Journal of The Korea Society of Computer and Information
Vol. 21 No. 3, pp. 65-71, March 2016
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.3.065

Target Prediction Based On PPI Network

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

Abstract

To reduce the expenses for development a novel drug, systems biology has been studied actively. Target prediction, a part of systems biology, contributes to finding a new purpose for FDA(Food and Drug Administration) approved drugs and development novel drugs. In this paper, we propose a classification model for predicting novel target genes based on relation between target genes and disease related genes. After collecting known target genes from TTD(Therapeutic Target Database) and disease related genes from OMIM(Online Mendelian Inheritance in Man), we analyzed the effect of target genes on disease related genes based on PPI(Protein-Protein Interactions) network. We focused on the distinguishing characteristics between known target genes and random target genes, and used the characteristics as features for building a classifier. Because our model is constructed using information about only a disease and its known targets, the model can be applied to unusual diseases without similar drugs and diseases, while existing models for finding new drug-disease associations are based on drug-drug similarity and disease-disease similarity. We validated accuracy of the model using LOOCV of ten times and the AUCs were 0.74 on Alzheimer's disease and 0.71 on Breast cancer.*

▸ Keyword : Systems biology, Target prediction, Target repositioning, Drug repositioning

I. Introduction

기존의 방법으로 새로운 약물을 개발하고 승인을 받기까지 10년 이상의 시간과 1조원 이상의 비용이 소모된다[1]. 이러한 개발 비용의 감소를 위하여 생물학적 데이터와 대량 신속처리가 가능한 컴퓨터를 기반으로 시스템스 바이올로지(Systems biology)가 부각되고 있다[2].

질병의 치료를 위하여 약물에 의해 선별적으로 조절되는 유전자를 표적유전자라고 한다. 2002년의 인간 게놈 분석에 따르면 약학적으로 흥미로운 6000-8000개의 표적유전자가 있다고 추정된다[3]. 그러나 이 중 현재 성공적으로 활용되거나 연구중인 표적유전자는 2000여개 정도이다[4]. 또한 제약 산업에서도 표적유전자를 찾는 작업은 산업 발전의 큰 장애물이다[5]. 따라서 새로운 표적유전자를 찾는 시스템적 모델의 구축은 제약 산업의 병목 현상을 해결 할 수 있다.

새로운 표적유전자를 찾는 기존의 모델은 PREDICT, TESS와 같이 약물-약물, 질병-질병 간 유사성을 바탕으로 구축되었다[6-7]. 이들 방법은 알려진 약물-질병 치료관계에서 작용하는 표적유전자의 기능만을 활용한다. 따라서 현재 존재하는 약물의 표적유전자로 예측의 범위가 한정될 뿐 아니라, 표현형이 유사한 질병이 없는 경우 새로운 표적유전자를 예측하기 어렵다는 한계가 있다.

이와 달리 본 연구에서는 네트워크를 기반으로 특정 질병에 대한 표적유전자의 모든 기능을 고려하여 새로운 표적유전자를 찾으며, 모든 표적유전자에 대하여 적용 가능한 모델을 구축하였다.

• First Author: Taekeon Lee, Corresponding Author: Youngmi Yoon
*Taekeon Lee(taekeon.m.lee@gmail.com), Dept. of Computer Engineering, Gachon University
**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. 02. 27, Revised: 2016. 03. 02, Accepted: 2016. 03. 18.
• 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).

따라서 표현형이 유사한 질병이 없는 경우에도 해당 질병의 새로운 표적유전자를 찾을 수 있으며, 화학적·유전적 구성이 유사하지 않은 약물에서도 체내에서 동일한 기능을 할 수 있는 표적유전자를 식별할 수 있다.

본 연구에서는 PPI 데이터를 기반으로 구성된 분자(Molecule) 단위 네트워크를 이용하여, 기존에 치료 관계에 있다고 알려진 표적유전자가 질병 관련 유전자에 미치는 영향을 거리기반으로 분석한다. 표적유전자와 질병 관련 유전자의 최단거리를 계산하였다. 최단거리의 경로가 다수인 경우에는 각각의 경로에 의한 영향을 모두 고려하였다. 이 후 검증된 약물의 표적유전자와 임의로 추출된 표적유전자가 질병 관련 유전자와의 거리에 있어 구분되는 특성을 토대로 질병에 유의한 표적유전자를 예측하는 모델을 구축하였다.

본 연구에서는 모델의 검증을 위하여 알츠하이머병(Alzheimer's disease)과 유방암(Breast cancer)에 대한 질병 관련 유전자와 각각의 질병에 대하여 치료 관계가 검증된 약물의 표적유전자를 사용하여 실험을 진행하였다. 이 후 모델의 유효성은 AUC(Area under the curve) 척도를 이용하여 검증하였다.

본 논문은 2절에서 연구와 관련된 표적유전자, 질병 관련 유전자, PPI 네트워크에 대한 데이터 자원의 소개 및 분류 모델 구축을 위한 분류도구를 설명한다. 3절에서는 본 연구의 전체적인 시스템 개요와 데이터 수집 및 정제, 네트워크 구축, 네트워크를 통한 분석과 분류 모델 구축에 대한 방법을 기술한다. 4절에서는 개발 및 실험 환경을 기술하고, 본 연구의 분류 모델의 정확도 측정 방법과 정확도를 제시한다. 5절에서는 분류 모델의 추가적인 활용에 대한 가능성을 제시하고, 추후 연구 방향에 대하여 기술한다.

II. Related works

본 연구는 컴퓨터 실험을 토대로 하여, 검증된 약물의 표적유전자가 무작위로 선택된 유전자와 구분되는 특징을 기반으로 새로운 표적유전자를 찾는 모델의 구축을 목표로 한다. 이를 위하여 약물의 표적이 되는 유전자와 유전자들의 상호 작용, 질병과 연관된 유전자에 대한 정보를 제공하는 데이터베이스가 필요하다. 따라서 다음과 같은 관련 연구를 바탕으로 본 연구를 실시하였다.

1. TTD(Therapeutic Target Database)

질병 치료의 표적이 되는 단백질, 핵산, 표적의 경로에 대한 정보를 가지고 있는 데이터베이스로 생체의학 연구와 약학 연구에 주로 이용된다[8]. 치료의 표적은 문헌으로부터 얻어 구성되었으며, 표적에 관련된 정보는 KEGG, MetaCyc/BioCyc, NetPath, PANTHER pathway, PathWhiz, PID, Reactome, WikiPathways와 같은 세계적으로 널리 이용되는 데이터베이스의 데이터를 정제하여 수록하였다.[9-16]

총 2,589개의 표적과 이들과 연결된 31,614개의 약물에 대한 정보와 ICD(International Classification of Diseases) 코드에 따른 2,537개의 질병이 포함되어 있다[4]. 이 중 본 연구에서는 알츠하이머병과 유방암에 각각 해당하는 질병 관련 유전자를 사용하였다.

2. OMIM(Online Mendelian Inheritance in Man)

OMIM은 인간 유전자와 유전 질환에 대한 권위 있는 공공 데이터베이스로 NCBI(National Center for Biotechnology Information)에 의해 웹으로 제공되며, 생의학 문헌으로부터 얻어진 데이터를 Johns Hopkins 대학에서 기록하고 편집한다[17-18]. 본 연구에서는 알츠하이머병과 유방암에 연관된 유전자를 각각 수집해서 사용하였다.

3. PID(Pathway Interaction Database)

미국 국립 암 연구소(US National Cancer Institute)와 네이처 출판 그룹(Nature Publishing Group)이 문헌을 기반으로 데이터를 수집하고 정제하여 공동으로 구성한 데이터베이스이다[9]. 분자, RNA, 단백질, 복합체의 상호작용에 관한 정보를 갖고 있으며, 본 연구에서는 표적유전자와 질병과 연관된 유전자 간의 관계와 경로의 수를 얻기 위하여 네트워크를 구성하는 데 사용하였다.

4. Random Forest

Random Forest는 분자의 양적 기술에 기초한 생물학적인 활동의 범주나 분자의 양적인 측면을 예측하기 위한 분류, 회귀 도구이다[19]. 이 도구는 무작위로 속성을 선택하여, 트리를 여러 개 만들고 Forest를 구성한다. 각각의 트리에서 속성 선택은 독립적이다. 예측의 결과는 앙상블 기법이 적용되어 Forest를 구성하는 트리에서 가장 많이 선택된 Class를 선택하게 된다[20]. 본 연구에서는 이미 질병을 치료할 수 있다고 알려져 있는 약물의 표적유전자와 무작위로 추출된 표적유전자를 구별하기 위한 규칙을 학습시키기 위하여 해당 모델을 사용하였다.

III. Method

1. System overview

본 연구의 전체 개요는 그림 1과 같다. 먼저, 연구에 사용될 데이터를 수집 및 정제하였다. 질병을 치료하는 것으로 알려진 표적유전자 집합을 수집, 정제하고 이를 Positive set으로 사용하였다. Negative set은 Positive set에 독립적인 유전자를 무

작위 추출하여 구성하였다. 질병 관련 유전자(Disease related genes)를 OMIM에서 수집, 정제한다.

유전자 데이터 정제 과정에서 데이터 형태의 일관성을 위하여 PPI 데이터를 구성하는 Molecules ID 형태로 사상(Mapping)하였다.

PPI 데이터를 기반으로 Molecule Network를 구성한 후, 네트워크 탐색 과정을 통하여 Positive set과 Negative set에서 질병 관련 유전자 사이의 최단거리의 간선 개수와 최단거리에 해당하는 경로의 수를 얻고, 이를 토대로 얻어진 벡터를 이용하여 분류 모델을 생성한다. 모델의 검증은 LOOCV(Leave One Out Cross Validation)에 따른 ROC curve의 AUC척도를 활용한다.

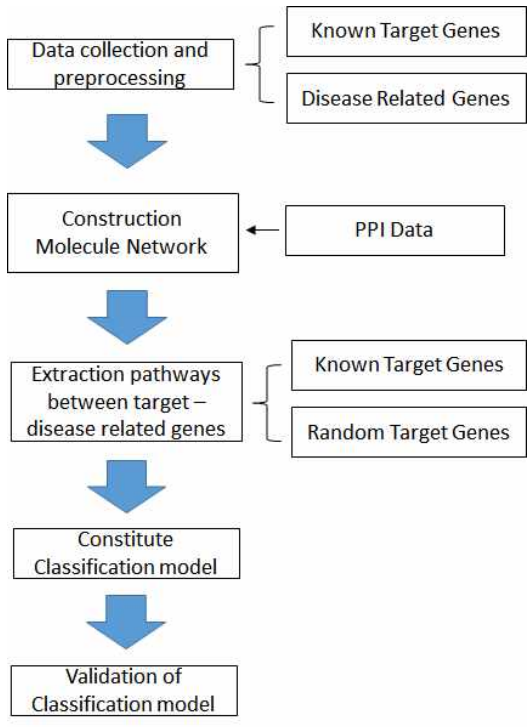

Fig. 1. System Overview

2. Data collection and preprocessing

2.1 Target genes

알츠하이머병과 유방암을 치료하는 것으로 알려진 약물의 표적유전자 집합을 TTD에서 각각 수집 후 정제하여 사용하였다. 약물의 승인 여부와 유전자의 임상 단계는 고려하지 않았으며, 일반적으로 적용 가능한 규칙의 생성을 위하여 질병의 특수한 아형에만 존재하는 유전자는 제외하고 일반형에 해당하는 유전자만 사용하였다. 여러 번의 검증을 위하여 독립적인 유전자를 알려진 표적유전자 수의 3배수로 10세트 무작위 추출 후 정제하였다.

2.2 Disease Related Genes

알츠하이머병과 유방암과 연관된 유전자 집합을 OMIM에서 수집 후 정제하였다. 표적유전자와 마찬가지로 질병의 일반형에 해당하는 표적유전자만을 추출, 정제하였다. 이 과정에서 분자로 사상하기 위하여 NCBI에서 제공하는 Entrez Gene ID가 있는 유전자만을 추출하였다.

2.3 PPI Data

Pathway Interaction Database에서는 NCI-Nature, Reactome, BioCarta에서 각각 얻어진 자료를 제공한다. Reactome은 온타리오 암 연구 협회와 콜드 스프링 하버 연구소, 뉴욕 의과대학, 유럽 바이오인포메틱스 협회에서 생화학 경로와 반응에 대한 공공 데이터를 제공하는 데이터베이스이다[10]. 또한, BioCarta는 생명과학 연구를 위하여 전문가 집단에 의해 구성되는 온라인 자원으로 유전자의 기능과 프로테오믹 경로(Proteomic pathway)에 대한 데이터를 포함하고 있다[21]. 본 연구에서는 PID에서 제공하는 모든 데이터를 분자 형태로 사상하여 네트워크를 구성하였다.

3. Exploring Network

본 연구에서 사용하는 표적유전자는 Uniprot ID, 질병관련 유전자는 Entrez ID, PPI 데이터는 PID의 Molecules ID로 각기 제공된다. 따라서 본 연구에서는 유전자에 의해 생성된 분자가 세포의 상태를 조절하는 단백질의 발현에 관여한다는 생물학적 지식에 기초하여, 유전자의 기능을 효과적으로 나타내기 위하여 모든 데이터를 분자 단위로 사상하였다.

본 연구에서는 표적유전자가 질병 관련 유전자에 미치는 영향을 분자 단위 네트워크상 거리기반으로 분석한다. 네트워크에서 분자 간의 거리가 가깝고 연결된 경로가 많을수록 표적유전자는 질병 관련 유전자에 보다 큰 영향을 미친다. 따라서 표적유전자와 질병 관련 유전자의 최단거리경로(Shortest path)와 최단거리를 갖는 경로의 수를 이용하였다. 가중치가 없는 PPI 데이터를 사용했으므로 최단거리는 표적유전자와 질병 관련 유전자를 연결하는 간선의 개수이다. 경로의 길이가 짧을수록 더욱 큰 영향을 미치는 것을 표현하기 위하여 최단거리의 역수를 영향력으로 정의하였다. 그림 2와 같이 표적유전자와 질병 관련 유전자 사이의 최단거리에 해당하는 경로가 여러 개 있는 경우, 단일 경로일 경우보다 큰 영향을 미치는 점을 고려하기 위하여 각 경로의 값을 합산하였다. 최종적으로 표 1과 같이 표적유전자가 질병 관련 유전자에 미치는 영향력으로 구성된 벡터(Vector)를 만든다. 분자 네트워크상에서 연결되지 않은 질병 관련 유전자는 벡터에서 제외하여 실제 약물에 의해 조절되는 질병 관련 유전자만을 사용한다.

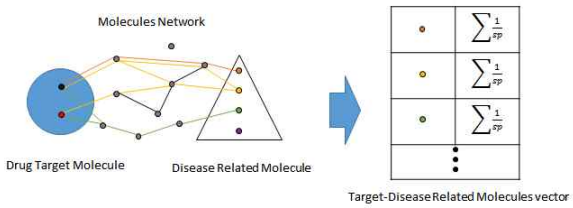

Fig. 2. Finding Shortest Paths & Generation of Target-Disease Related Molecules Vector

Table 1. Target-Disease Related molecules Vector

Molecules ID of Disease Related Genes	Influence
100052	0.25
100072	0.25
100186	0.17
100284	0.17
100368	0.25
100369	0.33
100384	0.17
100440	0.75
100591	0.20
100607	0.60
⋮	⋮

4. Classification model

본 연구 모델의 검증을 위하여 이미 알려진 표적유전자와 임의의 표적유전자에 대하여 각각 네트워크 탐색 과정을 수행하여 영향력 벡터를 만든다. 본 연구는 알려진 표적유전자의 작용을 바탕으로 새로운 표적을 찾는 연구이므로, 임의의 표적유전자 벡터의 질병 관련 유전자는 알려진 표적유전자의 영향력 벡터에서 등장한 유전자로 제한한다.

표2와 같이 얻어진 벡터 각각을 하나의 샘플로 하는 Feature Table을 구성하여, 알려진 표적유전자의 클래스(Class)에 TRUE, 임의의 표적유전자의 클래스에 FALSE를 부여한다. 이 과정에서 TTD의 이미 알려진 표적유전자를 Positive set으로 사용하고 이와 독립적으로 Positive set 유전자 수의 3배수를 무작위 추출하여 Negative set을 구성하였다. 이 후 구성된 Feature Table에 Random Forest method를 적용하여 분류 규칙(Classification rule)을 생성한다.

본 연구에서는 분류 모델의 검증을 위하여 LOOCV(Leave One Out Cross Validation) 방법을 사용한다. LOOCV는 하나의 샘플을 Test set으로 사용하고 나머지 데이터를 Training set으로 사용하는 교차검증법이다. Training set으로 분류 규칙을 학습하여 Test set의 Class를 맞추는 과정을 샘플 수만큼 반복하므로 데이터의 개수가 충분히 많지 않은 경우의 검증에 적합하다. 따라서 본 연구에서는 하나의 표적유전자를 제외한 나머지 표적유전자들로 규칙을 생성하고 이를 기반으로 제외되었던 하나의 표적유전자의 Class를 맞추는 검증을 전체 표적유전자 수만큼 반복한다.

Table 2. Feature Table Sample

Molecules ID / Target Gene	100052	100072	100186	100284	Class
ADAM10	0.25	0.25	0.17	0.17	TRUE
AKT1	1.00	0.67	0.33	0.33	TRUE
CCND1	1.20	0.25	0.20	0.20	TRUE
FYN	1.00	0.25	0.25	0.25	FALSE
TNFSF12	2.00	0.20	0.17	0.33	FALSE
UBE2A	1.25	1.00	0.20	0.40	FALSE

IV. Results

1. Experimental environment and Data

1.1 Experimental environment

본 연구는 Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz CPU, 32GB RAM, 64비트 운영체제의 머신으로 수행되었으며 개발도구는 Microsoft visual studio 2012를 사용하였다.

본 연구에서는 Weka 3.6을 사용하여 Random Forest 분류 분석 알고리듬으로 분류 모델을 구축하였다[19][22].

1.2 Data set

질병 관련 유전자 정보는 OMIM에서 알츠하이머병과 유방암에 연관된 유전자만을 각각 추출하였다[17]. 본 연구에서는 표 3과 같이 알츠하이머병48개, 유방암 56개의 유전자를 사용하였다.

알려진 표적유전자와 질병의 치료 관계는 TTD에서 추출하였으며 임상단계는 고려하지 않았다. 본 연구에서는 표 4와 같이 알츠하이머병, 유방암 각각 92개의 유전자를 사용하였다.

Table 3. Disease Related Genes

Alzheimer's disease				Breast cancer				
ACHE	CHRNB1	GABRA5	PDE9A	ADAM10	CXCR4	HRAS	MAP3K4	PTK6
APH1A	CHRNE	GABRA6	PIN1	AKT1	CYP19A1	HSD17B1	MDM2	RELA
APH1B	CHRNG	GCG	PLA2G7	ANGPT2	CYP1B1	HSP90AA1	MFGE8	SHH
APOE	CRTC1	GSK3A	PPP2CA	BRCA2	EGFR	HSPA5	MMP2	SRC
BCHE	CRTC2	GSK3B	PRKCD	CCND1	ENG	IL12A	NCOA3	TYMP
BDKRB2	CTSD	KYNU	PSEN1	CD34	EPHA2	IL12B	NFKB1	TYMS
CDK5	FLT3	MAPK14	PSENEN	CD3G	ERBB2	JUN	NRG1	VDR
CES1	FPR2	MAPT	PTGS2	CDC25A	ESR1	KDR	PGR	WNT5A
CFLAR	GABRA1	MPO	PTPRC	CDK4	ESRRA	LHCGR	PIK3CA	
CHRM1	GABRA2	NCSTN	RAC1	COPS5	FNTA	MAP2K1	PLAUR	
CHRM3	GABRA3	NGF	SNCA	CTSD	FOS	MAP2K5	POR	
CHRNA1	GABRA4	NGFR	TNFRSF1A	CXCL12	HLA-DRB1	MAP3K1	PTGS2	

Table 4. Drug Target Genes

Alzheimer's disease					Breast cancer				
APH1B	ADRA2A	GABRA4	MGEA5	TTBK2	HSP90AA1	FOLR1	HLA-DRB1	CTSD	PTGS2
HRH3	ADRA2C	GABRA5	BET1	ENSA	FNTA	POR	CD3G	NCOA3	PGR
BDKRB2	CHRNA1	GABRA6	CRTC1	PTPRC	PF11_0483	CD34	TOP1	At1g60490	UL39
GRM3	CHRNA2	GABRB1	CRTC2	CFLAR	MMP2	HSPA5	ENG	ST14	SRC
GSK3B	CHRNA3	GABRB2	TNFRSF1A	KYNU	JUN	S100A4	GPNMB	NRG1	ESR2
APH1A	CHRNA4	GABRB3	PLA2G7	FGF8	CXCR4	SLC39A6	FOS	MFGE8	VDR
PSENEN	CHRNA5	GABRG1	CCR2	PRKCD	AKT1	SHH	PTN	TPBG	CYP19A1
NCSTN	CHRNA6	GABRG2	ECE1	PTGS2	CDH2	KDR	MAP3K4	MUC1	GNRH1
PSEN1	CHRNA7	GABRG3	ALOX12	HTR4	CLU	ERBB2	MAP2K1	CXCL12	LHCGR
NGFR	CHRNA9	GALR1	CES1	IDE	STS	CDK4	MAP2K5	SCGB2A2	TYMS
MPO	CHRNA10	GALR2	CDK5	ACHE	MDM2	ADAM10	CDC25A	PTK6	EGFR
MAPK14	CHRNB1	GALR3	GRM2	BCHE	ANGPT2	MAGEA1	CYP1B1	COPS5	TSPO
GCG	CHRNB3	MAPT	PIN1		PRLR	CTCFL	NFKB1	SNCG	
CHRNB2	CHRNB4	PDE9A	PPP2R5A		TOP1MT	CD46	NFKB2	SERPINB5	
PPP2CA	CHRND	SNCA	APOE		ESR1	S1PR3	RELA	DNMT3B	
CHRM1	CHRNG	NGF	MKC7		IL12A	IKBKE	TYMP	BRCA2	
CHRM2	CHRNE	FLT3	APCS		IL12B	HSD17B1	ESRRA	STC1	
CHRM3	GABRA1	GSK3A	CTSD		NMBR	WNT5A	PLAUR	Smad9	
CHRM4	GABRA2	RAC1	FPR2		GRPR	MAP3K1	CCND1	EPHA2	
CHRM5	GABRA3	CFP	CTSC		BRS3	PIK3CA	PRL	HRAS	

2. Experimental Results

본 연구에서는 모델의 구축과 검증에 사용되는 데이터의 수가 알려진 표적유전자의 수에 의해 정해진다. 알려진 표적유전자의 수가 많지 않으므로 분류 모델의 검증을 위하여 LOOCV를 수행하였다.

본 연구에서는 모델 성능의 객관적 평가를 위하여 알츠하이머병과 유방암에 대한 LOOCV를 10회씩 반복하였다. 각각의 Negative set은 Positive set의 3배수로 추출하였으며, 10회 검증의 평균 AUC를 모델 유효성 평가에 사용하였다. AUC는 False Positive Rate를 X축, True Positive Rate를 Y축으로 하여 그려지는 ROC curve의 아래 면적이다. AUC의 값이 1에 가까울수록 성능이 우수함을 뜻하며, 일반적으로 0.5가 넘는 AUC는 해당 모델이 유의함을 뜻한다.

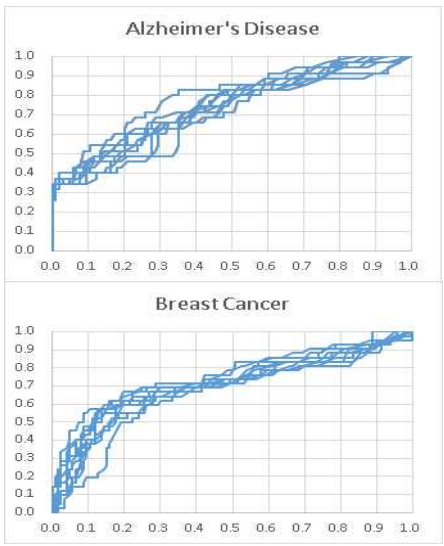

Fig. 3. ROC Curve

ROC curve는 그림 3과 같이 그려지며 각각의 ROC curve에 해당하는 AUC 결과는 표 5와 같다. 고른 AUC 결과를 통하여 데이터에 따른 편향성이 적음을 확인할 수 있으며, 평균 AUC 또한 알츠하이머병 0.74, 유방암 0.71로 본 연구의 분류 모델이 치료 가능한 표적유전자와 그렇지 않은 표적유전자를 효과적으로 분류할 수 있음을 보여준다.

Table 5. AUC Results

Validation #	Alzheimer's disease	Breast cancer
1	0.76	0.71
2	0.74	0.71
3	0.72	0.72
4	0.72	0.69
5	0.74	0.69
6	0.78	0.67
7	0.73	0.74
8	0.74	0.71
9	0.76	0.71
10	0.75	0.73
AVG	**0.74**	**0.71**
STD	**0.02**	**0.02**

Ⅴ. Conclusion

본 연구에서는 치료 관계가 검증된 표적유전자와 질병 유전자 간의 분자 네트워크상 거리를 분석하여 기존의 표적유전자를 대체할 수 있는 새로운 표적유전자를 찾는 모델을 구축하였으며, 분류 모델의 성능은 ROC curve에 따른 AUC 척도를 이용하여 검증하였다. 표적유전자는 특정 질병 치료에 영향을 미치는 작용 외에 다른 질병에 대하여 새로운 작용을 할 수 있다. 따라서 본 모델은 표적유전자의 알려진 기능이 아닌 새로운 질병에 대하여 작용할 수 있는 기능을 분석하였기 때문에 기존의 약물-약물, 질병-질병 유사성을 활용한 약물-질병관계 중심 연구보다 폭넓은 범위에서 새로운 표적유전자를 발견할 수 있다.

본 연구의 분류 모델의 검증과 별개로 질병 유전자 중 약물 치료에 중요한 유전자를 식별하기 위하여, 유방암의 10회 검증에서 각각 Best First 기반의 CFS 알고리즘으로 속성 부분 집합 선택을 수행하였다. 속성 부분 선택을 했을 경우의 평균 AUC 또한 0.71로 모든 속성을 사용한 경우와 흡사한 결과를 얻었으며, 전체 10회의 속성 부분 집합 선택 중 유전자 AKAP13이 9회 선택되었다. 이는 유전자 AKAP13이 유방암 치료에 사용될 수 있는 표적과 그렇지 않은 표적을 구분하는데 가장 많은 관여를 한다고 해석할 수 있다. ATAP13은 Rho GTPases signalling network의 중심적인 역할을 하는 유전자이다[23]. 많은 연구에서 Rho GTPases는 암의 초기 진행에서 중요한 역할을 한다고 밝혀졌다[24].

향후 연구에서는 치료에 보다 중요한 영향을 미치는 질병 관련 유전자를 식별하고, 식별된 유전자의 발현 상태와 표적유전자가 해당 유전자의 발현에 미치는 영향을 분석하여 효율적인 약물을 찾는 연구를 진행할 예정이다.

REFERENCES

[1] DiMasi, Joseph A., Ronald W. Hansen, and Henry G. Grabowski. "The price of innovation: new estimates of drug development costs." Journal of health economics, Vol. 22, No. 2, pp. 151-185, Mar. 2003.

[2] Kitano, Hiroaki. "Computational systems biology." Nature, Vol. 420, No. 6912, pp. 206-210, Nov. 2002.

[3] Landry, Yves, and Jean-Pierre Gies. "Drugs and their molecular targets: an updated overview." Fundamental & clinical pharmacology, Vol. 22, No. 1, pp. 1-18, Feb. 2008.

[4] Yang, Hong, et al. "Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information." Nucleic acids research, Nov. 2015.

[5] Iorio, Francesco, et al. "Discovery of drug mode of action and drug repositioning from transcriptional responses." Proceedings of the National Academy of Sciences, Vol. 107, No. 33, pp. 14621-14626, Aug. 2010.

[6] Gottlieb, Assaf, et al. "PREDICT: a method for inferring novel drug indications with application to personalized medicine." Molecular systems biology, Vol 7, No. 1, p. 496, Jan. 2011.

[7] Sawada, Ryusuke, et al. "Target-Based Drug Repositioning Using Large-Scale Chemical-Protein Interactome Data." Journal of Chemical Information and Modeling, Vol. 55, No. 12, pp. 2717-2730, Nov. 2015.

[8] Chen, Xin, Zhi Liang Ji, and Yu Zong Chen. "TTD: therapeutic target database." Nucleic acids research, Vol. 30, No. 1, pp. 412-415, Jan. 2002.

[9] Schaefer, Carl F., et al. "PID: the pathway interaction database." Nucleic acids research, Vol.37, suppl. 1, pp. D674-D679, Jan. 2009.

[10] Croft, David, et al. "Reactome: a database of

reactions, pathways and biological processes." Nucleic acids research, Nov. 2010.

[11] Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research, Vol. 28, No. 1, pp. 27-30, Jan. 2000.

[12] Caspi, Ron, et al. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." Nucleic acids research, Vol. 38, suppl. 1, pp. D473-D479, Jan. 2010.

[13] Kandasamy, Kumaran, et al. "NetPath: a public resource of curated signal transduction pathways." Genome biology, Vol. 11, No. 1, pp.1-9, Jan. 2010.

[14] Mi, Huaiyu, and Paul Thomas. "PANTHER pathway: an ontology-based pathway database coupled with data analysis tools." Protein Networks and Pathway Analysis, Humana Press, Vol. 563, pp.123-140, May. 2009.

[15] Pon, Allison, et al. "Pathways with PathWhiz." Nucleic acids research, May. 2015.

[16] Pico, Alexander R., et al. "WikiPathways: pathway editing for the people." PLoS biology, Vol. 6, No. 7, Jul. 2008.

[17] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {1. 18. 2016}. World Wide Web URL: http://omim.org/

[18] Hamosh, Ada, et al. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic acids research, Vol 33, suppl. 1, pp. D514-D517, Jan. 2005.

[19] Svetnik, Vladimir, et al. "Random forest: a classification and regression tool for compound classification and QSAR modeling." Journal of chemical information and computer sciences, Vol. 43, No. 6, pp. 1947-1958, Nov. 2003

[20] Breiman, Leo. "Random forests." Machine learning, Vol. 45, No. 1, pp. 5-32, Oct. 2001.

[21] Nishimura, Darryl. "BioCarta." Biotech Software & Internet Report: The Computer Software Journal for Scient, Vol. 2, No. 3, pp. 117-120, Jul. 2004.

[22] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter, vol. 11, No. 1, pp. 10-18, Jun. 2009.

[23] Wirtenberger, Michael, et al. "Association of genetic variants in the Rho guanine nucleotide exchange factor AKAP13 with familial breast cancer." Carcinogenesis, Vol. 27, No. 3, pp. 593-598, Oct. 2005.

[24] Ellenbroek, Saskia IJ, and John G. Collard. "Rho GTPases: functions and association with cancer." Clinical & experimental metastasis, Vol. 24, No. 8, pp. 657-672, Nov. 2007.

Authors

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.

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.