Disruption of gene co-expression network along the progression of Alzheimer's disease

Yurika Upadhyaya*, Linhui Xie†, Paul Salama†, Kwangsik Nho‡, Andrew J. Saykin‡, and Jingwen Yan*
*Department of BioHealth Informatics
Indiana University, Indianapolis, Indiana, 46202
Email: {yupadhya, jingyan}@iu.edu
†Department of Electric and Computer Engineering
Purdue University, Indianapolis, Indiana, 46202
Email: {linhxie, psalama}@iupui.edu
‡Department of Radiology and Imaging Sciences
Indiana University, Indianapolis, Indiana, 46202
Email: {knho, asaykin}@iu.edu

***Abstract*—Alzheimer's disease (AD) is one of the most common brain dementia characterized by gradual deterioration of cognitive function. While it has been affecting an increasing number of aging population and become a nation-wide public health crisis, the underlying mechanism remains largely unknown. To address this problem, we propose to investigate the gene co-expression network changes along AD progression. Unlike extant work that focus on cognitive normals (CNs) and AD patients, we aim to capture the network changes during the full range of disease progression, from CN, early mild cognitive impairment (EMCI) to late MCI (LMCI) and AD. In addition, many existing differential co-expression network analyses estimate the network of each group independently, which may possibly lead to sub-optimal results. Assuming that the gene co-expression patterns should be largely similar in consecutive disease stages, we propose to apply a modified joint graphical lasso model to estimate the networks of multiple diagnostic groups simultaneously. The permutation results shows that JGL model is much less likely to generate false positives with the similarity constraint. By comparing the estimated gene co-expression networks of all disease stages, we identified 8 clusters showing gradual changes during the progression of AD.**

***Index Terms*—Alzheimer's disease, differential co-expression, early detection.**

I. Introduction

Alzheimer's disease (AD) is a major neurodegenerative disorder that has been characterized by gradual memory loss and brain behavior impairment. According to the latest report [1], an estimated number of 5.7 million aging Americans are living with Alzheimer's and this number is expected to escalate in coming years given the rapid increase of aging population. To prevent this public health crisis, tremendous effort has been dedicated to discovery of effective AD biomarkers. In addition to *APOE* e4 alleles known as major genetic determinant [2], large-scale genome-wide association studies (GWAS) have led to identification of many novel genetic risk loci [3]. However, extant work largely investigated genetic variations or individual genes associated with AD. Very few studies paid attention to the interactions and associations among the gene products and how they are gradually disrupted during AD progression [4], [5].

To bridge this gap, we propose to perform a differential co-expression analysis across all the stages of AD, including cognitively normal (CN), early mild cognitive impairment (EMCI), late MCI (LMCI) and AD. One common method for generating co-expression networks is through pairwise Pearson's correlation. Though easy to implement, these simple strategies are very likely to generate false positives and many links are from indirect interactions. Other techniques such as least absolute error regression, Bayesian approach and graphical lasso model have also been used in construction of co-expression networks [6]. However, all of these methods can only estimate one network at a time. When applied to differential co-expression analysis, they estimate the network for each group separately by treating them as independent. This assumption clearly does not hold in disease study since disease formation is a progressive procedure. Co-expression networks in consecutive disease stages should be largely similar. For example, co-expression networks in CN group and EMCI group are expected to be largely similar with critical differences. Toward this, we propose to employ joint graphical Lasso [7] for simultaneous estimation of co-expression networks in multiple disease stages. We modified the traditional JGL algorithm to better model our assumption of similarity between consecutive disease stages.

In the present study, we focused our analysis on top AD-enriched pathways [8]. We estimated the co-expression networks of all diagnosis groups using modified JGL model and subsequently performed a comprehensive comparison analysis of co-expression networks using edge-level, node-level and network-level metrics. We used global clustering coefficient to identify structural property of each network. Node and edge centrality were calculated to allow comparison of individual interactions present in the network and identification of critical network entities. Finally, we were able to identify eight gene clusters showing gradual changes during the progression of AD. Five of them shows significant change from CN to EMCI and therefore have the potential to serve systems biomarkers for early screening of AD.

978-1-7281-0848-3/19/$31.00 ©2019 IEEE

II. METHODS

A. Dataset

Quality controlled plasma microarray data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) project (http://adni.loni.usc.edu). Detailed preprocessing steps can be found in [9]. We focused our analysis on genes involved in AD-enriched pathways highlighted in [8]. In total, 75 genes were included in this study. In the microarray data, if there are multiple probes corresponding to the same gene, we chose the probe with the maximum mean expression to represent the gene. Gene level expression was adjusted for RNA integrity Number (RIN), baseline age and sex with the weights derived from cognitive normals. Finally, 662 subjects without missing gene expression values were included (**Table.** I).

TABLE I
DEMOGRAPHIC INFORMATION OF PARTICIPANTS

Groups	Total(N)	Gender (M/F)	Age (Mean± Std)
CN	225	113/112	76.65±6.16
EMCI	193	105/88	79.26±7.35
LMCI	202	127/75	76.38±7.89
AD	42	27/15	75.69±9.46

Joint Graphical Lasso

Let us denote the gene expression data as $\mathbf{X} = [x_1, x_2, \cdots, x_n]$, where $x_n \subseteq \Re^p$. n is the number of subjects and p is the number of genes. To reconstruct a co-expression network among p genes, we assume that the $x_1, x_2, \cdots, x_n$ are independent and identically distributed with the positive definite $p \times p$ covariance matrix Σ. The inverse covariance matrix Σ^{-1} indicates the conditional independence between pairs of genes. Joint graphical lasso (**Eq.**1) estimates inverse covariance matrix of multiple groups together through maximizing penalized log likelihood [7]. Here, S is the empirical covariance matrix.

$$\min_{\{\boldsymbol{\Theta}\}} -\sum_{k=1}^{K} n_k \left(\log\det \boldsymbol{\Theta}^{(k)} - trace\left(\mathbf{S}^{(k)}\boldsymbol{\Theta}^{(k)}\right)\right) + P(\{\boldsymbol{\Theta}\}) \quad s.t. \quad \boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(K)} > 0 \tag{1}$$

Here, $\{\boldsymbol{\Theta}\} = \{\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(K)}\}$ and $\boldsymbol{\Theta}^{(k)}$ is the estimated inverse covariance matrix of $k-th$ group. The sparsity within each matrix and the similarity across matrices in K groups are encouraged by penalty $P(\{\boldsymbol{\Theta}\})$. λ_1 and λ_2 are two non-negative parameters to control the enforcement of sparsity and similarity. In [7], it is assumed that the networks across all pair of groups should be similar. However, this assumption does not always hold, especially for discovery of disease stage-specific networks. AD is a slowly progressive brain disorder that networks are expected to gradually dissolve or rewire during the progression. Gene co-expression network in the AD patients may have become very different compared to that of cognitive normals after years of progression. Therefore, we modified the penalty term $P(\{\boldsymbol{\Theta}\})$ to **Eq.** 2 such that the similarity among networks is only enforced for consecutive disease stages. For example, networks between CN and EMCI are encouraged to be similar.

$$P\left(\{\boldsymbol{\Theta}\}\right) = \lambda_2 \sum_{k<k'} \sum_{i\neq j} \left|\theta_{ij}^{(k)} - \theta_{ij}^{(k')}\right| + \lambda_1 \sum_{k=1}^{K} \sum_{i\neq j} \left|\theta_{ij}^{(k)}\right| \tag{2}$$

To solve the modified JGL model, we followed the steps in [7] using alternating directions method of multipliers (ADMM) algorithm. With the constraints $\boldsymbol{\Theta}^{(k)} = \boldsymbol{Z}^{(k)}$, the dual variables $\boldsymbol{U}^{(k)}$ are introduced to form scaled augmented Lagrangian [10].

$$\begin{aligned} L\left(\{\boldsymbol{\Theta}\}, \{\boldsymbol{Z}\}, \{\boldsymbol{U}\}\right) = &-\sum_{k=1}^{K} n_k \left(\log\det \boldsymbol{\Theta}^{(k)} - tr\left(\boldsymbol{S}^{(k)}\boldsymbol{\Theta}^{(k)}\right)\right) \\ &+ \frac{\rho}{2}\sum_{k=1}^{K} \left\|\boldsymbol{\Theta}^{(k)} - \boldsymbol{Z}^{(k)} + \boldsymbol{U}^{(k)}\right\|_F^2 \\ &+ P\left(\{\boldsymbol{Z}\}\right) \end{aligned} \tag{3}$$

The exact solution steps can be referred to [7]. We modified the step to update $\{\boldsymbol{Z}\}$ (**Eq.**4).

$$\min_{\{\boldsymbol{Z}\}} \left[\frac{\rho}{2}\sum_{k=1}^{K} \left\|\boldsymbol{Z}^{(k)} - \boldsymbol{A}^{(k)}\right\|_F^2 + P\left(\{\boldsymbol{Z}\}\right)\right] \tag{4}$$

We found that this minimization problem is completely separable for each elements(i,j) in the matrices,

$$\begin{aligned} \min_{Z_{ij}^{(1)},\ldots,Z_{ij}^{(K)}} \quad &\frac{\rho}{2}\sum_{k=1}^{K} \left\|Z_{ij}^{(k)} - A_{ij}^{(k)}\right\|_F^2 + \lambda_1 \sum_{\substack{k=1 \\ i\neq j}}^{K} \left|Z_{ij}^{(k)}\right| \\ &+ \lambda_2 \sum_{k=1}^{K-1} \left|Z_{ij}^{(k)} - Z_{ij}^{(k+1)}\right| \end{aligned} \tag{5}$$

The last penalty term is known as 1-d fused lasso and penalizes the absolute differences in adjacent values of β = $[\mathbf{Z}_{ij}^{(1)}, \ldots, \mathbf{Z}_{ij}^{(K)}]$. If we first consider λ_1 as zero and set $\mathbf{y}$=$[\mathbf{A}_{ij}^{(1)}, \ldots, \mathbf{A}_{ij}^{(K)}]$. The convex optimization problem becomes a simple 1-d fused lasso problem and can be easily solved with soft-thresholding technique [11].

B. Performance Evaluation

We compared the performance of JGL with graphical lasso, which estimates the co-expression network for each group separately. Using the permuted data, we generated 1000 co-expression networks using JGL and derived a frequency network for each group, where each edge has a value between 0 and 1000 indicating the times it is observed to be nonzero. Similarly, we generate another frequency network for each group using the results from graphical lasso. Since all the gene expressions are random, all edges should be zero. Any nonzero values will be considered as false positives. To ensure fair comparison, all the parameters in both models are tuned to achieve the best performance using Akaike information criterion (AIC).

C. Construction of Co-expression Network

We applied the modified JGL to estimate co-expression network for 4 disease stages. To evaluate the significance, we permuted the data 1000 times for each group. Permuted data was fed into the modified JGL model to estimate 4000 co-expression networks. For each group, edges with empirical P value >0.05 were considered insignificant and were removed from subsequent network comparison analysis. All edges were found to be significant except 5 edges in AD group.

D. Network Comparison

We compared co-expression networks across stages using edge level, node level and network level metrics. For each edge, its weight indicates the partial correlation between two connecting genes. Edges with absolute differences >0.01 between consecutive stages (e.g., EMCI -LMCI) were considered as potential biomarkers. We manually clustered these edges based on their patterns of change across disease stages. For every node, four centrality values were calculated: degree, betweenness, closeness and clustering coefficient. Finally, arithmetic mean measure was used to calculate the global clustering coefficient of the network. For all of these measures, we compared four networks by looking for values with continuous increasing or decreasing patterns from CN to AD.

III. Results

A. Comparison of JGL and graphical lasso

With the permuted data sets, it is expected that none of the genes are correlated. So all the edges in the estimated networks should have zero weight. However, with networks estimated using graphical lasso, we observed that 419 edges in CN, 420 edges in EMCI, 381 edges in LMCI and 250 edges were frequently (i.e., $\geq$ 100 out of 1000) estimated with non zero weights (i.e., false positives). With JGL, the weight of all estimated edges in CN, EMCI and LMCI are zero (**Fig.**1). In AD, 299 edges were occasionally (i.e., $<$ 100 out of 1000) estimated to be nonzero. This indicates that the similarity constraints on co-expression networks of consecutive stages can help effectively control the false positives. Therefore, the differential co-expression patterns identified through JGL will provide more accurate information of the altered biological system during disease formation.

B. Nodal Centrality Change during AD progression

For clustering coefficient, *C3*, *MAPK1*, *PAK1*, *PRKCH* and *SLA* showed continuously increasing pattern from CN to AD. In contrast, *CALM3*, *MAPK8* and *RASA1* showed a decreasing pattern. For degree centrality, *MAPK8* and *RHOA* were found to increase and *GRB2*, *PAK1*, *PRKCD*, *PRKCH*, *PRKCI*, *SORT1* were found to decrease when subjects progress to a more severe stage. For betweenness, *CR1*, *GRB2* and *RPS6KB1* demonstrated a decreasing pattern. For closeness, *DLG4* and *SRC* were observed to increase while PRKCD, RAP1A and RPS6KA1 showed a decreasing pattern. However, most of these are minor changes, except for *RPS6KB1* whose betweenness drops notably from CN to EMCI.

C. Pairwise Co-expression Change during AD progression

There are 66 gene pairs that demonstrated a continuously decreasing or increasing pattern from CN to AD and their patterns fall into eight clusters. Among these, five of them showed significant change from CN to EMCI. We combined these five clusters and formed a gene module as shown in **Fig.** 2, which is represented as early network. *RPS6KB1* was found to be the hub gene, followed by *GRINA*, *MAPK3*, *PRKCD*, *MAPK8* and *SORT1*. Many of these genes have been previously associated with AD already as individuals [12], [13]. However, this is the first study to reveal how their relationship changes during AD progression. Given that the co-expression patterns in this early network start to change in the CN stage, this network has great potential to serve as a systems biomarker to capture the biological alterations in the very early stage of AD.

D. Network Topology and Cluster identification across groups

When comparing the global network structure, network level clustering coefficient remains relatively stable from CN to LMCI, but drops significantly when progressing from LMCI to AD. This strong structural property indicates the resilience of the co-expression network in the prodromal stage of AD.

IV. Conclusion

We employed a modified JGL to borrow the strength of the relatedness across disease stages and jointly estimated multiple co-expression networks. Our results on permuted data sets showed that JGL is less likely to generate false positives. In the subsequent differential co-expression analysis, we found a significant change of clustering coefficient when subjects progress from LMCI to AD, but not in early stages. Node wise and edge wise comparison have led to eight gene clusters that demonstrated continuous changes from CN to EMCI, LMCI and AD. Particularly, five of them showed pairwise co-expression changes notably from CN to EMCI. Genes in these modules could be used as systems biomarkers for early screening in AD. More efforts are warranted to validate their function.

Acknowledgment

This research was supported by NIH grants R01 EB022574, R01 AG019771, P30 AG010133, NSF CRII 1755836 and Indiana University Collaborative Research Grant (IUCRG). This project was also funded, in part, with support from the Indiana Clinical and Translational Sciences Institute funded, in part by Grant Number UL1TR001108 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI

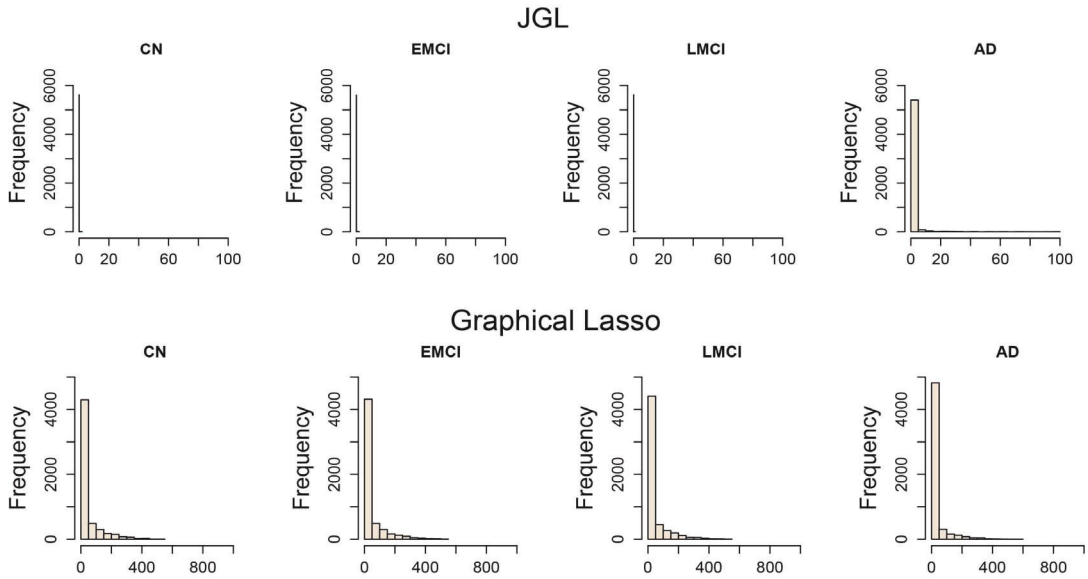

Fig. 1. Performance comparison of JGL and Graphical lasso on permuted data set.

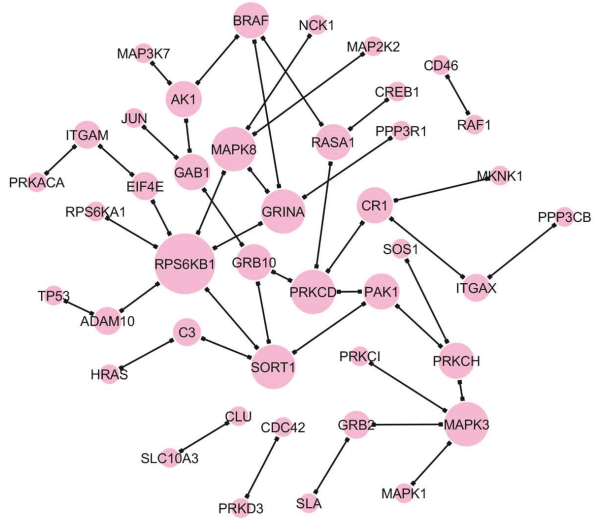

Fig. 2. Subnetwork showing early changes from CN to EMCI

investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

REFERENCES

[1] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, "Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy," *Nature Reviews Neurology*, vol. 9, no. 2, p. 106, 2013.

[3] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, G. W. Beecham *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature genetics*, vol. 45, no. 12, p. 1452, 2013.

[4] H. Kang, J. Lee, and S. Yu, "Differential co-expression networks using rna-seq and microarrays in alzheimer's disease," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1907–1908.

[5] Z. Wang, X. Yan, and C. Zhao, "Dynamical differential networks and modules inferring disrupted genes associated with the progression of alzheimer's disease," *Experimental and therapeutic medicine*, vol. 14, no. 4, pp. 2969–2975, 2017.

[6] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in bioinformatics*, p. bbw139, 2017.

[7] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[8] L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, T. M. Foroud, R. C. Green, X. Hu, M. J. Huentelman, S. Kim *et al.*, "Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers," *Brain imaging and behavior*, vol. 8, no. 2, pp. 183–207, 2014.

[9] A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho, S. L. Risacher, V. K. Ramanan, T. M. Foroud, K. M. Faber, N. Sarwar *et al.*, "Genetic studies of quantitative mci and ad phenotypes in adni: Progress, opportunities, and plans," *Alzheimer's & Dementia*, vol. 11, no. 7, pp. 792–814, 2015.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[11] R. J. Tibshirani, *The solution path of the generalized lasso*. Stanford University, 2011.

[12] J. L. Vázquez-Higuera, I. Mateo, P. Sánchez-Juan, E. Rodríguez-Rodríguez, A. Pozueta, M. Calero, J. L. Dobato, A. Frank-García, F. Valdivieso, J. Berciano *et al.*, "Genetic variation in the tau kinases pathway may modify the risk and age at onset of alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 27, no. 2, pp. 291–297, 2011.

[13] C.-H. Andersson, O. Hansson, L. Minthon, N. Andreasen, K. Blennow, H. Zetterberg, I. Skoog, A. Wallin, S. Nilsson, and P. Kettunen, "A genetic variant of the sortilin 1 gene is associated with reduced risk of alzheimers disease," *Journal of Alzheimer's Disease*, vol. 53, no. 4, pp. 1353–1363, 2016.