central nervous system. *Neuron*, **14**, 467–475.

Jenkins, D.I.H. et al. (1992) Impaired activation of the supplementary motor area in Parkinson's disease is reversed when akinesia is treated with apomorphine. *Ann. Neurol.*, **32**, 749–757.

Jiang, B. et al. (2017) AptRank: an adaptive PageRank model for protein function prediction on bi-relational graphs. *Bioinformatics*, **33**, 1829–1836.

Jie, B. et al. (2014) Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification. *Hum. Brain Mapping*, **35**, 2876–2897.

Karas, G.B. et al. (2004) Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *Neuroimage*, **23**, 708–716.

Lambert, J.C. et al.; European Alzheimer's Disease Initiative (EADI). (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, **45**, 1452–1458.

Laurens, V.D.M. and Hinton, G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Li, H. et al. (2021) BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. *Nucleic Acids Res.*, **49**, e129.

Lin, D. et al. (2014) Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Med. Image Anal.*, **18**, 891–902.

Liu, B. et al. (2019) BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.*, **47**, e127.

Liu, B. et al. (2020) HITS-PR-HHblits: protein remote homology detection by combining PageRank and hyperlink-induced topic search. *Brief. Bioinf.*, **21**, 298–308.

Melzer, T. et al. (2003) Appearance models based on kernel canonical correlation analysis. *Pattern Recognit.*, **36**, 1961–1971.

Parkhomenko, E. et al. (2009) Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–34.

Peng, X. et al. (2016) Deep subspace clustering with sparsity prior. In: *International Joint Conference on Artificial Intelligence, New York, USA*, pp. 1925–1931.

Sarma, A.D. et al. (2015) Fast distributed PageRank computation. *Theor. Comput. Sci.*, **561**, 113–121.

Shen, L. and Thompson, P.M. (2020) Brain imaging genomics: integrated analysis and machine learning. *Proc. IEEE*, **108**, 125–162.

Smith, S.M. et al. (2011) Network modelling methods for FMRI. *NeuroImage*, **54**, 875–891.

Sporns, O. (2014) Contributions and challenges for network models in cognitive neuroscience. *Nat. Neurosci.*, **17**, 652–660.

Thompson, P.M. et al. (2013) Genetics of the connectome. *Neuroimage*, **80**, 475–488.

Wan, J. et al. (2011) Hippocampal surface mapping of genetic risk factors in AD via sparse learning models. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, Toronto, Canada*, pp. 376–383.

Wang, J. et al. (2013) Disrupted functional brain connectome in individuals at risk for Alzheimer's disease. *Biol. Psychiatry*, **73**, 472–481.

Wang, M. et al. (2019) Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease. *Bioinformatics*, **35**, 1948–1957.

Wang, M. et al. (2021a). Deep self-reconstruction sparse canonical correlation analysis for brain imaging genetics. In: *IEEE International Symposium on Biomedical Imaging, Nice, France*, pp. 1790–1793.

Wang, M. et al. (2021b) Identify complex imaging genetic patterns via fusion self-expressive network analysis. *IEEE Trans. Med. Imaging*, **40**, 1673–1686.

- Wang,M. *et al.* (2021c) Identify consistent cross-modality imaging genetic patterns via discriminant sparse canonical correlation analysis. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **18**, 1549–1561.
- Ward,A.M. *et al.* (2015) Relationships between default-mode network connectivity, medial temporal lobe structure, and age-related memory deficits. *Neurobiol. Aging*, **36**, 265–272.
- Winkler,A.M. *et al.* (2010) Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage*, **53**, 1135–1146.
- Witten,D. *et al.* (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.
- Xia,M. *et al.* (2013) BrainNet viewer: a network visualization tool for human brain connectomics. *PLoS One*, **8**, e68910.
- Yan,E. and Ding,Y. (2011) Discovering author impact: a pagerank perspective. *Inf. Process. Manag.*, **47**, 125–134.
- Yan,J. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative. (2014) Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics*, **30**, i564–i571.
- Yip,A.G. *et al.* (2005) Apoe, vascular pathology, and the ad brain. *Neurology*, **65**, 259–265.
- Zhang,D. and Raichle,M.E. (2010) Disease and the brain's dark energy. *Nat. Rev. Neurol.*, **6**, 15–28.