Predicting disease-gene associations using weighted gene network and literature data

Sangwon Shin[1] and Youngmi Yoon[1]
[1]Dept. of Computer Engineering, Gachon University, Korea
tkddnjs1203@gmail.com

Introduction

It is important for identifying disease-gene associations to understand disease mechanisms and develop new drugs. However, identifying all associations through wet-lab is costly. Therefore, predicting associations based on computational methods are increasing. We propose a method to predict disease-gene associations using literature data and weighted biological network of HumanNet. Using biomedical literature, co-occurrence of genes was obtained from sentences that mentioned a specific disease. Out of gene pairs from literature, we use gene pairs that overlap with HumanNet that represent the probability of gene interactions. For each gene, we calculate a score using co-occurrence and HumanNet weight. We compare the precision of proposed method with existing methods.

Datasets

Biomedical literature

PubMed is a database that provides access to over 29 million biomedical literature from MEDLINE, life science journals and online books. We obtained biomedical abstracts from PubMed.

Disease names and synonyms

Disease Ontology is a database that provides standard terminology for diseases and detailed data related to diseases. We obtained names and synonyms for 4 diseases (Breast Cancer, Prostate Cancer, Lung Cancer, Colorectal Cancer) from Disease Ontology.

HumanNet

HumanNet is a probabilistic functional gene network of human genes. We obtained gene pairs and log-likelihood score(LLS) that represent their interaction degree from HumanNet.

Gene symbols

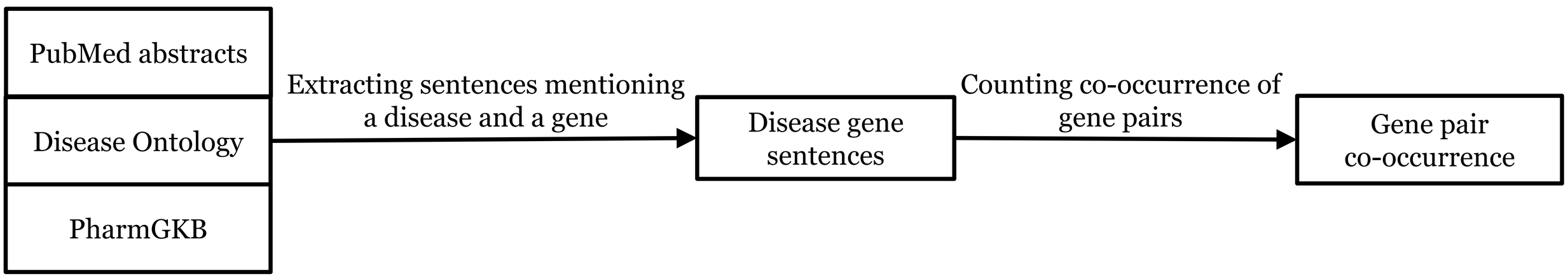
PharmGKB is a database that provides information on the effects of drugs on genes and their treatment relationships. We obtained 27003 gene symbols to identify the genes in the sentences from PharmGKB.

eDGAR

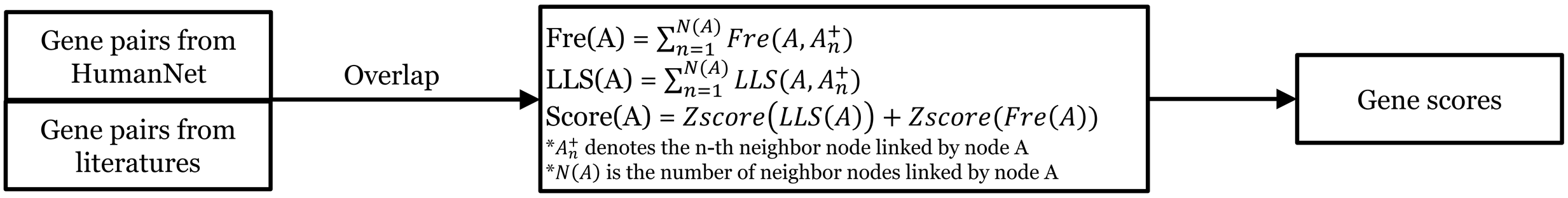
eDGAR is a database for collecting and organizing the data on gene-disease relationships as derived from the current versions of OMIM, HUMSAVAR and CLINVAR. We obtained disease related genes for each disease from eDGAR. (Breast Cancer: 27 genes, Prostate Cancer: 15 genes, Lung Cancer: 17 genes, Colorectal Cancer: 30 genes)

Methods

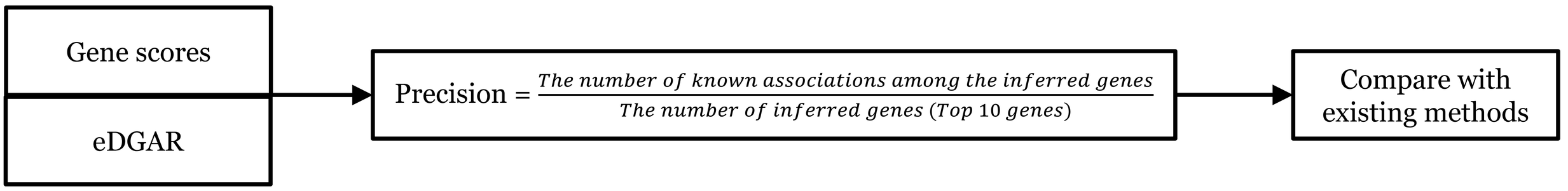
* Counting co-occurrence of gene pairs



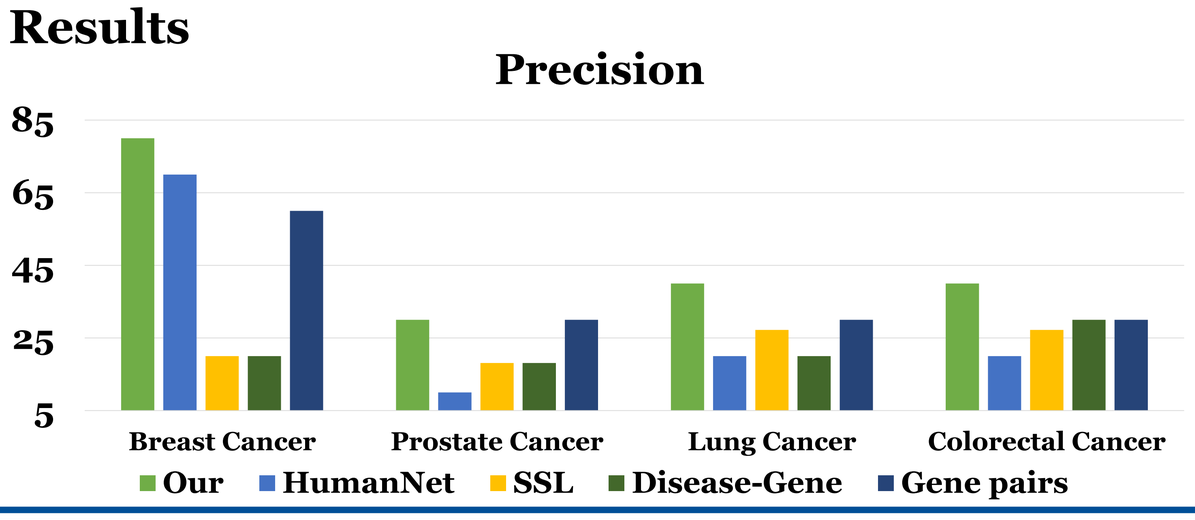
* Calculating gene scores



* Measuring precision and comparing with existing methods



Results



HumanNet

HumanNet is a method using only LLS in proposed method.

SSL

SSL is a method of predicting disease-related genes using disease gene sentences containing an auxiliary verb (may or might). [1]

Disease-Gene

Disease-Gene is a method for predicting disease-related genes based on the frequency of the gene in a sentence containing the disease.

Gene pairs

Gene pairs is a method of using the value from $Fre(A)$ as a gene score using the gene pairs from literatures.

Conclusion

We confirmed that proposed method shows higher precision than existing methods in 4 diseases Existing methods such as SSL only consider disease-gene relationships. The proposed method can consider both gene interactions general state and gene interactions in a specific disease. Therefore, the proposed method can help to more accurately identify disease-gene relationships by considering both relationships.

[1] J Kim *et al*, "SSL: Inferring disease-related genes using sentence structure and literature data", *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2017, 100-107.