Journal of The Korea Society of Computer and Information
Vol. 22 No. 3, pp. 131-138, March 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.03.131

Identifying literature-based significant genes and discovering novel drug indications on PPI network

Minseok Park*, Giup Jang**, Taekeon Lee***, Youngmi Yoon****

Abstract

New drug development is time-consuming and costly. Hence, it is necessary to repurpose old drugs for finding new indication. We suggest the way that repurposing old drug using massive literature data and biological network. We supposed a disease-drug relationship can be available if signal pathways of the relationship include significant genes identified in literature data. This research is composed of three steps-identifying significant gene using co-occurrence in literature; analyzing the shortest path on biological network; and scoring a relationship with comparison between the significant genes and the shortest paths. Based on literatures, we identify significant genes based on the co-occurrence frequency between a gene and disease. With the network that include weight as possibility of interaction between genes, we use shortest paths on the network as signal pathways. We perform comparing genes that identified as significant gene and included on signal pathways, calculating the scores and then identifying the candidate drugs. With this processes, we show the drugs having new possibility of drug repurposing and the use of our method as the new method of drug repurposing.

▸ Keyword : Systems biology, Network biology, Text mining, Drug repositioning

I. Introduction

최근에 의학 기술은 빠르게 발전해왔지만, 여전히 신약이 개발되고 시장에 나오기까지 많은 시간과 비용이 소요된다[1]. 이러한 이유로 약물 재창출 (Drug Repositioning)이 신약 개발의 대안으로 제시되고 있다. 약물 재창출이란 기존에 존재하는 약물이나, 안정성 문제로 시장 출시에 실패한 약물의 알려지지 않은 치료범위에 대하여 새로운 치료법을 찾는 것을 말한다[2]. 약물 재창출을 통한 신약 개발은 상대적으로 적은 시간과 비용이 소요된다. 이러한 이유로 최근에는 약물 재창출을 위한 다양한 연구들이 수행되고 있다. 약물 재창출의 대표적인 예로 Sildenafil 이 있다. Sildenafil은 최초에 고혈압 치료를 목적으로 개발이 시작되었다. 하지만 신약 개발 도중 발기부전 치료라는 새로운 적응증이 발견되었고, 이를 바탕으로 재창출되어 시장에 출시되었다.

본 연구에서는 학술지나 여러 문헌을 기반으로 텍스트 데이터 마이닝 (Text Data Mining)을 통하여 중요 유전자를 선별한다. 텍스트 데이터 마이닝이란 구조화되지 않은 텍스트에서 의미를 찾아내는 기술이다[3]. 구조화되고 패턴이 존재하는 다른 데이터 마이닝과는 다르게 텍스트 데이터 마이닝은 자연어와 같이 구조화 되어 있지 않기 때문에 구조 분석, 정보 추출 등과 같은 과정이

• First Author: Minseok Park, Corresponding Author: Youngmi Yoon
*Minseok Park (iamminseokpark@gmail.com), Dept. of Computer Engineering, Gachon University
**Giup Jang (giupjang0207@gmail.com), Dept. of IT Convergence Engineering, Gachon University
***Taekeon Lee (teakeon.m.lee@gmail.com), Dept. of Computer Engineering, Gachon University
****Youngmi Yoon (ymyoon@gachon.ac.kr), Dept. of Computer Engineering, Gachon University
• Received: 2017. 01. 26, Revised: 2017. 02. 13, Accepted: 2017. 02. 24.
• This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(Ministry of Science, ICT & Future Planning) (No. 2015R1A2A2A03004088).

동반된다. 텍스트 데이터 마이닝을 이용한 연구의 예로써 Jang 의 연구가 있다. Jang 의 연구에서는 문헌을 이용하여 질병-약물과 유전자-질병 간의 상호정보량을 이용하여 약물 간의 유사도를 측정하고 유사한 약물을 식별하는 새로운 방법을 제안하였다. 텍스트 데이터 마이닝을 이용하여 상호정보량 유사도를 계산하고 WHO에서 의약품을 분류하는 식별 코드인 ATC (Anatomical Therapeutic Chemical)를 이용하여 유사한 약물을 식별하는 방법을 제시한다[4].

본 연구에서는 네트워크상의 최단 거리 경로를 추출하여 최적의 작용 경로로써 사용한다. 오랜 시간 생물학적 데이터 축적으로 인하여 약물 재창출에 여러 컴퓨터를 이용한 (In Silico) 네트워크 기반의 연구가 시도되고 있다[5]. 기존의 약물 재창출 방법은 새로운 적응증을 우연하게 발견하여 내부 작용 경로는 알 수 없다는 단점이 존재한다. 그러나 네트워크를 기반으로 하는 약물 재창출 연구는 약물의 작용 경로를 파악할 수 있다. 네트워크를 이용한 약물 재창출 관련 연구로는 Cheng 의 연구가 존재한다. Cheng 의 연구에서는 다른 방식과 네트워크 방식을 비교하면서 약물과 표적 유전자 간의 상호작용 예측과 기존 약물의 재창출을 시도하였다[6]. 이 연구에서는 약물과 약물의 표적 유전자 그리고 그 사이의 상호작용으로 구성된 네트워크 상에서의 별도의 계산식을 통하여 새로운 관계를 발견한다. 또한, 이 연구에서는 약물 기반의 DBSI (Drug-Based Similarity Inference)와 표적 유전자 기반의 TBSI (Target-Based Similarity Inference) 그리고 네트워크 기반의 NBI (Network-Based Inference) 간의 비교와 함께 네트워크를 이용한 방식이 가장 우수한 성능을 나타냄을 보인다.

그러나 앞서 언급한 사례들은 기존 약물을 이용한 약물 재창출에 대한 연구들이지만 각 연구의 특징에 따른 어려움이 존재한다. 텍스트 데이터 마이닝 관련 연구는 약물 재창출을 시도하였지만, 재창출한 관계에 대한 작용 경로를 알 수 없는 단점이 존재하며, 네트워크 관련 연구의 경우에는 네트워크를 통하여 작용 경로를 알 수 있지만 이 경로가 생물학적으로 존재하는 작용 경로인지 알 수 없는 단점을 포함하고 있다. 그래서 우리는 텍스트 데이터 마이닝의 장점과 네트워크의 장점을 이용한다. 본 연구에서는 네트워크를 통해 작용 경로를 계산하고, 출판된 문헌을 통하여 생물학적 문헌을 통해 검증하는 과정을 포함한다. 이를 통하여 텍스트 데이터 마이닝을 이용한 방식과 네트워크를 이용한 방식에서 보이는 한계점을 해결한다.

우리는 문헌에서의 동시 출현 단어 분석과 상호작용 기반 네트워크상에서의 최단 거리 경로 추출을 함께 사용하여 연구를 수행한다. 본 연구에서는 방대한 문헌을 기반으로 동시 출현 단어 분석을 사용한다. 동시 출현 단어 분석이란 한 범위에서 특정 단어들이 동시에 출현하는 경우를 추출하는 방법이다. 우리는 유전자와 질병에 대하여 문헌에서의 동시 출현 단어의 빈도를 추출하고, 빈도를 통하여 추출된 유전자 중 상위 10%를 중요 유전자로써 식별한다. 또한 우리는 PPI (Protein-Protein Interaction)을 기반으로 하는 생물학적 네트워크를 사용한다. 네트워크에서의 노드 (Node)는 유전자를 의미하며, 각 연결선 (Edge)은 유전자 간의 상호작용을 의미한다. 가중치 (Weight)는 유전자 사이에서 발생하는 상호작용의 정도를 의미한다.

이 네트워크는 가중치로써 다른 데이터베이스에서 제공하는 생물학적 정보를 포함한 데이터를 사용한다. 생물학적 네트워크는 노드 간 거리가 짧으면 효율적인 경로로 가정한다. 생물학적 지식과 네트워크의 위상적 특징을 사용하는 최단 거리 경로를 사용하여 질병-약물 관계에 대한 최적의 경로 (Pathway)를 식별한다. 식별한 최적의 경로들을 문헌에서 추출한 중요 유전자와 비교하여 스코어를 계산한다. 계산된 스코어를 바탕으로 질병-약물 관계들을 서열화 시켜 새로운 적응증을 가질 수 있는 후보 약물을 식별한다. 본 연구에서는 유방암 (Breast Cancer)과 전립선암 (Prostate Cancer) 그리고 천식 (Asthma)에 대하여 검증을 위한 기존 약물과 질병 간의 실험을 수행한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 연구에 앞서 관련 있는 선행 연구와 본 실험에서 사용한 데이터베이스에 대한 소개를 기술한다. 3장에서는 본 연구에서 진행하는 문헌에서의 질병-유전자 동시 발생 빈도 측정, 중요 유전자 식별, 상호작용 기반 네트워크의 최단거리 경로 추출 방법과 스코어 계산 방식에 관해 기술한다. 4장에서는 연구 환경 및 결과를 기술한다. 5장에서는 본 연구의 정리와 앞으로 나아가야 할 방향을 제시한다.

II. Related works

1. Related works

본 연구는 문헌을 기반으로 하고 컴퓨터를 이용한 실험이기 때문에 실험을 위한 다양한 데이터의 필요성이 존재한다. 우리는 관련 연구들을 통하여 문헌과 약물, 질병, 유전자로 구성된 네트워크 등과 같은 데이터들을 사용하며 다음과 같은 관련 연구를 기반으로 본 연구를 진행한다.

1.1. MEDLINE

MEDLINE (Medical Literature Analysis and Retrieval System Online)은 의학 관련 문헌을 제공하는 데이터베이스이다[7]. 1946년부터 현재까지의 문헌을 포함하며 5,600개 이상의 과학 저널로부터 문헌을 수집한다. MEDLINE의 문헌은 학술지와 신문, 매거진 등을 포함하며, 각 문헌은 ISSN, 게시 날짜, 카테고리, 저자명, 초록 등을 포함한다.

본 연구에서는 MEDLINE으로부터 1946년도부터 2015년도 사이에 23,249개의 다양한 저널에 출판된 24,299,252편의 문헌을 사용한다. 이 데이터 중 초록을 문장 단위로 구분하여 사용한다. 구분된 각 문장에서 질병과 유전자의 동시 출현 빈도를 측정한다.

1.2. DrugBank

DrugBank는 생물정보공학 (Bioinformatics)과 화학정보공학 (Cheminformatics) 분야에서 약물에 대한 정보를 제공하는 데이

터베이스이다[8]. 이 데이터베이스는 약물의 표적 유전자와 유전자의 시퀀스 구조, 약물 작용 경로와 같은 약물의 정보를 제공한다. 또한 약물과 약물 (Drug-Drug)이나 식품 (Drug-Food)과의 상호작용과 약물의 작용 효소와 수송체에 대한 의학, 약리학, 분자학과 같이 전문적이고 상세한 데이터를 제공하고 있다.

본 연구에서는 DrugBank에서 제공하는 약물과 약물 표적 유전자와 FDA에서 승인된 약물 리스트를 사용한다. 승인된 약물 리스트는 약물과 특정 질병 간의 스코어 계산과 비교를 위하여, 약물 표적 유전자는 PPI (Protein-Protein Interaction) 네트워크와의 연결을 위한 사상 테이블을 구축하기 위하여 사용한다.

1.3. PharmGKB

PharmGKB (The Pharmacogenomics Knowledgebase)는 다양한 유전적 약물 반응에 대한 정보들을 선별하여 제공하는 데이터베이스이다[9]. 이 데이터베이스는 유전자, 약물, 약물 작용 경로 그리고 약물-유전자 관계 같은 약리유전학과 관련된 문헌 기반 데이터를 제공한다. 또한 약물 투약 지침이나 잠재적으로 작용 가능성이 존재하는 약물-유전자 관계, 유전자형-표현형 관계와 같이 의학적으로 관련 있는 다양한 데이터를 제공한다.

본 연구에서 동시 출현 빈도 계산 시에 약물은 유전자 심볼을 사용하기 때문에, 문헌의 문장 추출을 위하여 유전자 심볼 (GeneSymbol) 리스트가 필요하다. 또한 유전자와 네트워크의 사상을 위하여 Entrez Gene ID 를 사용한다. Entrez Gene ID는 NCBI (National Center for Biotechnology Information)의 데이터베이스에서 사용되는 유전자 식별 인덱스이다[10]. 본 연구에서 이 데이터베이스의 유전자 심볼 데이터를 사용하여 질병-유전자 동시 출현 빈도 계산과 데이터를 사상하기 위한 GeneSymbol-EntrezGeneID 사상 테이블을 구축한다.

1.4. Disease Ontology

Disease Ontology는 인간 질병의 표준 온톨로지를 제공하는 데이터베이스이다[11]. 문헌에서 질병을 명시하는 방법은 표준화되어 있지 않고 다양한 명칭으로 사용된다. 이 데이터베이스는 MeSH, ICD, OMIM 과 같은 다양한 외부 온톨로지를 통하여 인간 질병에 대한 표준화된 설명이나 용어 등을 제공한다. 또한 인간 질병에 대한 유전적 변형, 표현형, 단백질, 약물 그리고 항원 결정기 (Epitope) 같이 인간 질병에 대한 상세 데이터를 제공한다.

본 연구에서는 텍스트 마이닝을 이용하여 질병과 유전자의 동시 출현을 사용해야 하므로, 다양하게 불리우는 질병의 명칭들을 모두 사용할 필요가 있다. 예를 들어, 유방암은 'Breast Cancer', 'Breast Neoplasms' 등과 같이 다양하게 표현할 수 있다. 본 연구에서는 이러한 경우를 모두 포함하는 방법으로써 더 정확한 질병-유전자 동시 출현 빈도 계산을 위하여 이 데이터베이스에서 각 질병에 대한 동의어를 사용한다.

1.5. CTD

CTD (Comparative Toxicogenomics Database)는 환경적 노출이 인간 건강에 어떠한 영향을 미치는 가에 대한 이해를 증진시키고자 제공되는 공공 데이터베이스이다[12]. 이 데이터베이스는 약물, 질병, 유전자, 유전자 표현형, 작용 경로에 대한 정보를 제공한다. 또한 전문 큐레이터가 직접 문헌으로부터 화학구조-유전자 (Chemical-Gene)과 화학구조-질병 (Chemical-Disease) 그리고 질병-유전자 (Disease-Gene) 상호작용과 같이 인간의 건강과 관련된 다양한 의학적 데이터를 수집하여 제공한다.

본 연구에서는 CTD로부터 질병 연관 유전자와 기존에 알려진 질병-약물 관계를 사용한다. 질병 연관 유전자는 본 실험에서 사용하는 네트워크와의 연결을 위한 사상 테이블 구축을 위하여 사용한다. 기존에 알려진 질병-약물 관계는 본 실험에서 질병에 대한 새로운 약물로 식별한 후보 약물을 검증하기 위하여 사용한다.

1.6. HumanNet

HumanNet은 18,714개의 유전자와 476,399개의 연결선로 구성된 확률기반 기능적 유전자 네트워크이다[13]. 이 네트워크에서는 21가지 유형의 생물학적 데이터 수치들이 존재한다. 이 수치들은 각 데이터 유형에 대하여 함께 작용한다고 알려진 관계인 두 유전자 간의 연결에 대한 정도를 의미한다. 베이지안을 기반으로 이 수치들을 통합하여 하나의 네트워크를 구성한다. HumanNet에서의 각 상호작용은 두 유전자 사이의 기능적 연결을 나타내는 상호작용이 존재할 가능성을 나타내는 Log-Likelihood Score (LLS)를 가진다.

본 연구에서는 이 네트워크에서 제공하는 유전자 간의 상호작용과 통합된 가중치로 네트워크를 구축하며 사용한다. 구축된 네트워크를 이용하여 약물의 표적 유전자와 질병과 연관된 유전자 사이의 최소 거리 경로를 구하여 실험에 사용한다.

1.7. TTD

TTD (Therapeutic Target Database)는 실제 치료와 관련 있는 약물들에 대한 정보를 제공하는 데이터베이스이다[14]. 이 데이터베이스에서는 단백질, 핵산의 표적, 표적이 되는 질병, 표적에 영향을 미치는 네트워크상의 경로와 각각에 직접 영향을 주는 약물들에 대한 정보를 제공한다. 또한 약물에 대한 구조, 표적 유전자의 기능, 시퀀스, 화학식의 3차원 구조, 효소에 대한 정보와 치료 기법 임상실험 현황 등과 함께 실제 의학적인 치료에 대한 다양한 데이터를 제공한다.

본 연구에서는 TTD로부터 알려진 질병-약물 관계를 사용한다. 실험에서는 이 데이터베이스가 제공하는 질병-약물 관계를 기준 (Gold Standard)으로써 사용한다. 기준과 식별한 후보 약물의 비교를 통하여, 본 연구의 유효성을 검증한다.

1.8. Clinicaltrials.gov

Clinicaltrials.gov는 다양한 질병에 대하여, 다양한 조건에서 진행되는 임상 실험 정보를 제공하는 웹 기반 데이터베이스이

다[15]. 이 데이터베이스는 정부나 기업에서 진행하는 임상 실험에 대한 상세정보를 제공한다. 각 임상 실험 정보들은 실험의 목적과 설계와 적격 기준 그리고 실험이 진행되는 위치들을 포함하고 있다. 이 데이터베이스에 등록되는 실험들은 등록 전에 실험에 대한 품질 평가가 이루어지기 때문에 신뢰성 있는 데이터를 제공한다.

본 연구의 유효성을 검증하기 위하여 Clinicaltrials.gov에서 제공하는 질병-약물 관계를 사용한다. 또한 식별된 약물에 대하여 후보 약물을 선별하기 위하여 이 질병-약물 관계를 Evidence로 사용한다.

III. The Proposed Scheme

1. System Overview

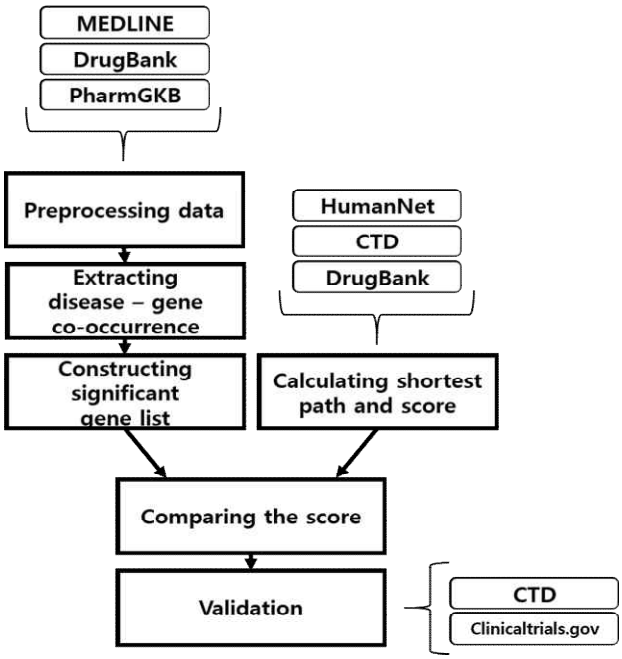

Fig. 1. System Overview

본 연구의 시스템은 그림 1 과 같이 구성되어 있다. MEDLINE에서 제공하는 문헌을 사용하며, 문헌에서의 초록을 한 문장 단위로 분리한다. 텍스트 마이닝 기법을 통하여 각 문장에서 나타나는 질병과 유전자의 동시 출현 빈도를 측정한다. 유전자와 질병의 동시 출현 빈도가 상위 10%인 유전자를 중요 유전자로써 선별한다. DrugBank에서 제공하는 약물 목록과 약물 표적 유전자와 CTD에서 제공하는 질병 연관 유전자를 사용하여, 약물과 질병이 HumanNet 네트워크상에서 연결될 수 있는 가장 짧은 경로를 추출한다. 문헌으로부터 추출한 중요 유전자들과 네트워크에서 추출한 최단 거리 경로가 포함하는 유전자들의 비교하여 스코어를 계산한다. 기존에 알려진 질병-약물 관계와 비교하여, 질병을 치료할 수 있는 새로운 후보 약물을 식별한다.

2. Method

2.1. Constructing Important Gene List with Co-occurrence

질병과 유전자의 동시 출현 빈도를 측정하기 위하여 MEDLINE을 사용한다. 본 연구에서는 1948년부터 2015년도까지의 현재 MEDLINE에서 제공하는 모든 문헌을 사용하였으며, 총 14,234개의 저널과 24,178,134개의 문헌을 사용한다. 동시 출현의 빈도를 측정하기 위하여 문헌의 초록을 사용하며, 초록을 문장 단위로 나누어 사용한다. 각 질병과 유전자의 동시 출현 빈도 추출 시, 질병은 Disease Ontology에서 제공하는 질병의 이름과 동의어를 사용하며 유전자는 PharmGKB에서 제공하는 총 27,007개의 유전자 심볼을 사용한다. 중요 유전자 리스트는 MEDLINE으로부터 질병-유전자의 동시 출현 관계를 수집하여 각 질병에서 중요하게 추측되는 유전자들을 따로 선별하여 구성한다. 문헌으로부터 언급된 각 질병-유전자의 동시 출현 빈도를 이용하여 유전자를 선별하고 이를 중요 유전자로써 사용한다.

2.2. Calculating Shortest Path and Score between a Drug and a Disease

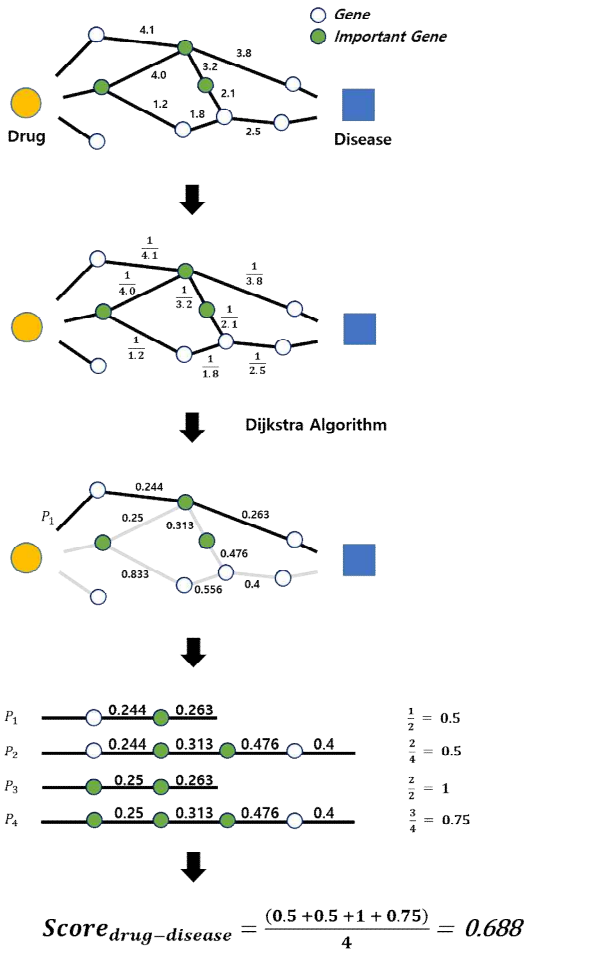

Fig. 2. Calculating a Disease-Drug Relationship Score Process

본 연구에서 사용하는 스코어 계산 과정은 그림 2 와 같이 구성 되어있다. 우리는 약물이 질병에 영향을 미치는 경로를 네트워크 거리 기반으로 계산한다. 약물은 표적 유전자를 시작으로 여러 유전자와 상호작용하며 영향을 미친다. 각 유전자와 유전자들은 상호작용을 하며, 상호작용 끝에 질병 연관 유전자에 영향을 미친다. 각 유전자의 상호작용 가능성은 거리로써 표현할 수 있다. 본 연구에서는 유전자의 상호작용을 네트워크로 표현한 HumanNet 데이터를 사용하며, 가능성이 높은 약물 작용 경로를 찾기 위하여 최단 거리 경로 (Shortest Path)를 사용한다.

질병과 약물 사이의 최단 거리 경로를 구축하기 위하여, 약물의 표적 유전자와 질병과 연관된 유전자를 이용한다. 전체 약물과 표적 유전자는 DrugBank에서 제공하는 2,218개의 FDA로부터 승인된 약물과 7,142개의 약물 표적 유전자를 사용한다. 이를 이용하여 약물과 표적 유전자의 Uniprot-GeneSymbol 사상 테이블을 구축하고 표적 유전자를 HumanNet 네트워크에 사상한다. 비교 질병들과 연관된 유전자들은 CTD에서 제공하는 19,846,714개의 질병 연관 유전자 사용한다. 질병들과 연관된 유전자는 유전자 심볼로 표현되며, 네트워크와의 사상을 위하여 CTD으로부터 GeneSymbol-EntrezGeneID 사상 테이블을 사용한다.

본 연구에서 사용하는 HumanNet의 각 노드는 유전자를 나타내고, 각 연결선은 유전자 간의 상호작용을 나타내며, 각 연결선의 가중치는 상호작용을 하는 정도를 의미한다. 약물과 질병 사이에서 상호작용하는 유전자를 이용하여 경로들을 만들고, 네트워크상에서 약물과 질병 관계 사이가 가까울수록 더 큰 영향이 미치기 때문에 경로 중 최단 거리 경로를 사용한다. HumanNet의 가중치는 높을수록 상호작용의 가능성이 높다는 것을 의미하기 때문에, 최단 거리 경로를 이용하기 위하여 각 가중치에 역수를 취하여 사용한다. 최단 거리 경로를 구축하기 위하여 다익스트라 알고리즘 (Dijkstra Algorithm)을 사용한다. 질병 연관 유전자는 반드시 경로 상에서 출현하기 때문에 최단 거리 경로 추출 이후 경로에 포함된 유전자 리스트에서 제거 후 사용한다.

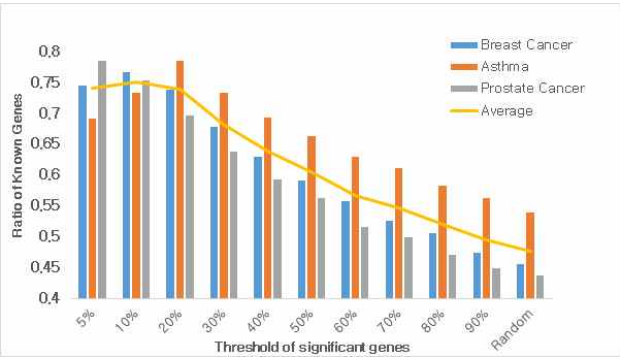

Fig. 3. Changes in the Ratio of Identified Genes to Known Genes Based on Threshold

2.3. Calculating Score between Drug and Disease Relationship

각 약물과 질병의 관계가 문헌으로부터 추출한 동시 출현 빈도와 연관성의 정도를 비교하기 위하여 식을 이용하여 스코어를 계산한다. 스코어 계산 방식에서는 승인된 약물과 질병의 관계를 사용한다. 약물과 질병이 네트워크에 사상 되면 다수의 약물 표적 유전자들과 질병 연관 유전자에 의하여 질병-약물 관계 사이에서 여러 경로가 발생한다. 질병-약물 관계에서 발생할 수 있는 모든 약물의 표적 유전자와 질병 연관 유전자 사이의 경로에서 나타날 수 있는 최단 경로들에 대해서 스코어 계산을 진행한다. 스코어 계산 방식은 다음과 같다.

$$Score = \frac{1}{p}\sum_{n=1}^{p}\frac{I_n}{G_n} \qquad (1)$$

식 (1)에서의 p는 약물 질병 관계의 최단 거리 경로의 수, G_n는 최단 거리 경로 내의 유전자의 개수를 나타내며, I_n는 G_n와 중요 유전자 집합의 교집합을 이루는 유전자의 개수를 의미한다. 각 질병-약물 관계에 대하여 스코어 계산 시 질병-약물 관계 사이에서 나타날 수 있는 모든 최단 경로에 대하여 경로에 포함된 유전자 집합과 중요 유전자로 선별된 집합을 비교하여 비율로 계산한 후 질병-약물 관계의 모든 최단거리에 대한 비율의 평균으로 해당 질병-약물 관계에 스코어를 계산한다.

IV. Result

1. Experimental Environment and Data

1.1 Experimental Environment

본 연구는 Intel(R) Core™ i7-4790 CPU @ 3.60GHz CPU, 32GB RAM, 64비트 운영체제의 머신으로 수행되었으며 개발도구는 Microsoft visual studio 2015를 사용하였다.

1.2 Data Set

본 연구는 유방암, 전립선암, 천식에 대하여 진행하였다. 문헌을 사용하기 위하여 MEDLINE에서 제공하는 14,234개의 저널에서 게시된 24,178,134개의 초록을 사용하였다. 질병마다 MEDLINE에서 동의어를 포함한 질병의 명칭과 유전자 심볼의 동시 출현 빈도를 측정하여 중요 유전자를 선별하였다.

그림 3 은 특정 임계값 (Threshold)에 따라 선별된 유전자 중 알려진 유전자의 변화하는 비율을 보여준다. 퍼센트는 특정 임계값을 나타내며 Random은 질병명과 동시 출현한 모든 유전자를 사용한 경우를 의미한다. 세 질병의 선별한 유전자에 대한 알려진 유전자 비율은 상대적으로 Random보다 높은 수치를 보

인다. 또한, 특정 임계값이 낮아질수록 알려진 유전자의 비율은 점차 증가하며, 10%에서 최고치를 보이고, 다시 감소하는 경향을 보인다. 이를 통하여 본 연구에서 사용하는 중요 유전자 선별에 대한 특정 임계값을 10%로 설정하였다.

2. Experimental Result

2.1 Identifying Significant Gene

문헌에서 선별한 중요 유전자에 대한 신뢰성을 판단하기 위하여 중요 유전자의 TOP 20 Gene을 선별하고 이를 다른 데이터베이스인 DisGeNET[16], Kegg[17], GHR[18]와 함께 사용하여 검증하였다. 이 데이터베이스 제공하는 질병과 질병에 대한 연관 유전자를 사용하며, 본 연구에서 선별한 중요 유전자에서의 Top 20 Gene에 대한 유효성을 얻기 위하여 이 데이터베이스에서 질병과 질병에 대한 연관 유전자를 Evidence로 사용하였다.

본 실험에서 선별한 중요 유전자의 Top 20 Gene는 다음과 같다. 표에서의 Evidence는 해당 유전자와 질병 간의 연관성을 발견 할 수 있는 데이터베이스를 의미한다. PC, AR과 같이 2글자 이하의 심볼을 가지고 있는 유전자는 유전자의 의미보다 다른 의미로써 사용될 수 있으므로, 노이즈의 가능성이 존재한다[19]. 하지만 본 연구에서는 이점을 고려하지 않았다. 그럼에도, 표 1 을 통하여 선별한 중요 유전자 중 다수의 유전자들이 해당 질병과 관련이 있음을 확인 할 수 있었다.

2.2 Identifying Novel Disease-Drug Relationship

본 연구에서는 유방암, 전립선암, 천식 세 가지 질병과 승인된 전체 약물 간의 관계를 대상 집합으로 사용하였다. TTD에서 제공하는 알려진 관계들을 사용하여 대상 집합과 비교 실험을 진행하였다. 약물은 2,223개를 사용하였으며, 알려진 관계는 유방암에 대하여 23개, 전립선암에 대하여 15개, 천식에 대하여 27개를 사용하였다. TTD에서 제공하는 모든 알려진 관계의 약물들은 승인된 2,223개에 포함되어 있다.

$$Zscore = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \qquad (2)$$

데이터의 신뢰성을 판단하기 위하여 식 (2)을 이용하여 스코어에 대한 표준정규분포의 임의 샘플링을 수행하였다[20]. 식 (2) 의 $\overline{X}$는 표본 집단의 평균, μ는 모집단의 평균, σ는 모집단의 분산을 의미하며 n은 추출할 표본 집단을 구성할 표본의 개수를 의미한다. 네트워크상에서 약물의 표적 유전자가 존재하지 않거나 경로가 그려지지 않은 약물은 제외하였고, 임의 샘플링에서는 1,449개의 약물을 사용하였다. 표준 집단 정규화를 이용하여 알려진 관계의 개수를 임의로 추출하였을 때의 집합과 알려진 관계들의 집합의 스코어를 비교한다. 식 (2)을 이용하여 알려진 관계들과 임의의 약물을 이용한 샘플링을 비교하였을 때, 유의미한 차이를 보였다. (p-value<0.001)

본 연구에서는 약물 재창출 가능성이 존재하는 후보 약물을 식별하기 위하여 알려진 관계에 있는 약물 중 스코어가 가장 높게 측정된 약물의 스코어를 특정 임계값으로 사용하였다. 이 임계값을 사용하여, 특정 임계값보다 스코어가 높게 측정된 약물을 후보 약물로 식별하였다. 각각 유방암, 전립선암, 천식과 모든 1,449개의 약물 관계에 대해서 스코어를 측정하였다. 특정 임계값으로 유방암에 대해서 스코어는 0.520를 부여하였다. 전립선암에 대해서는 0.434을 부여하였으며, 천식에 대해서 0.330를 부여하였다. 각 질병에 대한 식별된 후보 약물은 표 2

Table 1. Top 20 Genes Co-occurred with Diseases

Rank	Prostate Cancer		Asthma		Breast Cancer	
	Gene Symbol	Evidence	Gene Symbol	Evidence	Gene Symbol	Evidence
1	PC	None	T	None	BRCA1	DisGeNET, Kegg, GHR
2	AR	DisGeNET, PGDB, Kegg, GHR	AR	None	MB	DisGeNET
3	T	None	AHR	DisGeNET	T	None
4	ERG	DisGeNET	FEV	DisGeNET	EGFR	DisGeNET
5	HR	None	CD4	None	HR	None
6	PTEN	DisGeNET, PGDB, Kegg, GHR	TNF	DisGeNET	IV	None
7	PCA3	DisGeNET, PGDB	IV	None	SLN	None
8	EGFR	DisGeNET, PGDB	ADAM33	DisGeNET, Kegg, GHR	EGF	DisGeNET
9	AMACR	DisGeNET	CRS	None	BRCA2	DisGeNET, Kegg, GHR
10	EGF	DisGeNET	CS	None	AR	DisGeNET
11	IV	None	HR	None	PRL	DisGeNET
12	BRCA1	DisGeNET, PGDB, GHR	GC	DisGeNET	ERBB2	DisGeNET, Kegg
13	VDR	DisGeNET, PGDB	TSLP	DisGeNET	ATM	DisGeNET, GHR
14	GSTP1	DisGeNET, PGDB, Kegg	ACE	DisGeNET	CD24	DisGeNET
15	BCR	None	CRP	DisGeNET	PTEN	DisGeNET, Kegg, GHR
16	CD44	DisGeNET, PGDB	CD14	DisGeNET, Kegg	MUC1	DisGeNET
17	DCE	None	ADRB2	DisGeNET, Kegg, GHR	CD44	DisGeNET
18	DES	None	GSTM1	DisGeNET	CXCR4	DisGeNET
19	TNF	DisGeNET, PGDB	NGF	DisGeNET	TP53	DisGeNET, GHR
20	EZH2	DisGeNET, GHR	PC	DisGeNET	TNF	DisGeNET

Table 2. Candidate Drugs for Diseases

Breast Cancer	Prostate Cancer	Asthma
Azacitidine	Conjugated Equine Estrogens	Drostanolone
Flucytosine	Toremifene	Nandrolone phenpropionate
Decitabine	Estrone	Bicalutamide
Maraviroc	Clomifene	Testosterone Propionate
Conjugated Equine Estrogens	Fulvestrant	Cyproterone acetate
Toremifene	Mestranol	Methyltestosterone
Estrone	Ospemifene	Nandrolone decanoate
Dienestrol	Dienestrol	Enzalutamide
Fulvestrant		Flutamide
Mestranol		Maraviroc
Ospemifene		Cefdinir
Bazedoxifene		Gefitinib
Clomifene		Panitumumab

와 같다.

표 3-4에서 식별된 후보 약물들에 대하여, 질병-약물 관계를 제공하는 데이터베이스인 CTD, Clinicaltrials.gov 등이 예측한 후보 약물을 기준으로 사용하여 검증하였다.

Table 3. Comparison Candidate with Gold Standard from CTD

Disease	Percentage (Gold standard / Candidate)
Breast Cancer	71.4%
Prostate Cancer	70.0%
Asthma	40.9%

Table 4. Comparison Candidate with Gold Standard from Clinicaltrials.gov

Disease	Percentage (Gold standard / Candidate)
Breast Cancer	71.4%
Prostate Cancer	30.0%
Asthma	40.9%

우리는 다른 데이터베이스에서 식별되지 않은 약물을 새로운 질병-약물 관계로 식별하여 최종 후보 약물을 선정하였다. 표 5 는 식별된 최종 후보 약물 리스트이다.

Table 5. Candidate Drug List for Drug Repositioning

Disease	Candidate Drug
Breast Cancer	Maraviroc
Prostate Cancer	Conjugated Equine Estrogens
Asthma	Panitumumab
	Osimertinib
	Necitumumab
	OspA lipoprotein
	Nilutamide
	Drostanolone
	Nandrolone phenpropionate
	Enzalutamide

Ⅴ. Conclusion

본 연구에서는 MEDLINE에서의 문헌을 분석하고, 각 질병과 유전자의 동시 출현 빈도를 계산하여, 질병에 대한 중요 유전자를 식별하였다. 유전자 간 상호작용의 정도를 가중치로 가지고 있는 네트워크상에서, 질병-약물 관계의 최단 거리 경로를 구축하였다. 중요 유전자 리스트와 기존의 약물과 질병의 작용 경로에 포함된 유전자를 비교하여 각 약물에 대하여 스코어를 계산하였다. 각 질병-약물 관계에서 계산된 스코어를 통하여 기존의 질병-약물 관계를 대체 할 수 있는 새로운 관계를 발견하였고, 이를 통하여 새로운 약물 재창출 후보를 식별하였다. 본 연구에서는 문헌에서 추출한 중요 유전자를 식별하였고, 이 중요 유전자가 유효하다는 것을 보였으며, 네트워크와 문헌간의 관련성의 이용과 함께 유방암, 전립선암 그리고 천식에 대하여 후보 재창출 후보 약물을 제시하였다.

기존 연구에서는 문헌에서의 유전자와 질병의 동시 출현 시 두 단어의 관계에 대한 성격을 고려하지 않는다. 향후 질병과 유전자의 동시 출현의 빈도를 추출 시, 동시 출현의 여부와 함께 질병과 유전자 사이의 성질까지 고려할 예정이다. 또한 문헌은 대부분 사람을 통해 작성되기 때문에, 자연어에 대해 분석할 수 있는 NLP 기술을 통하여 한 문헌의 초록을 구성하는 문장들의 각 성질을 식별하고, 언급되는 질병과 약물 그리고 유전자의 관계를 분석하여, 기존 연구보다 효율적인 질병-약물 관계를 찾는 연구를 진행할 예정이다.

REFERENCES

[1] Oprea, T. I., and J. Mestres. "Drug repurposing: far beyond new targets for old drugs." The AAPPS journal Vol.14, No. 4, pp.759-763, Jul. 2012

[2] Liu, Zhichao, et al. "In silico drug repositioning–what we need to know" Drug discovery today, Vol.18, No. 3, pp. 110-115, Feb. 2013.
[3] Andronis, Christos, et al. "Literature mining, ontologies and information visualization for drug repurposing" Briefings in bioinformatics, Vol.12, No. 4, pp. 357-368, Jul. 2011.
[4] Jang et al. "Novel Drug Similarity Measuring Method based on Text Mining for Predicting Similar Drugs." Journal of KIIT, Vol.4, No. 7, pp. 127-137, Jul. 2016.
[5] Wu, Zikai, Yong Wang, and Luonan Chen. "Network-based drug repositioning" Molecular BioSystems, Vol.9, No. 6, pp. 1268-1281, Sep. 2013.
[6] Cheng, Feixiong, et al. "Prediction of drug-target interactions and drug repositioning via network-based inference." PLoS Comput Biol Vol.8, No. 5, pp. ppe1002503, May, 2012.
[7] Canese, Kathi, and Sarah Weis. "PubMed: The bibliographic database" The NCBI Handbook. 2013.
[8] Law, Vivian, et al. "DrugBank 4.0: shedding new light on drug metabolism" Nucleic acids research, Vol. 42, No. D1, pp. D1091-D1097, Jan. 2014.
[9] Whirl-Carrillo, Michelle "Pharmacogenomics Knowledge for Personalized Medicine" Clinical Pharmacology & Therapeutics, Vol.92, No. 4, pp. 414-417, Oct. 2012.
[10] Maglott, Donna, et al. "Entrez Gene: gene-centered information at NCBI" Nucleic acids Research, Vol.39, suppl 1, pp. D52-D57, Jan. 2011.
[11] Kibbe, Warren A., et al. "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data" Nucleic acids research, Vol.43, No. D1, pp. gku1011, Oct. 2014.
[12] Davis, Allan Peter, et al. "The Comparative Toxicogenomics Database's 10th year anniversary: update 2015" Nucleic acids research, Vol.43, No. D1, pp. D914-D920, Jan. 2015.
[13] Lee, Insuk, et al. "Prioritizing candidate disease genes by network-based boosting of genome-wide association data" Genome Research, Vol.21, No. 7, pp. 1109-1121, Jul. 2011.
[14] Yang, Hong, et al. "Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information" Nucleic acids research, Vol.44, No. D1, pp. D1069-D1074, Jan. 2016.
[15] Zarin, Deborah A., et al. "The ClinicalTrials.gov results database—update and key issues" New England Journal of Medicine, Vol.364, No. 9, pp. 852-860, Mar. 2011.
[16] Piñero, Janet, et al. "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants." Nucleic Acids Research, pp. gkw943, 2016.
[17] Kanehisa, Minoru, et al. "KEGG: new perspectives on genomes, pathways, diseases and drugs." Nucleic Acids Research, Vol.45, No. D1, pp. D353-D361, Jan. 2017.
[18] Genetics Home Reference, https://ghr.nlm.nih.gov/resources
[19] Kim et al "Inferring Disease-related Genes using Title and Body in Biomedical Text" KIISE Transactions on Computing Practices, Vol. 23, No. 1, pp. 28-36, Jan. 2017.
[20] Butcher, J. C. "Random sampling from the normal distribution" The Computer Journal, Vol.3, No. 4, pp. 251-253, Jan. 1961.

Authors

Minseok Park is an undergraduate student of Computer Science and Engineering at Gachon University, Korea. Minseok Park is currently an undergraduate researcher in Data Mining & Bioinformatics laboratory, Gachon University. He is interested in network biology and data mining.

Giup Jang received the B.S. degrees in Computer Science and Engineering from Gachon University, Korea, in 2016. Giup Jang is currently a master researcher in the Department of IT Convergence Engineering in Gachon University. He is interested in text mining, bioinformatics.

Taekeon Lee is an undergraduate student of Computer Science and Engineering at Gachon University, Korea. Taekeon Lee is currently an undergraduate researcher in Data Mining & Bioinformatics laboratory, Gachon University. He is interested in network biology and data mining.

Youngmi Yoon received the B.S. degree from Seoul National University in 1981; the M.S. degrees in statistics and computer science from Stanford University in 1984 and 1987 respectively, and the Ph.D. degree in computer science from Yonsei University in 2008. Youngmi Yoon worked as a software engineer from 1987 to 1993 at IntelliGenetics Corp. in Mountain View, CA, USA. She's been a professor at Gachon University from 1995. Her research interest includes database, data science, data mining, and bioinformatics.