Integrating Gene Expression and Protein Interaction Data for Signaling Pathway Prediction of Alzheimer’s Disease

Wei Kong, Jingmao Zhang, Xiaoyang Mou, and Yang Yang

1 Information Engineering College, Shanghai Maritime University, Shanghai 201306, China
2 DNJ Pharma and Rowan University, Glassboro, NJ 08028, USA
3 Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Correspondence should be addressed to Wei Kong; weikong@shmtu.edu.cn

Received 23 February 2014; Accepted 18 March 2014; Published 9 April 2014

Abstract

Discovering the signaling pathway and regulatory network would provide significant advance in genome-wide understanding of pathogenesis of human diseases. Despite the rich transcriptome data, the limitation for microarray data is unable to detect changes beyond transcriptional level and insufficient in reconstructing pathways and regulatory networks. In our study, protein-protein interaction (PPI) data is introduced to add molecular biological information for predicting signaling pathway of Alzheimer’s disease (AD). Combining PPI with gene expression data, significant genes are selected by modified linear regression model firstly. Then, according to the biological researches that inflammation reaction plays an important role in the generation and deterioration of AD, NF-kB (nuclear factor-kappa B), as a significant inflammatory factor, has been selected as the beginning gene of the predicting signaling pathway. Based on that, integer linear programming (ILP) model is proposed to reconstruct the signaling pathway between NF-kB and AD virulence gene APP (amyloid precursor protein). The results identify 6 AD virulence genes included in the predicted inflammatory signaling pathway, and a large amount of molecular biological analysis shows the great understanding of the underlying biological process of AD.

1. Introduction

Alzheimer’s disease (AD) is a progressive and fatal neurodegenerative disorder manifested by cognitive and memory deterioration. The characteristic pathology changes in AD are fibrin deposition in cerebral cortex; it is the deposition of beta-amyloid (Aβ) in cell space and poly-Tau protein in cell. In pathomorphism, the expression is senile plaques (SP) and neurofibrillary tangles (NFT).

Many studies have investigated the mechanism of AD from various perspectives of its complexity. Recent researches show that a more accepted hallmark of AD is brain inflammation. Inflammation clearly occurs in pathologically vulnerable regions of AD brain and it does so with the full complexity of local peripheral inflammatory responses [1–3]. In the periphery, degenerating tissue and the deposition of highly insoluble abnormal materials are classical stimulants of inflammation. Likewise, in the AD brain damaged neurons and neurites and highly insoluble Aβ peptide deposits and neurofibrillary tangles provide obvious stimuli for inflammation [4–7].

To give insight to the AD mechanisms, high-throughput gene expression data has received extensive attention and made substantial progress in reconstructing the gene regulatory network. However, due to the underlying shortcomings of microarray technology such as small sample size, measurement error, and information insufficiency, unveiling disease mechanism has remained a major challenge to the AD research community. To overcome these problems, pathway information and network-based approaches [8] have been applied and become more informative and powerful for discovering disease mechanism.

Protein–protein interaction (PPI) networks are reconstructed from protein domain characteristics, gene expression data, and structure-based information with other evidence, for example, gene homology, function annotations,
and sequence motifs [9]. PPI data contain structure information among different genes while gene expression data do not. In our study, PPI network data as a priori pathway information is introduced for predicting the inflammatory signaling pathway in AD. Many literatures have given outstanding achievements by integrating gene expression data and PPI data, such as identification of protein complexes [10], small subnetworks [11], and biomarkers [12]. Zhao et al. presented an integer linear programming (ILP) method to uncover pathways among the given starting proteins, ending proteins, and some transduction factor proteins [13]. However, how to select the transduction factor proteins is a great problem. In our study, a modified network-constrained regularization method [14] is proposed for linear regression analysis to select appropriate number of significant genes. Simulation results show that this method can lead to an efficiently global smoothness of regression coefficients.

Based on that, ILP model is presented to reconstruct the inflammatory signaling pathway by integrating PPI data with the AD gene expression data. In the ILP model, the starting and ending proteins of the predicting pathway need to be arranged in advance. Nuclear transcription factor NF-kB (nuclear factor-kappa B) as one of the most important inflammatory factors is selected as the starting gene of the signaling pathway. As we know that NF-kB plays a key role in regulating the immune response to infection, therefore incorrect regulation of NF-kB has been linked to cancer, inflammatory and autoimmune diseases, septic shock, viral infection, and improper immune development. NF-kB has also been implicated in processes of synaptic plasticity and memory [15]. On the other hand, APP (amyloid precursor protein) as the most important AD virulence gene and precursor protein of Aβ is arranged as the ending protein of the predicting pathway.

The experiment results show that 6 AD virulence genes are identified being included in the predicted inflammatory signaling pathway, and a large amount of inflammation related genes and pathways has been found by molecular biological analysis and they show the great understanding of the pathogenesis of AD.

2. Methods

2.1. Linear Regression Model. Linear regression model is widely used in estimation and variable selection. In our study, the model is applied to selected subset of significant genes which are important for AD and are going to be the transduction factors of reconstructing pathway. In the next prediction step, gene expression data and PPI data will be integrated by ILP model. After all of the above, a pathway could be identified between NF-κB and APP. The usual linear regression model can be expressed as

\[
\mu = \sum_{j=1}^{p} x_j \beta_j = x_1 \beta_1 + x_2 \beta_2 + \cdots + x_p \beta_p,
\]

where \(\mu = (\mu_1, \mu_2, \ldots, \mu_n)\) is response vector, \(n\) is the sample number, \(x_j = (x_1, x_2, \ldots, x_n)^T\), \(j = 1, 2, \ldots, p\), is the predictor, represented by the \(j\)th gene’s expression data in all samples, and \(\beta_j\) is the \(j\)th gene’s weight vector. Assume that the predictors are standardized and the response is centered, we get

\[
\sum_{i=1}^{n} \mu_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 0,
\]

for \(j = 1, 2, \ldots, p\).

Gene expression data has the characteristic of less sample and great noise. As a simple model, linear regression model has significant performance in handling less sample and great noise data. The significant genes will get a larger coefficient while the nonsignificant genes will get a smaller coefficient.

2.2. Network-Constrained Regularization for the Linear Regression Model. Before using linear regression model, coefficient \(\beta\) needs to be estimated. Many methods have been proposed which focused on addressing high-dimensionality genomic data such as LASSO, LA-SEN, and LARS. Here, a modified network-constrain regularization analysis method by C. Li and H. Li [14] is applied to estimate the coefficient since it has been proved to perform better than other methods. This method is a lasso-type problem. It defines a normalized Laplacian matrix \(L\) as

\[
L = \begin{cases} 
1 - \frac{w(u, v)}{d_u} & \text{if } u = v \text{ and } d_u \neq 0, \\
-\frac{w(u, v)}{\sqrt{d_u d_v}} & \text{if } u \text{ and } v \text{ are adjacent,} \\
0 & \text{otherwise,}
\end{cases}
\]

where \(w(u, v)\) represents the weight of edge between linked genes \(u\) and \(v\). \(d_u = \sum_{u \to v} w(u, v)\) represents all the adjacent genes of \(v\) on the network. Then the definition of the network-constrained regularization criterion is

\[
L(\lambda_1, \lambda_2, \beta) = (\mu - X \beta)^T (\mu - X \beta) + \lambda_1 |\beta_1| + \lambda_2 \beta^T L \beta,
\]

where \(X = (x_1 | K | x_p); |\beta_j| = \sum_{j=1}^{p} |\beta_j|; \lambda_1, \lambda_2\) are nonnegative turning parameters. And then we estimate \(\beta\) by minimizing (4):

\[
\beta = \arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\}.
\]

Minimizing (4) is equivalent to solving a lasso-type optimization problem. Turning parameters are estimated by 10-fold cross-validation (CV). Genes in gene interaction network are selected by PubGene; we chose genes related to Alzheimer.

2.3. Integer Linear Programming (ILP). The ILP model formulates signaling network detection as an optimization problem and treats a signaling network as a whole entity as described in its original publication [13]. PPI network is a weighted undirected graph, that can be described as
\( G(V, E, W) \), where \( V \) is vertices in the graph, representing protein; \( E \) is edge between proteins; and \( W \) represents the weight of edges. \( W \) can be calculated by gene expression data. The ILP model can be described as follows:

\[
\text{Minimize } S = -\sum_{i=1}^{\left| V \right|} \sum_{j=1}^{\left| V \right|} w_{ij} y_{ij} + \lambda \sum_{i=1}^{\left| V \right|} y_{ij}
\]

Subject to:

\[
y_{ij} \leq x_i, \\
y_{ij} \leq x_j, \\
\sum_{j=1}^{\left| V \right|} y_{ij} \geq 1, \text{ if } i \text{ is either a starting or ending protein,} \\
\sum_{j=1}^{\left| V \right|} y_{ij} \geq 2x_i, \text{ if } i \text{ is not a starting or ending protein,} \\
x_i = 1, \text{ if } i \text{ is a protein known in STN,} \\
x_i \in \{0, 1\}, \text{ } i = 1, 2, \ldots, \left| V \right|, \\
y_{ij} \in \{0, 1\}, \text{ } i, j = 1, 2, \ldots, \left| V \right|,
\]

where \( w_{ij} \) is the weight between proteins \( i \) and \( j \) in weighted undirected graph \( G \), \( y_{ij} \) is a binary variable to denote whether the edge \( E(i, j) \) is a part of the STN. \( x_i \) is also a binary variable to denote whether protein \( i \) is a component of the STN. \( \lambda \) is a positive penalty parameter. \( \left| V \right| \) includes all proteins in the PPI network. \( y_{ij} \leq x_i \) and \( y_{ij} \leq x_j \) mean that only if proteins \( i \) and \( j \) are both components of STN, the edge \( E(i, j) \) should be considered. \( \sum_{j=1}^{\left| V \right|} y_{ij} \geq 1 \) represents at least one protein contact with starting protein or ending protein. \( \sum_{j=1}^{\left| V \right|} y_{ij} \geq 2x_i \) makes sure that if \( x_i \) is selected as a component of STN, there are at least two proteins link to the vertex.

The starting protein and ending protein have confirmed above that the genes selected by linear model were treated as transduction factors. The method of chosen parameter \( \lambda \) can be found in its original publication. Then we detected protein pathway by the ILP model.

### 3. Results and Discussion

To evaluate ILP model, AD dataset, series GSE1297, was used which were human hippocampal gene expression downloaded from GEO DataSets from the National Center for Biotechnology Information (NCBI) offered by Blalock et al. [16]. The hippocampal specimens they used are obtained through the Brain Bank of the Alzheimer’s Disease Research Center at the University of Kentucky. The human Gene Chips (HG-U133A) of Affymetrix and Microarray Suite 5 are used in analyzing the microarray data. There are a total of 9 control, 7 incipient, 8 moderate, and 7 severe AD samples included in this dataset with 22283 gene expressions in each sample. The PPI data we used is downloaded from website BioGRID (http://thebiogrid.org/) with 12466 proteins and 40323 interactions in total.

The file format of microarray data downloaded from NCBI is CEL. The probe data needs data processing like background correct, normalization, probe correct, and so on. Then ANOVAs were used on preliminary select genes and removed all genes whose \( P \) value was less than 0.05. After processing, 7030 genes for each sample were left. Then taking linear regression model with modified network-constrained regularization and AD biological information, the coefficients \( \beta = (\beta_1, \beta_2, \ldots, \beta_p) \) and \( p = 7030 \) were obtained. Among them, 7017 values of \( \beta \) were zeroes and the other 13 \( \beta \)s were nonzero values. So these 13 genes with nonzero values of \( \beta \) were considered as significant genes to AD phenotype and they are denoted in green circles with “(1)” and gene names in Figure 1. The 13 selected genes can be mapped to PPI network to get the corresponding proteins and the interactions between them and other proteins. Each selected gene was connected with some other genes in PPI network by the edges. In ILP algorithm, edge between \( i \) and \( j \) was represented by \( y_{ij} \). When ILP chose this edge, \( y_{ij} = 1 \), otherwise \( y_{ij} = 0 \).

ILP tries to assign 0 or 1 for \( y_{ij} \) to ensure the result network has the largest weight. For the weight of edge, \( w_{ij} \), here we use the Pearson coefficient of the gene expression values to represent the weight between proteins \( i \) and \( j \).

Then using NF-\( \kappa \)-\( \beta \) as starting protein and APP as ending protein ILP model was applied to formulate the signaling network. In the ILP algorithm, penalty parameter \( \lambda \) is a size control parameter that needs to be adapted manually. If its value is too large, the predicted signaling network will be enormous, otherwise it will be too small to catch the useful biological information. In our simulation experiment, after adapting from small value to large value, \( \lambda \) was determined as 0.65.

We finally got a signaling pathway with several small subnetworks. This network is reconstructed by 45 genes including 13 selected significant genes and is shown in Figure 1. In Figure 1, “(0) NF-\( \kappa \)-\( \beta \)” represents the starting protein NF-\( \kappa \)-\( \beta \), “(3) APP” represents the ending protein APP, “(1)” with the protein names denote the corresponding selected genes by the regress model, and “(2)” with protein names are selected by ILP to reconstruct the signaling pathways between NF-\( \kappa \)-\( \beta \) and APP. In order to analyze the biological functions of the pathways and subnetworks, the predicted result was mapped into its coding gene pathway network, and the online analysis website DAVID (http://david.abcc.ncifcrf.gov/home.jsp) was utilized to further understand their molecular biological functions to AD. Table 1 shows the KEGG pathway analysis result.

First of all, among the prediction results, there are 5 genes that have been confirmed as the AD virulence genes such as SNCA, CALM1, GSK3B, PSEN1, and APP which have been biologically demonstrated playing crucial roles in AD. Based
on Table 1, T cell receptor signaling pathway, B cell receptor signaling pathway, the Notch signaling pathway, NOD-like receptor signaling pathway, Toll-like receptor signaling pathway, MAPK signaling pathway, neurotrophin signaling pathway, insulin signaling pathway, and so on were found to include a major part of important genes derived from the regression model. Specially, the main predicted pathway in Figure 1 includes NFKB1, NOTCH1, PSEN1, CTNNB, COP5, MAPK14, CENPC1, UBC, and APP; the molecular biological analysis shows that they have close correlation between inflammatory response and AD.

It was found that inflammation is a major mechanism of acute brain injury and chronic neurodegeneration [17]. During the onset of an inflammatory response, signaling pathways are activated for translating extracellular signals into intracellular responses converging to the activation of NF-κB, the central transcription factor in driving the inflammatory response [18]. NF-κB has long been considered a prototypical proinflammatory signaling pathway, largely based on the activation of NF-κB by proinflammatory cytokines, such as interleukin-1 (IL-1) and tumor necrosis factor α (TNFα), and the role of NF-κB on the expression of other proinflammatory genes including cytokines, chemokines, and adhesion molecules, which has been extensively reviewed elsewhere [9].

Recent studies have also found that Notch receptors in Notch signaling pathway regulate cell differentiation and function, and Notch1 has been shown to induce glia in the peripheral nervous system [19, 20]. NF-κB, Notch, MAPK, and PSEN1 included in the main pathway of Figure 1 were observed to have strong regulating functions between each other, since interleukin-1 (IL-1) activates NF-κB via interleukin-1 receptor-associated kinase (IRAK) and mitogen-activated protein kinase (MEKK1(MAP3K)) dependent inhibition of NF-κB inhibitor (I-κB) [21, 22]. V-rel reticuloendotheliosis viral oncogene homolog (c-Rel (NF-κB subunit)) can trigger Notch homolog 1 translocation-associated (NOTCH1 receptor) signaling pathway by inducing expression of Jagged1, ligand for Notch receptors [23, 24]. NOTCH1 receptor activated by Jagged1 or Delta-like 1 (DLL1) is cleaved by ADAM metallopeptidase domain 17 (ADAM17) and PSEN1 to intracellular domain of NOTCH1. NOTCH1 is transported to nucleus and participates in recombination signal binding protein for immunoglobulin kappa J region (RBP-J kappa (CBFI)) mediated transcription [24, 25].

It was also found that β-catenin- (CTNNB1-) dependent WNT signaling pathways have crucial roles in the regulation of diverse cell behaviours, including cell fate, proliferation, survival, differentiation, migration, and polarity [26, 27]. It is interesting to note that loss of TNFα function would inhibit Wnt/β-catenin signaling [28]. Recently studies show that Wnt/β-catenin and NF-κB are independent pathways; cross-regulation between the Wnt and NF-κB signaling pathways has emerged as an important area for the regulation of a diverse array of genes and pathways active in chronic inflammation, immunity, development, and tumorigenesis.

Figure 1: Protein signaling pathway predicting result between NF-κB and APP.
Table 1: KEGG pathway analysis of the predicted pathways and subnetworks in Figure 1.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate cancer</td>
<td>9</td>
</tr>
<tr>
<td>Notch signaling pathway</td>
<td>8</td>
</tr>
<tr>
<td>Neurotrophin signaling pathway</td>
<td>6</td>
</tr>
<tr>
<td>Pathways in cancer</td>
<td>6</td>
</tr>
<tr>
<td>Chronic myeloid leukemia</td>
<td>6</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>5</td>
</tr>
<tr>
<td>Melanogenesis</td>
<td>9</td>
</tr>
<tr>
<td>T cell receptor signaling pathway</td>
<td>5</td>
</tr>
<tr>
<td>Acute myeloid leukemia</td>
<td>5</td>
</tr>
<tr>
<td>NOD-like receptor signaling pathway</td>
<td>5</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>5</td>
</tr>
<tr>
<td>Insulin signaling pathway</td>
<td>6</td>
</tr>
<tr>
<td>Pancreatic cancer</td>
<td>4</td>
</tr>
<tr>
<td>B cell receptor signaling pathway</td>
<td>4</td>
</tr>
<tr>
<td>Small cell lung cancer</td>
<td>4</td>
</tr>
<tr>
<td>Progesterone-mediated oocyte maturation</td>
<td>4</td>
</tr>
<tr>
<td>MAPK signaling pathway</td>
<td>4</td>
</tr>
<tr>
<td>Toll-like receptor signaling pathway</td>
<td>5</td>
</tr>
<tr>
<td>Endometrial cancer</td>
<td>6</td>
</tr>
<tr>
<td>Spliceosome</td>
<td>4</td>
</tr>
<tr>
<td>Glioma</td>
<td>5</td>
</tr>
<tr>
<td>Adipocytokine signaling pathway</td>
<td>3</td>
</tr>
<tr>
<td>Epithelial cell signaling in <em>Helicobacter pylori</em> infection</td>
<td>3</td>
</tr>
<tr>
<td>RIG-I-like receptor signaling pathway</td>
<td>3</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>3</td>
</tr>
</tbody>
</table>

Both β-catenin and NF-κB activate inducible nitric oxide synthase (iNOS) gene expression [29].

In addition, the regulatory network between COPS5 and CENP CI has been extracted from our algorithm which is also observed to be implicated in the pathogenesis of AD. The COP9 (constitutive photomorphogenesis 9) signalosome (COPS), a large multiprotein complex that resembles the 19S lid of the 26S proteasome, plays a central role in the regulation of the E3-cullin RING ubiquitin ligases (CRLs). The catalytic activity of the COPS complex, carried by subunit 5 (COPS 5/Jab1), COPS-dependent COPS 5, displays isopeptidase activity; it is intrinsically inactive in other physiologically relevant forms [30]. Increased APP and accumulation of neurotoxic Aβ in the brain are central to the pathogenesis of AD. COPS5 is found to be a novel RANBP9-binding protein that increases APP processing and Aβ generation [31]. COPS5 regulates the stability of the inner kinetochore components CENP-T and CENP-W, providing the first direct link between COPS5 and the mitotic apparatus and highlighting the role of COPS5 as a multifunctional cell cycle regulator [32]. CENP-T interacts with both centromeric chromatin and microtubule binding kinetochore complexes. Transient targeting of CENP-C to a noncentromere LacO locus induces the recruitment of some outer kinetochore proteins, similar to CENP-T [33]. Our result exhibited that COPS5 regulates CENP-C in the main pathway and ubiquitin and NF-κB were found to be associated with them. Ubiquitin can degrade the IκB which is the inhibitor of NF-κB, processing of precursors, and activation of the IκB kinase (IKK) through a degradation-independent mechanism [34]. On the other hand, COPS5 functions through CDK2 to control premature senescence in a novel way, depending on cyclin E in the cytoplasm [35].

4. Conclusions

Although many efforts have been done several decades of AD, it is still difficult to uncover its phenotype-pathway relationship and pathogenesis. Recent studies show that the pathology of AD has an inflammatory component that is characterized by upregulation of proinflammatory cytokines, particularly in response to Aβ. However, the signaling pathways and regulatory networks of the inflammation in AD pathogenesis are very difficult to reconstruct due to the complexity.

To discover the inflammation signaling pathway and regulatory network of AD, in our study, protein interactive network data, PPI was introduced to overcome the information insufficiencies of DNA microarray gene expression data by integer linear programming (ILP) method. Two stages had been used in predicting inflammatory pathway for AD. Firstly, significant genes had been selected by linear regression analysis with the modified network-constrained regularization analysis. Then ILP model was applied to reconstruct the signaling pathway between NF-κB and AD virulence gene APP since NF-κB has long been considered a prototypical proinflammatory signaling pathway. From the molecular biology analysis, we found that genes on the main pathway of the reconstruction results play crucial roles in inflammatory response and APP which give more biological insight for AD pathogenesis, such as NF-κB, NOTCH1, CTNNB1, COPS5, and their signaling pathways. Even more, the pathogenic contribution of the inflammatory response in AD is supported by our finding of the regulating and functions of the genes and subnetworks in the predicted signaling pathways. In general, our studies on combining PPI and gene expression data discover the signaling pathways of inflammatory response on AD and help for deeply understanding the pathogenesis of AD.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (nos. 61271446 and 61003093).
References


