2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010)
IPCSIT vol. 53 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V53.No.1.09

A Integrated Computational Approach for Protein Sub-network Detection in Parkinson's Disease

Yue Huang[1+] and Yunying Huang[2]

[1]Institute of Signal and Information Processing, Department of Communication Engineering, Xiamen University, Xiamen, Fujian, 361005, CHINA

[2]Department of Electronic Engineering, Xiamen University, Xiamen, Fujian, 361005, China

Abstract. Parkinson's disease (PD) is a typical case of neurodegenerative disorder, which often impairs the sufferer's motor skills, speech, and other functions. Combination of protein-protein interaction (PPI) network analysis and gene expression studies provides a better insight of Parkinson's disease. A computational approach was developed in our work to identify protein signal network in PD study. First, a linear regression model is setup and then a network-constrain regularization analysis was applied to microarray data from transgenic mouse model with Parkinson's disease. Then protein network was detected based on an integer linear programming model by integrating microarray data and PPI database.

Keywords: Parkinson's disease; microarray data; linear regression model; integer linear programming; protein network detection

1. Introduction

Parkinson's disease (PD) is a chronic degenerative neurological disorder that affects one in one hundred people over age 60. It is characterized as a disorder of the central nervous system that results from the loss of cells in various parts of the brain, including a region called the substantia nigra. Loss of dopamine causes neurons to fire randomly, leaving patients less able to direct or control their movement.

Many studies investigate the mechanism of the disease from various perspectives because of its complexity. Development of microarray technology provides biologists with the ability to measure the expression levels of thousands of genes in experiments, which offer molecular clues about underlying mechanisms of the disease. Intense researches have contributed to analyzing microarray data for PD for years [1, 2]. Despite gene expression data, protein-protein interaction (PPI) network is also a hot topic in understanding mechanism of PD. Since proteins perform their functions by interacting with one another, PPI network provides a framework for the exploiting biological processes and give insights into mechanisms of diseases. Analysis of protein and pathway interactions might indicate common properties of good candidates to be targeted by therapy. According to literatures, PPI networks is always detected from protein domain characteristics, gene expression data, structure-based information with other evidence, e.g. gene homology, function annotations, and sequence motifs[3-4]. Zhao and others presented an integrated linear programming (ILP) method to uncover pathways among given starting proteins, ending proteins and some transduction factor proteins [5]. However, how to select transduction factor proteins were not discussed. In this paper, a linear regression model was utilized and a modified LA-SEN method was applied to address the model to select limited number of genes, which were considered as key genes related to PD. These genes are applied as transduction factors in the following ILP procedures. Next, a proteins signal transduction network as a

[+] Corresponding author.
E-mail address: yhuang2010@xmu.edu.cn.

biomarker is grown among key proteins of PD, integrating information from gene expression data and PPI network database. Important pathways and module functions are also identified after analysis.

2. Data Acuqistion

Microarray data from a PD study was used in the analysis, the brief description is shown as follow: substantia nigra tissue from postmortem brain of normal and Parkinson disease patients were used for RNA extraction and hybridization on Affymetrix microarrays: 9 replicates for the controls and 16 replicates for the Parkinson's disease patients were used. Both cohorts included males and females. The data was obtained from the gene expression omnibus database (GEO database), which are available in http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7621 Affymetrix microarray suite 5.0 was used to scan and analyze the relative abundance of each gene from the intensity signal value. A detailed description of the experiment assays can be found in the original publication [6].

3. Methods

3.1 Data preprocessing

Microarray data from different samples were copied into one file in Excel. All preprocessing was performance in R programming language. As described in Huber's work, 'Expresso' function in the 'affy' library of 'Bioconductor', including methods for background correction (*mas*), normalization (*quantile*) and so on, was applied to preprocess the data. Probes were rated as 'present' if it was detected on at least four chips in the study and were involved in the further analysis. To address genes with more than one probe, the probes with highest significance through different conditions were identified to each gene. ANOVAs were performed on the microarray data and the probes whose p-value were less than 0.05 in the test were removed from further analysis, since they had small variance over different conditions. After preprocessing, a new microarray matrix containing 7000 genes was left for further analysis.

3.2 Processing

Detail description of the proposed approach is presented in this part. Firstly, important genes could be selected by addressing linear regression model. And then, protein pathway could be identified by an integrated linear programming (ILP) method.

1) Linear Regression Model.

Regression model is widely used in key gene selection. Here, our goal is to look for a subset of genes which are important to AD and are going to be the 'seed point' of growing protein pathway in the next step. Therefore, a simple linear regression model was utilized to perform this procedure.

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{n} \text{,} \tag{1}$$

where $\mathbf{Y} = (y_1, y_2, \cdots y_j, \cdots y_M)^T$ is the response vector; M is total sample number; $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_\mathbf{j}, \cdots \mathbf{X}_\mathbf{N}]$ is the model matrix, where $\mathbf{X}_\mathbf{j} = (x_{1j,} \cdots x_{ij}, \cdots x_{nj})^T$, $j = 1, \cdots M$ is the jth gene expression value through all the samples, N is total gene number; x_{ij} is the expression value of jth gene in ith sample; $\mathbf{b} = [b_1, \cdots, b_j, \cdots b_N]^T$ is the weight vector of genes. b_j represents the weight of j th gene in the prediction. **n** is the noise, which is not our focus in the paper. The assumption is that X are standardized and response in and **Y** is centered, so that:

$$\sum_{i=1}^{n} y_i = 0 \text{, } \sum_{i=1}^{n} x_{ij} = 0 \text{, and} \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for } j = 1, \ldots, p$$

2) Network-constrain regularization analysis

In order to estimate vector **b** in the regression model in (1), several studies have developed many regularized algorithms for addressing high-dimensionality genomic data, especially some regularized methods such as LASSO, LA-SEN and LARS. A modified LA-SEN method, called Network-constrain regularization analysis is employed since it has better performance than other regularized methods [7]. This method tries to smooth the efficient regression coefficients and incorporate gene interaction network information. In addition, the method can lead to desirable grouping effects for predictors that are correlated or linked on the network. Firstly, the method induced a Laplacian matrix **I** describing gene interaction network, with uvth element defined as:

$$I(u,v)=\begin{cases}1-w(u,v)/d_u & \text{if } u=v \text{ and } d_u\neq 0\\ -w(u,v)/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are ajacent}\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

where $w(u,v)$ denotes the weight of edge between linked genes u and v; $d_v=\sum_{u\sim v} w(u,v)$; $\sum_{u\sim v}$ denotes the sum over all unordered pairs $\{u,v\}$ for which u and v are adjacent genes on the network. Next, the network-constrained regularization criterion is defined as

$$L(\lambda_1,\lambda_2,\mathbf{b})=|\mathbf{Y}-\mathbf{Xb}|^2+\lambda_2\mathbf{b}^T\mathbf{Ib}+\lambda_1|\mathbf{b}|_1 \tag{3}$$

where $|b|_1=\sum_{j=1}^{M}|b_j|$.

And then the method estimates weight vector $\mathbf{b}=[b_1,b_2,\ldots b_j,\ldots b_N]'$ by satisfying

$$\mathbf{b}=\arg\min_b\{L(\lambda_1,\lambda_2,\mathbf{b})\} \tag{4}$$

λ_1 and λ_2 are non-negative turning parameters.

Minimizing procedure in (4) is equivalent to solving a LASSO-type optimization problem, and enjoying the computational advantage of the LASSO. A detailed description of this Network-constrained regularization analysis method is available in its original publication [7].

Before application, scale transformation was required for the microarray data to make the response centered and predictors standardized according to assumption. In the paper, gene interaction network is derived from KEGG database by commercial software Ingenuity and open software Cytoscape. Ingenuity defines gene interactions according to stacks of literatures. One plug-in function in Cytoscape provide PPI network database. Gene interaction network is mapped from the protein-protein interaction network. In addition, parameters λ_1 and λ_2 involved in this criterion are determined by a 10-fold cross validation (CV) as described in original publication of 'LA-SEN' [8].

After applying network-constrained regularization analysis, the vector **b** in equation (1) can be determined. A group of genes whose coefficients in vector **b** are nonzero are detected and considered as important genes contributing to AD development.

3) Protein pathway identification

Many studies have contributed to protein signal transduction pathway research. Most of them identify separate linear pathways and then heuristically assemble them into a signaling network. In our work, a computational method based on an integer linear programming model was chosen to detect protein pathway. We followed the steps of Zhao's work since it treats a signaling network as a whole entity rather than heuristically ranking and assembling individual linear pathways [6].

The method integrates PPI database and gene expression data to uncover signal transduction networks (STNs) by formulate the problem as an integer linear programming (ILP) model. Given starting proteins, usually membrane protein, ending proteins, and some transduction factor proteins, the ILP model uncover the signal pathway by minimize an energy function:

$$S=-\sum_{i=1}^{|V|}\sum_{j=1}^{|V|}w_{ij}y_{ij}+\lambda\sum_{i=1}^{|V|}\sum_{j=1}^{|V|}y_{ij}\ , \tag{5}$$

where w_{ij} is the weight of edge between protein i and j in the PPI network; y_{ij} represents if the interaction between protein *i* and *j* is part of the signal pathway; and $v_i\in V$ represents the proteins in PPI network. Detail description of the method could also be found in its original publication. For application, PPI network database has been described in regression model solution; gene expression data has been described in the material section.

All the genes selected from regression model are mapped into their corresponding proteins, and then these proteins are treated as transduction factors in constructing proteins network.

4. Results and Discussion

All the algorithms in the method section were implemented by Matlab and R. After regression model, 23 genes were selected for patients' model as feature genes and were involved into further analysis, as shown in Table.I. After ILP modeling, a subset of proteins were selected from the PPI network, signal transduction pathway was also constructed among these proteins, and then was drawn by Cytoscape 3.0. Figure.1 shows the results derived from proposed method. Implication of protein signal network, e.g. function module identification, is described in discussion section below. In order to look for certain modules contained in the proteins signal pathway which are essential to PD studies, a web-based software 'WebGestalt' was utilized to analyze the results in the paper. This open software incorporates information from different public resources (e.g. KEGG, Gene Ontology) in order to provide an easy way to make sense out of a sets of genes (http://bioinfo.vanderbilt.edu/webgestalt) .

Table 1 genes identified from linear regression model

SNCA	LRRK2	PARK7	PINK1	UCH-L1
ATP13A2	PARK6	GBA	NR4A2	CHRNB2
MAPK	PLBD2	GRIA3	CLDN1	EIF4E
HTRA2	DJ-1	CASP3	UBA1	PRKN
AK7	MPP5	PARK3		

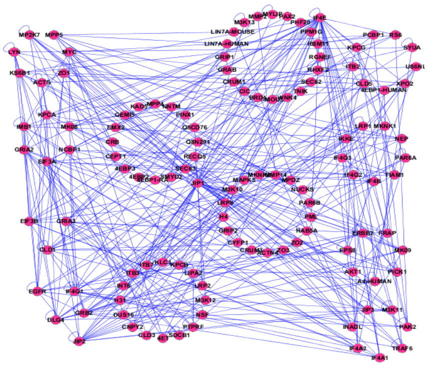

Fig. 1 Protein Pathway identification from ILP model

In order to implement the results, the protein pathways were mapped into its coding gene pathway network. Genes were classified into different types of categories based on three GO term function definitions based on the 'Build GO Tree' function in WebGestalt. Categories titles based on different definitions are shown in Table.Ⅱ.

Table 2 GO tree analhysis for the key genes

Biological Process	**Molecular Function**	**cellular component**
Cellular process	Binding	Cell part
Development	Catalytic activity	Envelope
Physiological process	Enzyme regulator activity	Organele
Regulation of biological process	Transporter activity	Protein complex

Also, important pathways were also identified from gene network based on the function in WebGestalt. Table.Ⅲ organized the important pathway from PD patients' model.

Table 3 Pathway identified from ILP model

MAPK signaling pathway	Neuroactive ligand-receptor interaction
Insulin signaling	Cell adhesion pathway
Leukocyte transendothelial migrational	Neurodegenerative disorders
EPEC pathway	Tight junction
Starch and sucrose metabolism	mTor signaling pathway
Glycan structures-degradation	Long term depression

Changes in MAPK-signalling may thus be common to PD pathophysiology, regardless of aetiology. Such changes may also be shown in blood samples during the preclinical stage of LRRK2-associated PD, which could be particularly important for the development of neuroprotective strategies to delay onset, or slow progression of PD [9].

Leucine-rich repeat kinase 2 (LRRK2) is a large, complex, multidomain protein containing kinase and GTPase enzymatic activities and multiple protein-protein interaction domains. Mutations in the gene encoding LRRK2 have been linked recently with autosomal-dominant Parkinsonism that is clinically indistinguishable from typical, idiopathic, late-onset PD [10]. Mutations linked to autosomal dominant forms of Parkinson's disease result in amino acid changes throughout the protein and alterations in both its enzymatic properties and interactions [11].

PD and AD are two most famous mental disorders in neurodegenerative disease. Compared with results from Ref [12], it is clear that the PD shares some pathways with Alzheimer's disease (AD), including insulin signaling pathway, MAPK signaling pathway, and pathway of neuroactive ligand-receptor interaction.

It has been studied in several literatures that Insulin pathway Study in Ref [13] proposed a cell model of PD, to validate the hypothesized that PD cybrids would exhibit deficits in insulin signaling. It also has concluded that neurodegenerative diseases, including Parkinson's disease (PD), have been linked to Type 2 Diabetes (T2D), and impaired mitochondrial function occurs in both diseases [13].

5. Conclusions

The paper presented a computational method to identify important genes and detect protein transduction pathway for PD studies. The presented method integrates information from PPI network and gene expression data and is implemented to uncover pathway among key proteins of PD, including SNCA PARK7 and so on. Some cancer related modules are identified according to the results. The proposed method can be widely used in better understanding other biological processes for its simple algorithm and efficient.

6. References

[1] G.Sutheland, Nicholas Matigian, Alistair Chalk, Matthew Anderson, and et.al. 'A cross study transcriptional analysis of Parkinson's disease.' Plosone, 2009, vol 4, pp.e4955-4962

[2] M.Renee, and H.Federoff. 'Microarrays in Parkinson's Disease: A Systematic Approach.' Neurotherapeutics, 2006, vol 3, pp.319-326.

[3] H.Suzuki. 'Protein-protein interactions in the mammanlian brain.'The journal of Physiology. 2006,vol 575, pp.373-377

[4] V. Limviphuvadh, S. Tanaka, S. Goto, K. Ueda and M. Kanehisa, The commonality of protein interaction networks determined in neurodegenerative disorders (ndds). Bioinformatics 2007, vol. 23, 2129-2138C

[5] X. Zhao, R.Wang, L.Chen and K.Aihara.'Uncovering signal transduction networks from high-throughput data by integer linear programming.' Nucleic Acids Research, 2008, vol.36, pp.e48

[6] TG.Lesnick, S.Papapetropoulos, D. Mash, J. Ffrench-Mullen et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet 2007, vol 3, pp. e98-e106.

[7] C. Li and H. Li, Network-constrained regularization and variable selection for analysis of genomic data, Bioinformatics, 2008, vol.24, pp. 1175-1182

[8] Zou, H Trevor. Regularization and variable selection via the elastic net. J.R.Statist.Soc.B. 2005, vol 67, pp.301-320.

[9] LR White, M. Toft, S.Kvam, MJ. Farrer, JO. Aasly. 'MAPK-pathway activity, Lrrk2 G2019S, and Parkinson's disease' J Neurosci Res. 2007, vol.85, pp.1288-94.

[10] VS.Anand, SP and Braithwaite. 'LRRK2 in Parkinsons's disease: biochemical functions.' FEBS J. 2009, vol. 276, pp.6428-35. LRRK2 in Parkinson's disease: biochemical functions.

[11] I. Mata, WJ. Wedemeyer, MJ. Farrer, JP.Taylor, KA. Gallo. 'LRRK2 in Parkinson's disease: protein domains and functional insights.' Trends Neurosci. 2006, vol.29, pp.286-93.

[12] Y. Huang, X.Zhou, Z.Xia, et.al. 'An image based system biology approach for Alzheimer's disease pathway analysis,' IEEE LISSA 2009, p128-132, Bethesda, U.S.A.

[13] J.Morris, A. Esteves, G.Bomhoff, R. Swerdlow, J. Stanford and P. Geiger. ' Investigation of Insulin Signaling in Parkinson's Disease Cytoplasmic Hybrid Cells.' FASEB J. 2010, vol. 21,pp.1053-6