Journal of KIIT. Vol. 16, No. 11, pp. 1-9, Nov. 30, 2018. pISSN 1598-8619, eISSN 2093-7571 1
http://dx.doi.org/10.14801/jkiit.2018.16.11.1

유전자를 중간 매개로 고려한 동시발생 기반의 약물-질병 관계 추론

신상원*[1], 신예은*[2], 장기업**, 윤영미***

Co-occurrence Based Drug-disease Relationship Inference with Genes as Mediators

Sangwon Shin*[1], Yeeun Sin*[2], Giup Jang**, and Youngmi Yoon***

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(No. 2018R1A2B6006223)

요 약

신약 재창출은 현재 사용되는 약물의 새로운 용도를 발견하는 방법이다. 텍스트 마이닝은 정형화되지 않은 문서로부터 의미 있는 지식을 획득하는 과정을 의미한다. 본 논문에서는 약물-유전자와 유전자-질병에서 동시에 측정된 유전자 출현 빈도의 비율을 고려하여 새로운 약물-질병 관계를 추론하는 방법을 제안한다. 생물학적 문헌으로부터 약물-유전자와 유전자-질병의 동시출현 빈도를 측정하고 각 약물과 질병에 대하여 유전자의 출현 비율을 계산한다. 약물-질병 관계의 가중치는 동시에 측정된 유전자 출현 비율의 평균을 이용하여 계산되고 이를 이용하여 각 질병의 분류 정확도를 측정한다. 약물-질병 관계를 추론하는 것에서 동시출현 빈도를 문장 단위로 측정하고 여러 관계를 고려하는 방법이 기존 방법보다 더 정확히 식별해내는 것을 보였다.

Abstract

Drug repositioning is to discover new uses of drugs. Text mining derives knowledge from unstructured text. We propose a method to predict new drug-disease relationships by taking into account the rate of frequency of genes simultaneously measured in disease-gene and gene-drug. Co-occurrence of drug-gene and gene-disease in the biological literature is counted and calculate the rate of the gene for each drug and disease. Weights of drug-disease relationships are calculated using the average of the rates of genes that are measured and used to measure the accuracy for each disease. In measuring drug-disease relationships, a more accurate identification of relationships was shown by measuring the frequency on a sentence and considering multiple relationships than existing method.

Keywords
text mining, data mining, co-occurrence, drug repositioning, bioinformatics

* 가천대학교 컴퓨터공학과
- ORCID[1]: http://orcid.org/0000-0002-2879-889X
- ORCID[2]: http://orcid.org/0000-0003-2127-0598
** 가천대학교 IT융합공학과
- ORCID: http://orcid.org/0000-0003-0027-3925
*** 가천대학교 컴퓨터공학과 교수(교신저자)
- ORCID: http://orcid.org/0000-0002-7420-4968

· Received: Aug. 28, 2018, Revised: Nov, 02, 2018, Accepted: Nov. 05, 2018
· Corresponding Author: Youngmi Yoon
Dept. of Computer Engineering Gachon University, Korea
Tel.: +82-31-750-4755, Email: ymyoon@gachon.ac.kr

Ⅰ. 서 론

새로운 약물이 개발되고 시장에 출시되기까지 평균적으로 15년의 기간과 10억 달러의 비용이 소요된다고 알려져 있다[1][2]. 또한 약물 개발 과정 초기에 제안되는 후보 약물 10,000개 중 한 개 정도만 약물의 효능과 안정성을 인정받아 시장에 출시된다[3]. 이러한 한계점을 극복하기 위하여 등장한 방법이 신약 재창출(Drug Repositioning)이다. 신약 재창출은 기존에 사용되어 온 약물을 다른 질병에 사용될 수 있게 만드는 방법을 의미한다[4]. 신약 재창출의 대표적인 사례로는 항우울제로 사용되어 온 부프로피온과 말라리아 치료제로 사용되어 온 아모디아퀸이 있다. 부프로피온은 항우울제로 개발이 되었으나 현재는 금연보조제로도 사용되고 있다[5]. 또한 Kim 등은 말라리아 치료제인 아모디아퀸이 당뇨병과 비만과 같은 대사성 질환에도 효과가 있음을 증명하였다[6]. 신약 재창출은 이미 FDA에 승인된 약물이나 시장에 출시된 약물을 대상으로 진행하기 때문에 약물의 안정성 측면에서 이점을 가지고 있으며 신약 개발 시간과 비용을 획기적으로 줄일 수 있다.

컴퓨터를 이용한 신약 재창출 접근법에는 경로(Pathway) 기반, 네트워크 기반, 분자 기반 등 다양한 방법이 있다[7]. 그 중 텍스트 마이닝(Text mining)은 정형화되지 않은 문서로부터 새로운 의미 있는 지식을 발견하는 과정을 의미한다[8]. 텍스트 마이닝은 Kveler 등이 생물 시스템의 복잡성과 그 폭을 파악하는 중요한 수단이라고 말할 만큼 생물정보학 분야에서 최근 주목받고 있는 접근법이다[9]. 텍스트 마이닝은 이전의 다른 방법들에 비하여 빠르고 다른 방법으로는 찾을 수 없는 정보를 찾아낼 수 있다[10]. Cohen과 Hunter는 논문의 양이 매우 빠르게 증가하고 있다고 언급하였으며[11], 이는 사용 가능한 데이터의 양이 많다는 것을 의미한다. 이처럼 사용 가능한 텍스트 데이터의 양이 기하급수적으로 증가함에 따라 이로부터 기존에 알려지지 않은 새로운 정보를 발견하기 위한 다양한 연구들이 진행되었다.

Frijters 등은 생물학적 문헌으로부터 서로 다른 두 키워드가 동시에 출현하는 빈도를 측정하는 도구를 개발하였으며 이를 이용하여 약물과 유전자, 질병 간의 밝혀지지 않은 관계를 발견하기 위한 방법을 제안하였다[12][13]. 먼저 그들은 생물학적 문헌의 초록으로부터 약물과 유전자, 질병이 상호 간 동시에 출현(Co-occurrence)하는 빈도수를 측정하였으며 이를 데이터베이스로 구축하였다. 동시출현은 텍스트에서 두 키워드가 동시에 출현하면 두 키워드가 관련이 있다는 가정을 기반으로 연구를 하는 방법이다[14]. 약물과 질병 간의 관계를 밝혀내기 위하여 그들은 R-scaled score를 이용하였다. R-scaled score는 약물과 질병에서 동시에 출현하는 유전자들을 대상으로 ABC 모델을 이용하여 약물-유전자와 유전자-질병의 동시출현 값 중 작은 값들의 평균을 계산한다. ABC 모델은 Swanson이 제안한 모델로서 A와 C의 관계가 알려지지 않은 상태에서 A와 B가 서로 연관이 있고 B와 C가 관련이 있다면 A와 C도 밀접한 관계가 있다는 전제조건을 기반으로 하는 관계 발견 방법이다[15]. 앞선 과정을 통하여 계산된 값은 분류 특징으로 사용하여 분류 정확도를 측정하였다.

Kim 등은 생물학적 정보와 문헌을 활용하여 약물의 유사도를 측정하고, 이를 기반으로 약물의 새로운 효과를 추론하는 방법을 제안하였다[16]. 그들은 약물과 관련된 데이터베이스로부터 약물의 효과 및 부작용에 대한 데이터를 추출하였으며 이를 기반으로 특정 효과 및 부작용이 나타나는 약물들을 하나의 군집으로 형성하였다. 하나의 군집과 그 군집에 포함되지 않는 후보 약물과의 유사도를 측정하기 위하여, 먼저 그들은 생물학적 문헌에서 군집이 공통으로 가지고 있는 작용과 후보 약물이 가지고 있는 작용이 동시출현하는 문장의 빈도를 측정하였다. 이를 군집에서 후보 약물의 작용이 나타나는 약물의 수를 측정한 값과 곱한 후 이들의 합을 계산하였다. 앞선 과정을 통하여 계산된 값이 클수록 후보 약물에 비교 군집에서 공통으로 나타나는 작용이 있을 것이라고 추론하였다.

그러나 그들이 제안한 방법에는 다음과 같은 한계점이 존재한다. Frijters 등은 생물학적 문헌으로부터 초록 단위로 키워드 간의 동시출현 빈도를 측정

하였다. 서로 다른 키워드가 한 문장이 아닌 초록에서 빈번하게 출현한 것이 키워드 간에 밀접한 관계가 있다는 것과 동일하다고 보기 어렵다. 또한 약물과 질병 간의 가중치를 계산할 때 동일한 유전자에 대하여 약물-유전자의 빈도와 유전자-질병의 빈도 중 작은 값을 이용한 것은 약물과 질병의 빈도 중 하나의 값을 선택하여 대표적인 값으로 사용한 것으로 다른 한 쪽의 나머지 값은 전혀 고려되지 못한다는 한계가 있다. Kim 등은 단순히 약물의 특정 효과와 부작용만을 이용하여 후보 약물을 제안하였다. 이는 유전자가 치료나 부작용에 어떠한 영향을 미치는지 확인할 수 없다.

본 연구에서 우리는 질병에서 측정된 유전자의 출현 빈도의 비율을 고려하여 새로운 약물-질병 관계를 추론하는 방법을 제안한다. 제안하는 방법에서는 생물학적 문헌의 초록을 문장 단위로 정제한 후, 한 문장에서 약물-유전자 또는 유전자-질병의 동시 출현 빈도를 각각 측정한다. 각 약물과 질병에 대하여, 각 유전자의 출현 빈도와 모든 유전자의 출현 빈도 합을 이용하여 해당 유전자의 출현 비율을 계산한다. 약물-질병 관계의 가중치는 약물-유전자와 유전자-질병에서 동시에 측정되는 유전자의 출현 비율의 평균을 이용하여 계산한다. 본 연구에서 제안하는 방법은 문장 단위의 실험이기 때문에 약물과 유전자, 질병 간의 관계를 더 자세하게 파악할 수 있으며, 동일한 유전자에서 측정된 약물과 질병의 출현 빈도를 모두 고려하여 약물-질병 관계를 추론할 수 있다.

본 연구의 전체적인 구성은 다음과 같다. 2장에서는 본 연구에서 사용된 데이터에 대한 설명과 관련 연구들을 소개한다. 3장에서는 본 연구의 시스템 개요와 실험 과정을 기술하며, 4장에서는 실험 환경과 결과를 기술한다. 5장에서는 본 연구에 대한 결론을 기술한다.

II. 실험 데이터

2.1 생물학적 문헌 데이터베이스

텍스트 마이닝을 기반으로 약물-질병 관계를 추론하기 위해서는 방대한 양의 생물학적 문헌 데이터를 필요로 한다. MEDLINE은 생물정보학 분야에서 널리 사용되고 있는 공공 문헌 데이터베이스로서, 생명공학과 생물학적 정보에 대한 내용을 포함하는 문헌을 제공한다[17]. MEDLINE은 2,500만개의 논문 데이터를 제공하고 있다. 본 연구에서는 Kim 등의 연구에서 정제된 9,803,245개의 MEDLINE 초록 데이터(http://embio.yonsei.ac.kr/files/hjkim/LDDSim/Kim_LDDSim.zip)를 사용하며[18], 초록을 문장 단위로 정제한 후 약물과 질병, 유전자를 식별하여 약물-유전자와 유전자-질병 동시출현 빈도를 측정한다.

2.2 약물 정보 데이터베이스

문장에서 약물을 식별하기 위해서는 약물의 명칭에 대한 정보가 필요하다. 본 연구에서는 약물의 명칭과 동의어에 대한 정보를 수집하기 위하여 DrugBank(http://www.drugbank.ca/)와 KEGG DRUG (http://www.genome.jp/kegg/)를 사용한다. DrugBank는 상세한 약물 데이터와 복합적인 약물 표적에 관한 정보를 포괄적으로 제공하는 공공 데이터베이스이다[19]. DrugBank에서는 FDA 승인된 약물에 대한 공식적인 명칭을 제공한다. 약물의 이름은 국가와 제약 회사마다 사용하는 명칭이 다르기 때문에, 약물의 동의어를 고려하는 과정이 필요하다. 우리는 약물 명칭의 동의어를 고려하기 위하여 KEGG DRUG로부터 데이터를 수집한다. KEGG DRUG는 일본과 미국, 유럽에서 승인된 약물에 대한 포괄적인 정보를 제공하는 데이터베이스이다[20]. 본 연구에서 우리는 DrugBank와 KEGG DRUG에서 동시에 제공하는 1,763개의 약물에 대하여 약물의 공식적인 명칭과 동의어를 수집한다.

2.3 질병 정보 데이터베이스

약물과 동일하게 질병의 이름 또한 다양하게 불리기 때문에 이를 표준화하는 과정이 필요하다. 이를 위하여 우리는 Disease Ontology(http://disease-ontology.org)를 이용하여 질병의 공식적인 이름뿐만 아니라 동의어까지 고려한다. Disease Ontology는 질

병에 대한 표준 정보를 제공하는 데이터베이스로서, 인간 질병에 대한 표준적인 용어와 질병과 관련된 상세한 데이터를 제공한다[21]. Disease Ontology는 질병의 개념과 질병에 기여하는 유전자, 관련 증상, 결과 및 징후를 제공한다. 본 연구에서는 Disease Ontology로부터 10가지 질병(유방암, 전립선암, 폐암, 천식, 고혈압, 저혈압, 당뇨, 에이즈, 알츠하이머, 대장암)에 대한 명칭과 동의어를 수집하여 문장에서 유전자 심볼과의 동시출현 빈도를 측정한다.

2.4 유전자 정보 데이터베이스

문장에서 유전자를 식별하기 위해서는 유전자 심볼이 필요하다. 이를 위하여 PharmGKB(http://www.pharmgkb.org/)로부터 유전자 심볼 정보를 수집한다. PharmGKB는 인간 유전자의 약물 반응에 대한 변이 정보를 제공하는 데이터베이스로서 약물이 유전자에 미치는 영향과 약물과 유전자의 치료 관계에 대한 정보를 제공한다[22]. PharmGKB는 약물 유전체 정보가 포함된 약물 라벨을 제공한다. 본 연구에서는 PharmGKB로부터 27,000개의 유전자 심볼을 수집하여 MEDLINE의 문헌 문장에서 동의어를 고려한 질병 또는 약물과의 동시출현 빈도를 측정하는데 사용한다.

2.5 약물-질병 관계 데이터베이스

약물-질병 관계 추론을 위한 분류기를 학습하기 위해서는 기존에 알려진 관계 정보가 필요하다. 이를 위하여 CTD(http://ctdbase.org/)로부터 약물과 질병의 치료 관계를 수집한다[23]. CTD는 약물과 유전자, 질병 간의 관계를 수동으로 정제한 데이터베이스로서 이들 간의 관계 정보를 제공한다. 본 연구에서는 실험에 사용한 10가지 질병에 대한 약물의 치료 관계를 수집한다. 수집된 정보는 각 질병에 대하여 분류기의 클래스를 지정할 때 사용된다. 한 질병에 대한 해당 약물의 치료 관계가 이미 밝혀져 있다면 TRUE를, 그렇지 않다면 FALSE를 부여한다. 이를 기반으로 AUC를 계산하여 10가지 질병에 대한 분류 성능을 측정한다.

III. 동시발생 기반 약물-질병 관계 추론

본 논문에서는 유전자를 중간 매체로 사용한 약물과 질병의 관계 추론을 위하여 MEDLINE 데이터에서 문장 단위로 유전자가 약물 또는 질병과 동시에 발생하는 빈도를 측정한다. PharmGKB에서 제공하는 27,000개의 유전자 정보, 약물의 동의어를 고려하기 위하여 DrugBank와 KEGG DRUG에서 동시에 제공하는 1,763개의 약물 이름과 동의어, Disease Ontology에서 제공하는 질병의 이름과 동의어 정보를 사용한다. 본 논문에서는 10가지의 질병에 대하여 실험을 진행한다. 그림 1은 전체적인 시스템 개요이다.

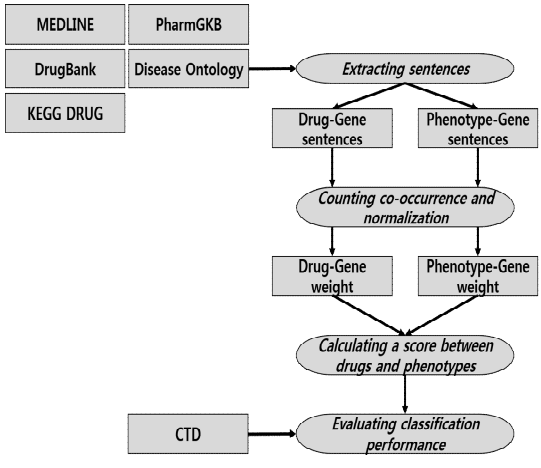

그림 1. 시스템 개요
Fig. 1. System overview

3.1 동시출현 측정

본 논문에서는 각각의 약물과 질병에 대하여 문장 단위에서 유전자와의 동시출현 빈도를 측정한다. 한 문장에서 질병 이름과 유전자 심볼이 동시에 언급되었다면, 해당 질병과 유전자의 동시출현 빈도를 1 증가시킨다. 동시출현 빈도를 측정 시 불용어는 고려하지 않고 Disease Ontology, DrugBank, KEGG DRUG에서 제공하는 질병의 이름과 유전자의 심볼이 동시에 언급이 될 경우에만 빈도를 증가시킨다. 불용어는 영어에서 대명사, 전치사, 관사와 같이 빈번하게 사용되는 단어로, 정보 검색 시 색인어로 사

용되지 않는 용어를 말한다. 불용어는 문서에서 불특정하게 출현하고 의미가 명확하지 않는 경우가 많아 동시출현 빈도 측정 시 제외한다. 한 약물에서 모든 유전자에 대한 동시출현 빈도가 전부 0이라면 해당 약물은 질병과의 관계를 측정할 때 유전자를 중간 매체로서 사용할 수 없으므로 실험에서 제외한다. 유전자-질병과 약물-유전자의 동시출현 빈도를 정규화하기 위하여 약물과 질병에 대하여 유전자의 비율 값을 부여한다. 한 질병 혹은 약물에서 모든 유전자에 대한 동시출현 값의 합을 구한 후, 각각의 동시출현 값을 합으로 나눈다. 정규화 된 동시출현 값을 각 관계의 가중치로 부여한다. 정규화 과정을 통하여 모든 값은 0에서 1사이의 값을 가지게 된다.

3.2 질병과 약물 간의 스코어 계산

유전자를 중간 매개로 하는 스코어를 계산하기 위하여 질병과 약물에서 공통으로 동시출현 값을 가지는 유전자들을 사용한다. 질병과 약물이 같은 유전자에 대해 모두 0이 아닌 동시출현 값을 가진다면 두 값의 평균을 계산하고, 모든 유전자에 대해 시행하여 나온 평균값들의 평균을 구한다.

$$Score = \frac{\sum_{k=1}^{n} \frac{d_k + p_k}{2}}{n} \quad (1)$$

식 (1)에서 d_k는 약물과 K번째 유전자의 정규화 된 동시출현 값, p_k는 질병과 K번째 유전자의 정규화 된 동시출현 값이다. n은 질병과 약물 사이에서 공통으로 동시출현 값을 가진 유전자의 개수이다. d_k와 p_k가 둘 다 존재하지 않거나 하나만 존재하는 유전자는 n에서 제외한다. 식 (1)의 방법으로 계산된 값을 해당 질병과 약물의 스코어라 명명한다. 스코어의 값이 0이 나올 경우 해당 약물과 질병은 겹치는 유전자가 없어 유전자를 중간 매체로 두지 않는다는 것을 의미하기 때문에 실험에서 제외한다. 그림 2는 스코어 계산 예시이다.

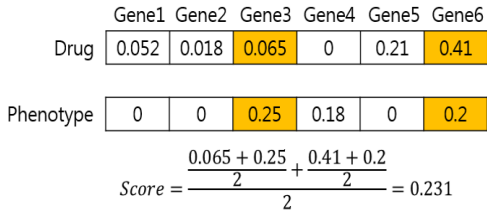

그림 2. 약물-질병의 스코어 계산
Fig. 2. Drug-disease score calculation

한 질병과 약물에 대하여 유전자의 동시출현 값이 그림 2와 같다면, 유전자3, 유전자5와 같이 질병과 약물이 공통으로 값을 가지는 유전자를 사용하여 스코어를 계산한다. 0.052, 0.018, 0.18, 0.21과 같이 질병과 약물에서 동시에 값을 갖지 않을 경우 스코어 계산에서 제외한다. 그림 2에서는 0.065와 0.25, 0.41과 0.2의 평균을 각각 구한 후 사용한 유전자의 개수인 2로 나눈 결과인 0.231을 질병과 약물의 스코어가 된다.

각 질병에 대하여 모든 약물에 대한 스코어를 계산한다. 관련이 없는 약물을 제외하는 과정을 거치면서 질병에 따라 사용되는 약물의 개수가 달라진다. 본 연구에서 사용된 해당 질병에서 사용할 수 있는 약물의 데이터의 개수는 표 1과 같다.

표 1. 질병에 대한 약물 개수
Table 1. Number of drugs for diseases

Disease	The number of drugs
Breast Cancer	1,347
Prostate Cancer	1,329
Lung Cancer	1,317
Asthma	1,331
High Blood Pressure	1,347
Low Blood Pressure	1,182
Diabetes	1,354
HIV	1,358
Alzheimer's Disease	1,303
Colon Cancer	1,309

3.3 AUC 측정

본 논문에서 제시하는 약물-질병 관계 추론을 위한 분류기의 정확도를 측정하기 위하여 약물과 질병의 치료 관계를 제공하는 데이터베이스인 CTD에서 약물-질병 관계 정보를 사용한다. 해당 약물과

질병의 치료 관계가 밝혀진 것은 TRUE, 밝혀지지 않은 것은 FALSE로 클래스 값을 부여한다. 각 질병에 대하여 랜덤 포레스트, 로지스틱, 나이브 베이즈를 시행하고 각 방법에 따른 AUC를 측정한다.

Ⅳ. 성능 평가

본 논문에서 제시하는 분류기의 성능 비교를 위하여 총 세 가지의 비교실험을 시행하였다. 첫 번째 비교실험으로 코사인 측정 방법을 사용하였다. 벡터 내적 계산에서 코사인 값은 그 값이 클수록 두 벡터의 방향성이 비슷하고 값이 작을수록 두 벡터의 방향성이 다르다는 것을 의미한다. 약물-유전자와 유전자-질병 동시출현 값을 각각의 벡터로 놓고 코사인 값을 계산하여 약물-질병의 스코어로 사용하였으며 이를 Cos1이라 하였다. 두 데이터의 연관성이 클수록 높은 스코어를 가지고 그렇지 않다면 낮은 스코어를 가진다. 코사인 계산은 각 벡터 원소에 0이 많으면 값이 작게 나올 수밖에 없다. 이 문제를 해결하기 위하여 두 번째 비교실험으로 같은 유전자에 대해 유전자-질병, 약물-유전자의 동시출현 값이 모두 0이 아닌 유전자만을 사용하여 코사인 계산을 하였으며 이를 Cos2라 하였다. 세 번째 비교실험은 Frijters 등이 사용한 R-scaled score를 사용하여 진행하였다. 정규화를 진행한 약물-유전자와 유전자-질병에서 같은 유전자에 대해 양 쪽 모두의 값이 0이 아닐 때 두 값 중에서 작은 값을 취하고, 그 값들의 평균을 약물-질병의 스코어로 사용하였다. 비교실험에서도 본 실험과 같이 CTD 데이터를 사용하여 클래스를 부여하고, 각 질병에 대해 랜덤 포레스트, 로지스틱, 나이브 베이즈를 시행한 후 각 방법에 대해 AUC를 측정하였다.

그림 3은 랜덤 포레스트를 이용하여 본 연구에서 제안된 방법과 기존의 방법의 분류 정확도를 비교한 그래프이다. AUC 평균은 Cos1이 0.589, Cos2이 0.586, R-scaled이 0.613, 본 연구의 방법이 0.664이다. 본 연구에서 제안된 방법의 AUC 평균이 Cos1 평균에 비해 0.075, Cos2 평균에 비해 0.078, R-scled 평균에 비해 0.051 더 높았다. 특히 에이즈의 경우 본 논문의 방법이 Cos2에 비하여 0.138 더 높은 AUC를 보였다.

그림 4는 로지스틱을 이용하여 본 연구에서 제안된 방법과 기존의 방법의 분류 정확도를 비교한 그래프이다. AUC 평균은 Cos1이 0.588, Cos2이 0.674, R-scaled이 0.551, 본 연구의 방법이 0.734이다. 본 연구에서 제안된 방법의 AUC 평균이 Cos1 평균에 비해 0.146, Cos2 평균에 비해 0.06, R-scled 평균에 비해 0.183 더 높았다. 특히 에이즈의 경우 본 논문의 방법이 R-scaled에 비하여 0.236 더 높은 AUC를 보였다.

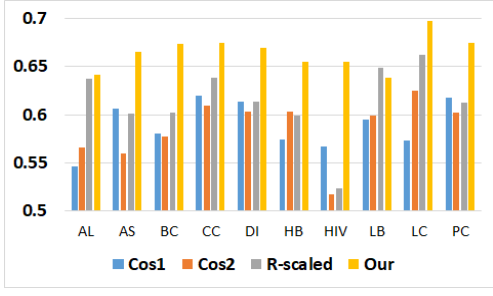

그림 3. Random Forest 기반의 분류 성능 비교 결과
Fig. 3. Random Forest based classification performance comparison results

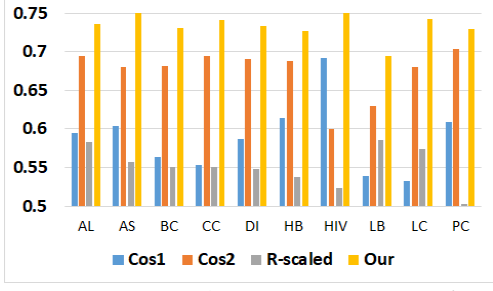

그림 4. Logistic 기반의 분류 성능 비교 결과
Fig. 4. Logistic based classification performance comparison results

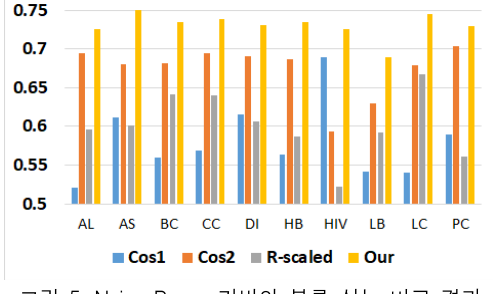

그림 5. Naive Bayes 기반의 분류 성능 비교 결과
Fig. 5. Naive Bayes based classification performance comparison results

그림 5는 나이브 베이즈를 이용하여 본 연구에서 제안된 방법과 기존의 방법의 분류 정확도를 비교한 그래프이다. AUC 평균은 Cos1이 0.58, Cos2이 0.673, R-scaled이 0.601, 본 연구의 방법 0.73이다. 본 연구에서 제안된 방법의 AUC 평균이 Cos1 평균에 비해 0.15, Cos2 평균에 비해 0.057, R-scled 평균에 비해 0.129 더 높았다. 특히 폐암의 경우 본 논문의 방법이 Cos1에 비하여 0.205 더 높은 AUC를 보였다.

랜덤 포레스트를 제외한 로지스틱과 나이브 베이즈에서는 모든 질병에 대하여 본 연구에서 제안된 방법의 AUC가 비교실험 결과에 비해 높았다. 랜덤 포레스트에서는 저혈압이 본 연구에서 제안된 방법보다 R-scaled가 0.011 높은 AUC를 보였다. 랜덤 포레스트에서는 저혈압을 제외한 모든 질병의 AUC가 기존 방법에 비해 본 연구에서 제안된 방법이 높았다.

Ⅴ. 결론 및 향후 과제

본 논문에서는 방대한 생물학적 문헌 데이터를 문장 단위로 정제한 후 한 문장에서 약물-유전자와 유전자-질병의 동시출현 빈도를 측정하였다. 각 유전자의 출현 빈도와 모든 유전자의 출현 빈도 합을 이용하여 해당 유전자의 출현 비율을 계산하였다. 이를 이용하여 약물과 질병에서 동시에 출현하는 유전자의 출현 비율의 평균을 이용하여 약물-질병 관계를 측정하였다. 이를 기존의 방법인 코사인 유사도와 Frijters 등이 사용한 R-scaled score를 이용하여 구한 약물-질병 관계와 AUC를 비교하였을 때 제안하는 방법이 더 높은 AUC를 가지는 것을 보였다. 제안한 방법으로 계산된 스코어가 높은 약물들 중 기존에 사용되고 있지 않는 약물은 신약 재창출을 시도할 수 있을 것으로 기대된다. 하루에도 수많은 텍스트 데이터들이 생성되고 있는 만큼 향후 연구에서는 최신 데이터 및 다른 논문의 데이터베이스를 추가하여 더욱 정확하게 스코어를 측정할 예정이다. 약물은 어떤 유전자, 혹은 질병에 효과가 있거나 부작용을 유발한다. 약물은 어떠한 질병을 치료하는 것을 목적으로 함으로 효과가 있는 것은 긍정적인 관계로 부작용을 유발하는 것은 부정적인 관계로 볼 수 있다. 본 연구에서는 동시출현 빈도를 측정 시 약물-유전자 또는 유전자-질병의 관계가 부정적인 의미인지 긍정적인 의미인지를 구분하지 않아 부정적인 관계를 가지는 약물이 상위 스코어에 위치할 가능성이 존재한다. 향후 연구에서는 동시출현 빈도를 측정 시 관계의 의미를 분석하여 긍정적인 관계에는 높은 가중치를 부여하여 더 정확한 스코어를 측정할 예정이다.

References

[1] Bruce Booth and Rodney Zemmel, "Prospects for productivity", Nature Reviews Drug Discovery, Vol. 3, No. 5, pp. 451-456, May. 2004.

[2] DiMasi Josph A, "New drug development in the United States from 1963 to 1999", Clinical Pharmacology & Therapeutics, Vol. 69, No. 5, pp. 286-296, May. 2001.

[3] Tokens Ross, "An overview of the drug development process", Physician executive, Vol. 31, No. 3, pp. 48-52, May/June. 2005.

[4] Jiao Li, Si Zheng, Bin Chen, Atul J.Butte, S. Joshua Swamidass, and Zhiuong Lu, "A survey of current trends in computational drug repositioning", Briefings in Bioinformatics, Vol. 17, No. 1, pp. 2-12, Jan. 2016.

[5] Douglas E. Jorenby, Scott J. Leischow, Mitchell A. Nides, Stephen I. Rennard, J. Andrew Johnston, Arlene R. Hughes, Stevens S. Smith, Myra L. Muramoto, David M. Daughton, Kimberli Doan, Michael C. Fiore, and Timothy B. Baker, "A Controlled Trial of Sustained-Release Bupropion, a Nicotine Patch, or Both for Smoking Cessation", New England Journal of Medicine, Vol. 340, No. 9, pp. 685-691, Mar. 1999.

[6] Hoe-Yune Jung, Bobae Kim, Hye Guk Ryu, Yosep Ji, Soyoung Park, Seung Hee Choi, Dohyun Lee, In-Kyu Lee, Munki Kim, You Jeong Lee, Woojin Song, Young Hee Lee, Hyung Jin

Choi, Chang-Kee Hyun, Wilhelm H Holzapfel, and Kyong-Tai Kim, "Amodiaquine improves insulin resistance and lipid metabolism in diabetic model mice", Diabetes, Obesity and Metabolism, Vol. 20, No. 7, pp. 1688-1701, Mar. 2018.

[7] Rachel A. Hodos, Brian A. Kidd, Shameer Khader, Ben P. Readhead, and Joel T. Dudley, "Computational Approaches to Drug Repurposing and Pharmacology", Wiley Interdisciplinary Reviews: Systems biology and medicine, Vol. 8, No. 3, pp. 186-210, May. 2016.

[8] Ah-Hwee Tan, "Text Mining: The state of the art and the challenges", In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Vol. 8, pp. 65-70, 1999.

[9] Ksenya Kveler, Elina Starosvetsky, Amit Ziv-Kenet, Yuval Kalugny, Yuri Gorelik, Gali Shalev-Malul, Netta Aizenbud-Reshef, Tania Dubovik, Mayan Briller, John Campbell, Jan C Rieckmann, Nuaman Asbeh, Doron Rimar, Felix Meissner, Jeff Wiser, and Shai S Shen-Orr, "Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed", Nature Biotechnology, Vol. 36, No. 7, pp. 651-659, July. 2018.

[10] Roger Hale and Chief Operating Officer, "Text Mining: Getting more value from literature resources", Drug Discovery Today, Vol. 10, No. 6, pp. 377-379, Mar. 2005.

[11] K. Bretonnel Cohen, Lawrence Hunter, "Getting Started in Text Mining", PLoS computational biology, Vol. 4, No. 1, e20, Jan. 2008.

[12] Raoul Frijters, Bart Heupersm pieter van Beek, Maurice Bouwhuis, René van Schaik, Jacob de Vlieg, Jan Polman, and Wynand Alkema, "CoPub: a literature-based keyword enrichment tool for microarray data analysis", Nucleic Acids Research, Vol. 36, No. suppl_2, pp. W406-W410, Apr. 2008.

[13] Raoul Frijters, Marianne van Vugt, Rubem Smeets, René van Schaik, Jacob de Vlieg, and Wynand Alkema, "Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases", PLoS computational biology, Vol. 6, No. 9, e1000943, Sep. 2010.

[14] Wilco W.M. Fleuren and Wynand Alkema, "Application of text mining in the biomedical domain", Methods, Vol. 74, pp. 97-106, Mar. 2015.

[15] DR Swanson, "Medical literature as a potential source of new knowledge", Bulletin of the Medical Library Association, Vol. 78, No. 1, pp. 29-37, Jan. 1990.

[16] Shin Kim, Jeongwoo Kim, and Sanghyun Park, "Inferring Hidden Drug Effects Using Similarity Drugs", Korea Computer Congress 2015, pp. 2074-2076, Jun. 2015.

[17] Kathi Canese and Sarah Weis, "PubMed: the bibliographic database", The NCBI Handbook [Internet]. 2nd edition., Mar. 2013.

[18] Hyunjin Kim, Youngmi Yoon, Jaegyoon Ahn, and SangHyun Park, "A literature-driven method to calculate similarities among disease", Computer methods and programs in biomedicine, Vol. 122, No. 2, pp. 108-122, Nov. 2015.

[19] David S. Wishart, Craig Knox, An Ghi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey, "DrugBank: a comprehensive resource for in silico drug discovery and exploration", Nucleic Acids Research, Vol. 34, No. suppl_1, pp. D668-D672, Jan. 2006.

[20] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima, "KEGG: new perspectives on genomes, pathways, disease and drugs", Nucleic Acids Research, Vol. 45, No. suppl_D1, pp. D353-D361, Nov. 2016.

[21] Warren A. Kibbe, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu,

Christopher J. Mungall, Janos X. Binder, James Malone, Drashtti Vasant, Helen Parkinson, and Lynn M. Schriml, "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. D1071-D1078, Jan. 2015.

[22] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein, "Pharmacogenomics knowledge for personalized medicine", Clinical Pharmacology & Therapeutics, Vol. 92, No. 4, pp. 414-417, Oct. 2012.

[23] Allan Peter Davis, Cynthia J. Grondin, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L. King, Thomas C. Wiegers, and Carolyn J. Mattingly, "The Comparative Toxicogenomics Database's 10th year anniversary: update 2015", Nucleic Acids Research, Vol. 43, No. D1, pp. D914-D920, Jan. 2015.

저자소개

신 상 원 (Sangwon Shin)

2014년 3월 ~ 현재 : 가천대학교 컴퓨터공학과 재학
관심분야 : 데이터 마이닝, 텍스트 마이닝, 바이오인포매틱스

신 예 은 (Yeeun Sin)

2014년 3월 ~ 현재 : 가천대학교 컴퓨터공학과 재학
관심분야 : 데이터 마이닝, 텍스트 마이닝, 바이오인포매틱스

장 기 업 (Giup Jang)

2016년 2월 : 가천대학교 컴퓨터공학과 졸업(학사)
2016년 3월 ~ 현재 : 가천대학교 IT융합공학과(석박통합)
관심분야 : 데이터 마이닝, 텍스트 마이닝, 바이오인포매틱스

윤 영 미 (Youngmi Yoon)

1981년 : 서울대학교 자연과학대학 졸업(학사)
1983년 : 오하이오 주립대학 수학과(학사 수료)
1984년 : 스탠포드대학교 컴퓨터과학과 졸업(이학석사)
2008년 : 연세대학교 컴퓨터과학과 졸업(공학박사)
1987년 ~ 1993년 : Software Engineer, IntelliGenetics, Mountain View, CA, USA
1995년 ~ 현재 : 가천대학교 IT대학 컴퓨터공학과 교수
관심분야 : 데이터베이스, 데이터 사이언스, 데이터 마이닝, 바이오인포매틱스