Developing an NLP and IR-based Algorithm for Analyzing Gene-disease Relationships

徐建業

Yen YT;Chen B;Chiu HW;Lee YC;Li YC;Hsu CY

Abstract

OBJECTIVES: High-throughput techniques such as cDNA microarray, oligonucleotide arrays, and serial analysis of gene expression (SAGE) have been developed and used to automatically screen huge amounts of gene expression data. However, researchers usually spend lots of time and money on discovering gene-disease relationships by utilizing these techniques. We prototypically implemented an algorithm that can provide some kind of predicted results for biological researchers before they proceed with experiments, and it is very helpful for them to discover gene-disease relationships more efficiently. METHODS: Due to the fast development of computer technology, many information retrieval techniques have been applied to analyze huge digital biomedical databases available worldwide. Therefore we highly expect that we can apply information retrieval (IR) technique to extract useful information for the relationship of specific diseases and genes from MEDLINE articles. Furthermore, we also applied natural language processing (NLP) methods to do the semantic analysis for the relevant articles to discover the relationships between genes and diseases. RESULTS: We have extracted gene symbols from our literature collection according to disease MeSH classifications. We have also built an IR-based retrieval system, "Biomedical Literature Retrieval System (BLRS)" and applied the N-gram model to extract the relationship features which can reveal the relationship between genes and diseases. Finally, a relationship network of a specific disease has been built to represent the gene-disease relationships. CONCLUSIONS: A relationship feature is a functional word that can reveal the relationship between one single gene and a disease. By incorporating many modern IR techniques, we found that BLRS is a very powerful information discovery tool for literature searching. A relationship network which contains the information on gene symbol, relationship feature, and disease MeSH term can provide an integrated view to discover gene-disease relationships.