다중 데이터 융합 기반 약물 부작용 예측 모델: 약물구조, 텍스트, 유전자 발현 데이터의 통합 분석

양아연 [1], 윤유진 [1], 고윤희 [2]
[1] 한국외국어대학교 바이오메디컬공학부 학부생
[2] 한국외국어대학교 바이오메디컬공학부 교수

bme22yangayeon@hufs.ac.kr, 202202277@hufs.ac.kr, younko@hufs.ac.kr

Multi-modal Integration of Enhanced Drug Side Effect Prediction using Structure, Text, and Gene Expression

A-yeon Yang[1], Yoo-Jin Yoon[1], Younhee Ko[1]
[1]Dept. of BioMedicalEngineering, HanKuk University of Foreign Studies

요 약

본 연구는 약물 부작용 예측의 정밀도 향상을 목표로, 약물 구조 정보, 부작용 텍스트 표현, 그리고 약물 처리에 따른 유전자 발현 데이터를 통합하는 다중 데이터 융합 접근법을 제안한다. 이질적인 정보를 통합해 입력 데이터로 구성한 후, 단일 예측 모델을 구축하고 특정 약물에 따른 부작용 발생 여부를 예측하는 딥러닝 모델을 개발했다. 특히, 다양한 데이터 조합을 통해, 모델 성능을 체계적으로 비교 분석하여, 각 데이터들이 약물 부작용을 예측에 미치는 영향을 분석했으며, 이를 통해 유전자 발현 데이터의 통합이 부작용 예측 성능에 중요한 영향을 미치는 것을 확인했다.

1. 서론

약물 부작용 예측은 신약 개발 시 개발 비용 절감과 환자 안전 강화를 위해 필수적인 과제로 부각된다.

초기 연구는 약물 구조식 유사도 기반 상호작용 예측[1] 및 약물-표적 네트워크 분석 등 화학 구조 기반 단일 모달리티에 집중되었다.

그러나 최근 약물 구조뿐만 아니라 전사체 및 세포 반응성 데이터를 통합하는 다중 데이터 융합 접근법이 주류로 자리 잡고 있다[2].

본 연구는 다양한 벡터화 기법과 데이터 조합을 적용해 약물 부작용 예측 모델을 구축하고, 각 조합별 성능의 비교 및 분석을 목표로 한다.

특히, 약물 구조 정보와 부작용 텍스트 표현 및 유전자 발현 변화를 통합해 예측 정밀도 향상을 모색했으며, 자연어 처리(NLP) 기반 텍스트 마이닝과 그래프 신경망(GCN)을 융합해 약물-부작용 간 복잡한 관계를 효과적으로 반영하는 모델을 제안하였다.

2. 데이터 수집 및 전처리

1) 약물의 구조 정보

본 연구에서는 MCF7 유방암 세포주에 대해 1728 개 약물의 구조 정보를 두 가지 방식으로 임베딩하여 사용했다. 첫 번째로, SMILES 를 기반으로 2048 비트 ECFP4 fingerprint 를 생성한 후, 약물 간 Tanimoto similarity 를 계산해 similarity matrix 를 구축하고, 이를 PCA 를 통해 464 차원 (95% 누적 분산)으로 축소했다. 두 번째로, Mol2vec 사전 학습 모델[3]을 이용해 300 차원의 연속적 벡터로 약물을 임베딩했다. 이렇게 구축된 구조 벡터들은 이후 부작용 예측 모델의 입력 특징으로 활용되었다.

2) 약물의 부작용 정보

SIDER(Side Effect Resource, 4.1 버전), FEARS(FDA Adverse Event Reporting system)의 공개 데이터베이스를 활용해 구축했다. SIDER 와 FAERS 는 약물과 부작용 간의 관계 정보를 제공했으며, 이 중 FEARS 는 Mendeley Data[4]에서 제공하는 노이즈가 제거된 정제된 데이터를 활용했다. 두 부작용 표기는 국제 의약 용어 사전인 MedDRA(Medical Dictionary for Regulatory Activities)에서 표준화된 용어만을 사용했다. 이렇게 수집한 부작용 정보는 2 가지의 NLP 기반 기법을 적용해 벡터화했다. Word2Vec 기반으로 200 차원의 부작용 임베딩 벡터를 형성했으며 이는 부작용 텍스트 간 의미적 유사성 반영을 위해 일반적인 코퍼스에서 사전 학습된 모델을 활용했다. 또한 BioBERT 기반 768 차원의 임베딩 벡터도 생성했다. 이는 생물학적 도메인에 특화된 문서들을 중심으로 사전 학습된 언어 모델로, 부작용 표현 간 더 깊은

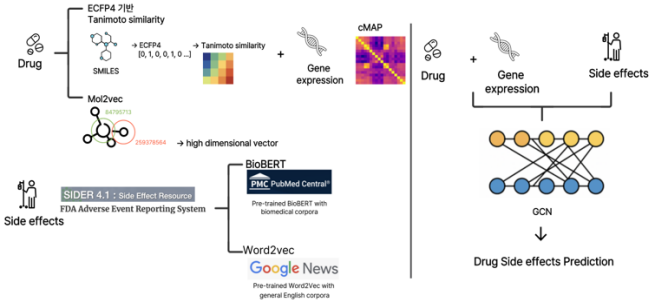
의미적 관계의 포착을 가능하게 했다.

3) 약물 유전자 발현 정보

약물 처리 후 세포 수준의 유전자 발현 변화를 반영하기 위해 LINCS L1000 프로젝트에서 제공하는 CMap(Connectivity Map) 데이터를 활용했으며, MCF7 세포주에 대한 유전자 발현 데이터(GSE70138)를 사용했다. 각 약물 처리 조건에서 측정된 주요 landmark 유전자들의 발현량으로, Incremental PCA 를 통해 전체 분산의 95%를 설명하는 총 294 차원의 입력 데이터로 차원 축소하여 사용했다.

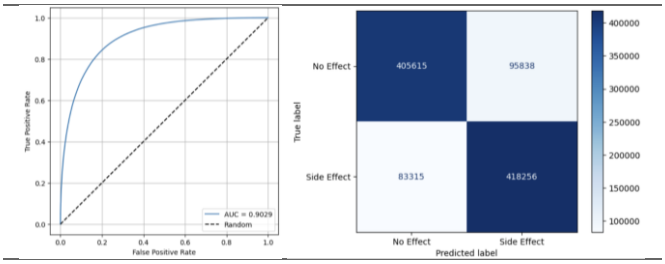
3. 모델 및 성능

본 연구에서는 약물 부작용 예측을 위해 다양한 입력 벡터를 조합한 모델을 구축했다. 입력 데이터는 약물 구조 기반 벡터, 부작용 임베딩 기반 벡터, 그리고 유전자 발현 벡터로 구성되고 이러한 이질적인 데이터들이 하나의 통합 입력으로 합쳐진 후, 다층 퍼셉트론(MLP)기반 분류 모델의 입력으로 들어가 학습에 이용되었다.



(그림 1) 모델 flow chart

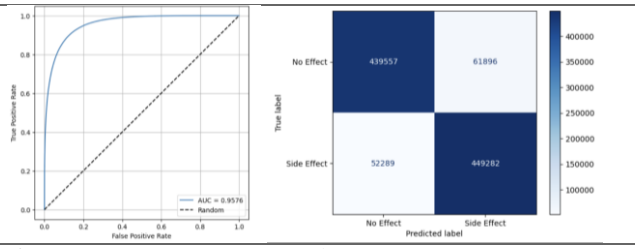
모델의 성능 평가는 ROC-AUC(Receiver Operating Characteristic Area Under the Curve)를 이용해 수행하였다. Train 및 test data 는 Positive:Negative = 1:1 비율로 구성하였으며, Nagative 의 경우 random sampling 을 통해 생성되었다. 여기서 Positive 는 약물과 부작용 간의 관계가 이미 알려진 경우를 의미하며, Negative 는 해당 관계가 알려지지 않은 약물-부작용 쌍으로 구성되었다. 약물에 대한 유전자의 발현 패턴을 나타내는 데이터를 포함하지 않은 모델의 성능의 ROC-AUC 는 0.9029 로 (그림 2), 약물에 따른 유전자들의 발현 데이터를 함께 이용한 모델에 비해 (그림 3) 상대적으로 낮은 값을 기록하였으며, 이를 통해, 약물 구조 및 부작용 임베딩 정보만으로는 약물의 부작용 예측이 쉽지 않음을 확인할 수 있었다.



(좌) ROC-AUC : 0.9029, (우) Confussion Matrix

(그림 2) 유전자 발현 벡터를 포함하지 않는 모델 성능, 약물 벡터 Mol2vec + 부작용 벡터 BioBERT 조합

그림 3 과 같이 유전자 발현 벡터를 추가로 포함한 모델 성능은 기존의 모델에 비해 약물 예측에 있어 ROC-AUC 는 0.9576 으로 크게 상승됨을 확인할 수 있었다. True Positive 와 True Negative 예측이 더욱 정확하게 이루어져 전반적인 예측 성능이 향상되었다. 이는 약물에 따른 유전자의 발현 정보가 약물의 실제 약리학적 기전 및 생물학적 반응을 반영함으로써, 약물의 부작용 예측의 정확도 향상에 크게 기여했음을 확인할 수 있었다.



(좌) ROC-AUC : 0.9576, (우) Confussion matrix

(그림 3) 유전자 발현 벡터를 포함하는 모델 성능, 약물 벡터 Mol2vec + 부작용 벡터 BioBERT + 유전자 발현 조합

4. 결론 및 향후 연구 방향

본 연구는 다중 데이터 융합을 통해 약물 부작용 예측 성능을 유의미하게 향상시켰으며, 유전자 발현 데이터가 성능에 크게 기여함을 확인하였다.

향후 연구에서는 약물의 물리화학적 특성까지 반영할 수 있는 벡터를 도입하여, 예측 모델의 성능을 한층 더 고도화할 예정이다.

5. 사사

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2023R1A2C1007756).

참고문헌

[1] Quang, D., Xie, X., "DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences", *Nucleic Acids Research*, 44(11), e107, 2016.

[2] Yeh, A., Wang, Z., "Explaining Protein Language Model Embeddings through Biological Label Space", *arXiv preprint*, arXiv:2503.02781v1, 2025.

[3] Sabrina Jaeger, Simone Fulle, Samo Turk, Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition, Journal of Chemical Information and Modeling, 58, 1, 27-35, 2018.

[4] Yi Zeng, Tuo Shi, Yifan Peng, Shiqi Wang, Suzhen Wang, Qi Li, Cleaned FAERS Adverse Events Data, Mendeley Data, V1, 2021.