그래프 신경망을 사용한 약물반응 예측의 동향과 챌린지

Trends and Challenges in Drug Response Prediction Using Graph Neural Networks

김찬영 [1], 신용민 [1], 표준희 [2,3], 신원용 [1,4]

[1] 연세대학교, [2]㈜디파이브테라퓨틱스, [3] 충북대학교 [4] 포항공과대학교

fred114@yonsei.ac.kr, jordan3414@yonsei.ac.kr, stdpjh@gmail.com, wy.shin@yonsei.ac.kr

Abstract

여러 종류의 암세포주 (cell line)에 대해 약물 치료제의 반응을 예측하는 약물반응 예측 (DRP: Drug Response Prediction) 연구는 의약학계에서 매우 중요한 주제이다. 최근 학계에서는 해당 문제를 그래프 신경망 (GNN: Graph Neural Network)을 사용하여 해결하고자 하는 노력이 대두되고 있다. 관련된 모델들은 다양한 유전체 데이터와 약물 데이터를 각각 그래프 형식으로 치환한 후에 GNN 방법론을 활용하여 학습시킨다. 그렇게 얻어진 임베딩 벡터를 활용하여 IC50 (inhibitory concentration) 혹은 AUC (Area Under the dose–response Curve) 수치를 예측한다. 본 논문에서는 해당 연구들에 관련된 데이터셋과 방법론에 관하여 소개하고 향후 모델의 해석력과 관련된 연구의 방향에 대해 제시하고자 한다.

Ⅰ. 서론

특정 질병에 대해 적절한 치료제를 사용하는 것은 의학 분야에서 굉장히 중요한 일이다. 특히 어떤 암세포주 (cell line)에 어떤 약물이 효과적인지에 대한 정보는 환자의 치료에 있어 큰 영향을 미친다. 약물예측 반응 (DRP: Drug Response Prediction)은 암세포주에 대한 약물의 반응 정도를 예측하는 연구이다. 최근에는 DRP 분야에 그래프 신경망 (GNN: Graph Neural Network)을 활용한 예측 연구가 활발히 진행되고 있다. 본 연구에서는 현재까지 DRP 연구에서 그래프 신경망의 활용 동향과 챌린지에 대해 살펴보고자 한다.

Ⅱ. 데이터셋

DRP 연구에서 사용하는 벤치마크 데이터셋은 크게 세 가지로 나누어진다. 첫번째는 암세포주 유전체 데이터셋, 두번째는 약물에 관한 약 데이터셋, 그리고 마지막은 각 암세포주에 대한 약물의 반응 정도를 가지고 있는 반응 데이터셋이다. 각 데이터셋들은 모두 테이블 형식으로 주어진다.

각각의 암세포주들의 정보를 담고 있는 유전체 데이터셋이 있다. 해당 데이터는 암세포주의 다양한 유전체 및 후생 유전학 정보 (gene expression profile, gene mutation, copy number variation, DNA methylation 등)를 담고 있다 [1]. 또한, 단백질들의 기능적 상호 작용 관계를 나타내는 Protein-Protein Interactions (PPI) 데이터셋, 유전체의 sequential 한 생화학/대사 경로 관계를 나타내는 pathway 데이터셋 등이 있다. 해당 데이터들은 GDSC, CCLE, STRING, KEGG 등의 데이터베이스에서 검색 및 다운로드 할 수 있다 [2].

약물 데이터셋에는 약물을 구성하는 분자의 구조를 나타내는 데이터와 각 분자를 구성하는 원자의 특징에 대한 정보를 담고 있는 데이터 셋이 있다. 해당 데이터들은 Pubchem, Deepchem 등에서 검색 및 다운로드할 수 있다 [3].

반응 데이터셋은 행에 암세포주 정보, 열에는 약물 정보가 있는 매트릭스 데이터이다. 각각의 행과 열이 만나는 셀에는 해당 암세포주와 약물의 반응정도에 대한 값인 IC50 (inhibitory concentration; 반수 최대 억제 농도) 혹은 AUC (Area Under the dose–response Curve)가 있다. IC50 은 특정 억제 물질이 특정 생물학적 과정 혹은 성분을 50% 억제하는데 필요한 양이다. 통상 항암제 기초 연구에서는 암세포의 성장을 50% 억제하는 약물의 농도로 정의한다. AUC 는 분류 문제에 사용되는 AUC-ROC 측도이다. 해당 데이터들은 GDSC 등에서 검색 및 다운로드할 수 있다.

Ⅲ. 그래프 신경망 모델

DRP 분야에 그래프 신경망을 활용한 연구들의 모델들은 위의 데이터셋들이 가진 특징들을 선택적으로 사용하여 학습이 된다. 그러나 모델들의 구조는 큰 틀에서 다음의 방법을 따른다 [3]. 유전체 데이터셋을 활용하여, gene expression, gene mutation, copy number variation 등의 암세포주내 각 유전체의 발현 및 변이에 대한 특징과 PPI, gene pathway 등과 같은 각 유전체 및 단백질 간의 상호작용을 나타내는 특징들을 이용하여 그래프 구조를 만들어 벡터 형태로 임베딩한다. 이 때, 필요한 특징들만 선택하여 사용한다. 아래 표 1은 각 모델 별로 사용한 데이터의 종류에 대한 예시를 보여준다 [2]. 또한, 약물 데이터셋을 활용하여 약물내의 분자의 구조와 각 분자를 이루는 원자들의 특징을 이용하여 벡터 형태로 임베딩한다. 그 후, 임베딩 된 각 암세포주와 각 약물의 벡터들을 이용하여 반응 데이터셋의 IC50 혹은 AUC 값과 비교하여 학습시킨다.

Method	Input
DeepDSC	△
CDRScan	○
tCNNs	○ □
GraphDRP	○ □
DeepCDR	○ △ □
MOLI	○ △ □

○ : Gene expression △ : Gene mutation □ : Copy Number Variation

표 1. 약물반응 예측 모델 별 사용 데이터 특징

이렇게 각 암세포주에 대한 약물의 반응정도를 잘 예측하는 것은 컴퓨터 과학적으로도 생물학적으로도 유의미한 결과이다. 하지만 여기서 더 나아가 신약 및 바이오마커 (biomarker) 개발과 임상 단계로의 translation 을 포함한 의약학 분야에서 정밀한 판단과 유의미한 결과를 가져다주는 것은 모델의 해석력 즉 모델이 반응 정도를 특정 값으로 예측한 이유에 대한 해석 가능여부에 있다. 즉 각 암세포주에 대한 약물 반응성이 차이를 보이는 원인을 알아내는 것이다. 이는 현재 딥러닝 분야의 XAI (explainable AI) 트렌드에도 부합한다. 따라서 향후 연구에는 DRP 모델의 해석에 관한 연구가 필요한 이유이다.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345).

REFERENCES

[1] T. Chu, T. T. Nguyen, B. D. Hai, Q. H. Nguyen, T. Nguyen, “Graph transformer for drug response prediction,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, to appear.

[2] R. Feng, Y. Xie, M. Lai, D. Z. Chen, J. Cao, and J. Wu, “AGMI: Attention-guided multi-omics integration for drug response prediction with graph neural networks,” in *Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Virtual Event, December 2021, pp. 1295-1298.

[3] F. A. Moughari and C. Eslahchi, “A computational method for drug sensitivity prediction of cancer cell lines based on various molecular information,” *PLoS One*, vol. 16, no. 4, pp. 1-31, April 2021.