FEBS Open Bio 5 (2015) 251–256

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

Mining disease genes using integrated protein–protein interaction and gene–gene co-regulation information

Jin Li [a,b,c,*,1], Limei Wang [c,d,1], Maozu Guo [a,*], Ruijie Zhang [c], Qiguo Dai [a], Xiaoyan Liu [a], Chunyu Wang [a], Zhixia Teng [a], Ping Xuan [a], Mingming Zhang [c]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China
[b] School of Life Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China
[c] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang, China
[d] School of Basic Medical Sciences, Harbin Medical University, Harbin, Heilongjiang, China

ARTICLE INFO

Article history:
Received 13 January 2015
Revised 19 March 2015
Accepted 24 March 2015

Keywords:
eQTL
Co-regulation network
Disease gene mining
Protein–protein interaction
Random walk with restart

ABSTRACT

In humans, despite the rapid increase in disease-associated gene discovery, a large proportion of disease-associated genes are still unknown. Many network-based approaches have been used to prioritize disease genes. Many networks, such as the protein–protein interaction (PPI), KEGG, and gene co-expression networks, have been used. Expression quantitative trait loci (eQTLs) have been successfully applied for the determination of genes associated with several diseases. In this study, we constructed an eQTL-based gene–gene co-regulation network (GGCRN) and used it to mine for disease genes. We adopted the random walk with restart (RWR) algorithm to mine for genes associated with Alzheimer disease. Compared to the Human Protein Reference Database (HPRD) PPI network alone, the integrated HPRD PPI and GGCRN networks provided faster convergence and revealed new disease-related genes. Therefore, using the RWR algorithm for integrated PPI and GGCRN is an effective method for disease-associated gene mining.

1. Introduction

In humans, despite the rapid increase in the discovery of disease–associated genes, the molecular basis of many diseases is still known. Even for diseases for which the molecular basis is partially understood, a large proportion of the associated genes are still unknown. The known disease-associated genes have been reported to represent only a very small proportion of the actual number of disease-associated genes [1,2]. Hence, mining for disease genes remains important.

Network-based approaches to human disease have multiple biological and clinical applications [3,4]. Many molecular networks have been constructed experimentally to characterize the physical and/or functional interactions between biomolecules [4,5]. There are many methods for disease gene mining using molecular networks, such as the direct neighborhood [6–13], Shortest path length [13–16], Diffusion kernel [8], random walk with restart [8,9,17], propagation flow [18], and clique backbone [19] methods. The random walk with restart (RWR) method has been reported to have the best performance in terms of precision and recall, while both the random walk and propagation flow methods are superior to the clustering and neighborhood methods [20,21]. The most useful network is the protein–protein interaction (PPI) network [6,8,9]. Some other resources are also used in disease gene mining, such as gene ontology, gene co-expression network, KEGG, structure, and TRANSFAC [7,10,15,16].

Expression quantitative trait loci (eQTLs) analyses of DNA use hundreds of thousands of single-nucleotide polymorphism (SNP) markers that capture human genetic variation [22]. This strategy has been successfully applied to several diseases, such as celiac disease [23], asthma [24] and type 2 diabetes [25]. An eQTL is a locus that regulates a gene expression phenotype [26]. If two genes are regulated by one or more of the same SNPs, they are considered to be co-regulated. Obviously, this co-regulation is only one type of gene interactions. We constructed a gene–gene co-regulation

Abbreviations: AD, Alzheimer disease; eQTLs, expression quantitative trait loci; GGCRN, gene–gene co-regulation network; HPRD, Human Protein Reference Database; KEGG, Kyoto Encyclopedia of Genes and Genomes; PPI, protein–protein interaction; RWR, random walk with restart; SNP, single-nucleotide polymorphism

* Corresponding authors at: School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China. Tel./fax: +86 451 86667543 (J. Li). Tel./fax: +86 451 86402407 (M. Guo).
E-mail addresses: lijin@hrbmu.edu.cn (J. Li), maozuguo@hit.edu.cn (M. Guo).
[1] Co-first authors.

http://dx.doi.org/10.1016/j.fob.2015.03.011
2211-5463/

network (GGCRN) using eQTL data and believe that it will be useful for disease gene mining.

In this study, we developed a GGCRN, integrated it with the PPI network, and used the RWR method to mine for candidate disease genes. Using Alzheimer disease (AD) as an example, we demonstrated that this newly developed GGCRN is an effective resource for disease gene mining.

2. Materials and methods

2.1. Materials

2.1.1. Protein–protein interaction data

The Human Protein Reference Database (HPRD) describes interaction networks in the human proteome [27]. All information in the HPRD has been manually extracted from the literature by expert biologists who read, interpreted and analyzed the published data. For this study, we used HPRD, release 9, which contains 38,989 protein–protein interactions among 9605 proteins.

2.1.2. EQTL data

We used human brain tissue data for this disease gene mining study. The data were obtained from a series of 193 neuropathologically normal human brain samples using the Affymetrix GeneChip Human Mapping 500 K Array Set and Illumina HumanRefseq-8 Expression BeadChip platforms [28]. The eQTLs were determined by Matrix eQTL [29]. In this study, the cis-eQTL definition was a SNP within the gene body +1 Mb up/down stream of the gene body. We calculated cis-eQTLs and trans-eQTLs and performed FDR adjustment (q value < 0.1) separately; then we combined the cis-eQTLs and trans-eQTLs. Finally, we obtained 25,866 significant SNP-gene association pairs of 3709 genes. The results can be downloaded from the seeQTL database [30].

2.1.3. AD-related genes

Online Mendelian Inheritance in Man (OMIM) is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. AD is classified as a neurodegenerative disorder, and it is associated with plaques and tangles in the brain [31]. We obtained 29 AD-related terms from OMIM [32]. After removing the terms with no approved gene symbol, we obtained 15 AD-related genes. Of these 15 genes, 14, 4 and 14 genes were present in the HPRD PPI, the GGCRN, and the HPRD PPI and GGCRN integrated network (Union network). We used the 14 genes that were present in the Union network for the subsequent analyses. See Supplemental Table 1.

2.2. Methods

2.2.1. Gene–gene co-regulation network construction

The human brain data included 25,866 significant SNP-gene association pairs of 3709 genes. For each gene, we first extracted the SNPs that regulate it, and we called these significant related SNPs. If a SNP regulated two genes, we called it a common SNP of the two genes. We considered two genes to be co-regulated if a specific proportion of SNPs regulated both genes. Mathematically, for any 2 genes (G_i and G_j), there are n_1 and n_2 significant related SNPs, respectively. The gene–gene co-regulation coefficient is defined as

$$\mathrm{coreco}(G_i, G_j) = \frac{\#(\text{SNPs in } G_i \cap \text{SNPs in } G_j)}{\#(\text{SNPs in } G_i \cup \text{SNPs in } G_j)},$$

Where # (A) is the element number in set A. In other words, $\#(\text{SNPs in } G_i \cap \text{SNPs in } G_j)$ is the number of common SNPs that regulate both gene G_i and G_j; and $\#(\text{SNPs in } G_i \cup \text{SNPs in } G_j)$ is the number of SNPs that regulate gene G_i or G_j. For example, if G_i and G_j have 100 and 80 significant related SNPs, respectively, and 30 of them are common SNPs for G_i and G_j, then the co-regulation coefficient is $30/(100 + 80 - 30) = 0.2$. After calculating all co-regulation coefficients for all gene pairs, a reasonable threshold value for filtering the significant co-regulated gene pairs had to be established; for this purpose, we used the clustering coefficient difference maximization method [33]. The main function of this method is the determination of the difference in the maximum clustering coefficient difference between a real network and a random network if the real network is highly credible. Finally, we obtained the GGCRN using the significant co-regulated gene pairs.

2.2.2. Random walk with restart algorithm

In this paper, we focused on the genetic data resources rather than on statistical methods. Therefore, we adopted a classic and efficient method. The random walk algorithm (RW) for graphs is defined as an iterative walker's random transition from its current node to a neighboring node, and this is initiated at a given source node [34,35]. The random walk with restart algorithm (RWR) [36] is a variant of the random walk that allows for the restart of the walk at every time step at source node s with probability r. Formally, the RWR is defined as:

$$p^{t+1} = (1 - r)Wp^t + rp^0$$

where W is the column-normalized adjacency matrix of the graph and p^t is a vector in which the ith element holds the probability of being at node i at time step t. A special case is the initial probability vector, p^0, which is the probability of being at source node, s. In our application, p^0 was constructed such that equal probabilities were assigned to the known disease genes, with the sum of the probabilities equal to 1. Genes were ranked according to the values in the steady-state probability vector p^N. This was obtained by performing the iteration until the change between p^t and p^{t+1} fell below 10^{-6}. The results of RWR are affected by the restart probability, r. We perform a numerical experiment to select the proper r value.

The flow chart is illustrated in Fig. 1.

2.2.3. Direct neighborhood algorithm

We also compared the RWR method with the direct neighborhood (DN) method. In the DN method, the interaction partners in the network were determined for each known disease gene. The more linkages that exist between a gene and known disease genes for a particular disease, the greater the possibility that it is related

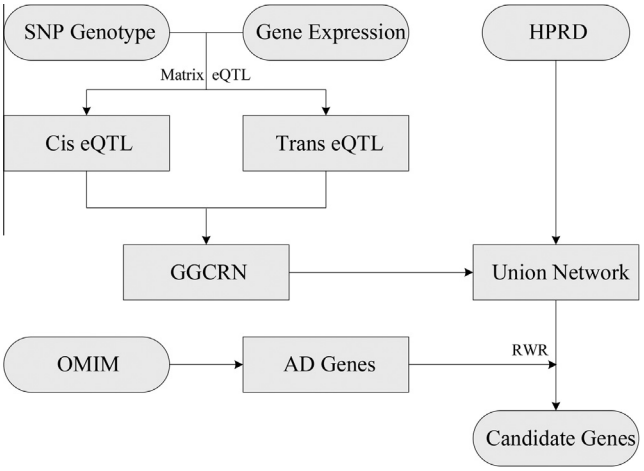

Fig. 1. The flow chart of RWR method using the Union network.