텍스트마이닝과 주경로 분석을 이용한 미발견 공공 지식 추론

- 췌장암 유전자-단백질 유발사슬의 경우 -

Inferring Undiscovered Public Knowledge by Using Text Mining Analysis and Main Path Analysis: The Case of the Gene-Protein 'brings_about' Chains of Pancreatic Cancer

안 혜 림 (Hyerim Ahn)*
송 민 (Min Song)**
허 고 은 (Go Eun Heo)***

초 록

본 연구에서는 췌장암의 유전자-단백질 상호작용 네트워크를 구성하고, 관련 연구에서 주요하게 언급되는 유전자-단백질의 유발관계 사슬을 파악함으로써, 췌장암의 원인을 규명하는 실증적인 연구로 이어질 수 있는 미발견 공공 지식을 제공하려 하였다. 이를 위하여 텍스트마이닝과 주경로 분석을 Swanson의 ABC 모델에 적용해 중간 개념인 B를 방향성을 가진 다단계 모델로 확장하고 가장 의미 있는 경로를 도출하였다. 본 연구의 주제가 된 췌장암의 사례처럼 시작점과 끝점조차 한정할 수 없는 미발견 공공 지식 추론에서 주경로 분석은 유용한 도구가 될 수 있을 것이다.

ABSTRACT

This study aims to infer the gene-protein 'brings_about' chains of pancreatic cancer which were referred to in the pancreatic cancer related researches by constructing the gene-protein interaction network of pancreatic cancer. The chains can help us uncover publicly unknown knowledge that would develop as empirical studies for investigating the cause of pancreatic cancer. In this study, we applied a novel approach that grafts text mining and the main path analysis into Swanson's ABC model for expanding intermediate concepts to multi-levels and extracting the most significant path. We carried out text mining analysis on the full texts of the pancreatic cancer research papers published during the last ten-year period and extracted the gene-protein entities and relations. The 'brings_about' network was established with bio relations represented by bio verbs. We also applied main path analysis to the network. We found the main direct 'brings_about' path of pancreatic cancer which includes 14 nodes and 13 arcs. 9 arcs were confirmed as the actual relations emerged on the related researches while the other 4 arcs were arisen in the network transformation process for main path analysis. We believe that our approach to combining text mining analysis with main path analysis can be a useful tool for inferring undiscovered knowledge in the situation where either a starting or an ending point is unknown.

키워드: 문헌 기반 발견, 미발견 공공 지식, 텍스트마이닝, 주경로 분석, 생물학 네트워크
Literature Based Discovery, Undiscovered Public Knowledge, Text Mining, Main Path Analysis, Bio Network

---

* 연세대학교 일반대학원 문헌정보학과 박사과정(hrahn@yonsei.ac.kr)
** 연세대학교 문헌정보학과 부교수(min.song@yonsei.ac.kr)
*** 연세대학교 일반대학원 문헌정보학과 박사과정(goeun.heo@yonsei.ac.kr)
논문접수일자 : 2015년 2월 16일 논문심사일자 : 2015년 3월 16일 게재확정일자 : 2015년 3월 19일
한국비블리아학회지, 26(1): 217-231, 2015. [http://dx.doi.org/10.14699/kbiblia.2015.26.1.217]

1. 서 론

정보 기술의 발전과 함께 폭발적으로 증가하고 있는 정보 자원을 처리하기 위해 발전한 텍스트마이닝은 방대한 양의 텍스트 데이터에서 자동적으로 유의미한 정보들을 추출할 수 있도록 하는 기술이다. 특히 Swanson의 연구(Swanson 1986a, 1986b 등)를 기반으로 문헌 기반 발견 연구가 다수 수행되어 온 생의학 영역에서는 생의학적 개념들간의 유의미한 연관성을 발견하고 문맥과 상호작용 관계를 이해하여 새로운 지식을 발견해 내기 위해 텍스트마이닝을 활용하고 있다(Blagosklonny and Pardee 2002).

췌장암은 그 원인이 밝혀지지 않았고, 다른 암에 비해 암 발생의 원인으로 작용하는 암 전 단계의 병변 역시 뚜렷하지 않아 조기 진단이 어려우며, 5년 생존율이 5% 이하로 예후가 무척 나쁜 암이다(서울대학교병원 의학정보). 이 때문에 췌장암의 발병 원인 및 그 과정을 규명하기 위한 연구가 다양하게 진행되고 있지만(He and Yuan 2014), 텍스트마이닝을 활용하여 지금까지 누적되어 온 연구 성과를 종합하고 문헌 기반의 미발견 공공 지식을 추론하려는 연구는 찾아보기 어렵다.

이에 본 연구에서는 텍스트마이닝과 주경로 분석(main path analysis)을 Swanson(1986a)의 ABC 모델에 접목하여 췌장암 관련 미발견 공공 지식을 추론하는 새로운 방법을 제안하였다. 최근 10년간 출판된 췌장암 관련 연구 논문을 대상으로 텍스트마이닝을 실시하여 방향성(directivity)을 반영한 유전자와 단백질의 방향성 상호작용 네트워크(directed interaction network)를 구성하였고, 특정 유전자 혹은 단백질을 중간 개념으로 지정하여 연결성을 찾아내는 대신 주경로 분석을 통해 가능한 모든 경로 중에서 가장 의미있는 주요 유전자-단백질 상호작용 사슬이라는 미발견 공공 지식을 추론하였다. 특히 유전자-단백질 상호작용 사슬 중 유발관계(relationship of 'brings_about'; NIH)에 집중하여 췌장암 관련 연구에서 주요하게 언급되는 유전자-단백질 유발사슬을 파악함으로써, 향후 췌장암의 발병 원인을 연구하는 단초를 제공하고자 하였다.

2. 이론적 배경

2.1 문헌 기반 발견

문헌 기반 발견은 명시적으로 연결되지 않은 문헌에 숨겨져 있었던 새로운 지식을 밝혀내는 과정으로, 미발견 공공 지식 추론이라 칭하기도 한다. 문헌 기반 발견을 처음으로 제안하였던 Swanson(1986a)은 특정 질병이나 질환인 C, 해당 질병에 대한 생리적인 조건이나 상태 또는 과정의 단어인 B, 해당 질병에 대한 치료제 또는 증상인 A간의 연결성을 문헌에서 찾아내는 ABC 모델을 개발하고 이를 기반으로 다수의 연구를 수행하였다(Swanson 1986b, 1988 등).

Swanson이 ABC 모델을 기반으로 수립한 가설이 임상의학자들에 의해 실제 유효한 것으로 검증되면서(DiGiacomo, Kremer, and Shah 1989) 많은 연구자들이 문헌 기반 발견에 관심을 갖게 되었고, 다양한 연구들이 수행되었다. 허고은과 송민(2014)은 이후 수행된 문헌 기반 발견 연구들을 개념 기반의 의미론적 접근 연

구와 술어 선정 기법 연구, 그래프 기반 연구로 분류하였다.

이중 그래프를 기반으로 하는 연구에서는 최근 A와 C 사이에 존재하는 중간 개념인 B를 한 가지로 제한하는 ABC 모델의 한계를 극복하려는 시도가 이어지고 있다. Wilkowski 등(2011)은 처음으로 B개념을 단일 개념이 아닌 여러 종류의 개념 집합으로 확장하여 구체화하였고, Cameron 등(2013)은 AnC 모델을 제안하여 B개념의 다수준 모델에 대한 가능성을 증명하였다. 허고은과 송민(2014)은 동시출현 개체와 동사 추출을 통해 B개념을 다수준으로 확장하였다.

본 연구에서는 B개념의 다수준 모델에 반영되는 동사의 추출 과정에서 방향성을 반영하여 미발견 공공 지식 추론의 정확성을 높이고, 주경로 분석을 통해 가장 의미 있는 경로를 검출하려 하였다. 생물학 네트워크(bio network)에서 방향성은 생물학적 프로세스의 역학을 이해하고 네트워크 분석을 통한 예측을 강화해준다는 측면에서 중요한 의미를 가진다(Vinayagam et al. 2011). 많은 상호작용을 간결하게 나타낼 수 있는 유전자 네트워크(Selga et al. 2009)에서는 방향성을 가지는 조절 네트워크(Regulatory Network; Gustafsson, Hornquist, and Lombardi 2005 등)와 전이 네트워크(Transfer Network; Popa et al. 2011 등) 등이 분화되어 발전하였으며, 상호작용 네트워크에 방향성을 더하려는 노력도 계속되고 있다(Mattiazzi et al. 2010; Selga et al. 2009 등). 특히, 본 연구에서는 Natarajan과 동료들의 연구(2005)에서 가능성이 확인된 바와 같이 유전자와 단백질을 동시에 표현하는 방향성 유전자-단백질 상호작용 네트워크를 구축하고 분석하여 미발견 공공 지식을 추론하고자 하였다.

2.2 주경로 분석

주경로 분석은 인용 네트워크에서 가장 의미 있는 경로를 추적할 수 있게 하는 네트워크 분석 방법으로, 일반적으로 특정 학문 분야의 발전 궤적을 추적하는 데 활용된다(Liu and Lu 2011).

주경로 분석에서는 내향 연결정도(indegree)가 0인 모든 노드를 경로의 시작점인 소스(source), 외향 연결정도(outdegree)가 0인 모든 노드를 경로의 끝점인 싱크(sink)라 가정하고 소스와 싱크 사이의 모든 노드와 아크에 관해 횡단 가중치(traversal weight)를 계산한 뒤 횡단 가중치가 가장 높은 경로를 주 경로를 결정한다(de Nooy, Mrvar, and Batagelj 2005). 횡단 가중치란 소스와 싱크 사이의 전체 경로에 대해 해당 노드나 아크를 지나가는 경로의 비율이다. 즉, 인용 네트워크를 과학적 지식이나 정보가 전달되는 전파 통로라 하면 가장 많은 줄기를 포함하고 있는 경로가 가장 중요한 경로가 되는 것이다.

본 연구에서는 다음과 같은 두 가지 이유에 따라 유전자-단백질 상호작용 네트워크를 주경로 분석으로 분석하였다. 첫째, 서론에서 언급한 바와 같이 췌장암은 그 원인이나 병변이 뚜렷하지 않으므로 특정 유전자나 단백질을 소스나 싱크로 특정하여 경로를 추론할 수 없다. 따라서 네트워크 내에서 가능한 모든 (즉, 내향 연결정도가 0이거나 외향 연결정도가 0인) 소스나 싱크에 대해 모든 경로를 반영하는 주경로 분석이 대안이 될 수 있다. 둘째, 인용 네트워크

에서의 인용의 누적이 해당 학문 영역에서의 영향력을 보여주듯(Yan, Ding, and Sugimoto 2011) 문헌에 기반한 상호작용 네트워크에서 특정 상호작용의 출현 빈도가 해당 상호작용의 중요도를 반영하고 있다고 가정하면, 주경로 분석을 통해 주요 상호작용 사슬을 확인할 수 있다.

3. 연구 방법

본 연구에서는 췌장암의 주요 유전자-단백질 상호작용 사슬이라는 미발견 공공 지식을 추론하기 위하여 (1) 관련 논문을 수집하고, (2) 수집한 논문 전문을 전처리한 후, (3) 개체를 추출하고, (4) 추출된 개체간의 관계성을 추출한 뒤, (5) 개체와 개체간의 관계성을 종합하여 상호작용 네트워크를 구축하였다. 이에 대한 상세 설명은 다음에서 순서대로 기술한다.

3.1 데이터 수집

데이터 수집은 PMC Central에서 이루어졌다. 2014년 10월 기준으로 최근 10년 이내 출판 논문 중 제목에 "pancreatic cancer"가 포함되어 있는 논문을 검색하였고, 이 중 전문이 제공되는 1,032건을 다운로드하여 분석에 활용하였다.

3.2 전처리

수집한 논문 전문의 전처리는 Stanford CoreNLP를 기반으로 일부 코드를 수정하여 진행하였고 UC Denver biolemmatizer, Lucene analyzer 등의 텍스트 처리 관련 자바 오픈 소스도 활용하였다. 각 논문의 전문은 문장 단위로 분할한 뒤 불용어(stopword, 523개 적용)를 삭제하였고, 품사 태깅 후 표제어 복원(lemmatization)을 수행하였다.

3.3 개체 추출

개체 추출(named entity recognition)은 사전을 기반으로 하였다. 먼저 유전자와 단백질의 추출을 위해 각각의 동의어 사전(synonym dictionary)을 작성하고 개체 추출에 적용하였다. 두 사전의 개요는 〈표 1〉과 같으며, 호모사피엔스에 해당하는 유전자와 단백질 항목만 적용하였다.

사전 검색은 최대 8개 단어로 구성된 구(phrase)에 대해 실시하였는데 해당 구에 품사 태깅이 되어 있으면 품사가 명사일 때만 사전

〈표 1〉 유전자-단백질 동의어 사전 개요

	유전자 동의어 사전	단백질 동의어 사전
출처	Gene database http://www.ncbi.nlm.nih.gov/gene/	UniProt http://www.uniprot.org/
포함 항목	synonyms, symbol from nomenclature authority, full name from nomenclature authority, other designations 포함	recommended name (full/short), alternative name (full/short) 포함
종류	47,779개	19,901개

을 검색하였다. 유사도 계산에는 TFIDF(Term Frequency Inverse Document Frequency) 가중치의 일종인 Soft TFIDF 알고리즘을 사용하였다(SecondString Project). 한계치는 0.99로 두고, 검색 결과가 중복 존재할 경우에는 점수가 더 높은 쪽을 선택하였다.

3.4 관계 추출

개체 간의 관계 추출은 387개 생물학 동사(bio verb)를 기반으로 진행하였으며, 다음과 같은 단순 규칙을 적용하였다.

- 동사는 개체 사이에 위치
- 부정문이 아닐 때에만 관계 저장
- 품사 태그를 통해 수동태 확인

또한 하나의 동사를 중심으로 좌우에 출현하는 모든 주체(subject) 개체와 객체(object) 개체에 대해 관계를 저장하였다.

3.5 네트워크 구성

본 연구에서는 네트워크의 분석에 네트워크 분석 도구인 Pajek을 활용하였다. Pajek에서는 네트워크를 확장자 'net' 형식의 파일로 저장하는데, 먼저 노드를 나열하고 이후 노드 사이에 존재하는 아크 또는 엣지만을 나열하여 밀도가 낮은 대용량 네트워크의 분석에 적합하다. 본 연구에서는 앞선 단계에서 추출된 개체를 노드로, 개체간의 관계를 아크로 나열하여 'net' 형식의 파일로 저장하고 Pajek에서 네트워크 분석을 실시하였다.

4. 결과 분석

4.1 개체/관계 추출 결과 후처리

유전자/단백질 개체 및 관계 추출 결과 파싱 중 에러가 발생한 43건을 제외하고 총 989건의 논문에 대해 중복을 제외한 10,068개의 관계가 추출되었다. 관계에 포함된 개체의 수는 1,769개였다.

개체 및 관계 추출 결과 검출빈도 상위에 오른 개체들은 〈표 2〉와 같이 일반 명사(예: cell, protein, type 등)인 경우가 많았다. 또한 전체 개체 목록을 확인한 결과 대문자 알파벳 약어 뒤 숫자가 따라 나오는 유전자/단백질 명명법(예. RPL12의 동의어 L12)과 달리 대문자 알파벳 약어보다 숫자가 먼저 출현하는 구(예. 12L)도 다수 개체로 추출된 것을 확인할 수 있었다. 이에 다음과 같은 규칙을 적용하여 개체 및 관계 추출에서 추출된 개체를 후처리(post-processing)하였다.

- 소문자만으로 표기된 한 단어짜리 일반명사는 개체가 아닌 일반명사로 간주한다.
- 대문자 알파벳 약어보다 숫자가 먼저 출현하는 구는 개체가 아니라고 간주한다.

후처리 결과 개체의 수는 1,341개, 이 개체들이 포함된 관계의 수는 중복을 제외하고 5,256개로 감소하였다. 검출 빈도에 따른 상위 10개의 개체와 관계를 후처리 전후로 비교하면 각각 〈표 2〉, 〈표 3〉과 같다. 〈표 2〉와 〈표 3〉에서 검색어 항목은 논문 원문에 출현하여 사전 검색에 사용한 구를 의미하며, 개체명 항목은 해당 구로 검색된 개체의 대표 명칭을 의미한다.

〈표 2〉 후처리 전후 검출 빈도 상위 10개 개체

후처리 전		후처리 후	
검색어	개체명	검색어	개체명
cell	CELP	EGFR	EGFR
protein	S52A1_HUMAN	MIA	MIA
type	SGCG	Kras	KRAS
MIA	MIA	tumor suppressor	TCHP_HUMAN
EGFR	EGFR	STAT3	STAT3
STAT3	STAT3	p53	TP53
test	PRSS21	Bcl-2	BCL2
Kras	KRAS	MS	METH_HUMAN
hr	HR	CD4	CD4
p53	TP53	A4	IGKV1D-27

〈표 3〉 후처리 전후 검출 빈도 상위 10개 relation

후처리 전			후처리 후		
주체	동사	객체	주체	동사	객체
CELP	supplement	C2	ZEB2	contain	CTBP2_HUMAN
CELP	culture	C2	ZEB2	contain	HDAC2
CELP	culture	AIRN	TNF	inhibit	PGH2_HUMAN
CELP	use	KIT	TXK	include	EGFR
S52A1_HUMAN	use	KIT	CTNNBIP1	compare	IBP2_HUMAN
S52A1_HUMAN	use	CI116_HUMAN	GLMN_HUMAN	compare	GLMN_HUMAN
S52A1_HUMAN	use	ALB	GLMN_HUMAN	produce	GLMN_HUMAN
S52A1_HUMAN	use	ALBU_HUMAN	IFNA1	induce	STAT1
CELP	show	S52A1_HUMAN	IGKV1D-27	peak	IBP2_HUMAN
CELP	supplement	AIRN	IGKV1D-27	peak	HSPG2

4.2 생물학 관계의 분류

일반적인 네트워크 분석에서는 한 종류의 아크 혹은 엣지를 가정한다. 네트워크에서 아크 혹은 엣지는 노드 간의 관계를 나타내는데, 가중치 혹은 화살표의 방향으로 그 차이를 반영하는 경우도 있지만 일반적으로는 한 네트워크에서 유사한 관계만을 나타낸다. 그러나 본 프로젝트에서 관계 추출을 위해 사용한 생물학 동사들은 그 의미가 다양하므로, 관계 추출에서 검출된 모든 관계를 하나의 네트워크로 구성해서는 분석이 쉽지 않다.

따라서 본 연구에서는 분석을 위해 의미가 유사한 동사를 그룹으로 묶고 이런 동사들이 포함되어 있는 관계를 한 종류의 관계로 가정하여 네트워크를 구성하기로 하였다. 생물학 관계(bio relation)의 분류에는 UMLS의 관계 구분을 적용하였다(NIH). 단, UMLS의 관계 구분

중 'affects'는 긍정적인 영향을 미치는 관계인 'positive'와 부정적인 영향을 미치는 관계인 'negative'로 추가 분류하였다. 관계 추출에 적용한 387개의 생물학 동사를 분류한 결과는 다음과 같았다.

- functionally_related_to - affects: 65개 (positive: 32개, negative: 33개)
- functionally_related_to - brings_about: 17개
- functionally_related_to - indicates: 11개
- functionally_related_to - performs: 294개

본 연구의 분석 대상이 되는 유발 관계에 포함되는 생물학 동사 17개는 〈표 4〉에 정리하였다.

〈표 4〉 bring_about으로 분류된 17개 생물학 동사

appear	emerge	originate
arise	generate	produce
begin	induce	result
cascade	lead	start
come	mediate	trigger
create	occur	

4.3 췌장암의 유전자-단백질 유발관계 네트워크 구성

후처리를 마친 5,256개의 관계 중 〈표 4〉에 정리한 유발관계의 생물학 동사를 포함하고 있는 관계는 531개였다. 531개의 관계를 반영하여 구성한 유발관계 네트워크의 개요는 〈표 5〉와 같다. 〈그림 1〉은 유전자-단백질 유발관계 네트워크를 네트워크 시각화 도구인 Gephi에서 시각화한 것이다.

〈표 5〉 췌장암의 gene-protein 유발관계 네트워크 개요

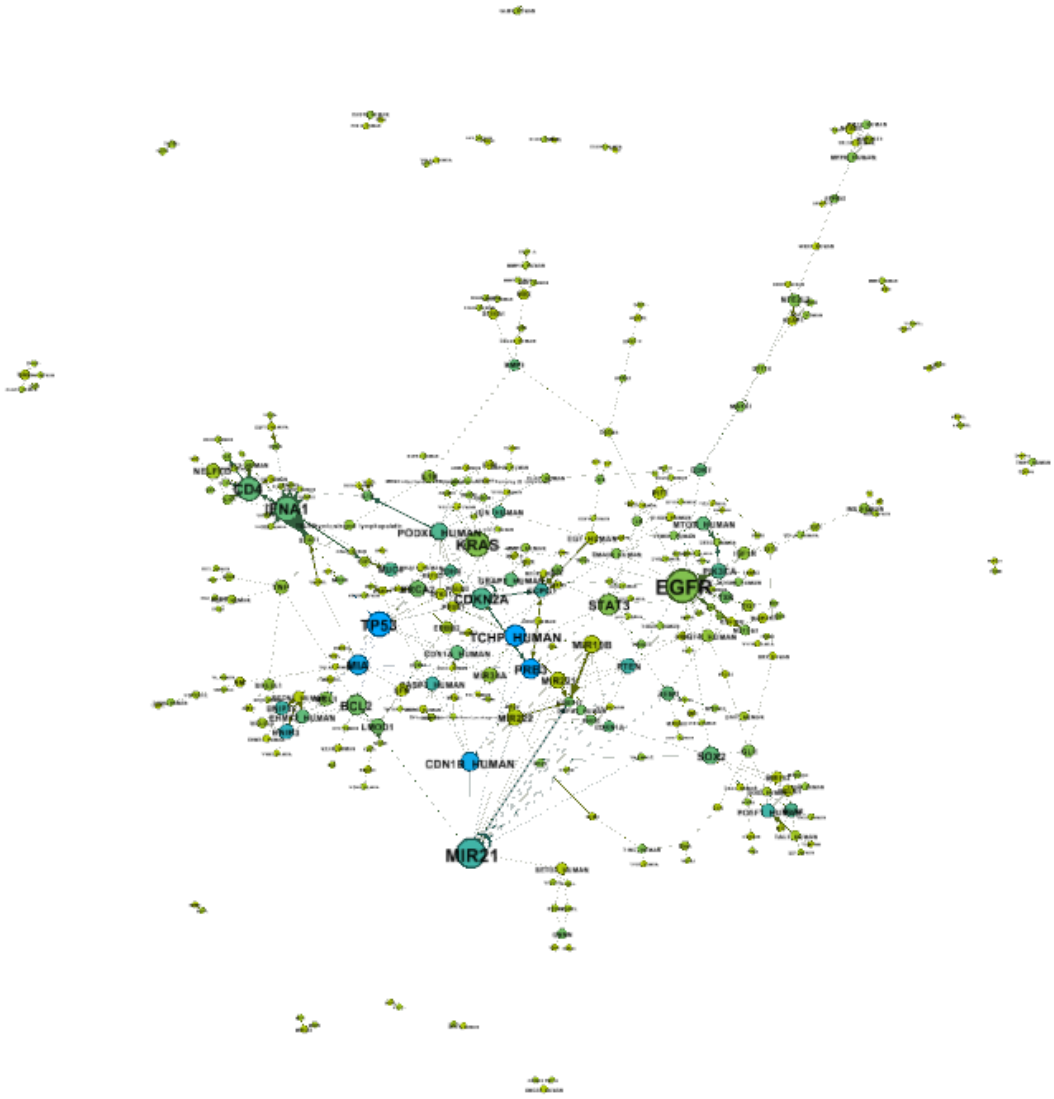
노드 수	332	밀도 (loop 허용)	0.0048
아크 수	531	평균 연결정도	3.1988

췌장암의 유전자-단백질 유발관계 네트워크는 총 20개의 약한 컴포넌트(weak component)로 구성되어 있으며 이 중 가장 큰 약한 컴포넌트에 288개의 노드가 속해 있다. 연결정도(degree) 기준 상위 10개 노드는 〈표 6〉과 같다.

4.4 주경로 분석

4.3절에 구성한 췌장암의 유전자-단백질 유발관계 네트워크는 사이클(cycle)이 존재하는 순환 네트워크(cyclic network)이므로 네트워크 내의 사이클을 제거하여 비순환 네트워크(acyclic network)로 변환한 뒤 주경로 분석을 실시하였다. 순환 네트워크를 비순환 네트워크로 변환하는 과정에서 네트워크의 노드 및 아크의 변화를 〈표 7〉에 정리하였다. 변환 과정에서 제거되는 노드의 목록은 〈표 8〉과 같다.

이렇게 얻은 비순환 유전자-단백질 유발관계 네트워크에서 주경로 분석을 실시한 결과 〈그림 2〉와 같은 주경로가 산출되었다. BHA15_HUMAN와 PALLD_HUMAN, RNF123을 소스로, MIR183와 MIR203A, CRTC2_HUMAN, LTMD1_HUMAN을 싱크로 하는 경로이다. 〈그림 2〉에서 각 아크 위에 적힌 숫자는 해당 아크의 횡단 가중치이다.



〈그림 1〉 췌장암의 gene-protein 유발관계 네트워크

〈표 6〉 유전자-단백질 유발관계 네트워크의 연결정도 기준 상위 10개 노드

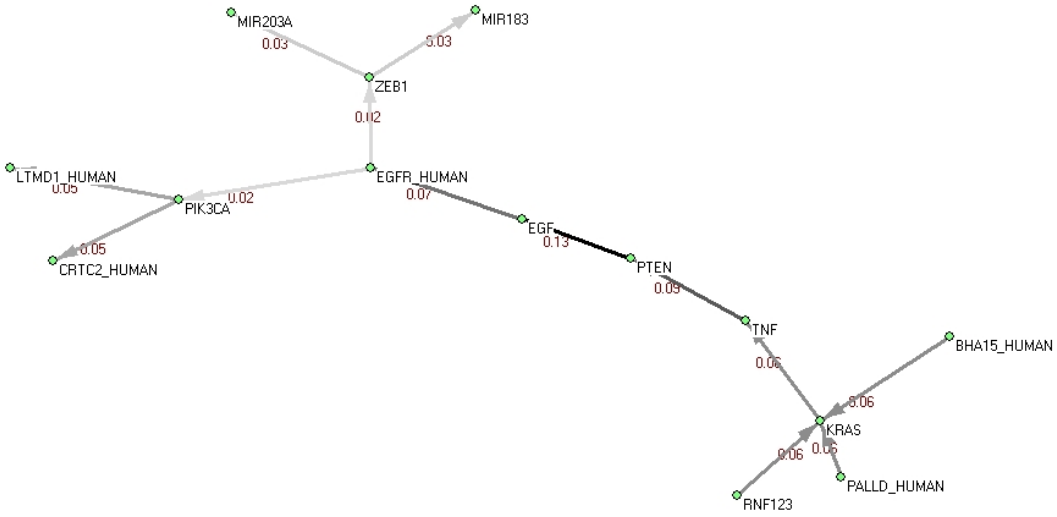
노드	연결정도	노드	연결정도
IFNA1	25	KRAS	16
EGFR	25	CDKN2A	15
MIR21	20	PRB3	14
CD4	19	TCHP_HUMAN	13
TP53	16	CCPG1	13

〈표 7〉 네트워크 변환 과정에서 네트워크의 변화

	순환 네트워크	비순환 네트워크
노드 수	332	315
아크 수	531	435

〈표 8〉 네트워크 변환 과정에서 삭제된 node 목록

노드	연결정도	노드	연결정도
BCL2	12	CDN1B_HUMAN	11
LMOD1	8	GRAP2_HUMAN	7
MIR21	20	EGFR	25
TP53	16	ERBB2	6
TCHP_HUMAN	13	FGFR1	2
IFNA1	25	IL1B	6
IL13_HUMAN	5	TSLPthymic stromal lymphopoietin	4
IL5	5	MIR34A	8
CDN1A_HUMAN	8		



〈그림 2〉 gene-protein 유발관계 네트워크의 주경로

유전자-단백질 유발관계 네트워크의 주경로에는 14개의 노드와 13개의 아크가 포함되어 있다. 주경로의 아크에 해당되는 13개의 유발관계 중 개체/관계 추출에서 검출된 관계 9개를 〈표 9〉에 정리하였다. 〈표 9〉에 포함되지 않은 KRAS -〉 TNF, TNF -〉 PTEN, PTEN -〉 EGF, EGFR_HUMAN -〉 ZEB1은 순환 네트워크를 비순환 네트워크로 연결하는 과정에

〈표 9〉 개체/관계 추출에서 검출된 관계

*주체	*객체	**PMC ID	년도	발견
RNF123 (KPC)	KRAS	3990331	2013	**KPC**, BRCA-KPC, and Mist1(KRasG12D/+) mice develop spontaneously arising **KRAS** mutant (TP53 and BRCA2 mutant, respectively) pancreatic tumors
PALLD_HUMAN (palladin)	KRAS	3919106	2014	Upregulation of alpha-SMA and **palladin** in fibroblasts induced by the cancer cells containing activated **KRAS**, the de novo CAFs can promote pancreatic tumorigenesis and progression through their morphologically forming invadopodia-like cellular protrusions which could secrete invadopodia proteins and proteolytic enzymes such as ADAM22, aminopeptidases, and cathepsins D and B used for decomposing the ECM and create an optimum premetastatic niche
BHA15_HUMAN (Mist1)	KRAS	3990331	2013	KPC, BRCA-KPC, and **Mist1**(KRasG12D/+) mice develop spontaneously arising **KRAS** mutant (TP53 and BRCA2 mutant, respectively) pancreatic tumors
EGF (epidermal growth factor)	EGFR_HUMAN (epidermal growth factor receptor)	3822297	2013	To overcome this resistance mechanism, downstream target of **epidermal growth factor** pathway, MEK inhibition was tested that induced apoptosis in **epidermal growth factor receptor** tyrosine kinase inhibitor-resistant lung cancer cells.
EGFR_HUMAN (epidermal growth factor receptor)	PIK3CA (PI3K)	4101748	2014	**The epidermal growth factor receptor (EGFR)** initiates a signal transduction cascade that leads to modulation of cellular functions through activation of a number of pathways, including the phosphatidylinositol 3-kinase (**PI3K**) and the mitogen-activated protein (MAP) kinase pathways.
PIK3CA (PI3K)	LTMD1_HUMAN (HCCR-1)	2880295	2010	EGF-induced **HCCR-1** over-expression is mediated by **PI3K**/AKT/mTOR signaling which plays a pivotal role in pancreatic tumor progression, suggesting that HCCR-1 could be a potential target for cancer therapeutics
PIK3CA (PI3K)	CRTC2_HUMAN (TORC2)	3717802	2012	NVP-BEZ235 is an orally bioavailable imidazoquinoline derivative that binds to the ATP-binding clefts of the class I **PI3K** and mTOR kinase, leading to inhibition of PI3K, as well as attenuation of TORC1 and **TORC2** activity
ZEB1	MIR203A (miR-203)	3837326	2010	**ZEB1** can inhibit expression of miR-200c, **miR-203** and miR-183, which in turn were shown to target BMI1, a known factor for stem cell renewal.
ZEB1	MIR183 (miR-183)	3837326	2010	**ZEB1** can inhibit expression of miR-200c, miR-203 and **miR-183**, which in turn were shown to target BMI1, a known factor for stem cell renewal.

* 괄호 안은 검색시 매칭되었던 유의어
** 부록에 PMC ID 순으로 저자 및 제목, 출처를 밝힘

서 연결 관계가 생성되었다.

주경로에 포함되어 있는 14개 노드 사이의 직접 외향 연결 관계를 사이클을 제거하기 전의 유전자-단백질 유발관계 네트워크에서 확인한 결과는 〈표 10〉과 같다. 주경로를 추출하는 과정에서 배제된 직접 연결이 존재하지 않으므로, 검출된 주경로가 우회 경로가 아닌 것을 알 수 있다.

〈표 10〉 주경로에 포함된 노드 사이의 직접 외향 연결 관계

	BHA15_HUMAN	CRTC2_HUMAN	EGF	EGFR_HUMAN	KRAS	LTMD1_HUMAN	MIR183	MIR203A	PALLD_HUMAN	PIK3CA	PTEN	RNF123	TNF	ZEB1
BHA15_HUMAN	-	×	×	×	○	×	×	×	×	×	×	×	×	×
CRTC2_HUMAN	×	-	×	×	×	×	×	×	×	×	×	×	×	×
EGF	×	×	-	○	×	×	×	×	×	×	×	×	×	×
EGFR_HUMAN	×	×	×	-	×	×	×	×	×	○	×	×	×	×
KRAS	×	×	×	×	-	×	×	×	×	×	×	×	×	×
LTMD1_HUMAN	×	×	×	×	×	-	×	×	×	×	×	×	×	×
MIR183	×	×	×	×	×	×	-	×	×	×	×	×	×	×
MIR203A	×	×	×	×	×	×	×	-	×	×	×	×	×	×
PALLD_HUMAN	×	×	×	×	○	×	×	×	-	×	×	×	×	×
PIK3CA	×	○	×	×	×	○	×	×	×	-	×	×	×	×
PTEN	×	×	×	×	×	×	×	×	×	×	-	×	×	×
RNF123	×	×	×	×	○	×	×	×	×	×	×	-	×	×
TNF	×	×	×	×	×	×	×	×	×	×	×	×	-	×
ZEB1	×	×	×	×	×	×	○	○	×	×	×	×	×	-

*두 노드 사이에 직접 외향 연결이 존재하면 ○, 존재하지 않으면 ×로 표기

5. 결론 및 제언

생의학 영역에 적용되는 바이오 텍스트마이닝은 생의학적 개념들간의 유의미한 연관성을 자동적으로 발견하고 그 문맥과 상호작용 관계를 이해하여 새로운 지식을 발견해 내는 것을 목적으로 한다. 본 연구에서는 췌장암의 발병 원인과 관련된 문헌 기반 미발견 공공 지식을 추론하기 위하여 텍스트마이닝과 주경로 분석을 Swanson의 ABC 모델에 접목하였다.

텍스트마이닝과 주경로 분석 모두 ABC 모델의 중간 개념인 B를 다단계 모델로 확장하고 가장 의미 있는 개념을 도출하는 데 적용되었다. 최근 10년간 출판된 췌장암 관련 연구 논문을 대상으로 텍스트마이닝을 실시하여 방향성을 반영한 유전자-단백질의 상호작용 네트워크를 구성하였으며, 생물학 동사를 기준으로 생물학 관계를 분류하여 유발관계 네트워크를 분리하였고, 주경로 분석을 통해 네트워크 내에서 가능한 모든 경로 중 주요 유발관계 사슬을 파악하였다.

그 결과 14개의 노드와 13개의 아크가 포함된 췌장암의 유전자-단백질 주요 유발관계 사슬이 검출되었다. 이 중 9개 아크는 실제 관련 연구에 출현한 관계를 반영하고 있었고, 나머지 4개 아크는 유발관계 네트워크를 순환 네트워크에서 비순환 네트워크로 변환하면서 생성된 것으로 확인되었다. 주경로에 포함된 14개 노드 사이에는 주경로에 포함된 연결 관계 외에 다른 연결 관계가 존재하지 않아 검출된 주경로가 우회경로가 아닌 것도 알 수 있었다.

본 연구에서 검출된 주경로는 췌장암 관련

연구에서 주요하게 언급되어 온 유전자-단백질 유전사슬로, 새로운 방법론을 통해 이전까지 발견되지 않았던 연결성을 확인할 수 있었다. 문헌에서의 언급 정도가 실제 상호작용의 중요도와 비례하는지는 실증적인 연구를 통해 검증되어야 한다. 그러나 본 연구의 주제가 된 췌장암의 사례처럼 시작점(소스)와 끝점(싱크)조차 한정할 수 없는 미발견 공공 지식의 추론에서 주경로 분석은 유용한 도구가 될 수 있을 것이다.

참 고 문 헌

서울대학교병원 의학정보. 췌장암 [online]. [cited 2015.2.5]. 〈http://terms.naver.com/entry.nhn?docId=926898&mobile&cid=51007&categoryId=51007#TABLE_OF_CONTENT1〉.

허고은, 송민. 2014. 텍스트 마이닝 기반의 그래프 모델을 이용한 미발견 공공 지식 추론. 『정보관리학회지』, 31(1): 231-250.

Blagosklonny, M. V. and A. B. Pardee. 2002. "Unearthing the gems." *Nature*, 416(6879): 373-373.

Cameron, D., O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, A. P. Sheth, and T. C. Rindflesch. 2013. "A graph-based recovery and decomposition of swanson's hypothesis using semantic predications." *Journal of Biomedical Informatics*, 46(2): 238-251.

De Nooy, W., A. Mrvar, and V. Batagelj. 2005. *Exploratory Social Network Analysis with Pajek*. Revised and Expanded Second Edition. New York, USA: Cambridge University Press.

DiGiacomo, R. A., J. M. Kremer, and D. M. Shah. 1989. "Fish oil dietary supplementation in patients with Raynaud's phenomenon: A doubleblind, controlled, prospective study." *American Journal of Medicine*, 8: 158-164.

Gustafsson, M., M. Hornquist, and A. Lombardi. 2005. "Constructing and analyzing a large-scale gene-to-gene regulatory network Lasso-constrained inference and biological validation." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3): 254-261.

Liu, J. S., and L. Y. Y. Lu. 2011. "An Integrated Approach for Main Path Analysis: Development of the Hirsch Index as an Example." *Journal of the American Society for Information Science and Technology*, 63(3): 528-542.

Mattiazzi, M., T. Curk, I. Krizaj, B. Zupan, and U. Petrovic. 2010. "Inference of the Molecular Mechanism of Action from Genetic Interaction and Gene Expression Data." *Omics-A Journal Of Integrative Biology*, 14(4): 357-367.

Natarajan, J., D. Berrar, W. Dubitzky, C. Hack, Y. Zhang, C. Desesa, J. R. Van Brocklyn, and E. G. Bremer. 2006. "Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line." *BMC bioinformatics*, 7: 373.

NIH. Current Relations in the Semantic Network [online]. [cited 2015.2.15]. 〈http://www.nlm.nih.gov/research/umls/META3_current_relations.html〉.

Popa, O., E. Hazkani-Covo, G. Landan, W. Martin, and T. Dagan. 2011. "Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes." *Genome Research*, 21(4): 599-609.

SecondString Project. Class SoftTFIDF [online]. [cited 2015.2.15]. 〈http://secondstring.sourceforge.net/javadoc/com/wcohen/ss/SoftTFIDF.html〉.

Selga, E., C. Oleaga, S. Ramirez, M. C. de Almagro, V. Noe, and C. J. Ciudad. 2009. "Networking of differentially expressed genes in human cancer cells resistant to methotrexate." *Genome Medicine*, 1: 83.

Swanson, D. R. 1986a. "Undiscovered public knowledge." *The Library Quarterly*, 56(2): 103-118.

Swanson, D. R. 1986b. "Fish oil, Raynaud's syndrome, and undiscovered public knowledge." *Perspectives in biology and medicine*, 30(1): 7-18.

Swanson, D. R. 1988. "Migraine and magnesium: eleven neglected connections." *Perspectives in biology and medicine*, 31(4): 526-557.

Vinayagam, A., U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, AM. A. ndrade-navarro, and E. E. Wanker. 2011. "A directed protein interaction network for investigating intracellular signal transduction." *Science signaling*, 4(189): rs8.

Xiang-Yi He and Yao-Zong Yuan. 2014. "Advances in pancreatic cancer research: Moving towards early detection." *World J Gastroenterol*, 20(32): 11241-11248.

Yan, E., Y. Ding, and C. R. Sugimoto. 2011. "P-Rank: An Indicator Measuring Prestige in Heterogeneous Scholarly Networks." *Journal of the American Society for Information Science and Technology*, 62(3): 467-477.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Heo, Go Eun and Song Min. 2014. "Inferring Undiscovered Public Knowledge by Using Text Mining-driven Graph Model." *Journal of the Korean Society for information Management*,

31(1): 231-250.

Seoul National University Hospital Medical Information. Pancreatic Cancer [online]. [cited 2015. 2.5].
〈http://terms.naver.com/entry.nhn?docId=926898&mobile&cid=51007&categoryId=51007#TABLE_OF_CONTENT1〉.

[부록] 주요 분석 논문 목록

PMC ID: 2880295. Xu, Z., Zhang, Y., Jiang, J., Yang, Y., Shi, R., Hao, B., Zhang, Z., Huang, Z., Kim, J., and Zhang, G. 2010. Epidermal growth factor induces HCCR expression via PI3K/Akt/mTOR signaling in PANC-1 pancreatic cancer cells. BMC Cancer, 10(1): 161.

PMC ID: 3717802. Venkannagari, S., Fiskus, W., Peth, K., Atadja, P., Hidalgo, M., Maitra, A., and Bhalla, K. 2012. Superior efficacy of co-treatment with dual PI3K/mTOR inhibitor NVP-BEZ235 and pan-histone deacetylase inhibitor against human pancreatic cancer. Oncotarget, 3(11): 1416-1427.

PMC ID: 3822297. Hamada, S., Masamune, A., and Shimosegawa, T. 2013. Novel therapeutic strategies targeting tumor-stromal interactions in pancreatic cancer. Frontiers in physiology, 4: 331.

PMC ID: 3837326. Wellner, U., Brabletz, T., and Keck, T. 2010. ZEB1 in Pancreatic Cancer. Cancers, 2(3): 1617-28.

PMC ID: 3919106. Xiao, Z., Luo, G., Liu, C., Wu, C., Liu, L., Liu, Z., Ni, Q., Long, J., Yu, X., and Takao, S. 2014. Molecular Mechanism Underlying Lymphatic Metastasis in Pancreatic Cancer. Journal of biomedicine and biotechnology, 2014.

PMC ID: 3990331. Mace, T. A., Pitaressi, J., Shakya, R., Frankel, W., Eubank, T., Bekaii-Saab, T., Bloomston, M., Phelps, M., Ludwig, T., Ostrowski, M., and Lesinski, G. B. 2013. Genetically engineered murine pancreatic cancer models approximate immunophenotypic properties of human patients. Journal for Immunotherapy of Cancer, 1(Suppl 1): P163.

PMC ID: 4101748. Cardin, D. B., Goff, L., Li, C., Shyr, Y., Winkler, C., Devore, R., Schlabach, L., Holloway, M., Mcclanahan, P., Meyer, K., Grigorieva, J., Berlin, J., and Chan, E. 2014. Phase II trial of sorafenib and erlotinib in advanced pancreatic cancer. Cancer Medicine, 3(3): 572-579.