Journal of Biomedical Informatics 54 (2015) 270–282

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

LGscore: A method to identify disease-related genes using biological literature and Google data

Jeongwoo Kim [a], Hyunjin Kim [a], Youngmi Yoon [b], Sanghyun Park [a,*]

[a] Department of Computer Science, Yonsei University, 50 Yonsei-ro, Sinchon-dong, Seodamun-gu, Seoul 120-749, South Korea
[b] Department of Computer Engineering, Gachon University, 1342 Sengnamdaero, Sujeong-gu, Seongnam-si, Gyeonggi-do, South Korea

ARTICLE INFO

Article history:
Received 3 June 2014
Accepted 5 January 2015
Available online 21 January 2015

Keywords:
Text-mining
Data mining
Gene
Disease
Google

ABSTRACT

Since the genome project in 1990s, a number of studies associated with genes have been conducted and researchers have confirmed that genes are involved in disease. For this reason, the identification of the relationships between diseases and genes is important in biology. We propose a method called LGscore, which identifies disease-related genes using Google data and literature data. To implement this method, first, we construct a disease-related gene network using text-mining results. We then extract gene–gene interactions based on co-occurrences in abstract data obtained from PubMed, and calculate the weights of edges in the gene network by means of *Z*-scoring. The weights contain two values: the frequency and the Google search results. The frequency value is extracted from literature data, and the Google search result is obtained using Google. We assign a score to each gene through a network analysis. We assume that genes with a large number of links and numerous Google search results and frequency values are more likely to be involved in disease. For validation, we investigated the top 20 inferred genes for five different diseases using answer sets. The answer sets comprised six databases that contain information on disease–gene relationships. We identified a significant number of disease-related genes as well as candidate genes for Alzheimer's disease, diabetes, colon cancer, lung cancer, and prostate cancer. Our method was up to 40% more accurate than existing methods.

1. Introduction

Since the human genome was sequenced, a large number of gene-based studies have been performed, and vast amounts of gene data have been generated. These data are stored in databases such as the Online Mendelian Inheritance in Man (OMIM) database [19]. Extracting hidden information from these databases offers new research opportunities and challenges. One of the best known tools with which to extract knowledge is text-mining.

In the biomedical area, text-mining has been used to identify biological entities such as protein and gene names in the literature. Furthermore, text-mining can reveal novel relationships among biological entities. Text-mining can provide opportunities to reduce the time and effort needed to extract relationships between biological entities from a large amount of publications. Interest in text-mining is increasing due to the increasing number of electronic publications stored in databases such as PubMed [26]. Furthermore, Swanson's ABC model [1,2] makes text-mining a feasible approach.

Network analysis also plays an important role in biological research. Gene networks, which describe gene–gene interactions, and protein networks, which describe protein–protein interactions, allow the visual relationships among biological entities in complex biological systems to be presented in a simple, clear manner. Network analysis also provides an opportunity to analyze which relationships are meaningful among various candidates. A network analysis provides several analysis measures as well, such as degree centrality, closeness centrality, and betweenness centrality to identify novel relationships among the large numbers of relationships in the network.

Several techniques have been developed to extract hidden information using text-mining and network analysis. Li et al. [16] tried to integrate both literature and microarray gene-expression data. They constructed a gene network using the co-occurrence-based text-mining method and then refined the network using microarray data. Their results showed that the network by Li et al. is more reliable than the co-occurrence-based network. Gonzalez et al. [10] presented a method which uses literature data and interactions. They extracted an initial set of genes and proteins from the literature and then integrated the set with interactions from the curated databases of BIND and DIP. They then constructed a network based on these data, ranking the genes and gene products using a combina-

* Corresponding author. Fax: +82 2 365 2579.
E-mail addresses: jwkim2013@cs.yonsei.ac.kr (J. Kim), chriskim@cs.yonsei.ac.kr (H. Kim), ymyoon0719@gmail.com (Y. Yoon), sanghyun@cs.yonsei.ac.kr (S. Park).

http://dx.doi.org/10.1016/j.jbi.2015.01.003
1532-0464/