# 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.

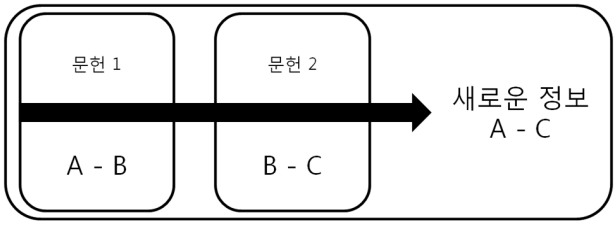


그림 1 스완슨 박사의 ABC 모델

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).