약물들의 유사성을 이용한 숨겨진 약물 효과 추론

김신*, 김정우*, 박상현*,†
연세대학교 컴퓨터과학과*
jcynico@gmail.com, jwkim2013@cs.yonsei.ac.kr, sanghyun@cs.yonsei.ac.kr

Inferring Hidden Drug Effects Using Similarity between Drugs

Shin Kim*, Jeongwoo Kim*, Sanghyun Park*,†
Dept. of Computer Science, Yonsei University *

요 약

신약재창출(drug repositioning) 기술은 큰 비용과 많은 시간을 요구하는 기존 신약개발 과정보다 효율적인 신약개발 전략이다. 많은 약물 정보와 의약학 문헌자료들이 공개적인 데이터베이스로(database) 제공됨에 따라 생물학적 정보를 기반으로 하는 신약재창출 기술 개발의 길이 넓어졌다. 본 논문에서는 이러한 생물학정보를 활용하여 약물 사이의 유사도(similarity)를 계산하고, 그 값을 기반으로 약물들의 알려지지 않은 효과를 추론하는 방법론을 제안한다. 특정 효과가 있는 약물들을 하나의 군집(cluster)으로 구성한다. 약물 데이터베이스와 생물학 문헌데이터를 이용하여, 효과(effect)/부작용(side effect) 측면에서 이 군집과 가장 유사한 약물을 추출한다. 위의 방법으로 추출된 약물과 약물 군집이 공유하는 특정 효과 사이의 중요한 생물학적 관련이 있을 것이라고 추론한다.

1. 서 론

신약개발은 연구 단계와 개발 단계라는 복잡한 과정을 거쳐야 하므로 많은 시간과 자본이 소요된다. 신약개발 관련 기술의 급격한 발전에도 불구하고 약물의 안정성과 예측 가능성에 대한 요구도 함께 증가하여 아직은 신약 개발에 드는 비용이 감소하고 있지 않았다. 이로 인해 이미 안정성을 검증받은 기존 약물들의 가치를 높이는 방법이 주목받고 있다. 이러한 방법을 신약재창출이라고 한다.

신약재창출 기술은 안정성이 검증된 약물들의 새로운 의학적 용도를 발굴하는 개발 전략이므로 시간과 자본을 절약할 수 있다. 신약재창출 기술은 기존 약물의 유용성을 증대시킬 뿐만 아니라 해당 약물에 대한 특허 독점 기간 연장에 도움을 주는 매우 유용한 전략이다.

하나의 약물은 다양한 질병에서 치료, 예방, 지연 효과를 가지기도 하고 다양한 부작용을 가지고 있기도 하다. 이처럼 약물이 가지고 있는 다양한 작용들을 활용하여 신약재창출 전략을 제시하고자 한다. 생물학 문헌데이터를 활용하여 작용 간 유사도를 계산하고. 이를 통해 약물 간 유사도를 계산하여 특정 작용을 가질 것이라 예상되는 약물을 추론하는 것이 연구의 목표이다.

특정 작용을 하는 약물들을 모아 하나의 약물 군집을 만들고, 이 약물 군집이 가지고 있는 작용들을 가장 많이 공유하는 새로운 약물은 높은 확률로 이 약물 군집에 포함될 수 있다는 가설을 기반으로 한다. 본 논문의 구성은 다음과 같다. 2장에서는 본 연구의 기반이 되는 관련 연구들에 관하여 기술하고, 3장에서는 본 논문이 제안하는 방법론을 설명한다. 마지막으로 4장에서 본 논문의 결론 및 발전 방향에 관하여 기술한다.

2. 관 련 연 구

유전자, 단백질, 질병 등의 생물학 개체들은 상호작용하므로, 생물학 분야에서 생물학적 개체들 사이의 관계를 찾는 작업은 매우 중요하다. 이러한 생물학 개체들 사이의 관계를 찾는 대표적인 방법으로 Swanson의 ABC model[1]이 있다.

Swanson은 의약학 문헌에서 A와 B의 연관성이 확인되고, B와 C의 연관성이 확인되었다면 A와 C의 연관성도 충분히 존재할 수 있음을 제안했다. 예를 들어, 작용 A를 하는 임의의 약물을 복용하면 작용 B가 나타나고 작용 B를 하는 또 다른 약물 C가 존재한다고 할 때, 작용 A와 약물 C의 연관 관계를 추론할 수 있다. 작용 A와 약물 C를 잇는 B와 같은 작용이 많을수록 더 큰 연관관계가 있다고 판단할 수 있다. 이러한 ABC model을 발전시켜 질병과 유전자의 관계를 추론하는 방법들에 대한 연구도 진행되고 있다.

Özgür[2]은 특정 질병과 연관이 있는 유전자와 같은 문헌의 같은 문장 안에 속해있는 유전자들을 추출하고, 두 유전자를 연결하여 유전자 네트워크를 구축하고, 구축된 네트워크에서 중심성(centrality)이 높은 유전자일수록 특정 질병과 관련이 깊을 것이라고 제안했다. 전립선암을 질병 데이터를 기반으로 실험하였고, 중심성을 기준으로 상위 20개의 유전자를 추출하였다. 추출된 20개의 유전자 중 최대 19개의 유전자가 질병과 관계가 있음을 검증하였다. 두 개의 개체가 문헌상에서 같은 문장에 동시에 출현하는 것을

† 교신저자(Corresponding Author)
※ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015R1A2A1A05001845).

활용하여도 충분히 의미 있는 연관성을 추출할 수 있다는 것을 위의 논문을 통해 확인하였다. 이처럼 생물학 문헌데이터를 활용하여 생물학 개체들 사이의 관계를 추론하는 다양한 방법론들이 연구되고 있다[3][4].

3. 후보 약물 추출 방법

3.1 사용 데이터

본 논문에서는 두 가지 데이터를 사용한다. 먼저 약물과 그에 대한 효과 및 부작용에 대한 데이터를 추출하기 위해 DrugBank[5]를 이용하였다. 또 효과 및 부작용들 사이의 관련성을 계산하기 위해서 생물학 문헌데이터를 활용하는데, 데이터는 PubMed[6]로부터 구축하였다. PubMed는 생물학과 관련된 문헌데이터를 가장 많이 보유하고 있으며, 누구나 손쉽게 사용할 수 있도록 잘 구축된 데이터베이스이다.

3.2 후보 약물 추출을 위한 가중치 계산 방법

먼저 특정 작용을 공통으로 가지고 있는 약물들을 하나의 군집으로 형성한다. 군집에 포함되는 모든 약물의 효과 및 부작용을 추출하고, 해당 효과 및 부작용 데이터를 군집에 포함되지 않는 약물들과의 유사도를 계산하기 위한 특성(Feature)으로써 사용한다. 후보 약물 A의 가중치는 다음과 같이 계산한다.

$$\mathrm{Drug}_{\mathrm{Weight(A)}} = \sum_{i \in N} Effect\_Weight(i) * n$$

N은 특정 작용을 나타내는 약물 군집과 약물 A에 공통으로 나타나는 작용들의 집합을 의미한다. n은 군집 내에서 작용 i를 가지는 Drug의 수를 의미한다. Effect_Weight(i)는 작용 i에 대한 가중치를 의미하며, 아래의 식으로 계산된다.

$$\mathrm{Effect\_Weight(i)} = frequency(Target\ Effect, i)$$

Target Effect: 약물 군집이 공통으로 가지고 있는 작용으로, 후보 약물들로부터 찾고자 하는 작용이다.

frequency(Target Effect, i)는 특정 작용(Target Effect)과 작용 i가 생물학 문헌상에서 동시에 등장하는 문장의 수를 의미한다. 두 작용이 함께 언급되는 문장의 수가 많을수록 두 작용 사이에는 더 많은 관련성이 있다고 판단한다.

특정 작용(Target Effecet)을 가지는 약물 군집과 공유하는 작용이 많고, 그 작용들이 특정 작용(Taget Effect)과 문헌상에서 많이 언급될수록 후보 약물은 더 높은 가중치를 부여받는다. 위의 식을 이용하여 특정 작용(Target Effect)을 가지고 있는 새로운 약물을 추론할 수 있다.

Drug 1	Drug 2	Drug 3	Drug 4	Drug 5
effect a	effect a	effect b	effect b	effect e
effect b effect_weight(b) = 50	effect c effect_weight(c) = 30	effect c effect e	effect d effect f	effect f
effect c effect_weight(c) = 30	effect d effect_weight(d) = 40			

그림 1. 약물 정보 예시

그림 1은 이미 알려진 약물들의 작용을 보여준다. 이해를 돕기 위해 그림 1에 있는 약물들로 작용 a를 가지고 있을 것이라 예상하는 약물을 어떻게 추론하는지 보여주고자 한다.

먼저 기존에 이미 작용 a를 가지고 있다고 알려진 약물들로 하나의 약물 군집을 형성한다. 약물 1과 약물 2가 이 약물 군집에 속하게 된다. 약물 군집에 나타난 작용이 약물 3에는 2개, 약물 4에도 2개가 나타나지만, 약물 5에는 나타나지 않는다. 이로 인해 약물 3과 약물 4는 약물 5보다 더 훌륭한 후보가 된다.

좀 더 정교한 추론을 위해 약물 군집에 나타나는 작용 간의 연관성을 활용한다. 두 작용이 같은 문헌의 같은 문장에서 빈번히 발견된다면 두 작용 사이에는 인과관계가 있거나 같은 원인이 있어 필연적으로 동시에 나타났다고 가정할 수 있다. 약물 군집 안에서 연관성이 있는 작용이 여러 번 나타난다면 더 큰 가중치를 주는 것이 합리적이기 때문에 등장 횟수만큼 가중치를 받는다.

약물 3과 약물 4의 가중치를 계산해보면 다음과 같다.

$$Drug\_Weight(3)$$
$$= effect\_weight(b) \times 1 + effect\_weight(c) \times 2$$
$$= 110$$
$$Drug\_Weight(4)$$
$$= effect\_weight(b) \times 1 + effect\_weight(d) \times 1$$
$$= 90$$

위 설명을 정리하면 우선 특정 작용이 나타난다고 이미 알려진 약물들로 하나의 약물 군집을 형성한다. 특정 작용과 약물 군집에 속한 작용들 사이의 연관성을 판단한다. 후보 약물들이 가지고 있는 작용들이 특정 작용과의 유사도가 높을수록, 약물 군집에서 자주 나타날수록, 후보 약물과 약물 군집 사이의 유사도가 높아진다. 또 군집과 공유하는 작용의 수가 많을수록 후보 약물의 가중치가 증가하게 된다. 결과적으로 약물 군집과 가장 유사도가 높은 후보 약물에 특정 작용이 있을 것이라 추론할 수 있다.

4. 발전 방향 및 결론

향후 훈련 집합(Training set)과 테스트 집합(Test set)을 나누어 본 논문에서 소개한 방법론의 효용성을 검증해 볼 예정이다. 이는 지도 학습(Supervised learning)에서 알고리즘의 성능을 평가하는 방법이다. 특정 작용을 가진 약물들로 형성된 군집을 훈련 집합으로 설정한 후, 훈련 집합에서 분리된 테스트 집합을 형성한다. 이는 군집의 일부 약물이 특정 작용을 하지 않는 것처럼 데이터를 조작하여 테스트 집합을 형성하는 것이다. 본 논문에서 소개한 방법론에 의해 테스트 집합이 훈련 집합으로 회귀하는지 검증할 것이다.

또한, 특정 작용으로부터 후보 약물을 추천해주는 웹 어플리케이션을 구축할 예정이다. 특정 작용에 대한 약물 군집과 이로써 추론된 후보 약물들을 보여주는 어플리케이션이다. 연구자들은 이 어플리케이션을 사용함으로써 신약재창출을 전략으로 하는 가설을 설정하는 데에 도움을 받을 수 있다.

약물을 사용하였을 때, 호르몬 혹은 단백질 등에 의하여 다양한 작용들이 필연적으로 나타나게 되지만 그에 비해 연구자가 어떤 작용이 나타날지 예측하기는 어렵다. 이미 알려진 약물의 정보를 이용한다면 작용 간의 연관관계를 밝히고 아직 발견되지 않은 약물의 작용을 찾을 수 있을 것이라 기대한다.

이로써 기존 약물이 가지고 있던 작용과 매우 흡사한 작용만을 가설로 설정해오던 현재의 한계에서 벗어나 연구자들은 다양한 작용에 대한 가설을 고려해볼 수 있게 될 것이다. 신약재창출을 전략으로 하는 연구가 활발해질 것이라 기대해본다.

5. 참 고 문 헌

[1] DrugBank: a knowledgebase for drugs, drug actions and drug targets. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M.Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6. PubMed ID: 18048412

[2] Swanson, Don R. "Fish oil, Raynaud's syndrome, and undiscovered public knowledge." *Perspectives in biology and medicine* 30.1 (1986): 7-18.

[3] Özgür, Arzucan, et al. "Identifying gene-disease associations using centrality on a literature mined gene-interaction network." *Bioinformatics* 24(13): i277-i285, 2008.

[4] Lee, “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] Adamic et al. “A literature based method for identifying gene-disease connections”, IEEE Computer Society Bioinformatics Conference, 2002

[6] PubMed National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/pubmed