Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Prioritization of candidate disease genes by combining topological similarity and semantic similarity

Bin Liu, Min Jin *, Pan Zeng

College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

ARTICLE INFO

Article history:
Received 26 December 2014
Revised 1 July 2015
Accepted 6 July 2015
Available online xxxx

Keywords:
Disease genes
Random walk
Topological similarity
Semantic similarity

ABSTRACT

The identification of gene–phenotype relationships is very important for the treatment of human diseases. Studies have shown that genes causing the same or similar phenotypes tend to interact with each other in a protein–protein interaction (PPI) network. Thus, many identification methods based on the PPI network model have achieved good results. However, in the PPI network, some interactions between the proteins encoded by candidate gene and the proteins encoded by known disease genes are very weak. Therefore, some studies have combined the PPI network with other genomic information and reported good predictive performances. However, we believe that the results could be further improved. In this paper, we propose a new method that uses the semantic similarity between the candidate gene and known disease genes to set the initial probability vector of a random walk with a restart algorithm in a human PPI network. The effectiveness of our method was demonstrated by leave-one-out cross-validation, and the experimental results indicated that our method outperformed other methods. Additionally, our method can predict new causative genes of multifactor diseases, including Parkinson's disease, breast cancer and obesity. The top predictions were good and consistent with the findings in the literature, which further illustrates the effectiveness of our method.

1. Introduction

Because many diseases, such as cancer, diabetes, and cardiovascular diseases, result from gene mutations, exploring the relationships between diseases and their causative genes has become an important topic in contemporary systems biology. These gene mutation-caused diseases are very common in developed countries and are becoming increasingly common in developing countries [1].

Linkage analyses and association studies have been proposed to identify disease genes [2–4]. However, the efforts of these methods result in genomic intervals of 0.5–10 cM that are composed of hundreds of genes [5,6]. Whether these genes are disease-causing requires further investigation.

In recent years, with the rapid accumulation of different types of genomic data, many calculation methods for prioritizing disease genes have been proposed. One remarkable advantage of these calculation methods is the reduction in manpower and material resources. Concretely, most of these methods are based on similarities between the genomic data of known disease genes and the genomic data of the candidate gene. The genomic data include sequence-based features [7,8], gene ontology (GO) annotation information [9,10], expression patterns [11–13], and protein interaction data [14,15]. In most cases, multiple sources of genomic data are combined to find causal genes, e.g., the combinations of GO annotation information with protein interaction data [16], GO annotation information with sequence-based features [17], and metabolic pathway data with protein interaction data [18].

Investigation of the interactions between the proteins that are encoded by genes in the human PPI network has become one of the primary and most powerful approaches for elucidating the molecular mechanisms that underlie complex diseases [19–21]. Such exploration has often been performed by comparing the network topology similarities of the nodes in the PPI network. There are many methods for measuring topological similarity, including calculating the number of common neighbors between two network nodes and calculating the distance between two network nodes. Due to incomplete data about the PPI network, some interactions between the proteins encoded by candidate gene and the proteins encoded by known disease genes are very weak. Thus, some candidate genes cannot be well identified. To achieve better prediction results, some studies have combined the PPI network with phenotype similarity information and reported good performances. However, we believe that the results could be further

* Corresponding author.
E-mail address: jinmin@hnu.edu.cn (M. Jin).

http://dx.doi.org/10.1016/j.jbi.2015.07.005
1532-0464/

improved. In biological data resources, large amounts of data describe the molecular function of genes or the biological processes in which the genes are involved. These data form the semantic information of a gene. If a candidate gene and a disease gene share a high level of semantic similarity, we can compensate for the weak interaction between the genes in the PPI network by adding the semantic similarity. Some studies have shown that GO annotation information, which is used to predict disease genes [22], is a very effective semantic resource. Based on these two types of data sources, i.e., protein interaction data and GO annotation information, this paper proposes a new method for inferring gene–phenotype relationships. We use the semantic similarity value between the candidate gene and known disease genes to set the initial probability vector of the random walk with restart (RWR) algorithm and apply this algorithm to the PPI network. When the final walk reaches a stable state, we predict new disease genes according to the candidate genes' rankings in the vector. We used leave-one-out cross-validation to demonstrate the effectiveness of our method. Compared with other methods, our method achieved better performance. Additionally, new causative genes of multifactor diseases, including Parkinson's disease, breast cancer, and obesity, are predicted with our method. The top predictions were good and consistent with the reports in the literature, which further illustrates the validity of our method.

2. Materials and methods

2.1. Data source

Gene ontology data (released in October 2013) and a human gene annotation dataset (released in October 2013) were from the Gene Ontology database [23]. The GO consists of three structured ontologies, i.e., biological process (BP), molecular function (MF), and cellular component (CC). The GO data contains 25,571 BP, 9661 MF, and 3386 CC terms. The gene annotation dataset contained 383,316 annotations of 18,911 genes.

In this paper, PPI data were downloaded from the Human Protein Reference Database (HPRD). All of the information in the HPRD has been manually extracted from the literature by expert biologists who have read, interpreted, and analyzed the published data.

Disease–gene association data were obtained from the Online Mendelian Inheritance in Man (OMIM) database [24].

2.2. Summaries of the RWR and HRSS algorithms

The RWR is a sorting algorithm [14] that simulates a random walker that either starts from a seed node, or from a set of seed nodes, and moves to its direct neighbors randomly at each step. Finally, based on the probability of the random walker reaching a specific node, we ranked all of the nodes in the graph. We used P_0 to represent the initial probability vector, and P_s is a vector that represents the probability of the random walker reaching all nodes on the graph at step s. The probability vector at step $s+1$ is given by

$$P_{s+1} = (1-\delta)MP_s + \delta P_0 \quad (1)$$

The row-normalized adjacency matrix of the graph is represented by parameter M.

The parameter $\delta \in (0,1)$ is the restart probability. At each step, the random walker can return to the seed nodes with probability δ. After some steps, the vector P_{s+1} will reach a steady state. This steady state is obtained by performing the iteration until the absolute value of the difference between P_s and P_{s+1} falls below 10^{-6}.

$$|P_{s+1} - P_s| < 10^{-6} \quad (2)$$

This paper used the HRSS algorithm (Fig. 1) that was developed by Wu [25] to measure the semantic similarity.

The information content (IC) is defined by Eq. (3),

$$\mathrm{IC}(c) = -\log\ p(c) \quad (3)$$

The probability of the occurrence of the term c in a specific corpus is represented by component $p(c)$.

The IC-based distance between the two terms u and v is defined in Eq. (4), where v is a descendant of u.

$$\mathrm{dist}_{\mathrm{IC}}(u,v) = \mathrm{IC}(v) - \mathrm{IC}(u) = \log\ p(u) - \log\ p(v) \quad (4)$$

Then, the IC-based specificity of the most informative common ancestor (MICA) of any two terms term_i and term_j is

$$\alpha_{\mathrm{IC}} = \mathrm{dist}_{\mathrm{IC}}(\mathrm{root}, \mathrm{MICA}) = -\log\ p(\mathrm{MICA}) \quad (5)$$

The $\mathrm{dist}_{\mathrm{IC}}$ between a term and the most informative leaf nodes (MIL) descending from the term refers to the generality of a term. Component β represents the average of the generality values of term_i and term_j.

$$\beta_{\mathrm{IC}} = \frac{\mathrm{dist}_{\mathrm{IC}}(\mathrm{term}_i, \mathrm{MIL}_i) + \mathrm{dist}_{\mathrm{IC}}(\mathrm{term}_j, \mathrm{MIL}_j)}{2} \quad (6)$$

The most informative leaf nodes of term_i and term_j are represented by MIL_i and MIL_j, respectively.

$$\mathrm{HRSS}(\mathrm{term}_i, \mathrm{term}_j) = \frac{1}{1+\gamma}\frac{\alpha_{\mathrm{IC}}}{\alpha_{\mathrm{IC}} + \beta_{\mathrm{IC}}} \quad (7)$$

where γ is defined as follows:

$$\gamma = \mathrm{dist}(\mathrm{MICA},\mathrm{term}_i) + \mathrm{dist}(\mathrm{MICA},\mathrm{term}_j) \quad (8)$$

Let $g1$ and $g2$ be two genes of interest and $tg1$ and $tg2$ the sets of all of the GO terms assigned to gene $g1$ and $g2$, respectively.

$$\mathrm{HRSS}^{\mathrm{GO}}_{\mathrm{MAX}}(g1,g2) = \max_{\substack{go_i \in tg1 \\ go_j \in tg2}}(\mathrm{HRSS}(go_i, go_j)) \quad (9)$$

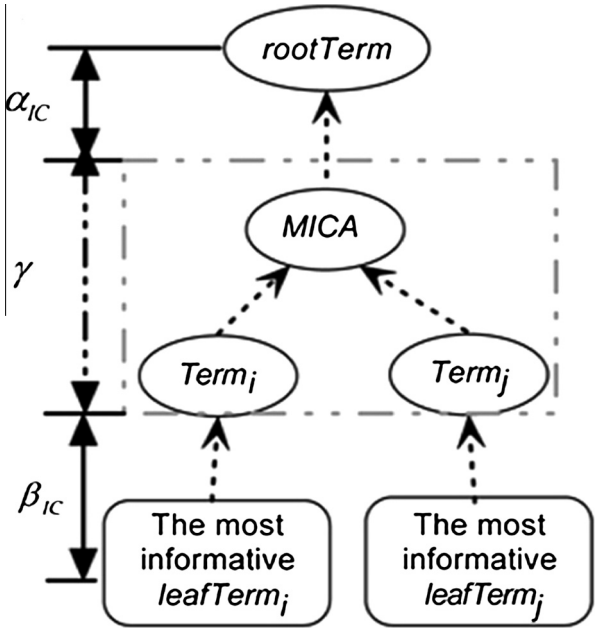

Fig. 1. A schematic illustration of the HRSS algorithm.