IDO: Inferring Describable Disease-Gene Relationships Using Opinion Sentences

Jeongwoo Kim
Yonsei University
Department of Computer Science,
Yonsei University, Seoul, Korea
+82-2-2123-7757
jwkim2013@cs.yonsei.ac.kr

Youngmi Yoon
Gachon University
Department of Computer Engineering,
Gachon University, Seongnam, Korea
+82-10-7261-2967
ymyoon@gachon.ac.kr

Sanghyun Park*
Yonsei University
Department of Computer Science,
Yonsei University, Seoul, Korea
+82-2-2123-5714
sanghyun@cs.yonsei.ac.kr

ABSTRACT
Text mining is widely used to infer relationships between biological entities. Most text-mining algorithms utilize a co-occurrence-based approach. The term co-occurrence denotes a relationship between two interesting entities if they appear in the same sentence. Using these approaches current studies have extracted relationships between biological entities such as disease-gene relationships. However, these approaches cannot provide specific information for inferred relationships such as the role of the gene in the disease. To overcome this limitation, we propose a novel approach for inferring disease-gene relationship that provides specific knowledge of the inferred relationships. To implement this method, we first built terms based on text analysis to extract opinion sentences that include disease-gene relationships. We then extracted these opinion sentences and inferred disease-gene relationships by using disease-related and gene-related terms in the opinion sentences. Using these extracted relationships and terms, we inferred disease-related genes and constructed a disease-specific gene network. To validate our approach, we investigated the top k (k = 20) inferred genes for prostate cancer and analyzed the constructed gene network using three network analysis measures. Our approach found more disease-gene relationships than comparable method, and inferred describable disease-gene relationships.

CCS Concepts
• Applied computing ➡ Life and medical sciences ➡ Bioinformatics

* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park)

SAC 2016, April 04-08, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2851613.2851616

Keywords
Text-mining; Relationship; Gene; Disease; Network; Analysis

1. INTRODUCTION
Biological research has been driven primarily by an interest in disease processes, and as such, large amounts of literature data have been generated. However, extraction of the knowledge from the data generated is important in biology. One of the best-known methods for extracting data from the literature is text mining. Text mining provides opportunities to reduce time and effort for extracting knowledge from the biological literature.

In biology, text mining can be used to infer relationships between biological entities such as proteins, genes, diseases, and drugs. Their relationships can be useful knowledge to identify the association between biological entities and disease.

Previous studies [2][4][13] have inferred relationships between biological entities by using various text-mining approaches. However, these approaches cannot describe these relationships concretely. For example, these approaches can confirm that the gene may be associated with the disease, but they cannot identify the role of the gene in the extracted relationship.

The goal of this study is to infer describable disease-gene relationships. Here, we propose a novel approach to addressing this goal that utilizes opinion sentences in PubMed [12] literature data. An opinion sentence is a sentence that describes the authors conclusions based on their experimental results. Our assumptions are as follows:

- Opinion sentences include useful knowledge to describe disease-gene relationships.

- Opinion sentences are stated in the conclusion section among the several sections in the literature.

- If the opinion sentence describes disease-gene relationships, the sentence includes disease-related terms (associated with diseases such as metastasis) and gene-related terms (associated with the gene such as mutation) as well as disease name and gene symbol.

Our aim in extracting disease-gene interactions is to identify disease-gene relationships by using disease-related and gene-related terms to confirm specific information. To achieve this, we extracted opinion sentences that contain useful information for disease-gene relationships.

The main contributions of this work include: