IMA: Identifying disease-related genes using MeSH terms and association rules

Jeongwoo Kim [a,*], Changbae Bang [a,*], Hyeonseo Hwang [a], Doyoung Kim [a], Chihyun Park [a],

Sanghyun Park [a,†]

Affiliations

[a] Department of Computer Science, Yonsei University 50 Yonsei-ro, Sinchon-dong, Seodamun-gu, Seoul 120-749, South Korea

jwkim2013@yonsei.ac.kr, bcb225@gmail.com, hyeonseo0129@gmail.com, arbc139@gmail.com, chihyun.park@yonsei.ac.kr, sanghyun@yonsei.ac.kr

[*] These authors equally contributed to this paper.

[†] corresponding author Tel: +82 2 2123 5714; fax: +82 2 365 2579

Abstract Genes play an important role in several diseases. Hence, in biology, identifying relationships between diseases and genes is important for the analysis of diseases, because mutated or dysregulated genes play an important role in pathogenesis. Here, we propose a method to identify disease-related genes using MeSH terms and association rules. We identified genes by analyzing the MeSH terms and extracted information on gene-gene interactions based on association rules. By integrating the extracted interactions, we constructed gene-gene networks and identified disease-related genes. We applied the proposed method to study five cancers, including prostate, lung, breast, stomach, and colorectal cancer, and demonstrated that the proposed method is more useful for identifying disease-related and candidate disease-related genes than previously published methods. In this study, we identified 20 genes for each disease. Among them, we presented 34 important candidate genes with evidence that supports the relationship of the candidate genes with diseases.

Keywords: Association rules, data mining, gene, disease

1. Introduction

A gene is a locus of DNA that consists of nucleotides and has genotypes made up of different DNA sequences. Through various biological experiments, researchers confirmed that genotypes determine the resulting phenotypes. The phenotype describes various biological or physical traits, such as disease, eye color, and height. For this reason, identification of disease-gene relationships is important in biology. However, the size and number of human genes is too large to analyze for all disease-gene pairs. The biological experimental cost is