ELSEVIER

Pathophysiology 22 (2015) 15–29

ISP
PATHOPHYSIOLOGY

www.elsevier.com/locate/pathophys

Analysis of Parkinson's disease pathophysiology using an integrated genomics-bioinformatics approach

Li M. Fu[a,*], Katherine A. Fu[b]

[a] *Biomedical Engineering Department, AHMC Healthcare, Los Angeles, CA, USA*
[b] *Keck School of Medicine, University of Southern California, Los Angeles, CA, USA*

Received 29 April 2014; received in revised form 11 October 2014; accepted 20 October 2014

Abstract

The pathogenesis and pathophysiology of a disease determine how it should be diagnosed and treated. Yet, understanding the cause and mechanisms of progression often requires intensive human efforts, especially for diseases with complex etiology. The latest genomic technology coupled with advanced, large-scale data analysis in the field known as bioinformatics has promised a high-throughput approach that can quickly identify disease-affected genes and pathways by examining tissue samples collected from patients and control subjects. Furthermore, significant biological themes indicative of genomic events can be recognized on the basis of affected genes. However, given identified biological themes, it is not clear how to organize genomic events to arrive at a coherent pathophysiological explanation about the disease. To address this important issue, we have developed an innovative method named "Expression Data Up-Stream Analysis" (EDUSA) that can perform a bioinformatics analysis to identify and rank upstream processes effectively. We applied it to Parkinson's disease (PD) using a genomic data set available at a public data repository known as Gene Expression Omnibus (GEO). In this study, disease-affected genes were identified using GEO2R software, and disease-pertinent processes were identified using EASE software. Then the EDUSA program was used to determine the upstream versus downstream hierarchy of the processes. The results confirmed the current misfolded protein theory about the pathogenesis of PD, and provided new insights as well. Particularly, our program discovered that RNA (ribonucleic acid) metabolism pathology was a potential cause of PD, which in fact, is an emerging theory of neurodegenerative disorders. In addition, it was found that the dysfunction of the transport system seemed to occur in the early phase of neurodegeneration, whereas mitochondrial dysfunction appeared at a later stage. Using this methodology, we have demonstrated how to determine the stages of disease development with single-point data collection.

Keywords: Functional genomics; Gene expression; Pathophysiology; Bioinformatics; Parkinson's disease

1. Introduction

Parkinson's disease (PD) is the second most frequent neurodegenerative disorder after Alzheimer's disease. Clinically, PD is characterized by bradykinesia, rigidity, and resting tremor, and the disease becomes more debilitating as it progresses. It is well known that the disease is caused by the loss of neurons in the substantia nigra pars compacta (SNpc), which in turns leads to dopamine deficiency in the nigrostriatal pathway [1]. As a pathological hallmark of PD, Lewy bodies are abnormal aggregates of proteins found inside nerve cells, containing mainly α-synuclein [2]. A small percentage of PD occurs in a familial or inherited form, while the majority of cases have no genetic linkage and are referred to as sporadic PD. Several genes causing PD have been found, notably, genes coding for α-synuclein, parkin, LRRK2, PINK1, and DJ-1 [3]. Mutations in some of these genes are found in familial PD and occasionally in sporadic PD.

The study of PD genes is essential for understanding inherited PD, but it also sheds light on sporadic PD because both forms of PD share some common pathophysiology, as illustrated by the fact that α-synuclein is a major component

[*] Corresponding author at: Biomedical Engineering Department, AHMC Healthcare, Alhambra, CA 91801, United States. Tel.: +1 9493314196.
E-mail address: lifu.usa@gmail.com (L.M. Fu).

http://dx.doi.org/10.1016/j.pathophys.2014.10.002
0928-4680/© 2014 Elsevier B.V. All rights reserved.

of the Lewy Body in both forms of PD, but mutations in the α-synuclein gene have not been found in sporadic PD [1]. The etiology and pathogenesis of sporadic PD is complex and cannot be explained by mutations in a handful of genes; its disease mechanisms are not entirely clear, though some genetic polymorphisms and environmental factors have been identified. In the post-genomic era, genome-wide gene expression profiling analysis based on microarrays identified many genes and pathways in connection with the disease pathophysiology [4,5].

The technology of next-generation sequencing (NGS) emerged in the last decade, and has found important applications in genomic research [6]. In gene expression profiling, it is believed that with digital profiling, NGS exhibits higher sensitivity, specificity, accuracy and a better dynamic range than the current microarray technology, though microarray analysis remains popular thanks to the support of strong bioinformatics. For research purposes, a more organized microarray database better meets the needs of researchers.

High-throughput genomic technology like microarrays has generated a large amount of data, which creates an enormous interest in data mining and modeling. To process the genomic data and understand the embedded information demands complex computational analysis in a field known as bioinformatics. The interpretation of genes one-at-a-time fails to capture the global meaning of the data. An important type of information structure underlying gene expression data is the gene regulatory network (GRN) that can test and infer relationships among biological concepts from the molecular to system levels.

Various approaches have been developed to determine the gene network based on gene expression data, such as the Boolean network [7], the differential equation, cluster analysis [8], the Bayesian network [9], regression [10] and the graphical Gaussian model [11]. Computational methods for constructing gene networks can be broadly divided into two categories: models with discrete variables (Boolean and probabilistic networks, etc.) and models with continuous variables (differential equations and Markov models, etc.) [12,13]. The system dynamics of GRNs can be better captured using time-series data (data with the time parameter included), which can be modeled through methods such as dynamic Bayesian networks and differential equations [14]. Notice that the causal modeling of gene relationships entails temporal precedence information underlying the expression data.

Graph theoretical models have been applied to both discrete and continuous models mainly for describing the topology of a GRN. Of particular interest is the directed acyclic graph (DAG), which means a directed graph (with direction specified in the link between two nodes) with no cycles. The DAG is important in the context of gene networks because of its role in causal modeling [10]. The Bayesian network is a DAG where each node represents a variable and each arc indicates conditional dependency (associated with a conditional probability) so that, given its parent, each variable is independent of its non-descendants. This network was initially demonstrated on the gene expression data of yeast cell cycles [9].

Functional analysis of large gene lists is an issue closely related to the modeling of gene interactions and regulations. For the same experiment, the list of genes identified as significantly expressed (up- or down-regulated) often varies with the gene-selection method. This observation motivates the theme-based approach that abstracts enriched biological themes from a large list of genes derived from genomic experiments [15–17]. This approach produces similar results despite different gene lists computed by different methods based on the same experimental data. The gene ontology (GO) has been developed to maintain a unified, controlled vocabulary as well as annotations for genes and gene products [18]. An important application of the GO is to perform the enrichment analysis that identifies which GO categories are over-represented on the gene sets of interest. Along this line, the use of GO graphs for coding the relationships among annotations was shown to further improve the enrichment analysis [19].

A central issue for building a pathophysiological theory about a disease is the organization of the involved pathophysiological processes according to their chronological development. Nevertheless, in practical circumstances, samples for genomic analysis are often samples collected at a single point in time, for example, from postmortem tissue. It would be questionable to construct a complete pathophysiology from such limited genomic data. Under the circumstances, we found that the relative developmental stages of the pathophysiological processes could be recognized if the disease progresses in a tree-like divergent manner. To perform this analysis, we devised an innovative computational method called the “Expression Data Up-Stream Analysis” (EDUSA). The basic principle behind this method was proven from statistical and engineering perspectives and applied to sporadic PD. Our results agreed with the best-known theory about PD, answered some controversial questions, and provided new insights into its pathophysiology.

2. Materials and methods

2.1. Identification of disease-related genes

In this study, the genomic data concerning PD were obtained from the NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/) in the public domain with the series accession number GSE8397 and the platform GPL96 (Affymetrix Human Genome U133A Array).[1] This data set consisted of data from 47 tissues samples, including 29 PD samples

[1] Two best known studies concerning the gene expression profiling of PD were published by Moran et al. [5] and Zhang et al. [4], respectively. An