텍스트 마이닝을 활용한 바이오 네트워크 구축의 동향

연세대학교 ▌김현진
가천대학교 ▌윤영미*

1. 서 론

텍스트 마이닝을 통한 생물 의학 연구(Biomedical text mining)는 1986년 스완슨(Swanson) 박사의 레이노 증후군(Raynaud's syndrome)에 관한 연구 [1] 이후로 많은 발전을 거듭해왔다. 그 동안 텍스트 마이닝은 생물 의학 연구 분야에서 주로 공개된 문헌들로부터 아직 공개되지 않은 정보를 얻어내는데 사용되었다. 그 중심에 있는 개념이 바로 스완슨 박스의 ABC 모델이다(그림 1).

ABC 모델은 A와 B가 관련이 있고 B와 C가 관련이 있으면 A와 C도 관련이 있을 수 있다는 간단한 모델이다. 이를 이용하여 실제적으로 생물 의학적 실험을 하지 않아도, 단순히 문헌들을 마이닝하는 것만으로도 새로운 정보를 얻을 수 있다는 것이 밝혀지면서 [2] 많은 연구자들이 생물 의학적 텍스트 마이닝 연구에 참여하기 시작하였다.

생물 의학적 텍스트 마이닝 연구는 주로 새로운 유전자의 기능(Gene function)이나 새로운 질병 관련 의약품 등 생물 의학적 개체 간의 연결 가능성을 찾는 것에 집중 되었다. 하지만 상대적으로 연결 가능성을 넘어, 해당 연결들에 가중치를 부여하고 이를 바탕으로 네트워크를 구축하는 연구는 아직 많이 이루어지지 않은 상태이다. 생물학적 프로세스(Biological process)는 하나의 바이오 개체에 의해 이루어지는 것이 아니라 여러 개의 바이오 개체에 의해 조직적이고 연쇄적으로 이루어지기 때문에 바이오 네트워크는 그러한 생물학적 프로세스를 파악하는데 매우 중요한 역할을 할 수 있다. 그 동안의 바이오 네트워크는 주로 임상적으로 도출된 데이터들을 이용하여 구축되었는데 텍스트 데이터로도 바이오 네트워크를 구축할 수 있고, 기존 바이오 네트워크를 보강함으로써 새로운 생물학적 프로세스의 메커니즘을 밝혀낼 수도 있다.

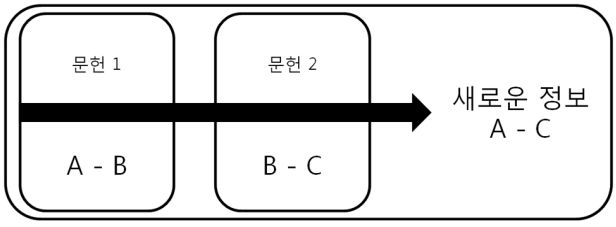

그림 1 스완슨 박사의 ABC 모델

이에 따라 본 고에서는 텍스트 마이닝을 활용하여 바이오 네트워크를 구축하는 것이 중요한 연구 주제라고 판단하여 생물 의학적 텍스트 마이닝과 바이오 네트워크에 대해 설명하고, 텍스트 마이닝을 활용하여 바이오 네트워크를 구축한 최신 연구 동향들을 소개하고자 한다.

2. 생물 의학적 텍스트 마이닝

기본적인 생물 의학적 텍스트 마이닝이라고 하면 역시 서론 부분에서 언급한 스완슨 박사의 ABC 모델이다. 그 이후에 나온 다른 방법들도 결국 ABC 모델의 확장된 형태이거나 ABC 모델을 응용한 것들이 대부분이다. ABC 모델의 궁극적인 목적은 텍스트 데이터를 이용하여 연결 가능성이 있는 새로운 생물 의학적 개체간의 연결을 찾는 것이다. 이를 위해서는 세가지 단계가 필요하다(그림 2). 먼저 바이오 개체 이름과 텍스트 데이터를 확보해야 한다. 그리고 확보한 텍스트 데이터에서 바이오 개체 이름을 이용하여 관계를 추출해야 하고, 마지막으로 추출해낸 관계들을 이용하여 기존에 없던 새로운 관계를 찾아내야 한다. 이는 이 세가지 각각의 단계에서 연구할만한 소지가 존재한다는 뜻이다. 명확한 바이오 개체 이름과 확실하게 검증된 결과들을 담고 있는 텍스트 데이터를 사용

* 종신회원
† 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2012R1A2A1A01010775).

한다면 최종적으로 찾아낸 관계들의 정확도가 높아질 수 밖에 없다. 또한 텍스트 데이터에서 기존 관계들을 어떻게 추출했는가와, 추출한 관계들에 어떤 방법을 적용하여 새로운 관계를 찾아냈는가도 결과에 영향을 미친다. 생물 의학적 텍스트 마이닝 방법들은 대부분 마지막 부분인 '추출한 관계들에 어떤 방법을 적용하여 새로운 관계를 찾아냈는가'에 초점을 맞춘 연구들이다.

텍스트 데이터에서 어떤 생물 의학적 관계를 추출하기 위해서는 바이오 개체 이름과 텍스트 데이터가 필요하다. 바이오 개체로는 질병 명, 유전자 명, 단백질 명, 의약품 명, 증상 이름, miRNA 이름 등이 있고, 텍스트 데이터로는 생물 의학 관련 연구 논문들이 주로 쓰인다. 현재 생물 의학 연구자가 사용할 수 있는 바이오 개체 데이터 및 텍스트 데이터는 여러 가지가 있다 (표 1).

텍스트 데이터와 바이오 개체 이름 리스트를 가지고 있다면 텍스트 데이터에서 바이오 개체 간에 관련이 있는지 없는지 알아낼 수 있다. 모든 텍스트 데이터들을 사람이 직접 읽어보며 관련 유무를 파악하는 것이 가장 정확하겠지만, 일반적으로 텍스트 데이터로 이용되는 생물 의학 문헌들은 그 개수가 매우 많기 때문에 일일이 문헌들을 읽어서 바이오 개체 관계를 추출하는 것은 사실 상 불가능 하다. 따라서 어떤 텍스트가 특정한 바이오 개체 관계에 대해 표현하는지를 알아내는 방법이 필요한데, 한 문헌 안에서 두 개의 바이오 개체 이름들이 동시에 등장하거나, 한 문헌의 한 문장 안에서 두 개의 바이오 개체 이름들이 동시에 등장한다면 해당 바이오 개체들 사이에 관계가 있다고 보는 것이 일반적이다. 텍스트에서 표현하고 있는 기존 관계들을 어떻게 추출할 것인지를 연구해도

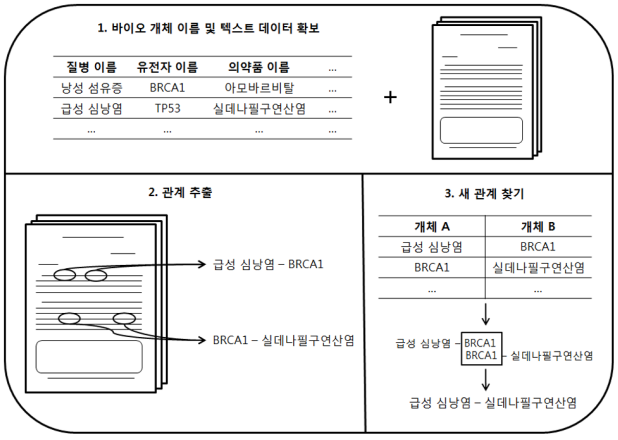

그림 2 생물 의학적 텍스트 마이닝에서 ABC 모델의 세 단계

바이오 개체 간의 연결 정확도를 향상시킬 수 있고, 최종적으로 바이오 개체들로 만드는 바이오 네트워크의 결과에도 영향을 미치게 된다.

마지막으로, 추출한 관계들에서 기존에 없는 새로운 관계를 추출하는 방법들에 대한 연구가 가장 많이 이루어지고 있다고 언급하였는데, 해당 부분의 몇 가지 방법들을 소개하겠다.

우선, 전통적인 방법으로 스완슨 박사의 ABC 모델이 있다. ABC 모델은 A – B 관계와 B – C 관계가 존재할 때, 기존 관계 리스트에 없는 A – C 관계를 찾는 것이다. Petric [3]은 ABC 모델을 발전시킨 알고리즘을 제안하였는데, 모든 문서들에서 적게 등장하는 희귀 단어(Rare term)를 중간 단계 역할인 B로 활용하는 것이다. 적게 등장하는 희귀 단어가 A와도 같이 등장하고 C와도 같이 등장하였다면, A와 C가 관련이 있을 확률이 더 높다는 가정 하에 제안된 방법이다. Li [4] 방법의 경우, MeSH(Medical Subject Headings; 미국 국립의학도서관(NLM)이 정하는 의학분야의 주제 명 인덱스로, 문헌의 내용을 나타내는 적절한 용어를 10~15개 용어들로 나타낸 것) 용어들을 가지고 용어-문헌

표 1 연구에 사용할 수 있는 관련 데이터베이스

데이터 종류	데이터베이스 이름
생물 의학 문헌	PubMed
	MEDLINE
	EMBASE
유전자 명	HGNC
	Gene Ontology
	CTD
	PharmGKB
	OMIM
의약품 명	CTD
	PharmGKB
	DailyMed
	PubPK
단백질 명	RCSB
	HPRD
	wwPDB
질병 명	Disease Ontology
	CTD
	PharmGKB
	OMIM
	MEDIC
miRNA 이름	miRBase
	miRecords

행렬(Term document matrix)을 만든다. 이 행렬을 기반으로 용어들 간의 상호정보량(Mutual information) 값을 구한다. 이 때, 상호정보량이 클수록 해당 두 용어들의 중요도는 높아진다. 여기에 ABC 모델을 적용하여 A – B 관계들 중 상호정보량이 큰 용어 B들을 고르고, 그 용어 B들과 연결된 용어 C들 중에서도 상호정보량 순으로 용어 C를 고른다. Li의 방법은 용어-문헌 행렬을 만들고, 그 행렬을 바탕으로 상호정보량을 이용했다는 점에서 의미가 있다. Tsuruoka [5]도 ABC 모델을 확장한 형태의 방법을 제안하였는데, 추출한 관계들에서 기존에 없는 새로운 관계를 추출할 때 용어 A와 용어 C 사이의 직접적 연결(Direct association)과 간접적 연결(Indirect association)의 강도와 신뢰도를 고려하는 순위화 방법을 사용하였다.

생물 의학 텍스트 마이닝을 이용하면 텍스트 데이터에서 바이오 개체 간의 관계를 얻을 수 있고 이를 기반으로 바이오 네트워크를 구축할 수 있다.

3. 바이오 네트워크

바이오 네트워크는 바이오 개체들을 노드(Node)로 하여 관련이 있는 바이오 개체들을 연결함으로써 바이오 개체들 사이의 관계를 표현한다. 바이오 네트워크의 종류에는 여러 가지가 있는데, 대표적으로 유전자 조절 네트워크(Gene regulatory network), 단백질 네트워크, 질병 네트워크, 그리고 통합 바이오 네트워크(Integrated bio-entity network) 등이 있다. 바이오 네트워크를 활용한다면 바이오 개체의 기능(Function), 바이오 개체 사이의 기작(Mechanism)이나 생물학적 프로세스의 패스웨이(Pathway)를 더 수월하게 연구할 수 있다.

유전자 조절 네트워크는 유전자 하나를 하나의 노드로 하여 유전자 발현의 조절 양상을 표현해주는 네트워크이다 (그림 3A). 유전자의 종류는 크게 세 가지로 나눌 수 있는데, 작동 유전자(Promoter 혹은 Operator), 조절 유전자(Regulator), 그리고 구조 유전자(Structural gene)이다. 실질적으로 발현되어 단백질이 되는 부분은 구조 유전자이고, 작동 유전자는 구조 유전자의 전사(Transcription)를 개시하는 유전자이다. 이 과정에서 조절 유전자가 끼어들어 특정 작동 유전자의 개시를 억제하고 결과적으로 해당 작동 유전자가 개시하는 구조 유전자 부분의 전사를 막을 수 있다. 유전자 조절 네트워크는 각 유전자가 다른 유전자들을 어떻게 조절하는지를 표현하므로 유전자 발현 연구에 많은 도움을 준다.

모든 생명체 내에서 일어나는 대사 활동들은 단백질들의 상호작용에 의해 이루어진다. 이는 질병 연구에서 단백질 네트워크가 얼마나 중요한 역할을 할 수 있는지를 나타낸다 [6]. 단백질 네트워크는 하나의 단백질이 하나의 노드로 표현되며, 단백질 사이에 연관이 있다면 연결선으로 이어진다 (그림 3B). 생명체 대사 활동의 근원이나 질병의 근본적인 원인은 유전자에서부터 시작되나 생명체 내에서 대사 활동을 수행하거나 질병 발병이나 치료에 실질적으로 관여되는 것은 단백질이기 때문에 단백질 네트워크도 유전자 네트워크 못지않게 중요하다. 단백질 네트워크를 연구하면 단백질의 기능을 예측할 수 있고, 유전자 단위의 연구에도 도움을 줄 수 있으며, 나아가 신약 개발에도 기여할 수 있다.

질병 네트워크 역시 하나의 노드가 하나의 질병을 표현하며 보통 노드의 크기는 해당 질병과 관련되어 있는 유전 정보의 개수를 나타내고 질병 사이에 공유하는 유전 정보가 있다면 연결선으로 이어진다 (그림 3C). 이 때, 연결선의 굵기는 두 질병이 공유하는 유전 정보가 많고 적음을 의미한다. 질병 네트워크에서 노드 사이 연결선의 중요성은, 하나의 유전적 이상(Genetic abnormality)이 단순히 하나의 유전자 혹은 단백질에서 끝나는 것이 아니라, 연결선을 타고 퍼져나감으로써 다양한 기능의 이상을 유발한다는 점에 있다 [7-9]. 이는 질병 네트워크에 대한 연구가 질병의 진단, 예후, 그리고 치료 및 신약 개발에 도움이 될 수 있다는 사실을 뒷받침 한다.

통합 바이오 네트워크는 다른 바이오 네트워크들이 모두 합쳐진 형태의 네트워크이다. 한 노드가 하나의 바이오 개체(유전자, 단백질, 질병, miRNA 등)를 표현하고, 바이오 개체들이 서로 연관이 있다면 연결선이 이어진다 (그림 3D). 바이오 개체들은 각각 독립적으로 특정한 기능을 담당하기 보다는 대부분 서로 간에 상호 작용을 함으로써 생명체 내에서 기능을 수행한다. 이러한 사실을 바탕으로 지역 가설(Local hypothesis)이 질병 연구에 많이 이용되고 있는데, 지역 가설이란 유전자 또는 단백질 등의 바이오 개체가 생명체 내의 기능이나 질병과 관련되어 있을 때, 해당 바이오 개체와 네트워크에서 연결되어 있는 다른 바이오 개체도 같은 기능이나 질병과 관련되어 있을 가능성이 높다는 가설이다 [7, 10]. 통합 바이오 네트워크는 모든 바이오 개체들 사이의 관련성을 포괄적으로 보유함으로써 세포의 기능이나 질병에 대해 보다 더 깊은 수준의 연구를 가능하게 한다.

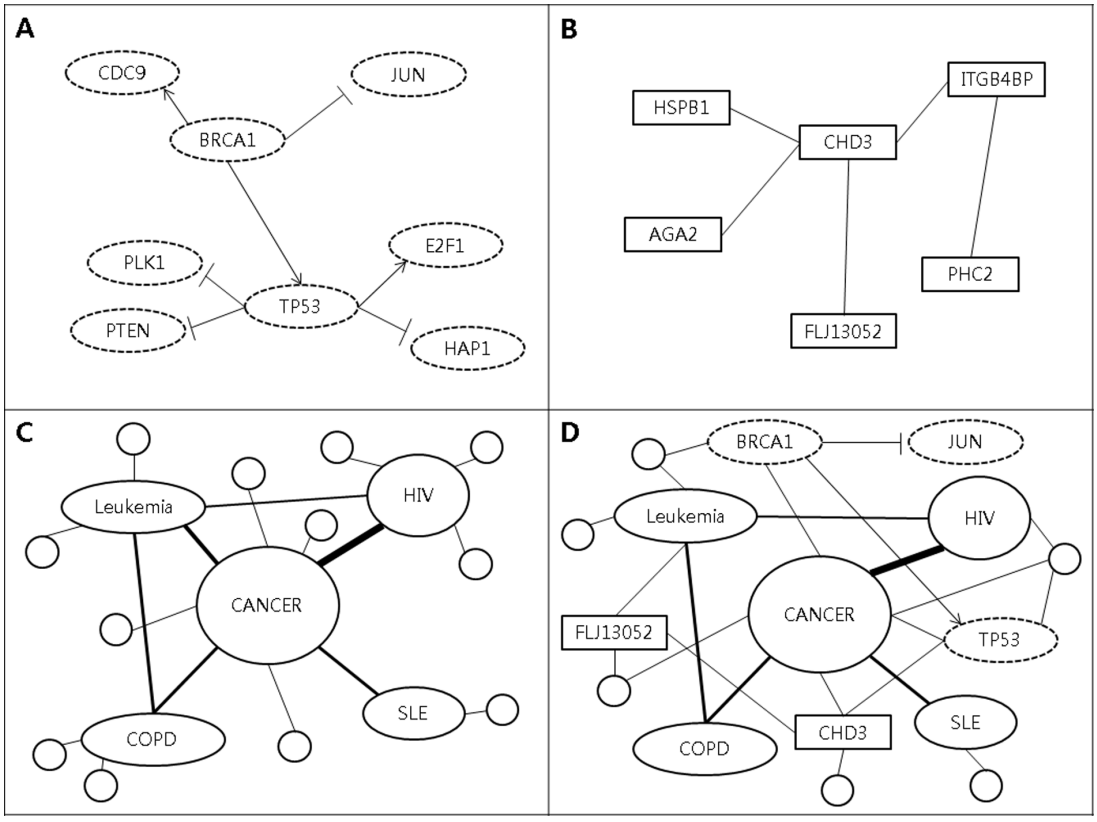

그림 3 바이오 네트워크의 종류. (A) 유전자 조절 네트워크. 방향성이 있는 네트워크로, 화살표는 발현을 돕는다는 의미이고 T 모양은 발현을 억제한다는 의미이다. (B) 단백질 네트워크. 서로 연관이 있는 단백질은 연결선으로 이어져 있다. (C) 질병 네트워크. 노드의 크기가 클수록 연관되어 있는 유전 정보가 많다. 두 질병이 공유하는 유전 정보가 많을수록 연결선도 굵어진다. 유전 정보를 적게 가지고 있는 질병들은 질병 명을 적지 않고 작은 노드로 표시하였다. (D) 통합 바이오 네트워크. 모든 바이오 개체들의 서로 간의 상호작용을 나타낸다.

4. 텍스트 마이닝과 바이오 네트워크

텍스트 마이닝이 새로운 바이오 개체 사이의 연결성을 찾아내고, 바이오 네트워크 연구의 중요성이 대두되면서 텍스트 마이닝으로 바이오 네트워크를 구축하는 연구들이 생기기 시작하였다. 생물 의학 관련 문헌들에서 바이오 개체 간의 관계를 추출할 수 있다면, 그 관계들을 서로 연결하여 원하는 바이오 네트워크를 만들 수 있다. 기존 유전 정보들만을 이용하여 만들어진 네트워크를 텍스트 기반 관계들로 보강할 수도 있고, 기존 네트워크를 검증하는 역할에 쓰일 수도 있다. 기본적인 관계들의 연결로 바이오 네트워크를 구축할 수도 있지만 연결의 신뢰도, 방향성 등의 추가적인 정보가 있다면 더욱 정확한 네트워크를 구축할 수 있다. 생물 의학적 텍스트 마이닝으로 바이오 네트워크를 구축하는 최근 연구들로는 다양한 텍스트 데이터들을 활용하는 방법이나, 새로운 바이오 개체 관계를 찾는 알고리즘의 보완, 바이오 개체 간의 유사도, 신뢰도, 방향성 등을 구하여 새로운 형태의 바이오 네트워크를 구축하는 방법 등이 있다.

Song [11]은 생물 의학 문헌 데이터베이스 중 하나인 PubMed로부터 자동적으로 유전자 조절 정보(Gene regulatory information)를 추출한 후, 유전자 조절 네트워크를 만드는 시스템을 제안하였다. Song의 시스템은 텍스트 데이터로부터 유전자 조절 관계(Gene regulatory relation)를 추출하여 유전자 조절 네트워크를 구축하며, 계통 명(Strain number), 유전자형(Genotype), 그리고 표현형(Phenotype) 정보들로 네트워크의 연결들에 신뢰성을 부여하였다. 이는 해당 네트워크를 사용하여 연구할 생물 의학 연구자들에게 큰 도움이 될 수 있다.

Subramani [12]는 생물 의학 문헌 데이터베이스 중 하나인 PubMed와 단백질 데이터베이스인 HPRD(Human Protein Reference Database), 그리고 패스웨이 데이터베이스인 KEGG로부터 단백질-단백질 연결 정보와 패스웨이를 추출하여 시각화하는 방법을 제안하였다. Subramani는 본인의 방법이 텍스트 데이터와 기존의 공개된 데이터베이스를 함께 사용하여 단백질 네트워크를 구축한 첫 번째 방법이며, 텍스트 데이터에서 새로운 단백질

관계를 찾았다는 점에서 연구의 의의가 있다고 주장하였다.

Zhang [13]은 표현형에 대한 텍스트 데이터인 표현형 기록(Phenotype records)을 이용하여 질병의 표현형과 관련된 속성 벡터(Feature vector)를 만들었다. 그리고 속성 벡터 간 코사인 유사도(Cosine similarity)를 이용하여 질병 표현형 사이의 유사도를 구하고 질병 네트워크를 구축하였다. 하지만 Zhang의 방법은 속성을 어떤 생물 의학적 단어들로 몇 개나 구성하느냐에 따라 결과가 달라지고 해당 생물 의학적 단어들을 선택하는 기준이나 속성 개수를 선정하는 과정이 모호하다는 점에서 한계점을 드러내고 있다.

생물 의학적 텍스트 마이닝을 이용하여 통합 바이오 네트워크를 구축한 연구들도 다수 존재한다. Saeys [14]는 기존 텍스트 마이닝 기법에 기계 학습(Machine learning)을 적용하여 더 다양한 바이오 개체 관계들을 추출하여 네트워크를 구축하였다. Saeys는 이벤트(Event)를 기반으로 바이오 개체 관계들을 추출하였는데, 이 과정에서 텍스트 데이터에서 이벤트를 구분하게 위해 SVM(Support Vector Machine)을 이용한 바이너리 분류(Binary classification)를 수행하였다. Bell [15]의 방법은 PubMed 데이터와 기존의 단백질, 유전자, 질병, 패스웨이 등을 가지고 베이지안 네트워크(Bayesian network)를 만든 후 각 연결이 사실(True)일 확률을 계산하여 통합 바이오 네트워크를 완성한다. Bell은 해당 베이지안 네트워크에서 BFSP(Breadth-First Search with Pruning)와 MPP(Most Probable Path)를 이용하여 새로운 바이오 개체 관계를 찾아내기도 하였다. Eronen [16]도 PubMed 데이터와 기존의 단백질, 유전자, 의약품, 표현형, 패스웨이, 유전자 위치(Locus) 등의 정보를 바탕으로 통합 바이오 네트워크를 구축하였다. Eronen은 네트워크를 구축하는 과정에서 연결선에 가중치를 주었다. 가중치는 관련성(Relevance), 유익성(Informativeness), 그리고 신뢰성(Reliability)을 기반으로 점수화하였다. 뿐만 아니라 해당 네트워크를 분석하여 사용자가 원하는 정보를 제공할 수 있는 시스템(Biomine)도 구축하였다. Katukuri [17]는 연도별로 정리되어 있는 생물 의학 문헌 데이터베이스인 Medline을 이용하여 바이오 개체 네트워크를 구축하였다. 이때 각 바이오 개체는 의미적 유형(Semantic type), 관련된 저자들(Related authors), 그리고 문헌에 나온 빈도수(Document frequency) 정보를 가지고 있고, 바이오 개체 간의 연결선도 문헌에서의 연결 빈번성이나 연도별 문헌에서의 등장 연속성을 계산한 연결의 강도(Strength)나 지속성(Duration) 정보를 가지고 있다. 해당 정보들은 바이오 개체 사이의 새로운 링크(Link)를 찾는데 이용된다.

그 밖에 텍스트 데이터로 의약품 네트워크를 구축한 Bui [18]의 방법, 질병-환경요인(Etiological factor) 네트워크를 구축한 Liu [19]의 방법, 질병-유전자 네트워크를 구축한 Quan [20]의 방법, 그리고 음식-질병-유전자 네트워크를 구축한 Yang [21]의 방법 등이 존재한다.

5. 결 론

생물학적 실험으로 정립되고 검증되던 바이오 네트워크 구축 분야에서 실질적인 실험 없이 생물 의학적 문헌들만을 이용하는 텍스트 마이닝은 매력적인 연구 기법이다. 현재 생물 의학 텍스트 마이닝으로 바이오 네트워크를 구축하는 연구들은 다양한 데이터들을 함께 활용하거나, 기존 바이오 개체 관계들에서 새로운 관계를 찾아 네트워크를 보강하는 것에 집중하거나, 바이오 개체 간의 유사도, 신뢰도, 방향성 등을 구하여 새로운 형태의 바이오 네트워크를 구축하는 연구들이 대부분이다. 다양한 데이터들을 활용할 때 구글(Google)의 검색(Search) 데이터나 SNS(Social Network Service) 데이터 등의 색다른 데이터들을 바이오 네트워크를 구축하는데 이용한다면 흥미로운 결과가 나올 수도 있다. 또한, 기존 바이오 개체 관계들을 조합하여 새로운 바이오 개체 관계들을 발견하고자 할 때나 바이오 개체들 사이의 유사도, 신뢰도, 방향성 등을 계산하고자 할 때, 생물 의학 분야에서 자주 쓰이는 기법들이 아닌 그래픽스나 사운드, 계산이론 등에서 쓰이는 기법을 적용시켜볼 수도 있다. 다만 전통적으로 생물 의학 텍스트 마이닝을 이용한 방법들은 검증이 어렵다는 단점이 있다. 그러한 문제의 해결을 위해 생물 의학적 텍스트 마이닝으로 구축된 바이오 네트워크 검증 방법에 대해 연구하는 것도 좋은 주제가 될 수 있다.

참고문헌

[1] Don R. Swanson, “Fish oil, Raynaud’s syndrome, and undiscovered public knowledge”, Perspectives in Biology and Medicine, vol. **30**(1), pp. 7-18, 1986.

[2] Don R. Swanson, “A second example of mutually isolated medical literatures related by implicit, unnoticed connections”, Journal of the American Society for Information Science, vol. **40**, no. 6, pp. 432-435, 1989.

[3] Ingrid Petric *et al.*, “Literature mining method RaJoLink

for uncovering relations between biomedical concepts", Journal of Biomedical Informatics, vol. **42**, no. 2, pp. 219-227, 2009.

[4] Guangrong Li *et al*., "Mining Biomedical Knowledge Using Mutual information ABC", IEEE International Conference on Granular Computing, 2011.

[5] Yoshimasa Tsuruoka *et al*., "Discovering and visualizing indirect associations between biomedical concepts", Bioinformatics, vol. **27**, issue 13, pp. i111-i119, 2011.

[6] Tuba Sevimoglu *et al*., "The role of protein interaction networks in systems biomedicine", Computational and Structural Biotechnology Journal, vol. **11**, issue 18, pp. 22-27, 2014.

[7] K. Goh *et al*., "The human disease network", Proc. Natl Acad. Sci., USA, vol. **104**, pp. 8685-8690, 2007.

[8] D. B. Goldstein *et al*., "Common genetic variation and human traits", N. Engl. J. Med., vol. **360**, pp. 1696-1698, 2009.

[9] E. E. Schadt *et al*., "Molecular networks as sensors and drivers of common human diseases", Nature, vol. **461**, pp. 218-223, 2009.

[10] M. Oti *et al*., "Predicting disease genes using protein-protein interactions", J. Med. Genet., vol. **43**, pp. 691-698, 2006.

[11] Yong-Ling Song *et al*., "Text Mining Biomedical Literature for Constructing Gene Regulatory Networks", Interdiscip Sci Comput Life Sci, vol. **1**, issue 3, pp. 179-186, 2009.

[12] Suresh Subramani *et al*., "HPIminer: A text mining system for building and visualizing human protein interaction networks and pathways", Journal of Biomedical Informatics, vol. **54**, pp. 121-131, 2015.

[13] Shi-Hua Zhang *et al*., "From phenotype to gene: Detecting disease-specific gene functional modules via a text-based human disease phenotype network construction", FEBS Letters, vol. **584**, pp. 3635-3643, 2010.

[14] Yvan Saeys *et al*., "Event based text mining for integrated network construction", Proceedings of JMLR, vol. **8**, pp. 112-121, 2010.

[15] Lindsey Bell *et al*., "Integrated Bio-Entity Network: A System for Biological Knowledge Discovery", PLoS ONE, vol. **6**, issue 6, pp. e21474, 2011.

[16] Lauri Eronen *et al*., "Biomine: predicting links between biological entities using network models of heterogeneous databases", BMC Bioinformatics, vol. **13**, pp. 119, 2012.

[17] Jayasimha R. Katukuri *et al*., "Hypotheses generation as supervised link discovery with automated class labeling on large-scale biomedical concept networks", BMC Genomics, vol. **13**, suppl. 3, pp. S5, 2012.

[18] Quoc-Chinh Bui *et al*., "A novel feature-based approach to extract drug – drug interactions from biomedical text", Bioinformatics, vol. **30**, no. 23, pp. 3365-3371, 2014.

[19] Yueyi Liu *et al*., "The "etiome": identification and clustering of human disease etiological factors", BMC Bioinformatics, vol. **10**, suppl. 2, pp. S14, 2009.

[20] Changqin Quan *et al*., "Gene – disease association extraction by text mining and network analysis", Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis, pp. 54-63, 2014.

[21] Hui Yang *et al*., "Mining Biomedical Text towards Building a Quantitative Food-Disease-Gene Network", Learning Structure and Schemas from Documents: Studies in Computational Intelligence, vol. **375**, pp. 205-225, 2011.

약 력

김 현 진

2010 연세대학교 컴퓨터과학과 졸업(학사)
2010~현재 연세대학교 컴퓨터과학과 통합과정
관심분야: 바이오인포매틱스, 데이터 마이닝, 텍스트 마이닝, 추천 시스템, 데이터베이스
Email: chriskim@cs.yonsei.ac.kr

윤 영 미

1981 서울대학교 자연과학대학 졸업(학사)
1983 오하이오 주립대학 수학과 졸업(학사수료)
1987 스탠포드대학교 컴퓨터과학과 졸업(석사)
2008 연세대학교 컴퓨터과학과 졸업(박사)
1987~1993 IntelliGenetics Inc., California, USA, 소프트웨어 엔지니어
1995~현재 가천대학교 컴퓨터공학과 교수
관심분야: 바이오인포매틱스, 데이터마이닝, 데이터베이스 시스템
Email: ymyoon@gachon.ac.kr