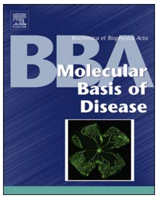
BBA - Molecular Basis of Disease xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

BBA - Molecular Basis of Disease

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



Identification of genes related to proliferative diabetic retinopathy through RWR algorithm based on protein–protein interaction network☆

Jian Zhang[a,b,c,1], Yan Suo[a,b,c,1], Min Liu[d], Xun Xu[a,b,c,*]

[a] Department of Ophthalmology, Shanghai General Hospital, School of Medicine, Shanghai JiaoTong University, Shanghai, China
[b] Shanghai Key Laboratory of Fundus Disease, Shanghai, China
[c] Shanghai Engineering Center for Visual Science and Photomedicine, Shanghai, China
[d] College of Information Engineering, Shanghai Maritime University, Shanghai, China

ARTICLE INFO

Keywords:
Proliferative diabetic retinopathy
Random walk with restart
Protein–protein interaction network

ABSTRACT

Proliferative diabetic retinopathy (PDR) is one of the most common complications of diabetes and can lead to blindness. Proteomic studies have provided insight into the pathogenesis of PDR and a series of PDR-related genes has been identified but are far from fully characterized because the experimental methods are expensive and time consuming. In our previous study, we successfully identified 35 candidate PDR-related genes through the shortest-path algorithm. In the current study, we developed a computational method using the random walk with restart (RWR) algorithm and the protein–protein interaction (PPI) network to identify potential PDR-related genes. After some possible genes were obtained by the RWR algorithm, a three-stage filtration strategy, which includes the permutation test, interaction test and enrichment test, was applied to exclude potential false positives caused by the structure of PPI network, the poor interaction strength, and the limited similarity on gene ontology (GO) terms and biological pathways. As a result, 36 candidate genes were discovered by the method which was different from the 35 genes reported in our previous study. A literature review showed that 21 of these 36 genes are supported by previous experiments. These findings suggest the robustness and complementary effects of both our efforts using different computational methods, thus providing an alternative method to study PDR pathogenesis.

1. Introduction

Diabetic retinopathy (DR) is an eye disease that is the most common complication of diabetes with the potential of causing blindness [1]. Most of the patients with diabetes will suffer from this disease at some extent [2]. Clinically, DR can be classified into non-proliferative DR (NPDR) and proliferative DR (PDR) [3]. In comparison with NPDR, PDR is more severe in threatening the human sight and can be characterized by the formation of retinal neovascularization.

To characterize the pathogenesis of PDR, proteomics studies have been applied and have uncovered a series of genes involved in the disease. For example, a LC-MS/MS-based study on a large cohort of patients discovered the upregulation of proteins involved in the recruitment of chemokines in the inflammatory response caused by PDR and oxidative stress [4]. Angiogenesis and vascular permeability were also regarded as PDR-associated progresses and the related genes were identified previously [5]. Besides, angiotensinogen, neuroserpin, extracellular superoxide dismutase, and pigment epithelium-derived factor were also identified as PDR-related genes by previous proteomic studies [6,7]. Although these studies together with many others have successfully revealed several PDR-related genes, the experimental methods, such as MS, 2D-DIGE, and SDS-PAGE, are time- and cost-consuming, which limit the discovery power and in turn promote the development of in silico prediction methods on identifying the PDR-related genes.

In recent years, several investigators study the pathogenesis of different human diseases with the help of advanced computational methods [8–11]. Network approach is deemed as a type of powerful methods for investigating various human diseases [9,12–18]. Several studies used the protein–protein interaction (PPI) networks to give

Abbreviations: PPI, protein–protein interaction; RWR, the random walk with restart; SP, shortest path; MES, maximum enrichment score; MIS, maximum interaction score; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes

☆ This article is part of a Special Issue entitled: Accelerating Precision Medicine through Genetic and Genomic Big Data Analysis edited by Yudong Cai & Tao Huang.

* Corresponding author at: Department of Ophthalmology, Shanghai General Hospital, School of Medicine, Shanghai JiaoTong University, Shanghai, China.
E-mail address: drxuxun@tom.com (X. Xu).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.bbadis.2017.11.017
Received 3 October 2017; Received in revised form 15 November 2017; Accepted 25 November 2017

investigations. The advances of building PPI networks allow us to predict novel disease genes based on the connections between target genes and known disease genes [15,17,19–22]. Previously, we proposed a computational method using the shortest-path (SP) algorithm and successfully identified 35 PDR-related genes with nine of them supported by previous experiments [23]. However, this algorithm was recently found to yield complementary results with others, such as random walk with restart (RWR) algorithm [24,25], in identifying novel disease genes [17]; this result suggests that more PDR-related genes can be discovered using different computational methods.

In this study, we applied another classic algorithm, the RWR algorithm [24,25], and developed a three-stage filtration strategy to identify novel PDR-related genes based on a PPI network and known PDR-related genes curated previously [23]. Thirty-six genes associated with PDR pathogenesis were yielded by the method, most of which are novel compared with our previous predictions [23]. A detailed analysis by searching experimental evidence from the literature demonstrated that 21 of them have supports of their roles in PDR pathogenesis, suggesting the robustness of the novel method. This study provides a new insight into further investigations of the molecular basis of PDR pathogenesis.

2. Materials and methods

2.1. Dataset

Thirty-four known PDR-related genes were curated in our previous study [23]; these genes were accessed from MalaCards (http://www.malacards.org/) [26] and a published review article [27]. The detailed information of these 34 genes can be found in Table 1 of our previous study [23]. In this study, we would use these genes to mine novel potential PDR-related genes.

2.2. PPI network

Human proteins play diverse functional roles through the formation of complexes or polymers by interacting with each other, and those proteins with interactions often share similar functions [28–31]. This knowledge has further been used to identify new disease-related genes [15–18,23,32] and predict protein functions [33–35]. In this study, we also constructed a PPI network for inferring novel PDR-related genes.

The human PPI information was retrieved from the STRING database (Version 10.0, http://string-db.org/) [36], consisting of 19,247 proteins and 4,274,001 interactions (direct (physical) and indirect (functional) interactions). These PPIs were obtained from five sources: (1) genomic context predictions, (2) high-throughput lab experiments, (3) (conserved) co-expression, (4) automated text mining, and (5) previous Knowledge in databases. These PPIs can measure the associations between proteins from their several aspects. Using this type of PPI information can provide more opportunities to mine useful novel knowledge.

Using the 19,247 proteins as nodes and 4,274,001 interactions as edges, a PPI network can be constructed. Given that the STRING database also assigns an interaction score ranging from 150 to 999 to each PPI, these interaction scores were added to the PPI network as the weights of the corresponding edges. For proteins p_a and p_b, the score of the interaction between them was formulated as $S(p_a, p_b)$. For convenience, the PPI network was denoted as G in the following text.

2.3. RWR algorithm

The RWR algorithm, as one of the classic ranking algorithms [24,25], always simulates a random walker starting from a seed node or several seed nodes and walking on a constructed or natural network. In this study, 34 Ensembl IDs of known PDR-related genes mentioned in Section 2.1 were picked up as seed nodes in the RWR algorithm. Before executing the algorithm, an initialization vector, P_0, was constructed and consisted of 19,247 components. Each component of the vector represented the probability that the corresponding node would be a PDR-related gene. Thus, in P_0, the components corresponding to 34 Ensembl IDs of PDR-related genes were set to 1/34, and the other components were set to zero. The RWR algorithm updated the vector in a reasonable way. In detail, let P_i be the vector after the ith updating procedure was complete. This vector was updated as follows:

$$P_{i+1} = (1 - r)A^T P_i + rP_0 \tag{1}$$

where A was the column-wise normalized adjacency matrix of the PPI network G and r was the probability of returning to the seed nodes, indicating the importance of the seen nodes (we tried 0.8 in this study). When the probability vector became stable, which was measured by using $\|P_{i\,+\,1} - P_i\|_{L_1} < 10^{-6}$, the RWR algorithm stopped and outputted $P_{i\,+\,1}$ as the result. Each component in this vector indicated the probability of a node being a PDR-related gene. A gene that was assigned a high probability was more likely to be a potential PDR-related gene. We set the threshold 10^{-5} for selecting potential PDR-related genes, i.e., genes that were assigned probabilities higher than 10^{-5} were picked up for further investigations.

2.4. Three-stage filtering strategy

As mentioned in Section 2.3, some potential PDR-related genes can be extracted by using the RWR algorithm. Among the obtained genes, having all of them as PDR-related genes is impossible. Some false positives and those with low likelihood of being PDR-related genes may also be included. The filtering strategy, which can filter out most possible genes, is necessary. In this study, we filtered potential PDR-related genes yielded by using the RWR algorithm from three aspects, which were described as below.

2.4.1. Permutation test

The utility of the RWR algorithm is highly related to the accuracy of PPI network G. Some false positives can be selected due to the structure of the network. To control this type of genes, a permutation test was proposed. In this test, we randomly produced 1000 Ensembl ID sets, each of which contained 34 Ensembl IDs. Then, Ensembl IDs in each set were set as seed nodes in the RWR algorithm, yielding a probability for each node in the network G. After testing all 1000 sets, each gene was assigned 1000 probabilities. For each potential PDR-related gene g yielded by the RWR algorithm, a measurement, regarded as p-value, can be calculated by

$$p - value(g) = \pi/1,\,000 \tag{2}$$

where π represented the number of Ensembl ID sets, which yielded a higher probability of the potential PDR-related gene g than that yielded by the known PDR-related genes. Clearly, a high p-value of a gene g meant that it was not special for known PDR-related genes given that it can be produced by several other gene sets and assigned a higher probability. Accordingly, we should select potential PDR-related genes with low p-values. Given that 0.05 is the standard cutoff value for evaluating statistical significance, this value was set as the threshold of p-value.

2.4.2. Interaction test

The permutation test can only exclude some false positives. This test cannot help us select essential genes. Thus, we further built the interaction test to do that. For each potential PDR-related gene g, another measurement, regarded as maximum interaction score (MIS), was calculated in this test, which was defined by

$$MIS(g) = \max\{S(g, g')\text{: } g' \text{ is a known PDR-related gene}\} \tag{3}$$

where $S(g,g^{'})$ was the interaction score between g and $g^{'}$. As mentioned in Section 2.2, proteins that can interact with each other are more likely to share similar functions. If the scores of interactions are considered,