약물 리포지셔닝 분야에서 단백질 위치 정보의 활용 가능성에 대한 연구

여윤구*, 윤영미**, 박상현*†
연세대학교 컴퓨터과학과*, 가천대학교 컴퓨터공학과**
yyk@cs.yonsei.ac.kr, ymyoon@gachon.ac.kr, sanghyun@cs.yonsei.ac.kr

A research on effects of protein localization in drug repositioning

Yunku Yeu*, Youngmi Yoon**, Sanghyun Park*†
Dept. of Computer Science, Yonsei University*, Dept. of Computer Engineering, Gachon University**

요 약

단백질의 위치 정보는 단백질의 특성을 파악할 수 있는 중요한 정보 중 하나다. 그러나 약물의 리포지셔닝 연구에서 아직까지 단백질의 위치 정보가 활용되지 않고 있다. 본 논문은 단백질의 위치 정보를 위치 벡터로 나타내고, 불완전한 위치 벡터를 단백질 상호작용 네트워크를 이용해 보완하였다. 단백질 위치 벡터를 비교 분석한 결과, 같은 약물의 타겟 단백질 사이에서 높은 위치 벡터 유사성이 나타났으며, 화학 구성이 비슷한 약물일수록 더 높은 벡터 유사성을 나타내었다. 이는 단백질의 위치 벡터가 약물 리포지셔닝 연구에서 활용되던 기존 예측 지표와 마찬가지로 유의미한 패턴을 갖고 있음을 의미하며, 추후 약물 리포지셔닝 연구에서의 활용 가능성이 높음을 나타낸다.

1. 서 론

약물 리포지셔닝(drug repositioning)은 신약 개발의 위험성을 줄일 수 있는 연구방법론 중의 하나다. Pammolli 등이 2000~2008년의 약물 개발 동향을 분석한 연구[1]에 따르면, 신물질이 임상실험까지 성공할 확률은 약 2.01%이며, 약물 개발에 평균 13.9년의 기간이 소요되는 것으로 나타났다. 따라서 생물정보학의 접근 방식을 이용해 가능성 있는 후보군을 예측하는 방법은 신약 개발 비용을 크게 경감할 수 있는 매우 중요한 연구 도구이며 신약 개발 과정에서 널리 활용되고 있다.

약물 리포지셔닝은 기본적으로 신물질을 개발하는 것이 아니라, 기존 약물의 새로운 용도를 탐색하는 연구 분야다. 이를 위해 약물 간의 유사성 또는 약물이 작용하는 타겟 단백질 간의 유사성에 기반하여 새로운 약물과 타겟의 관계를 예측한다. 예컨데, 어떤 질병에 듣는 약물과 유사한 성질을 갖는 다른 약물을 해당 질병에 적용하거나, 어떤 약물을 타겟 단백질과 비슷한 성질을 갖는 다른 단백질에 작용하는지 실험해 볼 수 있다. 이러한 예측(prediction)은 가능한 한 많은 데이터를 종합할수록 높은 정확도를 기대할 수 있다. Gottlieb 등은 약물의 화학 구조, 부작용(side-effect) 정보, 아미노산 서열(amino acid sequence), 단백질 상호작용 네트워크(protein-protein interaction network)에서의 거리 등을 이용하여 약물 간의 유사성을 비교 분석하였다[2].

본 논문에서는 약물의 특징을 파악하는 데 있어서 중요한 예측 지표(feature)로서 단백질의 위치(localization) 정보를 활용하고자 한다. 단백질의 위치 정보란 단백질이 작용하는 세포내 위치를 의미한다. 많은 수의 단백질이 특정한 작용 위치를 갖고 있으며, 작용하는 위치가 비슷하면 그 기능 또한 비슷할 가능성이 있다. 단백질의 위치 정보는 질병 간의 동시 이환성(comorbidity)을 연구하는 데에 활용된 바 있으나[3], 약물의 리포지셔닝 연구에서는 아직 활용된 바가 없다.

다만, 대부분의 생물 데이터처럼 단백질의 위치 정보 또한 완전한 데이터가 아니다. 위치 정보가 없는 단백질도 있으며, 거짓 긍정 오류(false-positive error) 또한 배제할 수 없다. 이를 극복하기 위하여 본 연구에서는 네트워크 이론과 단백질 상호작용 정보를 활용하여 단백질 위치 정보의 단점을 보완하였다. 보강된 단백질 위치 정보를 분석하여 약물 리포지셔닝 연구에서 단백질 위치 정보의 활용 가능성을 분석하였다.

2. 본 론

2.1 약물 및 단백질 관련 데이터의 수집

본 연구에서 사용한 데이터는 표 1과 같다.

표 1 데이터 출처 및 설명

구 분	출 처	데이터 종류
약물 관련	Drugbank[4]	약물 식별정보 타겟 단백질 정보
단백질 관련	UniProt[5]	단백질 식별정보 세포내 위치정보 단백질 상호작용(PPI)
	I2D[6], BioGrid[7]	단백질 상호작용

종합된 데이터는 7,677개의 약물과 136,871개의 단백질 및 231,790개의 단백질 상호작용 데이터를 포함한다.

2.2 단백질 위치정보 벡터의 구성

단백질의 위치 정보는 10개의 실수 값으로 구성된 위

† 교신저자(Corresponding Author)

※ 이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2012R1A2A1A01010775).

치 벡터로 표현하였다. 위치 벡터의 각각의 값은 해당 단백질이 세포 내의 위치 중 대표적인 10곳에서 작용할 가능성을 의미한다. 세포 내의 위치는 Park의 연구[3]에 따라 각각 시토졸(cytosol), 소포체(endoplasmic reticulum, ER), 세포 외부, 골지(Golgi), 페록시솜(peroxisome), 미토콘드리아, 세포핵(nucleus), 리소좀(lysosome), 원형질 막(plasma membrane), 기타 위치의 10곳으로 선정하였다. UniProt 데이터에 나타난 단백질의 위치 정보는 위치 벡터에서 1로 설정하였으며, 그 외의 위치는 0으로 설정하였다. 향후 실험에서 위치 벡터 간의 유사도를 계산할 때에는 두 위치 벡터의 코사인 유사도(cosine similarity)를 이용하였다.

2.3 단백질 상호작용 네트워크 구성

2.1에서 서술한 대로, UniProt과 DIP, BioGrid에서 단백질 상호작용 데이터를 수집하였으며, 이를 병합하여 단백질 상호작용 네트워크를 구성하였다. 단백질의 상호작용 데이터는 두 단백질이 물리적으로 결합하는지 여부를 나타내는 것으로서, 단백질 간의 관련성을 나타내는 중요한 데이터 중 하나다. 단백질 상호작용 네트워크는 가중치가 있는 무방향성(undirected) 단순 그래프 형태로 나타내었다. DIP와 UniProt에서 가져온 간선(edge)은 실험적으로 검증된 데이터로서 더 높은 정확도를 기대할 수 있기 때문에 BioGrid 데이터에서 가져온 간선에 비해 더 높은 가중치를 부여했다. (DIP, UniProt: 0.6, BioGrid: 0.5)

2.4 네트워크를 통한 단백질 위치정보 전파(propagation)

단백질의 위치정보가 완전하지 않기 때문에, 앞서 구축한 단백질 상호작용 네트워크를 통하여 단백질의 위치 정보를 주변 노드에 전파하도록 했다. 위치정보 전파는 Vanunu 등의 방법[8]을 이용하였으며, 이는 네트워크 구조를 통해 위치 정보를 전파하며 전체 네트워크의 변화 정도가 매우 작아질 때까지 반복적으로(iterative) 이루어진다. 시점 i에서 단백질 x의 위치 벡터를 x_i, 단백질 x의 초기 위치 벡터를 x_0, 단백질 x의 이웃 노드 집합을 X, 단백질 상호작용 네트워크에서 노드 x와 노드 y 사이 간선의 가중치를 w_{xy}, 학습 비율(learning rate)를 δ라고 할 때, 다음 시점 i+1에서의 단백질 x의 위치 벡터는 다음 식 (1)과 같이 결정된다.

$$x_{i+1} = \delta \sum_{y \in X} w_{xy} y_i + (1-\delta) x_0 \quad (1)$$

즉, 매 시점 단백질 x의 위치 벡터는 이웃 노드들의 위치 벡터와 x의 초기 위치 벡터간의 가중치 합으로 결정된다. 학습 비율은 0.6으로 설정하였으며, 위치정보 전파 작업은 모든 위치 벡터의 변화값의 합이 10^{-6} 미만이 될 때까지 반복 수행하였다.

3. 실험 및 결과 분석

약물의 리포지셔닝 분야에서 단백질의 위치 벡터의 활용 가능성을 평가하기 위하여 다음과 같은 두 가지 실험을 수행하였다.

먼저 여러 개의 타겟 단백질을 갖는 약물 데이터에서 타겟 단백질 사이에서 위치 벡터가 얼마나 유사한지를 분석하였다.(그림 1) 만약 동일한 약물의 이미 알려진 타겟 단백질 간에 위치 벡터가 유사하다면, 약물의 새로운 타겟 단백질을 예측하는 데에 위치 벡터를 활용해 볼 수 있을 것이다. 그림 1에서 짙은 선(all)은 본 연구에서 사용한 모든 단백질 쌍에서 위치 벡터의 유사도를 계산한 다음, 전체 단백질 쌍 중 가로축의 유사도 값을 갖는 쌍의 비율을 세로축에 표시하는 누적 꺾은선 그래프이다. 예를 들면 하위 50%에 위치한 단백질 쌍은 약 0.2의 위치 벡터 유사도를 갖고 있으며, 0.75 이상의 유사도를 갖는 단백질 쌍은 전체의 약 20%에 불과하다. 반면 그림 1에서 옅은 선(multi target)은 같은 약물의 타겟 단백질 사이에서 위치 벡터 유사도를 계산한 것이다. 모든 단백질 쌍에서 계산한 것보다 누적 꺾은선 그래프가 우측으로 크게 이동한 것을 관찰할 수 있다. 하위 50%에 위치한 단백질 쌍도 약 0.79의 위치 벡터 유사도를 갖고 있으며, 0.75 이상의 유사도를 갖는 단백질 쌍이 약 55%에 달했다. 이는 동일한 약물의 타겟 단백질 사이에서 명백

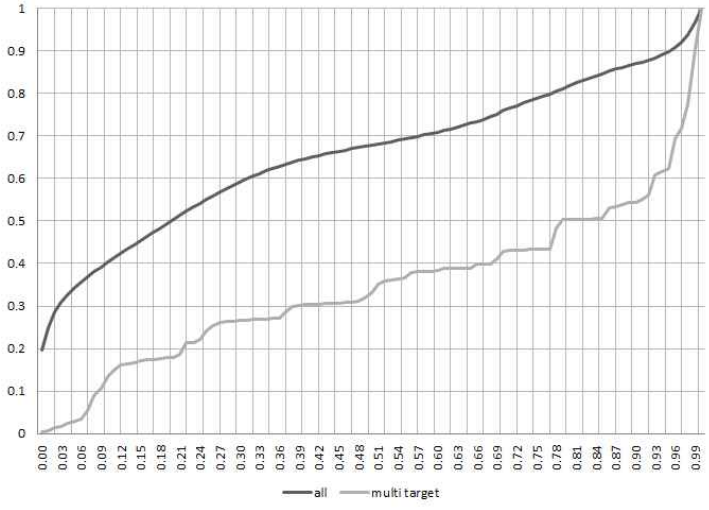

그림 1. 동일한 약물의 여러 타겟 단백질간 위치 벡터 유사도 분석

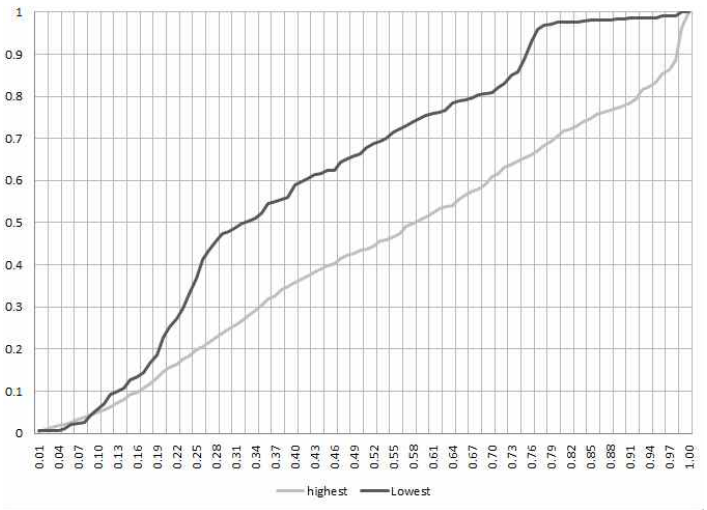

그림 2. 약물간 SMILES 유사도와 타겟 단백질간 위치 벡터 유사도 분석

한 위치 벡터의 유사도가 나타나고 있음을 보여 준다.

두 번째로는 SMILES (Simplified Molecular-Input Line-Entry System)의 유사성과 단백질 위치 벡터 간의 유사성을 분석하였다. SMILES는 약물의 화학 구조를 나타내는 문자열 형태의 데이터다. 즉, 두 번째 실험은 화학적 측면에서 유사한 약물과 그렇지 않은 약물 사이에서 위치 벡터의 유사도가 어떻게 나타나는지를 분석하는 것이다. 이를 위해서 CDK(Chemical Development Kit)을 이용해[9] SMILES 데이터를 특정 화학 구조의 유/무를 나타내는 부울(boolean) 벡터로 변환하였으며, 재카드 계수(jaccard coefficient)를 적용하여 SMILES 벡터 간의 유사도를 계산하였다. 이와 같이 모든 약물 쌍에서 SMILES 유사도를 위와 같이 계산한 다음, 전체 약물 쌍 중 SMILES 유사도 기준 상위 0.1%와 하위 0.1%를 샘플링하여 해당 약물들의 타겟 단백질간 위치 벡터 유사도를 비교하였다(그림 2). 그 결과 SMILES 유사도가 높은 단백질 쌍에서 더 높은 위치 벡터 유사도가 관찰되었다. 이는 약물의 특징을 나타내는 중요 예측 지표 중 하나인 화학 구조와 단백질 위치 벡터 간에서 유의미한 상관 관계가 존재함을 나타내는 것이다.

4. 결론 및 향후 연구

본 논문은 약물의 리포지셔닝 연구에서 사용할 수 있는 새로운 예측 지표로서 단백질의 위치 벡터를 제안하였다. 단백질의 위치 정보는 단백질의 특징을 예측할 수 있는 중요한 정보 중 하나지만 지금까지 약물 리포지셔닝 연구에서 활용되지 않았다. 본 논문에서는 단백질 상호 작용 네트워크를 이용하여 단백질 위치 벡터를 전파하여 데이터의 불완전성을 보완하였다. 단백질 위치 벡터의 활용 가능성을 검증하기 위하여, 본 연구에서는 동일한 약물의 타겟 단백질 사이에서 위치 벡터의 유사도를 관찰하였으며, 기존 예측 지표 중 하나인 SMILES의 유사도와 위치 벡터 유사도 간의 연관 관계를 관찰하였다. 그 결과 두 가지 지표 모두에서 단백질 위치 벡터가 유의미한 패턴을 나타내었다. 이와 같은 결과를 종합하여 볼 때, 단백질의 위치 정보가 약물의 리포지셔닝 연구에서 예측 지표의 하나로서 이용 가능성이 있음을 판단할 수 있었다.

향후 연구로서, 단백질의 위치 정보를 보다 많은 다른 예측 지표와 비교 분석함으로써, 활용 가능성을 더욱 심도 있게 분석할 예정이다. 향후 분석의 결과로 활용 가능성이 높게 판단될 경우 단백질 위치 정보를 이용한 약물 리포지셔닝 연구를 연속해서 수행할 예정이다.

5. 참고문헌

[1] Pammolli, Fabio, Laura Magazzini, and Massimo Riccaboni. “The productivity crisis in pharmaceutical R&D.“ Nature Reviews Drug Discovery, vol. 10, no. 6, pp. 428-438, 2011.

[2] Gottlieb, Assaf, et al. “PREDICT: a method for inferring novel drug indications with application to personalized medicine.“ Molecular systems biology, vol. 7, no. 1, 2011.

[3] Park, Solip, et al. “Protein localization as a principal feature of the etiology and comorbidity of genetic diseases.“ Molecular systems biology vol. 7, no. 1, 2011.

[4] Law, Vivian, et al. “DrugBank 4.0: shedding new light on drug metabolism.“ Nucleic acids research, vol. 42, no. D1, D1091-D1097, 2014.

[5] Magrane, Michele. “UniProt Knowledgebase: a hub of integrated protein data.“ Database: The Journal of Biological Databases & Curation, vol. 2011, no. 9, 2011.

[6] Brown, K.R., and Jurisica, I., “ Unequal evolutionary conservation of human protein interactions in interologous networks.“ Genome Biology, vol. 8, no. 5, R95, 2007.

[7] Chatr-aryamontri, Andrew, et al. “The BioGRID interaction database: 2013 update.“ Nucleic acids research, vol. 41, no. D1, D816-D823, 2013.

[8] Vanunu, Oron, et al. “Associating genes and protein complexes with disease via network propagation.“ PLoS computational biology, vol. 6, no. 1, e1000641, 2010.

[9] Steinbeck, Christoph, et al. “Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics.“ Current pharmaceutical design, vol. 12, no. 17, 2111-2120, 2006.