LIF: A method to infer disease–gene relationships using literature data and impact factor

Jeongwoo Kim [a, †]
jwkim2014@naver.com
Jinyoung Lee [a, †]
wlsdyd09@naver.com
Heechul Kang [a]
quietus999@gmail.com
Chunghun Kim [a]
jhoney7374@gmail.com
Sanghyun Park [a, *]
sanghyun@yonsei.ac.kr

ABSTRACT
Biological relationships are important in discovering the causes of disease. Therefore, a number of studies have been conducted to extract information regarding the relationships between biological entities. However, given the large number of journals and amount of literature that is available, it is difficult to assess data regarding biological relationships. In this study, we present a method called LIF, which infers disease–gene relationships using literature data and impact factor. Since the impact factor is influenced by a large number of researchers, we considered that the impact factor can be used as a measure to evaluate relationships that are extracted from literature data. To implement the LIF method, we extracted genes from disease-specific literature data. We then calculated the weight of the genes based on the impact factor of the literature in which the genes were described. For validation, we investigated the top N inferred genes for lung cancer, using an answer set. The answer set comprised several databases that contained information on disease–gene relationships. We demonstrated that the LIF is a useful method to infer disease–gene relationships compared with existing methods.

CCS Concepts
• **Applied computing → Life and medical sciences → Bioinformatics**

Keywords
Text-mining; Disease; Gene; Impact Factor

1. INTRODUCTION
After the Human Genome Project (HGP), sequences of genes were determined. Genes of the human genome were identified and mapped to physical chromosome locations. Following this biomedical advance, large amounts of genetic information are stored in databases such as PubMed [29]. This database offers new opportunities to researchers who attempt to find unknown relationships between genes and diseases. Extracting these relationships is important because they can contribute to discovering the causes and novel treatments of diseases. There have been many approaches for discovering the relationships between genes and diseases. One current approach that has been an area of recent interest is text-mining.

Text-mining is a useful approach for extracting information from large amounts of literature data. As the amount of literature data increases, text-mining is becoming increasingly popular with researchers because of its several advantages. Text-mining is an efficient method for discovering new information. Additionally, it is possible to find unexpected but meaningful information by considering several sources of literature that are generated from biological studies. Given these advantages, text-mining has been widely used in biomedical research to identify relationships such as gene–gene interactions, protein–protein interactions, and disease–gene relationships [2, 10, 13].

In this study, impact factor was considered as one of the most useful measures to determine the reliability and significance of literature data. The impact factor is calculated and utilized as a quantitative measure for evaluating journals. The measure represents the frequency with which the average article in a journal has been cited in a given period of time. We therefore assumed that the experimental results of high impact factor papers have been verified by several researchers. Therefore, we considered that extracted information in a journal with high impact factor is more reliable and significant than that from papers published in low impact factor journals. Applying this concept to biological text-mining, we proposed an LIF method to extract more trustworthy disease–gene relationships.

In this paper, we propose a novel method using impact factor and literature data to infer disease–gene relationships. We extracted literature about lung cancer, using MeSH tags in PubMed. Then, we collected gene data from the HGNC (HUGO Gene Nomenclature Committee) database [12] and obtained the impact factor list from the omics online organization [26]. After preprocessing the literature, we extracted gene data from the literature. We then calculated a score for each gene based on impact factors. Finally, we inferred the top N genes with the highest score.

The rest of the paper is divided into four sections. In Section 2, we describe previous studies related to our current work. We describe the proposed method in Section 3 and present our results and a discussion in Section 4 and 5, respectively. We conclude the paper by discussing the implications of our findings in Section 6.

[a] Department of Computer Science, Yonsei University, Seoul, Korea. Tel: +82-2-2123-7757
[*] Corresponding author
[†] These authors contributed equally to this work

SAC 2017, April 03-07, 2017, Marrakech, Morocco

DOI: http://dx.doi.org/10.1145/3019612.3019613