Expert Systems With Applications 85 (2017) 249–260

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Systematic identification of differential gene network to elucidate Alzheimer's disease

Chihyun Park[a], Youngmi Yoon[b], Min Oh[b,c], Seok Jong Yu[a], Jaegyoon Ahn[d,*]

[a] Biomedical HPC Technology Research Center, Korean Institute of Science and Technology Information, Daejeon, Republic of Korea
[b] Department of Computer Engineering, Gachon University, Seongnam-si, Gyeonggi-do, Republic of Korea
[c] Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[d] Department of Computer Science and Engineering, Incheon University, Incheon, Republic of Korea

ARTICLE INFO

Article history:
Received 7 October 2016
Revised 27 April 2017
Accepted 16 May 2017
Available online 17 May 2017

Keywords:
Genetic network construction
Omics data integration
Network analysis
Alzheimer's disease

ABSTRACT

Alzheimer's disease (AD) is a genetically complex neurodegenerative diseases and its pathological mechanism has not been fully discovered. The mechanism of AD can be inferred by elucidating how molecular entities are interacting on the pathway level and how some pathways collectively influence the occurrence of the disease. Such an analysis is considerably complex and cannot be manually performed by experts. It can be solved by integrating huge heterogeneous dataset and systematically building an intelligent system which model molecular network and analyze the causality. Here, we present a novel method to construct an optimized AD-specific differential gene network by integrating a high-confidence interactome and gene expression data. In order to consider an epigenetic factor, we identified differentially methylated genes in AD and the results were projected on the network for mechanism analysis. Through diverse topological analysis and functional enrichment tests, we experimentally demonstrated that the several potential genes and sub networks were significantly related with AD and they could be used to elucidate the molecular mechanism. Taken the experimental results and literature studies together, we newly discovered that ribosomal process-related genes and DNA methylation might play an important role in AD. The proposed system is applicable not only to AD but also to various complex genetic disease models that require new molecular mechanism analysis based on network.

1. Introduction

Alzheimer's disease (AD) is the most prevalent neurodegenerative disease that mostly affects the aging population (Santiago and Potashkin, 2014). Globally, the number of AD patients is expected to increase vastly, posing a huge threat to public health (Brookmeyer, Johnson, Ziegier-Graham, & Arrighi, 2007). AD has been characterized by the accumulation of amyloid-beta (Ab) plaques and protein tau in the form of neurofibrillary tangles (Santiago and Potashkin, 2014). Significant efforts have been made to date to identify potential disease mechanisms that could elucidate these observations (Zhang et al., 2013). Several studies have focused on biological risk factors with compounds targeting the amyloid-beta pathway (LaFerla, Green, & Oddo, 2007) and genetic susceptibility playing a critical role in the underlying pathophysiology of late-onset AD (Bertram, Lill, & Tanzi, 2010). It was revealed in mouse models that rare mutations in APP, PSEN1, and PSEN2 can have an effect on downstream genes and consequently cause AD (Bertram et al., 2010). Despite these efforts, the identified common and low-frequency variants to the heritability of AD are very small, and large fractions of the genetic variance remain hidden (Zhang et al., 2013). Genetic risk factors are still important, but this approach cannot elucidate the complexity of AD pathogenesis. The causal mechanisms of disease occurrence and progression have not been precisely revealed. For this reason, the only treatments available are those to relieve symptoms (Zhang et al., 2013).

For comprehensive understanding of the causal mechanisms of complex diseases such as cancer, there have been many attempts to apply network approaches (Ahn, Yoon, Park, Shin, & Park, 2011; Chen et al., 2008). In a network analysis, it is possible to investigate regulatory relationships among genes, how genes share biological functions, and how disease susceptibility propagates their information through interactions. Since the complex diseases generally do not originate from variations in a single gene (Santiago and Potashkin, 2014), it is important to consider a simultaneous alteration in complex biological pathways. Network modeling is appropriate to perform this function.

* Corresponding author.
E-mail addresses: chihyun.park@kisti.re.kr (C. Park), ymyoon@gachon.ac.kr (Y. Yoon), minoh@vt.edu (M. Oh), codegen@kisti.re.kr (S.J. Yu), jgahn@inu.ac.kr (J. Ahn).

http://dx.doi.org/10.1016/j.eswa.2017.05.042
0957-4174/

Recently, several network-based approaches have been proposed to uncover mechanisms underlying AD (Gaiteri et al., 2016). In order to construct networks, integrative approaches with multiple data sources such as combining gene expression and interactomes or inferring regulatory interactions with a weighted gene co-expression network from gene expression profiles were the representative and widely used methods until now (Aubry et al., 2014; Liang et al., 2012; Schulz et al., 2012; Talwar et al., 2014). After the construction of networks, topological analysis was performed in most methods in order to identify functional modules or informative sub-networks that provide evidence for upstream regulators. To enrich the networks, several studies attempted to integrate heterogeneous nodes and edges such as single-nucleotide polymorphisms (SNPs) and microRNA (miRNA) as an anchor node to investigate causes of AD (Zhang et al., 2013; Chandrasekaran & Bonchev, 2016).

Despite the previous work, further efforts are still needed to construct the integrated differential gene networks that represent the functional architecture of diseases (Hsu, Juan, & Huang, 2015; Rhinn et al., 2012). To build integrative networks, both gene expression profiles and interaction data have been used together (Ahn et al., 2011). Because the underlying interactions used in this process are condition independent, the final networks that formulate disease specificity or a certain phenotypic condition might include false interactions even if expression profiles are used. More specifically, many false interactions, which do not actually occur even though the expression pattern is differential in the disease compared to normal status, might be included because the interactome database was curated without considering specific disease conditions. Therefore, the network should be optimized by removing false edges.

Furthermore, additional sources such as epigenetic factors or already known information about the target disease can be included in the network in order to enrich the following analysis. Of many additional sources, it has been reported that epigenetic factors such as DNA methylation are significantly associated with neurodegenerative diseases (Coppedè, 2014; Jakovcevski & Akbarian, 2012). According to these studies, it has been strongly supposed that gene-specific epigenetic changes contribute to the onset and progression of neurodegenerative disorders, including in adult patients, through dynamic regulation of expression levels. There have been several attempts to develop epigenetic drugs that alleviate AD in animal models (Coppedè, 2014). In addition, recent studies have found that several genes that alter their RNA expression levels by DNA methylation may have a role in human AD. In conclusion, there have been continuous attempts to study AD susceptibility with epigenetic markers. However, there is little effort to systemically investigate functional mechanisms or the role of DNA methylation in AD on the gene network level (Jager et al., 2014). In other words, the connectivity among methylated genes and disease-related genes has been as yet insufficiently disclosed to explain the mechanism of AD.

In this study, we propose a novel method to construct AD specific differential gene network from publicly available datasets. To enhance the significance of the mechanism study, external knowledge such as DNA methylation profiles is integrated into the network and topological analysis is performed. This study utilizes heterogeneous set of data to get some patterns or rules in the context of the genetic network, with help of graph data mining techniques. Those patterns or rules can explain several aspect of AD, and medical expert and intelligent systems focus on finding those kinds of patterns for various diseases including cancer.

2. Methods

2.1. Data description

Four types of heterogeneous datasets were used to construct a network. The list of them is as follows: gene expression profiles, interactome database, pathway database, and DNA methylation profiles. The first two datasets were used to discriminate informative interactions, and the remaining ones were used as support jn analyzing disease networks.

We downloaded recently published and large sample-sized gene expression profiles. We integrated two independent array-based expression profiles that experimented with the same platform (Narayanan, 2014). We only used data obtained from the prefrontal cortex. Their GEO accession numbers were GSE33000 and GSE44770. They were composed of 467 samples (normal: 157, AD: 310) and 229 samples (normal: 100, AD: 129), respectively. As shown in Supporting Fig. 1, two expression profiles for each condition, normal and AD, are similar to each other. There were only a few sets of gene expression data publicly usable and large-sized, with the exception of these two sets of data.

An interactome database was used to identify connectivity between two genes. We used a recently published human protein interaction dataset (Rolland et al., 2014) in company with a widely used genetic interaction dataset, humanNet (Lee, Blom, Wang, Shim, & Marcotte, 2011). The protein interaction dataset we used consisted of 23,233 high-confidence interactions compiled from systematic screening with high-throughput yeast two-hybrid systems and validated using biological assays. Throughout the manuscript, we will call this interaction dataset bPPI (biophysical Protein-Protein Interaction). HumanNet is known to be appropriate to find disease-associated genes through the guilt-by-association approach. The dataset was composed of more than four hundred thousand genetic interactions including scores. To obtain the more accurate and meaningful interactions, we combined bPPI and the top scoring interactions of humanNet.

Pathways from Reactome (Croft et al., 2014) and The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) databases were used to emphasize biologically informative interactions. Finally, the recently published DNA methylation profiles corresponding to prefrontal cortex region was used to investigate how differentially methylated genes (DMGs) affect AD. The GEO accession number was GSE80970 and it composed of 142 samples (normal: 68, AD: 74). Furthermore, several DMGs which have been found in previous papers (Bakulski et al., 2012; Sanchez-Mut et al., 2015) were used together.

2.2. System overview

The proposed method consisted of two phases. Fig. 1 shows the detailed workflow of the proposed method. The first phase is to extract differentially expressed gene pairs. In this phase, the degree of differentially expression is measured for gene pairs that have already been found to have a relationship from the interactome database. And then, the statistical tests were performed to determine the optimal parameters which reflect AD specificity. Secondly, pathway information is integrated in the constructed network and AD related information was also integrated as a node properties. Finally, topological analysis and functional enrichment test were conducted to elucidate AD relativeness.

2.3. Identification of informative gene pairs

To measure the degree of differentially expressed patterns for normal versus AD tissue between two genes, we adopted the scoring scheme that outperformed others to identify informative