Journal of The Korea Society of Computer and Information
Vol. 21 No. 1, pp. 115-123, January 2016

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.1.115

Extraction of specific common genetic network of side effect pair, and prediction of side effects for a drug based on PPI network

Youhyeon Hwang *, Min Oh**, Youngmi Yoon ***

Abstract

In this study, we collect various side effect pairs which are appeared frequently at many drugs, and select side effect pairs that have higher severity. For every selected side effect pair, we extract common genetic networks which are shared by side effects' genes and drugs' target genes based on PPI(Protein-Protein Interaction) network. For this work, firstly, we gather drug related data, side effect data and PPI data. Secondly, for extracting common genetic network, we find shortest paths between drug target genes and side effect genes based on PPI network, and integrate these shortest paths. Thirdly, we develop a classification model which uses this common genetic network as a classifier. We calculate similarity score between the common genetic network and genetic network of a drug for classifying the drug. Lastly, we validate our classification model by means of AUC(Area Under the Curve) value.

▸ Keyword : Drug, Side effect, Genetic network, Bioinformatics, Data mining, PPI, Classification model

I. Introduction

약물 부작용(Side Effect)이란 임상에서 환자에게 약물을 투여하였을 때 예상치 못하게 나타나는 모든 반응을 의미 한다. 약물 부작용은 환자의 치료 기간을 더욱 증가 시킬 뿐만 아니라 환자의 건강에 심각한 악영향을 끼치기도 하며, 다양한 추가 비용을 필요로 한다[1]. 약물 부작용의 대표적인 사례로는 1948년부터, 오랜 기간 사용해 왔던 방사선 조영제인 트로트라스트가 암을 유발하는 부작용이 있다는 것 과 1957년 입덧 방지제로 임산부들에게 처방되었던 탈리도마이드가 기형아 출산에 큰 영향을 주었다는 것 등이 있다. 최근에 밝혀진 약물 부작용 사례로는 2004년 관절염 치료제 바이옥스를 복용한 다수의 환자들이 심장 질환을 일으킨 사례 및 2005년 통풍 치료 약물인 벤즈브로마론을 복용했던 환자들이 급성간염으로 사망하는 부작용 사례가 있었다[2]. 최근 약물 부작용 관련 보고는 점점 증가하여 미국 식품 의약품 안전국(FDA)에 매년 50만 건 이상 보고되고 있다. 이에 관련하여 FDA는 1969년부터 현재까지 보고된 9백만건 이상의 부작용 정보를 포함하고 있는 FAERS(FDA Adverse Event Reporting system)를 최근 공개하였다. 이와 같이 약물 부작용 사례가 매해 증가함에 따라 약물 부작용에 관련된 연구는 지속적으로 관심 있게 다루어져 왔다[3].

약물 부작용은 독립적으로 단일하게 나타나기 보다는 대부분 여러 부작용들이 함께 나타난다. 예를 들어, 특정 약물에서 A부작용이 나타나면 B부작용이 함께 나타나는 형태이다. 따라서 대부분의 기존 연구들도 약물 부작용을 독립적으로 다루기 보다는 다양한 부작용을 함께 다뤄왔다[4][5]. 본 연구에서는 이러한 점에 착안하여 다수의 약물에서 빈번하게 공동으로 출현하는 부작용 쌍(Side effect pair)에 집중하였다. 빈번하게 나타나는 부작용 쌍들 중 심각한 것으로 간주되는 부작용 쌍에 대하여 본 연구의 방법으로 두 부작용과 약물이 공유하는 공통 유전자 네트워크를 추출하였으며, 이 네트워크를 분류자로 사용하여 약물 분류 모델을 개발 및 검증하였다.

본 논문의 전체적인 구성은 다음과 같다. 2절에서는 연구와 관련된 약물 부작용 정보 설명, 연구에서 사용한 데이터 자원에

• First Author: Youhyeon Hwang, Corresponding Author: Youngmi Yoon
*Youhyeon Hwang(youhyeonhwang@gmail.com), Dept. of Computer Engineering, Gachon University
**Min Oh(minoh0201@gmail.com), Dept. of Computer Engineering, Gachon University
***Youngmi Yoon (ymyoon@gachon.ac.kr), Dept. of Computer Engineering, Gachon University
• Received: 2016. 01. 20, Revised: 2016. 01. 26, Accepted: 2016. 01. 29.
• 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).

대한 소개 및 분류분석에 사용한 식을 기술한다. 3절에서는 전체적인 본 연구의 시스템 개요, 데이터 수집, 부작용 쌍 선택 기준, 공통 유전자 네트워크 추출 방법, 분류 모델 방법을 기술한다. 4절에서는 실험 환경 및 정확도 측정 방법, 본 연구 분류 모델의 정확도를 기술한다. 마지막으로 본 연구가 앞으로 나아갈 방향에 대하여 5절에 기술하였다.

II. Preliminaries

1. Related works

본 연구는 컴퓨터 실험을 통하여 약물의 부작용을 예측하는 것을 골자로 하는 연구이므로, 약물 부작용 정보에 대한 선행 지식 및 약물 정보를 제공하는 데이터베이스에 대한 정보가 필요하다. 다음과 같은 관련 연구를 통하여 본 연구를 진행하였다.

1.1 Side effect information

본 연구에서는 기존에 연구된 부작용 유전자(Side effect gene) 정보 및 부작용의 심함 정도(Severity of Side Effect)를 사용하였다. 특정 부작용과 관련되어 있다고 알려진 단백질 혹은 유전자를 부작용 유전자라 한다. 이러한 부작용-유전자 관계 정보는 기존 문헌들을 조사하여 추출된 목록을 사용하였다[6].

다양한 부작용들 중 부작용의 심한(Severe) 정도를 판단하는 것은 개개인에 따라 판단의 기준이 다르기 때문에 매우 어렵다. 그러나 최근 약물 부작용의 상대적인 심함 정도(The relative severity of adverse drug reactions)를 판단하는 연구가 있었다[7]. 이 연구에서는 인터넷 기반의 크라우드소싱(Crowdsourcing)을 사용하여 약물 부작용의 심각함 순위 점수를 산출하였다. 총 126,512개의 부작용 쌍을 구성하였고, 각 쌍을 구성하는 두 부작용 중 심각하다고 간주 되는 부작용을 사용자가 선택하도록 하였다. 이에 대한 결과로 총 2929개의 부작용에 대하여 심한 정도 순위를 측정하였다.

1.2 DrugBank Database

DrugBank는 약물 정보를 포함하고 있는 데이터베이스로 생물정보학 및 화학정보학 연구에 자주 사용된다[8]. 약물의 세부적 정보인 화학적, 약학적 데이터를 포함하며 해당 약물의 표적(target)에 대한 종합적인 정보(단백질 염기서열, 구조, 작용경로)를 수록하고 있다. 대부분의 정보는 문헌으로부터 정제 되어 저장되기 때문에 약학 연구자, 약사, 임상연구가등 전문가들에게 높은 신뢰를 얻고 있다. 전 세계적으로 널리 알려져 있는 약물 데이터 자원인 PharmGKB, ChEBI, KEGG, 유전자 및 단백질 데이터 자원인 GeneCards, PDB, PubChem, UniProt 에서 데이터를 가져와 정제하여 수록한다. 전체 8,312개의 약물 및 2,269개의 FDA-승인 약물, 6,000개 이상의 실험 약물을 포함하며 약 4,000개의 단백질 염기서열에 연결된 약물 정보도 수록하고 있다.

이러한 방대한 양의 약물 데이터를 기반으로 사용자는 컴퓨터 실험(in silico)을 통한 약물 표적의 발견, 약물 디자인(design), 약 결합(docking) 및 스크리닝(Screening), 약물의 신진대사 예측 등에 이를 이용 할 수 있다. 본 연구에서는 DrugBank 데이터베이스에서 약물의 표적 유전자 정보를 수집하여 사용하였다.

1.3 Comparative Toxicogenomics Database(CTD)

CTD(Comparative Toxicogenomics Database)는 독성유전학 데이터베이스로서, 질병과 약물의 관계, 질병과 유전자 및 단백질간의 관계, 약물과 유전자간의 관계 정보 등을 포함하는 데이터베이스 이다[9]. CTD는 미국의 국립 환경 보건 과학원(National Institute of Environmental Health Sciences)에서 지원하며 노스캐롤라이나 주립 대학교(NCSU)의 연구자들이 개발 및 운용하고 있다. 본 연구에서는 CTD에 수록된 약물 관련 유전자(Drug related gene)를 수집하여 사용하였다.

1.4 SIDER 4.1 Database

SIDER는 약물의 부작용 정보를 포함하고 있는 데이터베이스로서 여러 자원에 흩어져있는 부작용에 대한 정보를 한 데 모아 통합하여 저장한 데이터베이스이다[10]. 최신 버전은 2015년 10월에 공개되었으며, 1,430개 약물과 5,868개의 부작용 정보를 포함하고 있고 139,756개의 약물-부작용 쌍을 포함하고 있다. 사용자는 컴퓨터 실험을 통하여 이러한 정보를 기반으로 약물의 알려지지 않은 부작용을 예측 할 수 있고, 체내의 약물 부작용 분자 생물학적 메커니즘을 추론 할 수 있다. 본 연구에서는 SIDER 4.1의 약물-부작용 정보를 수집하여 시스템에 적용하였다.

1.5 ORA, Over Representation Analysis

초기하 분포를 이용한 과출현 분석(ORA, Over Representation Analysis)은 One-tailed Fisher 정확검정(Fisher's exact test) 과 동일한 방법으로 특정 집합의 원소가 다른 집합에 통계적으로 유의하게 과출현 하였는지를 측정하는 방법이다. 그림 1의 2X2 분할표(Contingency table)를 통하여 계산되며, 초기하 확률 분포를 기반으로 한다. 예를 들어, 그룹 1의 원소가 그룹 2에서 유의하게 과출현 하였는지에 대한 결과는 초기하 확률 분포에 근거하여 그림 1의 분할표에서 a개 혹은 그 이상으로 얻어질 확률 값으로 계산된다. 이를 계산하는 식은 다음과 같다.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} \qquad (1)$$

과출현 분석의 P-value는 (1) 식을 이용하여 그룹1의 원소가 그룹2에서 a개 이상부터 최대 c개 까지 출현할 초기하분포 값의 합으로 계산된다. 본 연구에서는 분류 단계에서 과출현 분석을 사용하여 두 네트워크 사이의 유사도를 측정하였다.

	Group2 (O)	Group2 (X)	
Group1 (O)	a	b	a+b
Group1 (X)	c	d	c+d
	a+c	b+d	a+b+c+d

Fig. 1. 2X2 Contingency Table

III. The Proposed Scheme

1. System overview

본 연구의 전체적인 개요는 그림 2와 같다. 먼저 약물과 부작용의 관계를 포함하고 있는 정보와 약물 부작용에 관여한다고 알려진 유전자 목록, 약물 부작용의 심한 정도를 나타내는 순위 정보를 수집 및 정제한다. 이를 사용하여 약물 부작용 쌍을 구성하고, 구성된 약물 부작용 쌍 중 특정 기준에 따라 공통 유전자 네트워크를 추출할 부작용 쌍을 선택한다. 또한, 공통 유전자 네트워크를 추출하기 위하여 필요한 약물 표적 정보, 약물에 관련된 유전자 정보 및 PPI 데이터를 수집하여 정제한다. 각 부작용 쌍에 대하여 공통 유전자 네트워크를 추출한 후, 이를 분류자로 사용하여 검증하고자 하는 약물에 해당 부작용 쌍이 속하는지 예측하는 분류 모델을 개발한다.

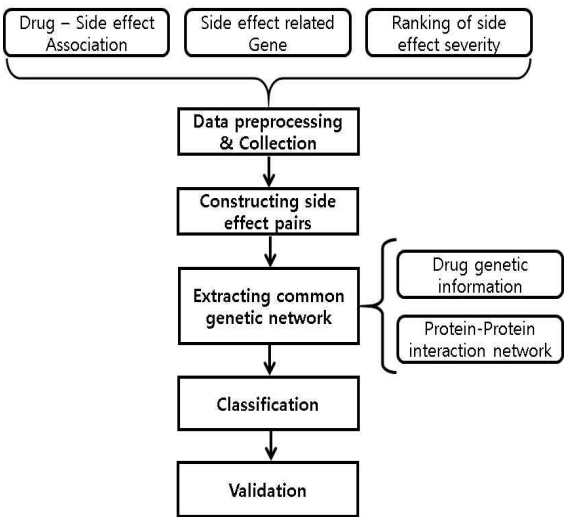

Fig. 2. System Overview

2. Data collection and preprocessing

2.1 Side effects and Drug target genes

본 연구에서는 약물과 부작용간의 관계 정보, 약물 부작용에 관여하는 유전자 정보, 약물 부작용의 심한 정도를 나타내는 정보를 수집하여 사용한다. 약물과 부작용 간의 관계 정보는 최근 새로 업데이트 된 SIDER 4.1에서 수집 및 정제하여 사용하고, 약물 부작용에 관여한다고 알려져 있는 유전자 목록의 수집은 최근에 진행된 약물 부작용 연구에서 수집해 놓은 목록을 사용한다[10][6]. 약물 부작용의 심한 정도를 판단하기 위한 기준을 채택하기가 쉽지 않았으나 최근 진행된 연구 중 사람들에게 설문 조사 형식으로 어떤 부작용이 가장 심각하다고 생각하는지 선택하게끔 하여 이에 따라 부작용 순위점수를 산출한 연구가 있었다[7]. 본 연구에서는 이 연구의 부작용 순위 점수를 수집하여 약물 부작용의 심각한 정도를 판단한다.

본 연구에서 사용된 약물 유전자 관련 정보는 약물의 표적 유전자(Drug target gene) 정보와 약물과 관련된 유전자(Drug related gene) 정보이다. 약물의 표적 유전자 정보는 DrugBank에서 수집하여 사용하고 약물에 관련된 유전자 정보의 수집은 CTD에서 수집한다[8][9].

2.2 PPI(Protein-protein interactions) Network

PPI(Protein-protein interactions) 정보는 PPI 네트워크 데이터베이스인 BioGrid[11] (version 3.3.124), Database of Interacting Proteins[12] (May 2015), IntAct[13] (May 2015), Molecular Interaction Database[14] (May 2015)에서 수집 및 정제한다.

3. Extracting side effect pairs

유의한 공통 유전자 네트워크를 추출할 부작용 쌍 후보들을 찾기 위하여, 본 연구에서는 실험에 사용 가능한 모든 부작용에 대하여 부작용 쌍을 구성한다. 구성된 부작용 쌍들에 대하여 두 부작용이 약물에서 함께 나타나는 빈도를 계산하고 이들 중 상위 1%에 속한 부작용 쌍을 추출한다. 또한, 이들 중 심각한 부작용이라고 판단되는 부작용이 하나라도 속한 쌍을 최종적으로 실험에 사용한다. 부작용의 심각성에 대한 기준은 기존 연구에서 사용된 순위 점수를 사용하여 임계값(Threshold) 0.6 상위에 속하는 부작용을 위험하다고 판단한다. 표 1은 부작용 쌍 추출의 예시 이다.

Table 1. Example of extracting side effect pairs

Side Effect 1	Side Effect 2	Severity Score of Side Effect 1	Severity Score of Side Effect 2	Frequency
Dyspep-sia	Rash	0.317	0.155	213
Rash	Hypoten-sion	0.155	0.372	194
Anorexia	Rash	0.427	0.155	187
Rash	Tachyca-rdia	0.155	0.603	187
Rash	Anxiety	0.155	0.236	181
…	…	…	…	…

4. Extracting common genetic network

3에서 추출한 부작용 쌍에 대하여 각 부작용 쌍에 유의한 유전자 공통 유전자 네트워크를 추출한다. 본 연구의 공통 유전자 네트워크란, 두 부작용을 함께 포함하고 있는 약물 중 대부분의 약물들이 공통으로 공유하고 있는 유전자 네트워크를 뜻한다. 즉, 두 부작용을 가진 약물을 투여하였을 때 체내에서 공통적으로 작용하는 유전자 네트워크를 말한다.

공통 유전자 네트워크를 발굴하는 방법은 그림 3과 같이 총 5가지 단계로 나눠진다. 먼저, PPI 네트워크를 기반으로 하여 두 부작용에 관여한다고 알려진 부작용 유전자와 두 부작용을 함께 포함하고 있는 약물 표적 유전자 사이의 최단 경로(Shortest path)를 구한다. 두 부작용을 함께 포함하고 있는 모든 약물에 대하여 첫 단계를 반복한다.

두 번째 단계에서는 첫 단계에서 구해진 약물 표적 유전자와 부작용 유전자 사이의 최단 경로 네트워크들을 4개의 노드(유전자)와 3개의 연결(Short Path:SP)로 잘라낸다. 전체 약물에 대하여 이와 같은 과정을 진행 한 후, 최종적으로 모든 SP를 취합하여 같은 SP가 몇 번 이나 출현하였는지 빈도를 계산한다. 전체 SP중 두 번 이상 출현한 SP만 남기고 나머지는 제거한다.

세 번째 단계에서 남겨진 SP들을 출현빈도에 따라 내림차순 정렬 한 후, 가장 많이 출현한 SP를 시작으로 Seed를 추출하는 작업을 진행한다. 가장 위의 SP를 Seed로 지정하고, 나머지 SP를 탐색하면서 Seed에 속한 노드(유전자)와 탐색하는 SP에 속한 노드가 얼마나 겹치는지 확인한다. 이중 하나의 노드라도 겹치면 해당 SP는 향후 Seed 후보에서 제외 되며 공통 유전자 네트워크 조립 시에 사용될 조립 후보 SP로 분류된다. 모든 SP에 대하여 이러한 과정을 반복하여 Seed SP 집합과 조립 후보 SP 집합을 추출한다.

네 번째 단계로는, 이전 단계에서 추출된 모든 Seed SP에 대하여 조립 후보 SP 집합을 사용하여 네트워크 조립을 진행한다. 그림 3의 Step 4와 같이 Seed SP 하나에 대하여 조립후보 SP 집합을 탐색하면서 3개의 노드와 2개의 연결이 겹치면 두 SP를 조립한다. 더 이상 겹치는 부분이 없을 때까지 계속해서 조립후보 SP 집합을 탐색한다. 모든 Seed SP에 대하여 위 단계를 진행 하면, 최종적으로 Seed별로 조립된 Seed 확장 네트워크가 추출된다.

마지막 단계에서는 Seed 확장 네트워크를 서로 비교하여 두 확장 네트워크의 노드가 75%이상 겹치면 두 네트워크를 통합시킨다. 이러한 과정을 반복하여 최종적인 공통 유전자 네트워크를 추출한다. 전체 단계 중 네트워크를 잘라내고 조립하여 통합하는 단계는 기존의 약물 공통 경로 네트워크를 추출하는 논문의 방법을 활용하였다[15].

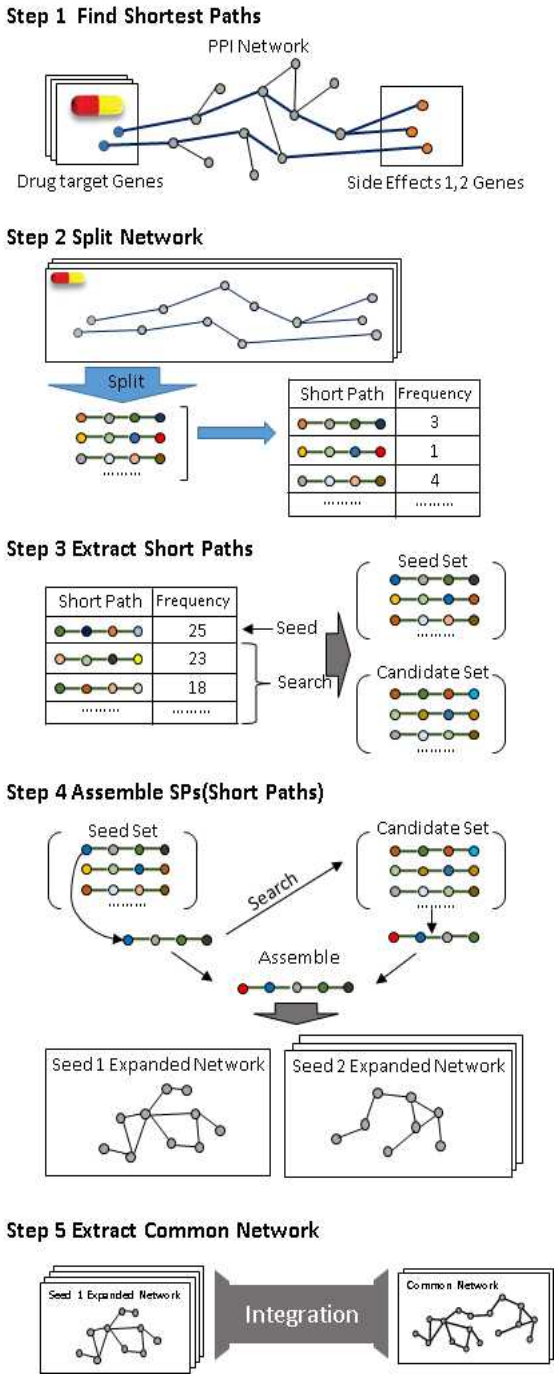

Fig. 3. Stages of extracting common genetic network

5. Classification model for predicting side effect pair

본 연구에서는 전 단계 에서 추출한 부작용 쌍 각각으로부터 얻어진 공통 유전자 네트워크를 분류자로 사용하여 특정 약물에 해당 부작용 쌍이 출현하는 지 예측하는 분류 모델을 개발한다. 부작용 쌍 별 공통 유전자 네트워크는 부작용 유전자와 약물 표적 유전자를 사용하여 추출되므로, 실험에 사용되는 모든 약물에 대하여 약물-부작용 관계 정보를 약물-부작용 유전자 관계로 사상(Mapping)시킨다. 다음으로, 부작용 쌍 별 공통 유전자 네트워크를 추출할 학습(Training) 데이터 세트와 모델 검증에 사용 될 검증(Test) 데이터 세트를 구성하기 위하여 전체 약물 샘플을 긍정 세트(Positive Set)와 부정 세트(Negative Set)로 구별한다. 긍정 세트와 부정 세트의 구성 방법은 다음과 같다. 긍정 세트는 두 부작용 유전자를 모두 포함하고 있는 약물로 구성 하고, 부정 세트는 두 부작용 유전자 전부를 포함하고 있지 않은 약물로 구성한다. 본 연구에서는 분류 모델의 검증을 위하여 10 Fold Cross-validation 방법을 사용하므로 각 Fold에 따라 학습 데이터 세트와 검증 데이터 세트의 샘플 구성이 달라진다. 학습 데이터는 전체 긍정 샘플 중 검증 데이터에 포함되지 않은 모든 긍정 샘플로 구성한다. 검증 데이터는 학습 데이터로 구성되지 않은 긍정 샘플과 전체 부정 샘플로 구성한다.

학습 데이터와 검증 데이터를 구별 한 후, 학습 데이터로부터 부작용 쌍 공통 유전자 네트워크를 추출하고, 이를 분류자로 사용하여 검증 데이터를 분류 한다. 검증 데이터를 분류하는 방법은 다음과 같다. 먼저, 그림 4와 같이 약물과 관련이 있다고 알려져 있는 유전자(Drug related gene)들과 약물 표적 유전자 사이의 최단 경로(Shortest Path)를 PPI 네트워크 기반으로 찾는다. 이렇게 구성한 약물 네트워크와 부작용 쌍 공통 유전자 네트워크 간의 유사도를 비교하여 유사하면 긍정으로 유사하지 않으면 부정으로 분류한다. 유사도를 측정하는 방법은 초기하 분포를 이용한 과출현분석(ORA, Over Representation Analysis)을 사용하였다. 이는 One-tailed Fisher 정확검정(Fisher's exact test) 과 동일한 방법으로, 비교하고자 하는 유전자 집합이 정답 유전자 집합에 통계적으로 유의하게 많이 포함되어 있는가를 측정하는 방식이다. 본 연구에서는 정답 유전자 집합을 부작용 쌍 공통 유전자 네트워크에 포함된 유전자로 지정하고, 비교하고자 하는 유전자 집합을 각 약물 네트워크에 포함된 유전자로 지정한다. 검정으로 산출된 P-value(유의 확률)가 작을수록 정답 유전자 집합에서 더 많이 출현 한다는 의미이다. 검증 데이터에 속한 모든 약물에 대하여 유의확률을 계산하고 계산된 유의확률이 임계값(Threshold) 보다 작으면 긍정, 크면 부정으로 분류한다.

본 연구에서는 분류모델을 검증하기 위한 척도로, AUC(Area Under the Curve) 값을 사용하였다. ROC Curve 위의 한 점은 하나의 임계값에 따라 분류된 긍정, 부정 샘플 수로 계산된다. 따라서 본 연구에서는, 모든 샘플을 부정으로 분류하는 가장 작은 임계값을 시작으로 하여, 지속적으로 임계값을 증가 시키면서 최종적으로는 모든 샘플을 긍정으로 분류 할 때 까지 총 10개의 임계값을 설정하였다. 임계값 설정을 위하여 전체 검증 샘플을 P-value에 따라 오름차순 정렬 하였고, 가장 작은 값에서부터 가장 큰 값까지 10% 씩 잘라, 각 10%에 속한 샘플들의 P-value 중 가장 큰 값을 각각의 임계값으로 설정하였다.

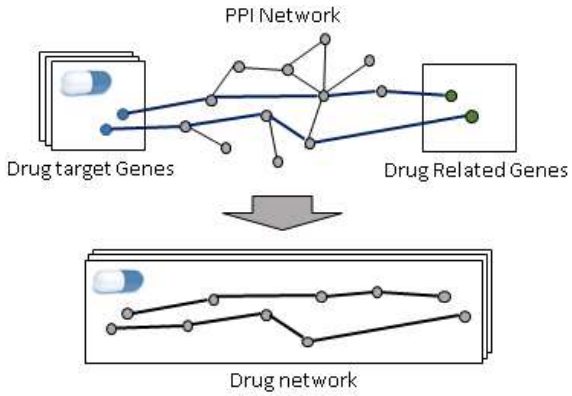

Fig. 4. Extracting drug network which are used in experiment

IV. Results

1. Experimental environment and Data

본 연구에서는 개발 환경으로 Microsoft visual studio 2012를 사용하였다. 실험 환경은 Inter(R) core(TM)i7-4770K CPU @ 3.50GHz, 32GB RAM, 64비트 운영체제이다. 실험 전반에 사용된 PPI 네트워크 데이터는 총 4개의 데이터베이스에서 수집하였으며 40만개 이상의 Protein-Protein 이진(Binary) 연결을 포함한다.

1.1 Drug data and Side effect data

본 연구에 사용된 약물 데이터는 Drugbank, CTD에서 수집하였다. 약물 표적 유전자(Target gene) 정보는 Drugbank에서, 약물에 관련되어 있다고 알려진 약물-관련유전자(Drug related gene)는 CTD에서 수집하였다. 총 384개의 약물에 대하여 약물-표적유전자 관계 정보, 약물-관련유전자 정보를 확보하였다.

실험에 사용된 384개의 약물에 대하여 약물-부작용 관계 정보는 SIDER 4.1에서 수집하였다. 전체 부작용 중 부작용-유전자 관계를 포함하고 있는 부작용만 사용하였다. 부작용-유전자 관계 정보는 기존연구에서 수집하여 총 146개의 부작용-유전자 관계를 사용하였다. 본 연구에서는 부작용 쌍의 공통 유전자 네트워크 추출부터 약물의 분류까지 유전자를 기반으로 하므로

약물-부작용 관계 정보를 약물-부작용유전자 관계로 사상(Mapping)하여 사용하였다. 표 2는 약물 하나에 포함된 표적 유전자 개수, 관련유전자 개수, 부작용 유전자 개수 분포 이다.

Table 2. Distribution of drug target genes, drug related genes and drug side effect genes

	AVG	STD	MIN	MAX	Median
Drug-Target gene	5.59	7.52	1	48	3
Drug related gene	10.71	10.96	1	49	6
Drug-side effect gene	14.20	8.69	1	41	13

2. Selected Side effect pairs

본 연구에서는 총 146개의 부작용에 대하여 모든 부작용 쌍을 구성하였다. 구성된 10,585개의 부작용 쌍 중 공통 유전자 네트워크를 추출할 부작용 쌍을 선택하기 위하여 각 부작용 쌍의 약물 출현 빈도와 부작용의 심각성을 고려하였다. 10,585개 부작용 쌍 전체에 대하여 총 384개 약물에서 두 부작용이 함께 출현하는 빈도를 계산하였고, 이 중 출현 빈도 상위 1퍼센트를 추출하였다. 추출된 부작용 쌍 중 심각성 정도가 0.6 이상인 부작용 쌍을 채택하였다. 표 3은 최종적으로 골라진 16개 부작용 쌍 이다.

Table 3. 16 Side effect pairs used in the experiment

Side Effect 1	Side Effect 2	Severity Score of Side Effect 1	Severity Score of Side Effect 2	Frequency
Rash	Tachycardia	0.155	0.603	187
Tachycardia	Hypotension	0.603	0.372	171
Tachycardia	Tremor	0.603	0.361	145
Tachycardia	Anxiety	0.603	0.236	144
Dyspepsia	Tachycardia	0.317	0.603	143
Tachycardia	Syncope	0.603	0.480	141
Anorexia	Tachycardia	0.427	0.603	129
Tachycardia	Vertigo	0.603	0.285	129
Tachycardia	Thrombocyto-penia	0.603	0.579	117
Cough	Tachycardia	0.150	0.603	112
Tachycardia	Flushing	0.603	0.124	111
Angioedema	Tachycardia	0.463	0.603	110
Alopecia	Tachycardia	0.216	0.603	95
Pancreatitis	Rash	0.631	0.155	94
Ataxia	Rash	0.634	0.155	93
Rash	Renal failure	0.155	0.856	92

3. Validation of classification model

본 연구에서는 분류 모델을 검증하기 위하여 10 Fold Cross-validation(교차 검증) 방법을 수행하였다. 또한, 모델의 정확성을 검증하는 척도로는 AUC(Area Under the Curve)를 사용하였다.

3.1 10 Fold Cross-validation

10 Fold Cross-validation은 분류 모델을 검증 하는 방법 중 하나로, 전체 데이터를 10세트로 나누어 9세트는 모델을 학습시키는데 사용하고 나머지 1세트는 모델 검증을 위하여 사용하는 방법이다. 따라서 각 Fold에 따라 해당 Fold의 샘플을 분류한 정확도가 측정되고, 전체 10개 Fold의 정확도를 평균한 값이 그 분류모델의 정확도가 된다. 본 연구에서는 무작위로 각 Fold에 속하는 약물 샘플을 추출하여 사용하였으며 각 Fold의 분류 정확도는 AUC 척도로 측정 하였다.

3.2 ROC(Receiver Operating Characteristic) Curve and AUC(Area Under the Curve)

ROC(Receiver Operating Characteristic) Curve는 FPR(False Positive Rate: $1-Specificity$)을 X축, TPR(True Positive Rate: Sensitivity)을 Y축으로 하여 그려지는 그래프이다. 그림 5의 혼돈행렬(Confusion Matrix)을 통하여 FPR, TPR 값이 계산되고, 이를 사용하여 그래프가 그려진다. 즉, 분류모델에서 긍정 샘플과 부정 샘플을 분류하는 기준인 특정 임계값(Threshold)에 따라 TP, TN, FP, FN 값이 계산되고, 그에 따른 FPR, TPR 값이 산출 되면 그 점이 그래프의 한 점이 된다. 최종적으로 다양한 임계값(Threshold)에 따라 ROC Curve가 그려지며, ROC Curve 아래의 면적을 AUC(Area Under the Curve)라 한다. AUC의 값이 1에 가까울수록 분류기의 성능이 뛰어나다는 의미이고, 무작위로 아무런 학습 조건 없이 분류했을 때의 AUC 값을 0.5 라 한다. 일반적으로 0.5 초과 1 이하의 AUC 값을 가지는 분류기를 의미 있는 분류모델이라고 판단한다. FPR과 TPR 값을 계산하는 식은 각각 (2), (3)번 식과 같다.

$$TPR = \frac{TP}{TP+FN} \quad (2)$$

$$FPR = \frac{FP}{FP+TN} \quad (3)$$

		실제(Actual)	
		Positive	Negative
예측 (Predict)	Positive	True Positive(TP)	False Positive(FP)
	Negative	False Negative(FN)	True Negative(TN)

Fig. 5. Confusion Matrix

Table 4. The number of genes of common networks

Side Effect 1	Side Effect 2	1Fold	2Fold	3Fold	4Fold	5Fold	6Fold	7Fold	8Fold	9Fold	10Fold	AVG
Rash	Tachycardia	246	292	282	303	265	279	328	286	300	300	288.1
Tachycardia	Hypotension	89	103	86	107	85	94	102	107	89	98	96
Tachycardia	Tremor	125	121	127	141	130	141	143	139	144	131	134.2
Tachycardia	Anxiety	141	145	149	165	152	168	168	158	161	160	156.7
Dyspepsia	Tachycardia	64	77	78	99	79	90	84	99	89	82	84.1
Tachycardia	Syncope	119	127	96	124	122	123	124	126	126	118	120.5
Anorexia	Tachycardia	87	87	95	113	84	103	107	100	103	94	97.3
Tachycardia	Vertigo	72	84	59	96	62	74	79	84	93	93	79.6
Tachycardia	Thrombocytopenia	96	105	93	104	95	97	105	105	104	98	100.2
Cough	Tachycardia	83	54	53	84	88	76	82	96	98	88	80.2
Tachycardia	Flushing	64	101	87	79	87	105	97	103	98	102	92.3
Angioedema	Tachycardia	151	162	176	160	166	165	161	170	175	169	165.5
Alopecia	Tachycardia	83	106	78	109	108	106	113	116	110	100	102.9
Pancreatitis	Rash	158	228	229	261	228	248	253	231	251	203	229
Ataxia	Rash	204	206	188	220	223	221	199	250	236	254	220.1
Rash	Renal failure	71	169	94	176	106	170	136	131	179	134	136.6

Table 5. AUC Results

Side Effect 1	Side Effect 2	1Fold	2Fold	3Fold	4Fold	5Fold	6Fold	7Fold	8Fold	9Fold	10Fold	AVG	STD
Rash	Tachycardia	0.76	0.77	0.68	0.61	0.81	0.83	0.70	0.72	0.64	0.76	**0.729**	0.066
Tachycardia	Hypotension	0.79	0.68	0.82	0.73	0.79	0.78	0.76	0.69	0.82	0.80	**0.766**	0.049
Tachycardia	Tremor	0.81	0.81	0.75	0.75	0.71	0.80	0.75	0.75	0.73	0.76	**0.762**	0.032
Tachycardia	Anxiety	0.82	0.79	0.69	0.71	0.83	0.82	0.80	0.72	0.78	0.73	**0.769**	0.050
Dyspepsia	Tachycardia	0.83	0.78	0.71	0.62	0.72	0.80	0.71	0.75	0.72	0.71	**0.734**	0.057
Tachycardia	Syncope	0.81	0.70	0.82	0.69	0.77	0.79	0.85	0.72	0.81	0.76	**0.772**	0.052
Anorexia	Tachycardia	0.70	0.69	0.67	0.67	0.66	0.79	0.75	0.74	0.61	0.74	**0.702**	0.049
Tachycardia	Vertigo	0.76	0.75	0.81	0.72	0.75	0.73	0.86	0.70	0.63	0.76	**0.747**	0.059
Tachycardia	Thrombocytopenia	0.78	0.78	0.82	0.64	0.73	0.81	0.72	0.73	0.65	0.72	**0.738**	0.060
Cough	Tachycardia	0.78	0.85	0.71	0.71	0.61	0.83	0.68	0.78	0.67	0.79	**0.741**	0.074
Tachycardia	Flushing	0.88	0.55	0.83	0.69	0.75	0.85	0.84	0.63	0.83	0.78	**0.763**	0.105
Angioedema	Tachycardia	0.81	0.80	0.68	0.77	0.79	0.82	0.82	0.70	0.82	0.73	**0.773**	0.050
Alopecia	Tachycardia	0.75	0.77	0.79	0.72	0.70	0.87	0.80	0.69	0.66	0.72	**0.748**	0.058
Pancreatitis	Rash	0.62	0.46	0.51	0.46	0.62	0.61	0.57	0.57	0.59	0.55	**0.557**	0.057
Ataxia	Rash	0.69	0.52	0.64	0.65	0.75	0.78	0.66	0.44	0.64	0.49	**0.626**	0.105
Rash	Renal failure	0.70	0.60	0.75	0.58	0.67	0.68	0.59	0.61	0.65	0.68	**0.653**	0.052
AUC Average and Standard Deviation												**0.724**	0.061

4. Experimental Results

본 절에서는 최종적으로 골라진 16개의 부작용 쌍에 대하여 본 연구의 분류 모델을 검증한 결과를 기술한다.

4.1 Common genetic network

분류 모델의 검증을 위하여 선택된 16개 부작용 쌍에 대하여 부작용 쌍 각각의 공통 유전자 네트워크를 추출하였다. 각 Fold별로 학습 데이터와 검증 데이터가 달라지므로 최종적으로 총 10개의 Fold에 대한 공통 유전자 네트워크가 추출된다. 각 부작용 쌍 별로 추출된 공통 유전자 네트워크에 속하는 유전자 개수 및 평균은 표 4와 같다.

4.2 AUC Result for each side effect pair

Fold별로 추출된 공통 유전자 네트워크를 분류자로 사용하여, 각 fold의 검증 샘플에 포함되어 있는 약물을 Ⅲ.5의 방법

으로 분류 하였다. 부작용 쌍 별 최종 AUC 결과는 표 5와 같다. 각각의 Fold에 대하여 AUC가 계산되며 전체 10개 Fold의 AUC값 평균이 해당 부작용 쌍의 최종 AUC 값으로 산출된다. AUC가 가장 높은 부작용 쌍(Angioedema-Tachycardia)의 10 Fold Cross-validation AUC 계산 결과를 표 6에 첨부하였다.

최종적으로 계산된 결과를 살펴보면 총 16개의 부작용 쌍 중 세 쌍을 제외하고 전부 0.70 이상의 고른 AUC를 얻었음을 알 수 있다. 또한 이 세 쌍을 포함하고도 전체 16개 부작용 쌍의 AUC 평균을 계산하였을 때 0.724의 AUC값을 가지는 것을 확인할 수 있다. 이는 본 연구의 분류기가 자주 발생하면서도 심각한 부작용을 포함하고 있는 부작용 쌍에 대하여 특정한 약물이 두 부작용을 포함하고 있는지를 잘 예측하였다고 판단할 수 있는 수치이다. 두 가지 부작용이 함께 발생하는 것에 대한 유전자 네트워크 기반의 기존 연구가 진행되지 않았기 때문에 기존 모델들과의 결과 비교는 불가능 하지만, 본 연구의 분류모델 AUC로 미루어보아 본 연구의 분류자로 사용된 유전자 공통 유전자 네트워크가 유의미 하게 추출되었음을 알 수 있고 분류 방법 또한 적절하였다는 것을 보여준다.

Table 6. 10 folds' AUC of Angioedema-Tachycardia side effect pair

Fold	Training Set	Test Set		AUC
		Positive	Negative	
1	99	11	48	0.81
2	99	11	48	0.80
3	99	11	48	0.68
4	99	11	48	0.77
5	99	11	48	0.79
6	99	11	48	0.82
7	99	11	48	0.82
8	99	11	48	0.70
9	99	11	48	0.82
10	99	11	48	0.73
Average	**99.0**	**11.0**	**48.0**	**0.773**
Stan. DIV				0.050
Total		**110**	**48**	

V. Conclusions

본 연구는 약물에서 흔히 함께 발생하는 두 가지 부작용에 대하여 부작용 쌍을 구성하고, 심각하다고 판단되는 부작용 쌍에 대하여, 이 두 부작용을 가진 약물들이 공통적으로 공유하는 공통 유전자 네트워크를 추출하였다. 추출된 공통 유전자 네트워크를 분류자로 사용하여, 특정 약물이 이와 비슷한 약물 네트워크를 가지면 긍정으로 그렇지 않으면 부정으로 분류하는 분류모델을 개발 및 검증 하였다.

본 연구 분류모델의 AUC 결과를 토대로, 각 부작용 쌍에 따른 공통 유전자 네트워크가 잘 발굴되었음을 확인 할 수 있다. 따라서 향후 연구에서는 추출된 공통 유전자 네트워크를 활용하여 약물과 부작용들 간의 유전자 작용 매커니즘 및 유전자 네트워크의 기능(Function)을 밝힐 예정이다.

본 연구에서는 약물에서 빈번하게 나타나는 두 개의 부작용에 대하여 공통 유전자 네트워크를 추출 하였지만, 향후 연구에서는 더 많은 부작용들을 활용하여 부작용 군집(Cluster)을 구성하고 각 부작용 군집에 대한 공통 유전자 네트워크를 발굴할 예정에 있다.

REFERENCE

[1] Classen D C, Pestotnik S L, Evans R S, Lloyd J F, Burke J P. “Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality.”, JAMA, Vol. 277, No. 4, pp. 301-6, Jan. 1997.

[2] Song mi Lee, “Drug leads to disease”, Sodam publisher, Oct. 2007.

[3] Kuhn, Michael, et al. "A side effect resource to capture phenotypic effects of drugs.", Molecular systems biology, Vol. 6, No. 1, pp. 343, Jan. 2010.

[4] Perucca, Piero, and Frank G. Gilliam. "Adverse effects of antiepileptic drugs.", The Lancet Neurology, Vol. 11, No. 9, pp. 792-802, Sep. 2012.

[5] Atias, Nir, and Roded Sharan. "An algorithmic framework for predicting side effects of drugs.", Journal of Computational Biology, Vol. 18, No. 3, pp. 207-218, March 2011.

[6] Kuhn, Michael, et al. "Systematic identification of proteins that elicit drug side effects.", Molecular systems biology, Vol. 9, No. 1, pp. 663, April 2013.

[7] Gottlieb, Assaf, et al. "Ranking Adverse Drug Reactions With Crowdsourcing.", Journal of medical Internet research, Vol. 17, No. 3, March 2015.

[8] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. “DrugBank: a comprehensive resource for in silico drug discovery and exploration.”, Nucleic Acids Res., Vol. 1 No. 34, pp. 668-72, Jan. 2006.

[9] Davis AP et al., “The Comparative Toxicogenomics Database's 10th year anniversary: update 2015.”, Nucleic Acids Res., Oct. 2014.

[10] Kuhn M, Letunic I, Jensen LJ, Bork P. “The SIDER

database of drugs and side effects.", Nucleic Acids Res., Oct. 2015.

[11] Chatr-Aryamontri A. et al., "The BioGRID interaction database: 2015 update." Nucleic Acids Research., Nov. 2014.

[12] Salwinski, Lukasz, et al. "The database of interacting proteins: 2004 update." Nucleic acids research, Vol. 32, D449-D451, Jan. 2004.

[13] Orchard, Sandra, et al. "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases." Nucleic acids research, Nov. 2013.

[14] Licata, Luana, et al. "MINT, the molecular interaction database: 2012 update." Nucleic acids research, Vol. 40, No. D1, pp. D857-D861, Jan. 2012.

[15] Silberberg, Yael, et al. "Large-scale elucidation of drug response pathways in humans.", Journal of Computational Biology, Vol. 19, No. 2, pp. 163-174, Feb. 2012.

Authors

Youhyeon Hwang received the B.S. degrees in Computer Science and Engineering from Gachon University, Korea, in 2015. Youhyeon Hwang is currently a researcher in the Department of Computer Science, Data Mining & Bioinformatics Laboratory, Gachon University. He is interested in data mining, bioinformatics, database.

Min Oh received the B.S. degrees in Computer Science and Engineering from Gachon University, Korea, in 2015. Min Oh is currently a research associate in the Department of Computer Science at Gachon University.
He is interested in translational bioinformatics 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.