ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 23, No. 1, pp. 28-36, 2017. 1
https://doi.org/10.5626/KTCP.2017.23.1.28

생물학 문헌 데이터의 제목과 본문을 이용한 질병 관련 유전자 추론 방법

(Inferring Disease-related Genes using Title and Body in Biomedical Text)

김 정 우 † 김 현 진 † 여 윤 구 † 신 민 철 † 박 상 현 ††
(Jeongwoo Kim) (Hyunjin Kim) (Yunku Yeo) (Mincheol Shin) (Sanghyun Park)

요 약 1990년대 게놈프로젝트 이후 유전자와 관련된 많은 연구가 진행되고 있다. 데이터 저장 기술의 발달로 연구의 결과물들은 다량의 문헌들로 기록되고 있으며, 이러한 문헌들은 새로운 생물학적 관계들을 추론하는 데이터로 유용하게 사용되고 있다. 이러한 이유로 본 연구에서는 생물학 문헌들을 활용하여 질병과 관련한 유전자를 추론하는 방법론에 대해서 제안한다. 문헌들을 제목과 본문으로 구분하고, 각 영역에서 등장한 유전자들을 추출한다. 제목 영역에서 추출된 유전자는 중심 유전자로 구분하고, 본문 영역에서 추출된 유전자는 제목에서 추출된 유전자와 관계를 갖는 주변 유전자로 구분한다. 이러한 과정을 각 문헌에 적용하여, 지역 유전자 네트워크를 구축한다. 구축된 지역 유전자 네트워크는 모두 연결하여 전역 유전자 네트워크를 구축한다. 구축한 네트워크를 분석하여 질병 관련 유전자를 추론하였으며, 비교 실험을 통해 제안하는 방법론이 질병 관련 유전자를 추론하는 유용한 방법론임을 입증하였다.

키워드: 유전자, 질병, 유전자-질병 관계, 텍스트 마이닝, 유전자 네트워크

Abstract After the genome projects of the 90s, a vast number of gene studies have been stored in online databases. By using these databases, several biological relationships can be inferred. In this study, we proposed a method to infer disease-gene relationships using title and body in biomedical text. The title was used to extract hub genes from data in the literature; whereas, the body of the literature was used to extract sub genes that are related to hub genes. Through these steps, we were able to construct a local gene-network for each report in the literature. By integrating the local gene-networks, we then constructed a global gene-network. Subsequent analyses of the global gene-network allowed inference of disease-related genes with high rank. We validated the proposed method by comparing with previous methods. The results indicated that the proposed method is a meaningful approach to infer disease-related genes.

Keywords: gene, disease, disease-gene relationship, text-mining, gene network

· 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015R1A2A1A05001845)

† 비 회 원 : 연세대학교 컴퓨터과학과
jwkim2014@naver.com
firadazer@yonsei.ac.kr
purplerainf@gmail.com
smanioso@yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
(Yonsei Univ.)
sanghyun@yonsei.ac.kr
(Corresponding author임)

논문접수 : 2016년 6월 24일
(Received 24 June 2016)
논문수정 : 2016년 9월 20일
(Revised 20 September 2016)
심사완료 : 2016년 10월 21일
(Accepted 21 October 2016)

정보과학회 컴퓨팅의 실제 논문지 제23권 제1호(2017. 1)

1. 서 론

1990년 게놈프로젝트 이후, 유전자와 관련된 많은 연구가 진행되고 있다. 이러한 연구의 결과들은 질병과 유전자는 밀접한 관계가 있음을 입증하였다. 또 데이터 저장 기술의 발달로 다량의 문헌정보가 데이터베이스에 기록되었다. 생물학 문헌 데이터를 관리하고 제공하는 대표적인 데이터베이스로 PubMed가 있다[1]. 이러한 문헌 데이터베이스를 활용하면 생물학적 개체들 사이의 정보를 추출할 수 있고, 추출된 관계들을 분석함으로써 새로운 후보 관계들을 추론할 수 있다. 하지만 이러한 문헌 데이터들은 그 양이 많고 산재하여 있다는 특성 때문에 사람이 모든 문헌데이터를 일일이 분석하는 것은 불가능하다. 이러한 이유로 생물학 문헌데이터를 분석하는 방법은 생물학 분야에서 새로운 도전과제 및 목표가 되었다.

생물학 분야에서 텍스트 마이닝은 생물학적 개체들 사이의 관계를 추론하는 방법론으로 사용되고 있다. 이 중 널리 이용되는 방법론 중 하나는 동시 출현이다. 동시 출현은 관심 있는 두 개의 단어가 한 문단 혹은 한 문장에 동시에 출현할 경우, 두 단어는 연관성이 있다고 가정하고 두 단어 사이의 관계를 추출하는 접근 방식을 의미한다. 동시 출현 접근법을 통하여 생물학 문헌을 분석하면, 생물학 문헌상에서 등장하는 다양한 생물학 개체들 사이의 관계를 추론할 수 있다.

질병의 발병은 단순히 하나의 요인에 의해서 발생하는 경우보다 다양한 생물학적 개체들 사이의 복잡한 관계들에 의해서 생성되는 경우가 대부분이다. 이러한 이유 때문에 생물학적 개체들 사이의 연관성을 추론하는 것은 중요한 연구 과제라 할 수 있다.

본 논문에서는 이러한 점에 착안하여, 문헌 데이터를 활용하여 질병 관련 유전자를 추론하는 방법론에 대한 연구를 진행하였다. 문헌 데이터를 제목과 본문, 두 가지 형식으로 분류하여 분석하였다. 제목 영역에서 추출된 유전자는 중심 유전자로 추출하였고, 해당 문헌의 본문에서 추출되는 유전자들은 중심 유전자와 관계를 가지는 주변 유전자로 추출하였다. 각 문헌에서 이러한 과정을 반복하여 중심 유전자와 주변 유전자를 연결하는 지역 유전자 네트워크들을 구축하였다. 구축된 지역 유전자 네트워크들은 하나로 통합되어 전역 유전자 네트워크를 구축한다. 전역 유전자 네트워크를 분석하여 각 노드(유전자)에 대한 점수를 계산한다. 점수가 높은 상위 N개의 유전자를 추출하고, 해당 유전자가 실제로 질병과 관련이 있는지에 대해 검증하였다.

2. 관련 연구

2.1 생물학 문헌 텍스트 마이닝

Chiang[2]은 생물학 문헌 데이터를 활용하여 유전자 정보를 제공하는 시스템을 구축하였다. 그들은 유전자 정보 시스템을 구축하기 위하여 4가지 형식(생물학적 기능, 유전자와의 연관성, 질병과의 연관성, 유전자-유전자 관계)으로 유전자 정보를 구분하였다. 문헌 데이터 속에서 4가지 형식의 정보를 활용하는 시스템을 구축하여, 문헌 분석에서 요구되는 시간과 노력을 감소시켜 편의성을 제공하였다.

Xie[3]는 문헌 데이터를 기반으로 microRNA 데이터를 추출하는 방법론을 제안하였다. 그들은 문헌상에서 microRNA 데이터 추출을 위해 microRNA 데이터를 기술하는 75가지의 규칙을 생성하였다. 구축한 규칙들을 기반으로 약 26,000개의 문헌상에서 후보 microRNA 데이터를 추출하였고, 추출된 관계를 직접 검토하였다. 그 결과 79개의 암과 236개의 microRNA 데이터, 그리고 878개의 암과 microRNA 관계를 추출하였다.

2.2 생물학 개체들 사이의 관계 추론

Lee[4]의 연구에서는 질병과 약물 사이의 관계를 추론하기 위하여 문헌 데이터를 활용하는 방법론을 제안하였다. 문헌상에서의 동시 출현을 기반으로 질병-약 사이의 관계를 추론하고, 해당 관계에 대한 가중치를 주기 위하여 context vector라는 것을 정의하고 사용하였다. context vector는 질병-약물 관계가 추출된 논문에서 등장하는 생물학적 개체들에 대한 정보를 의미한다. 그들은 알츠하이머 질병에 대해서 실험을 하였고, 의미 있는 질병과 약물의 관계를 추론하였다.

Oron[5]은 질병과 유전 사이의 관계를 추론하기 위하여 prioritization 함수를 정의하여 사용하였다. prioritization 함수는 유전자정보뿐만 아니라 단백질 복합체에 대한 데이터를 함께 활용하여 구성하였으며, prioritization 함수를 활용하여, 질병 관련 유전자를 추론하는 방법론을 제안하였다.

Li[6]는 생물학 문헌데이터와 microArray 유전자 발현 데이터를 통합하는 방법론에 대해서 연구를 진행하였다. 생물학 문헌 데이터에서 동시 출현을 기반으로 유전자 네트워크를 구축하였고, 문헌을 기반으로 구축된 유전자 네트워크를 microArray 유전자 발현 데이터를 활용하여 재구축하였다. 혈관 형성과 관련된 데이터를 실험 데이터로 하여 제안하는 방법론을 검증하였으며, 문헌에서 동시 출현을 기반으로 구축한 네트워크보다 더 신뢰성 있는 네트워크를 구축하였음을 입증하였다.

3. 방법론

이번 장에서는 제목과 본문을 활용하는 생물학적 문헌 텍스트 마이닝 기법에 대해 기술한다. 그림 1은 제안하는 방법론에 대한 전반적인 개요를 나타낸다. 우선 PubMed로부터 결장암과 유방암과 관련된 문헌 데이터를 각각 얻는다. 문헌들은 전처리 단계를 통해 저자 정보, 소속 기관, 논문지 정보 등 불필요한 정보를 제거하고, 제목과 본문만을 데이터로 사용한다. 제목과 본문을 분류한 후, 각 영역에서 유전자 심볼을 추출한다. 유전자 심볼 리스트에 대한 정보는 HGNC[7] 데이터베이스로부터 가져왔다.

그림 1에서 볼 수 있듯이, 문헌 데이터는 제목과 본문 부분으로 나누어지고, 각 위치에서 추출된 유전자들은 중심 유전자와 주변 유전자로 나누어진다. 각 문헌으로부터 지역 유전자 네트워크가 구축되고, 구축된 모든 지역 유전자 네트워크들은 하나의 전역 유전자 네트워크로 구축된다. 유전자 네트워크에서 두 노드 사이의 가중치는 추출된 유전자의 위치 정보와 빈도수에 의해서 계산된다. 전역 유전자 네트워크를 분석하고 가중치를 기반으로 각 유전자에 대한 점수를 계산하여, 이를 기반으로 질병 관련 유전자를 추론한다.

3.1 전처리

전처리 과정을 통해 먼저 문헌 데이터 중에 사용하지 않는 저자 정보, 논문지, 날짜 등의 데이터를 제외하고 제목과 본문만을 추출한다. 특수 기호 사이에 등장하는 유전자를 추출하기 위해, 문헌들에서 소괄호, 하이픈, 슬래시를 제거하였다. 또 문헌상에서 정확한 유전자 심볼을 추출하기 위하여, 약 37,000의 유전자 심볼 중 3글자 미만의 유전자 심볼 40개를 제외하였다. 3글자 미만의 유전자 심볼들은 유전자를 표현하는 경우보다 다른 의미로 사용되는 경우가 많다.

3.2 지역 및 전역 유전자 네트워크 구축

전처리 과정을 마친 문헌 데이터 제목과 본문 각각에서 유전자를 추출하고, 중심 유전자와 주변 유전자를 연결하는 지역 유전자 네트워크를 구축한다. 만약 제목 영역에서 2개 이상의 유전자가 등장할 경우, 모든 유전자를 서로 연결한다. 반면에 본문 영역에서 2개이 상의 유전자들이 등장할 경우, 주변 유전자 사이에는 관계를 형성하지 않는다. 즉, 본문 영역에서 등장하는 주변 유전자들은 제목 영역에 등장하는 중심 유전자에만 관계를 형성한다.

위의 과정을 통해 구축된 모든 지역 유전자 네트워크들을 연결하여, 전역 유전자 네트워크를 구축한다. 전역 유전자 네트워크에서 관계들에 대한 가중치는 지역 유전자 네트워크의 가중치를 사용한다. 만약 지역 유전자 네트워크 중에서 중복된 관계가 있으면, 각 지역 유전자 네트워크의 가중치들을 합산하여 가중치를 재설정한다. 지역 유전자 네트워크의 가중치는 유전자 점수 계산 단계를 거쳐 생성된다.

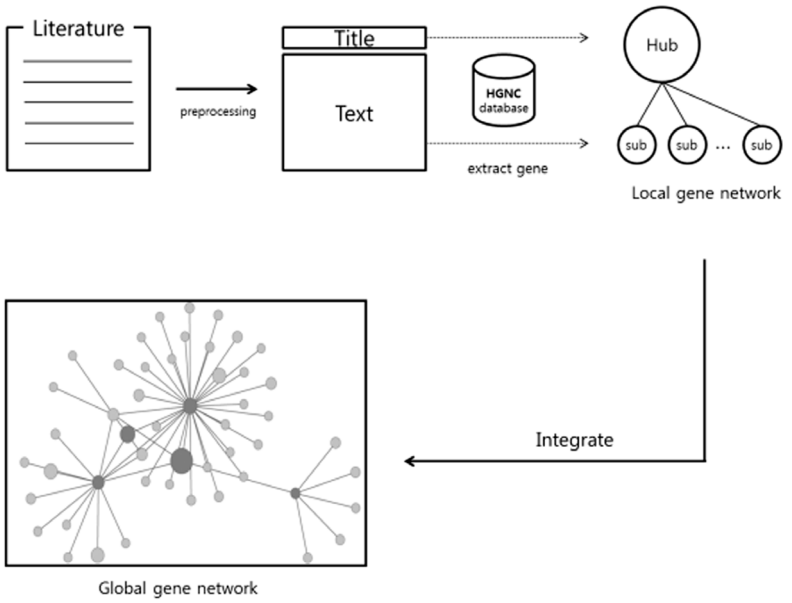

그림 1 제안하는 방법론의 개요

Fig. 1 The outline of proposed method

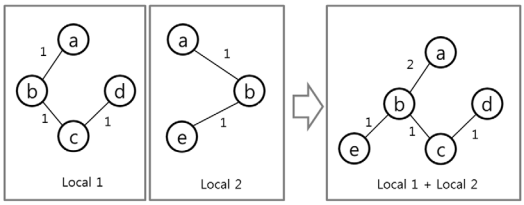

그림 2 네트워크 구축 예제
Fig. 2 Example for network construction

그림 2는 지역 네트워크를 병합하여 전역 네트워크를 구축하는 예제이다. 그림에서 볼 수 있듯이 중복된 관계의 경우 각 관계의 가중치를 합하여 새로운 가중치를 산정하고, 새로운 관계의 경우 네트워크를 확장한다.

3.3 유전자 점수 계산

우선 제목 영역에 등장하는 유전자가 본문 영역에 등장하는 유전자보다 더 중요하다는 가정을 기반으로 지역 유전자 네트워크의 가중치를 생성하였다. 가중치는 기본적으로 두 가지의 요소를 고려하여 생성한다. 첫 번째는 유전자가 문헌에서 등장 횟수를 의미하는 빈도수이고, 두 번째는 유전자가 출현한 영역이다. 아래는 지역 유전자 네트워크에서 가중치를 계산하는 식을 나타낸다.

$$\text{가중치}(\text{중심},\text{중심}) = 1$$

$$\text{가중치}(\text{중심},\text{주변}) = \frac{\textit{주변 노드의 빈도수}}{\textit{모든 주변 노드의 수}}$$

$$\text{가중치}(\text{주변},\text{주변}) = 0$$

위의 식에서와 같이 중심 유전자 사이의 가중치는 가장 큰 값인 1의 값을 가지게 된다. 반면에 주변 유전자 사이의 가중치는 0으로써, 관계가 형성되지 않는다. 중심 유전자와 주변 유전자 사이의 가중치는 주변 유전자의 출현 빈도수와 비례한다. 위의 식에서 모든 주변 노드의 수는 문헌 하나에서 등장하는 모든 유전자의 수를 나타내고, 주변 노드의 빈도수는 한 문헌에서 특정 주변 노드의 출현 횟수를 의미한다. 가중치(중심,주변)의 의미는 논문에 제목에 유전자가 포함된다면, 그 논문은 해당 유전자에 대한 분석 논문일 확률이 크고, 그렇다면 그 논문에서 추가로 등장하는 유전자들은 제목에 등장한 유전자와 관련이 있는 유전자라고 판단하였다.

위의 수식을 기반으로 각 지역 유전자 네트워크들은 가중치를 형성하게 되고, 이를 기반으로 하나의 전역 유전자 네트워크를 구축한다. 구축된 전역 유전자 네트워크의 노드에 대한 점수는 연결 중앙성을 기반으로 계산된다. 연결 중앙성이란 관계가 형성되어 있는 이웃 노드들과의 가중치를 모두 합산하여, 해당 노드의 점수로 사용하는 방법이다. 노드 점수에 대한 계산은 아래의 식을 이용한다.

$$\text{scoring function} = \sum_{n=1}^{N(A)} \textit{가중치}(A, A_n^+)$$

위의 식에서, N(A)는 노드 A와 연결된 모든 이웃 노드의 수를 의미한다. A_n^+는 n번째 이웃 노드를 의미하고, $\textit{가중치}(A, A_n^+)$는 노드 A와 n번째 이웃 노드 사이의 가중치를 의미한다. 유전자 네트워크의 노드는 이웃 노드의 수가 많을수록, 또 이웃 노드와의 가중치가 높을수록 더 많은 점수를 받게 된다.

4. 결과 및 평가

이번 장에서는 실험 결과를 통해 제안하는 방법론과 기존 연구와의 비교 실험 결과를 제시한다. 실험을 위해 결장암과 유방암의 데이터를 사용하였다. 검증을 위한 방법으로는 점수가 높은 상위 20개의 유전자를 추출하고, 해당 유전자가 실제로 질병과 관련이 있는지에 대한 여부를 확인하였다. 질병 관련 유전자의 대한 추론 방법은 상위 N개의 유전자가 얼마나 질병과 관련이 있는지 여부가 중요하다. 그 정확도가 높을수록 아직 알려지지 않은 추론 유전자들이 좋은 후보 군이 되기 때문이다. 질병과 유전자의 관련성 여부를 판단하기 위해서 질병 관련 유전자를 제공하는 데이터베이스들을 활용하였다.

4.1 데이터 및 네트워크 구성

결장암과 유방암에 대한 문헌 데이터는 PubMed로부터 해당 질병 명의 검색을 통해 얻었고, HGNC로부터 약 37,000개의 유전자 심볼에 대한 데이터를 받았다. 문헌 데이터 및 구축한 유전자 네트워크에 대한 구성 정보는 표 1을 통해 나타내었고, 구축한 네트워크는 그림 2를 통해 제시하였다.

표 1에서는 각 질병에 대해 사용한 문헌 데이터의 수와 문헌데이터를 사용하여 구축한 전역 유전자 네트워크의 구성 성분에 대한 값을 나타낸다. 표 1에서 제안하는 방법론의 네트워크 구성 요소가 더 작은 값을 보임을 확인할 수 있다. 이는 제안 하는 방법론은 제목에 유전자가 등장하지 않을 경우 해당 문헌은 사용하지 않기 때문이다.

표 1 데이터 및 유전자 네트워크의 구성
Table 1 Data and properties of gene networks

	Colorectal Cancer		Breast Cancer	
Method	Co-occurrence	Our	Co-occurrence	Our
Literature	57,648		173,515	
Node of Network	2,419	400	3,971	760
Edge of Network	10,275	278	22,336	570

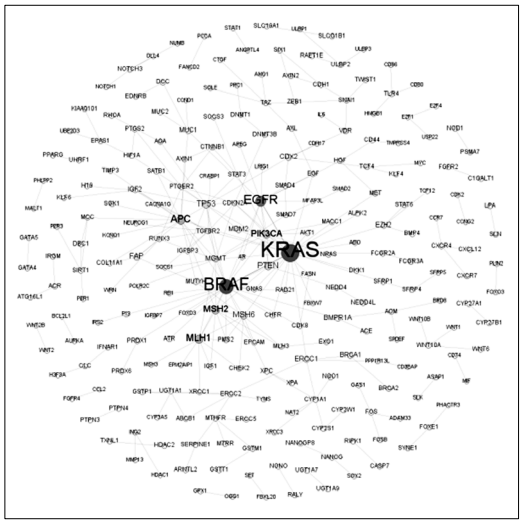

그림 3 결장암에 대한 유전자 네트워크
Fig. 3 Colorectal cancer-related gene network

표 2 정답 데이터 집합
Table 2 Answer Set Data

	Answer Set						
Disease	Colorectal Cancer				Breast Cancer		
Known genes	23	12	35	13	20	9	107
Reference	Sanger	KEGG	GHR	NCI	Sanger	KEGG	GHR

그림 3은 결장암 데이터에 제안하는 방법론을 적용하여 구축한 유전자 네트워크이다. 그림에서 확인할 수 있듯이, 노드와 간선의 수가 많지 않은 유전자 네트워크이다.

표 2는 질병 관련 유전자를 검증하기 위해 사용한 데이터베이스들에 대한 정보를 나타낸다. Sanger[8], KEGG[9], GHR[10], NCI[11] 모두 질병 관련 유전자에 대한 정보를 제공하는 데이터베이스이다.

4.2 동시 출현 기반 방법론과의 비교

가장 널리 사용되는 텍스트마이닝 방법론 중 하나인 동시출현 기반의 관계 추론 방법과 비교 실험을 하였다. 동시출현 기반의 관계 추론은 한 문장에 두 개 이상의 유전자가 등장할 경우, 유전자들 사이의 연관성이 있다고 판단하고 관계를 생성하는 방법론이다. 실험을 위해 문헌 데이터를 한 문장 단위로 나누고, 유전자가 두 개 이상 등장하는 문장을 추출하였다. 추출한 문장에서 가능한 모든 유전자와 유전자 사이의 관계를 만들고, 빈도수를 기반으로 네트워크를 구축하였다.

그림 4는 제안하는 방법론과 동시 출현 기반 방법론과의 비교 실험 결과를 보여준다. 가로축은 추론한 유전자의 수를 나타내고, 세로축은 실제로 질병과 관련이 있다고 알려진 유전자의 수를 나타낸다. 그림 4에서 확인할 수 있듯이, 결장암과 유방암에 대한 두 실험 모두에서 제안하는 방법론이 N이 작을수록 더 많은 수의 질병 관련 유전자를 추론하였다. 이를 통해 제안하는 방법론이 기존의 동시 출현 기반의 방법론보다 질병 관련 유전자 추론 방법론으로써 더 우수한 가중치를 사용함을 입증하였다.

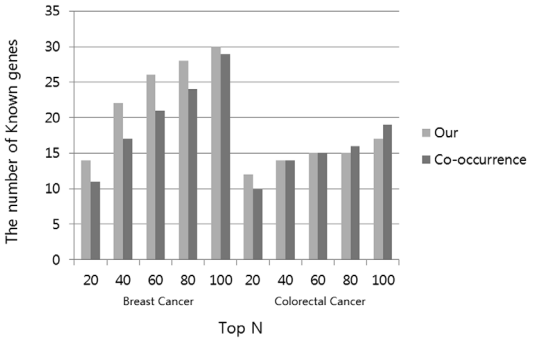

그림 4 동시 출현 방법과의 비교 실험 결과
Fig. 4 Results of comparison with co-occurrence based method

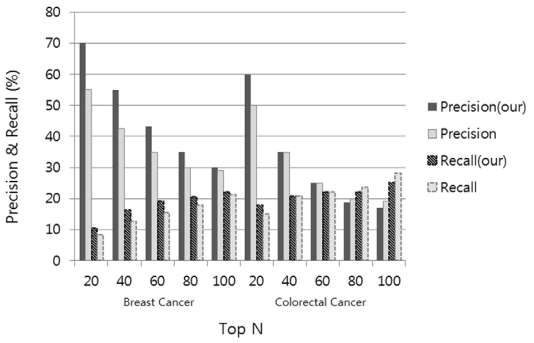

그림 5 동시 출현 방법과의 Precision과 Recall 비교
Fig. 5 Comparison precision and recall with co-occurrence based method

그림 5는 제안하는 방법론과 동시 출현 기반 방법론과의 Precision과 Recall 값을 나타낸다. 그림에서 회색은 제안하는 방법론을 나타내며, 검은색은 동시 출현 방법론의 결과를 나타낸다. 그림 5를 통해 확인할 수 있듯이, 제안하는 방법론이 N의 값이 작을수록 더 높은 Precision과 Recall 값을 나타냄을 확인할 수 있다.

또 기존의 동시 출현 기반의 방법론은 단순히 등장의 여부를 기반으로 관계를 생성하기 때문에, 유전자 네트워크에 너무 많은 관계가 포함된다. 이러한 점은, 너무 많은 질병 관련 후보 유전자를 추론하게 된다는 문제점을 발생시킬 수 있다. 하지만 제안하는 방법론은 제목에

반드시 유전자 이름이 등장하여야 관계를 생성하기 때문에, 기존의 방법론에 비해 적은 수의 관계를 생성한다.

제안하는 방법론은 같은 문헌 데이터에서 기존의 방법론보다 매우 작은 수의 유전자와 관계를 추출하는 것을 확인할 수 있다. 하지만 구축한 네트워크 분석을 통해 추출한 유전자는 더 유의미한 결과를 보임을 그림 3을 통해서 확인하였다. 이러한 실험 결과들을 통해서 제안하는 방법론이 후보 유전자를 추론하는 방법론으로써 더 적합하며, 네트워크 분석이 더 용이한 방법임을 입증한다.

4.3 PRINCE 알고리즘과의 비교

PRINCE[12] 알고리즘은 질병 관련 유전자를 추론하는 방법론 중 하나이다. 기본적으로 네트워크를 기반으로 하는 접근 방식이고, 질병 사이의 유사도와 단백질-단백질 상호 작용 정보를 사용한다. 아래는 PRINCE 알고리즘과의 비교 실험 결과를 나타낸다.

그림 6에서 보이듯이, 결장암과 유방암 데이터 모두에서 제안하는 방법론이 PRINCE 알고리즘보다 더 많은 질병 관련 유전자를 추론하였다. 유방암 데이터의 경우

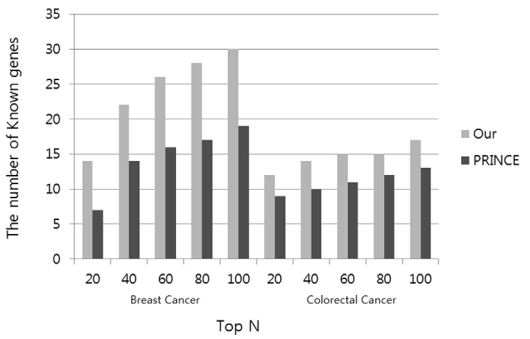

그림 6 PRINCE 알고리즘과의 비교 실험

Fig. 6 Comparison results with PRICNE algorithm

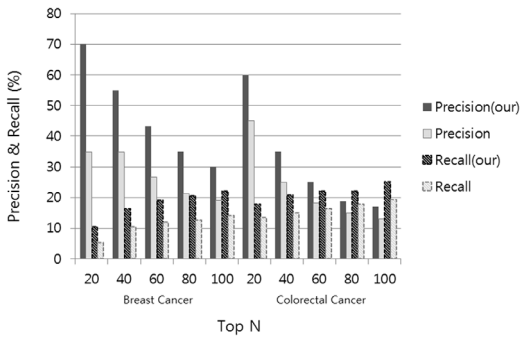

그림 7 PRINCE 알고리즘과의 Precision과 Recall 비교

Fig. 7 Comparison precision and recall with PRINCE algorithm

제안하는 방법론이 약 2배에 가까운 많은 질병 관련 유전자를 추론함을 확인할 수 있었다.

그림 7은 제안하는 방법론과 PRINCE 알고리즘과의 Precision과 Recall 값을 나타낸다. 회색이 제안하는 방법론을 나타내며, 검은색은 PRINCE 알고리즘을 나타낸다. 그림 7에서 확인할 수 있듯이, 제안하는 방법론이 전체 구간에서 더 높은 Precision과 Recall값을 보임을 확인할 수 있다.

4.4 Top 20 추론 유전자 분석

제안하는 방법론을 통해 추론한 상위 20개의 유전자에 대해서, 데이터베이스 검증 이외에 추가적인 문헌 검증을 통해 유의미성을 확인하였다. 아래의 표는 결장암 데이터에서 제안하는 방법론을 통해 추론한 상위 20개의 유전자를 나타낸다.

표 3에서 Evidence는 해당 유전자와 질병과의 연관성을 확인할 수 있는 참조 자료를 나타낸다. Sanger, KEGG, GHR, NCI는 데이터베이스를 의미하며, literature는 유전자와 질병에 대한 정보를 담고 있는 생물학적 문헌 자료를 의미한다. EGFR, MTHFR, CD44, CDX2, MGMT, CHEK2, UGT1A1, STAT3모두 질병 관련 유전자 정보를 제공하는 데이터베이스들을 통해 검증되지는 않았지만, 문헌 데이터를 통해서 해당 유전자들이 결장암과 관련이 있음을 확인할 수 있었다.

표 3 결장암에 대해 추론한 상위 20개의 유전자

Table 3 Top 20 genes involved in colorectal cancer inferred by our method

Rank	Gene	Evidence
1	KRAS	Sanger, KEGG, GHR
2	BRAF	Sanger
3	APC	Sanger, KEGG, GHR, NCI
4	EGFR	literature[13]
5	MLH1	Sanger, KEGG, GHR, NCI
6	MTHFR	literature[14]
7	PTEN	NCI
8	TP53	Sanger, KEGG, GHR, NCI
9	DCC	KEGG, GHR
10	PIK3CA	Sanger, GHR
11	CD44	literature[15]
12	MSH2	Sanger, KEGG, GHR, NCI
13	MUTYH	Sanger, GHR
14	CDX2	literature[16]
15	MGMT	literature[17]
16	MDM2	Sanger
17	MSH6	Sanger, KEGG, GHR, NCI
18	CHEK2	literature[18]
19	UGT1A1	literature[19]
20	STAT3	literature[20]

표 4 유방암에 대해 추론한 상위 20개의 유전자
Table 4 Top 20 genes involved in breast cancer inferred by our method

Rank	Gene	Evidence
1	BRCA1	Sanger, KEGG, GHR
2	BRCA2	Sanger, KEGG, GHR
3	EGFR	literature[21]
4	MUC1	GHR
5	PTEN	KEGG, GHR
6	CD44	literature[22]
7	TP53	Sanger, KEGG, GHR
8	PIK3CA	Sanger, GHR
9	CYP2D6	literature[23]
10	ATM	GHR
11	CXCR4	GHR
12	CD24	literature[24]
13	ERBB2	Sanger, KEGG, GHR
14	EZH2	GHR
15	CHEK2	Sanger, GHR
16	MDM2	literature[25]
17	EGF	literature[26]
18	ESR1	GHR
19	PALB2	Sanger, GHR
20	RAD51	GHR

표 4는 제안하는 방법론을 사용하여 유방암 데이터에서 추론한 상위 20개의 질병 관련 유전자를 의미한다. 데이터베이스를 통해 검증되지 않은 EGFR, CD44, CYP2D6, CD24, MDM2, EGF의 유전자들도 유방암과 관련성이 있음을 문헌 정보를 통해 확인할 수 있었다.

질병 관련 유전자를 추론하는 방법론은 기존에 연관성이 알려진 질병 관련 유전자를 추출하는 것도 중요하지만, 유의미한 후보 유전자를 추론하는 것이 중요하다. 이러한 관점에서 봤을 때, 제안하는 방법론은 이미 알려진 질병 관련 유전자를 추출하는데 있어서도 다른 기존의 방법론들보다 유의미한 결과를 보였으며, 데이터베이스를 통해 검증되지 않은 유전자들에 대해서도 해당 질병과의 연관성을 문헌 정보를 통해 확인하였다. 이러한 이유로 제안하는 방법론은 질병 관련 유전자를 추론하는 유용한 방법론이라고 할 수 있다.

5. 결론 및 향후 연구

본 연구에서는 문헌 데이터를 사용하여 질병 관련 유전자를 추론하는 방법론을 제안하였다. 기존의 방법들과의 차이점은 문헌 데이터를 제목과 본문으로 나누어 분석하였다는 점과, 문헌별로 지역 유전자 네트워크를 구축하고, 병합하는 과정을 통해 하나의 전역 유전자 네트워크를 구축하는 방식이다.

제안하는 방법론을 검증하기 위해서 결장암과 유방암 데이터에 대해 실험을 수행하였으며, 기존의 질병 관련 유전자 추론 방법인 동시 출현 기반 방법론과 PRINCE 알고리즘과의 비교 실험을 수행하였다. 각 실험에서 결장암과 유방암 데이터 모두에서 더 많은 질병 관련 유전자를 추론함을 확인할 수 있었다. 또 데이터베이스를 통해 검증되지 않은 유전자들에 대해서도 문헌 검증을 통해 질병과의 연관성을 확인할 수 있었다. 이러한 결과를 통해 제안하는 방법론이 질병 관련 유전자를 추론하는 유용한 방법론임을 확인하였다.

본 논문에서 제안하는 방법론이 결장암과 유방암뿐만 아니라 다른 모든 유전적 질병에 대해서 좋은 결과를 보임을 입증하기 위하여, 다른 유전적 질병에 대해서 추가적인 실험을 진행할 계획이다. 또 현재 방법론에서는 지역 유전자 네트워크를 구축할 때, 빈도수를 기반으로 유전자 사이의 가중치를 계산한다. 향후 연구에서는 좀 더 세세하게 문헌 데이터를 분석하여, 빈도수보다 더 유의미한 가중치 선정 방법에 대한 연구를 진행할 것이다.

References

[1] PubMed: MEDLINE Retrieval on the World Wide Web. DOI=http://www.sanger.ac.uk/
[2] Chiang, J.H., Yu, H.C., and Hsu, H.J. GIS: a biomedical text-mining system for gene information discovery. Bioinformatics. 20, 1, (2004), 120-121.
[3] Xie, B., Ding, G., Han, H., Wu, D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics. 2013. 29(6):638-644.
[4] Lee, S., Choi, J., Park, K.., Song, M., and Lee, D. Discovering context-specific relationships from biological literature by using multi-level context terms. BMC Medical Informatics and Decision Making 12(Suppl 1):S1 (2012).
[5] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R. Associating Genes and Protein Complexes with Disease via Network Propagation. PLoS Comput Biol 6(1): e1000641.
[6] Li, S., Wu, L., and Zhang, Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics. 22, 17 (2006), 2143-2150.
[7] HGNC Database, HUGO Gene Nomenclature Committee (HGNC), EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. DOI=http://www.genenames.org/
[8] Wellcome Trust Sanger Institute. DOI=http://www.sanger.ac.uk/
[9] KEGG: Kyoto Encyclopedia of Genes and Genomes. DOI=http://www.genome.jp/kegg/
[10] National Library of Medicine (US). Genetics Home Reference [Internet]. Bethesda (MD): The Library.

DOI=http://ghr.nlm.nih.gov/
[11] National Cancer Institute: Comprehensive Cancer Information. DOI=http://www.cancer.gov/
[12] Gottlieb, A., Magger, O., Berman, I., Ruppin, E., Sharan, R. PRINCIPLE: a tool for associating genes with diseases via network propagation. Bioinformatics. 2011. 27(23):3325-3326.
[13] Hong, L., Han, Y., Zhang, H., Zhao, Q., Yang, J., Ahuja, N. High expression of epidermal growth factor receptor might predict poor survival in patient with colon cancer: a meta-analysis. Genet Text Mol Biomarkers. 2013; 17(4) :348-51.
[14] Teng, Z., Wang, L., Cai, S., Yu, P., Wang, J., Gong, J., Liu, Y. The 677C>T (rs1801133) polymorphism in the MTHFR gene contributes to colorectal cancer risk: a meta-analysis based on 71 research studies. PLoS One. 2013; 8(2):e55332.
[15] Saito, S., Okabe, H., Watanabe, M., Ishimoto, T., Iwatsuki, M., Baba, Y., Tanaka, Y., Kurashige, J., Miyamoto, Y., Baba, H. CD44v6 expression is related to mesenchymal phenotype and poor prognosis in patients with colorectal cancer. Oncol Rep. 2013 Apr; 29(4):1570-8.
[16] Hinoi, T., Loda, M., Fearon, ER., Silencing of CDX2 expression in colon cancer via a dominant repression pathway. J Biol Chem. 2003 Nov 7;278(45): 44608-16.
[17] Park, JH., Kim, NS., Park, JY., Chae, YS., Kim, JG., Sohn, SK., Moon, JH., Kang, BW., Tyoo, HM., Bae, SH., Choi, GS., Jun, SH. MGMT -533G>T polymorphism is associated with prognosis for patients with metastatic colorectal cancer treated with oxaliplatin-based chemotherapy. J Cancer Res Clin Oncol. 2010 Aug;136(8):1135-42.
[18] Liu, C., Wang, QS., Wang, YJ. The CHEK2 I157T variant and colorectal cancer susceptibility: a systematic review and meta-analysis. Asian Pan J Cancer Prev. 2012;13(5);2051-5.
[19] Bajro, MH., Josifovski, T., Panovski, M., Jankulovski, N., Nestorovska, AK., Metevska, N., Petrusevska, N., Dimovski, AJ. Promoter length polymorphism in UGT1A1 and the risk of sporadic colorectal cancer. 2012 Apr;205(4):163-7.
[20] Wang, W., Zhao, C., Jou, D., Lu, J., Zhang, C., Lin, L., Lin, J. Ursolic acid inhibits the growth of colon cancer-initiating cells by targeting STAT3. Anticancner Res. 2013 Oct;33(10):4279-84.
[21] Tang, Y., Zhu, L., Li, Y., Ji, J., Li, J., Yuan, F., Wang, D., Chen, W., Huang, O., Chen, X., Wu, J., Shen, K., Loo, WT., Chow, LW. Overexpression of epithelial growth factor receptor (EGFR) predicts better response to neo-adjuvant chemotherapy in patients with triple-negative breast cancer. J Transl Med. 2012 Sep 19;10 Suppl 1:S4.
[22] Tulsyan, S., Agarwal, G., Lal, P., Agrawal, S., Mittal, RD., Mittal, B. CD44 gene polymorphisms in breast cancer risk and prognosis: a study in North Indian population. PLoS One. 2013 Aug 5;8(8):e71073
[23] Jung, JA., Lim, HS. Association between CYP2D6 genotypes and the clinical outcomes of adjuvant tamoxifen for breast cancer: a meta-analysis. Pharmacogenomics. 2014 Jan;15(1):49-60.
[24] Buck, K., Hug, S., Seibold, P., Ferschke, I., Altevogt, P., Sohn, C., Schneeweiss, A., Burwinkel, B., Jager, D., Flesch-Janys, D., Chang-Claude, J., Marme, F. CD24 polymorphisms in breast cancer: impact on prognosis and risk. Breast Cancer Res Treat. 2013 Feb;137(3):927-37.
[25] Piotrowski, P., Lianeri, M., Rubis, B., Knula, H., Rybczynska, M., Grodecka-Gazdecka, S., Jagodzinski, PP. Murine double minute clone 2,309T/G and 285G/C promoter single nucleotide polymorphism as a risk factor for breast cancer: a Polish experience. Int J Biol Markers. 2012 Jul 19;27(2):e105-10.
[26] Araujo, AP., Ribeiro, R., Pinto, D., Pereira, D., Sousa, B., Mauricio, J., Lopes, C., Medeiros, R. Epidermal growth factor genetic variation, breast cancer risk, and waiting time to onset of disease. DNA Cell Biol. 2009 May;28(5):265-9.

김 정 우
2013년 상명대학교 컴퓨터과학과(학사)
2013년~현재 연세대학교 컴퓨터과학과 통합과정. 관심분야는 바이오 인포매틱스, 데이터마이닝, 텍스트마이닝

김 현 진
2010년 연세대학교 컴퓨터과학과(학사)
2016년 연세대학교 컴퓨터과학과(박사)
관심분야는 데이터 마이닝, 생물정보학

여 윤 구
2009년 연세대학교 컴퓨터과학과(학사)
2011년 연세대학교 컴퓨터과학과(석사)
2016년 연세대학교 컴퓨터과학과(박사)
2016년 5월~현재 Postdoctoral researcher, Dept. of Clinical Sciences, UTSW. 관심분야는 생물정보학, 데이터 마이닝, 빅데이터 프로세싱

신 민 철

2011년 연세대학교 컴퓨터과학과(학사) 2011년~현재 연세대학교 컴퓨터과학과 통합과정. 관심분야는 데이터베이스 시스템, 분산처리 시스템, 빅 데이터

박 상 현

1989년 서울대학교 컴퓨터공학과 졸업(학사). 1991년 서울대학교 대학원 컴퓨터공학과(공학석사). 2001년 UCLA 대학원 컴퓨터공학과(공학박사). 1991년~1996년 대우통신 연구원. 2001년~2002년 IBM T. J. Watson Research Center, Post-Doctoral Fellow. 2002년~2003년 포항공과대학교 컴퓨터공학과 조교수. 2003년~2006년 연세대학교 컴퓨터과학과 조교수. 2006년~20011 연세대학교 컴퓨터과학과 부교수 2011년~현재 연세대학교 컴퓨터과학과 교수. 관심분야는 데이터베이스, 데이터마이닝, 바이오인포매틱스, 적응적 저장장치 시스템, 플래쉬메모리, 인덱스, SSD