약물 재창출을 위한 양방향 크로스 어텐션 기반 약물 반응 예측 모델

조윤주, 김정현
세종대학교
yoonju0301@sju.ac.kr, j.kim@sejong.ac.kr

Drug Response Prediction Model Based on Bidirectional Cross Attention for Drug Repurposing

Yoonju Cho, Junghyun Kim
Sejong Univ.

요 약

본 논문은 암 세포주에 대한 약물 반응을 정확하게 예측하기 위해 양방향 크로스 어텐션 메커니즘을 도입한 모델을 제안한다. 제안 모델은 암 세포주 잠재 표현과 약물 잠재 표현 사이의 양방향 크로스 어텐션을 통해 두 데이터 간의 복잡한 상호작용을 학습하도록 설계되었다. 실험 결과, 제안 모델은 기존 모델 대비 예측 성능과 일반화 능력이 전반적으로 향상되었으며, 모든 평가 지표에서 일관되게 개선된 성능을 보였다.

Ⅰ. 서론

암 치료를 위한 신약 개발은 중요한 과제이지만, 새로운 항암제를 설계하는 과정은 막대한 비용이 들고 오랜 시간이 걸린다. 이러한 문제를 해결하기 위해 약물 재창출이 신약 개발의 중요한 전략으로 주목받고 있다. 약물 재창출이란 이미 시판되었거나 임상 단계에서 안정성이 검증된 약물을 새로운 질병의 치료제로 사용하는 방법이다. 최근에는 약물 재창출 후보를 효율적으로 발굴하기 위해 딥러닝 기법을 활용한 약물 반응 예측 연구가 활발히 이루어지고 있다.

기존 연구 중 하나인 drug repurposing using multi-omics data integration with autoencoders (DeepDRA) [1]는 암 세포주 데이터와 약물 데이터를 활용한 오토인코더 기반 약물 반응 예측 모델을 제안하였다. DeepDRA [1]는 오토인코더를 통해 세포주 데이터와 약물 데이터의 차원을 축소하여 잠재 표현을 생성하고, 이를 연접한 뒤 분류기에 입력하여 약물 반응을 예측한다. 그러나 두 데이터를 독립적으로 처리한 뒤 단순 연접하는 방식은 두 데이터 간의 관계를 충분히 반영하지 못한다는 한계가 존재한다.

이러한 한계를 해결하기 위해 본 논문에서는 양방향 크로스 어텐션 (bidirectional cross attention) 메커니즘을 도입한 모델을 제안한다. 제안 모델은 세포주 잠재 표현과 약물 잠재 표현 간의 양방향 크로스 어텐션을 통해 세포주와 약물 사이의 상호작용을 효과적으로 학습하여 예측 성능을 향상시켰다. 실험 결과, 제안 모델은 기존 모델 [1]에 비해 복잡도가 증가하였으나, 신약 개발 전체 과정에 소요되는 시간을 고려할 때 모델 학습 시간은 유의미한 영향을 미치지 않는다. 또한 약물 반응 예측에서는 예측 정확도 향상이 중요한 요소이므로 본 논문에서는 성능 향상의 중요성을 우선적으로 고려하였다.

Ⅱ. 본론

본 논문에서는 주요 암 데이터베이스인 GDSC [2], CTRP [3], CCLE [4]에서 수집한 세포주 및 약물 반응 데이터를 활용하여 실험을 진행하였다. 세포주 데이터는 유전자 발현 (gene expression), 돌연변이 (mutation), 복제수 변이 (copy number variation) 특징을 포함하며, 약물 반응 데이터는 각 세포주와 약물 쌍에 대한 반응 값으로 구성되어 있다. 약물 데이터는 약물 반응 데이터에 포함된 각 약물의 simplified molecular input line entry system (SMILES) 표현을 PubChem 데이터베이스에서 수집하고, RDKit 화학정보학 라이브러리를 통해 분자 기술자 (molecular descriptors)와 분자 지문 (molecular fingerprints)을 추출하여 구성하였다. 최종적으로 데이터셋은 세포주, 약물, 그리고 해당 쌍에 대한 약물 반응 값을 포함한다.

기존 모델 [1]은 세포주 데이터를 처리하는 세포주 오토인코더, 약물 데이터를 처리하는 약물 오토인코더, 그리고 약물 반응 예측을 위한 MLP 분류기로 구성된다. 오토인코더는 인코더와 디코더로 구성되며, 인코더는 입력 데이터의 핵심적인 특징을 추출하여 잠재 표현을 생성한다. 디코더는 잠재 표현을 원본 데이터로 재구성하며, 이 과정에서 재구성 손실을 계산한다. 인코더를 통해 생성된 세포주 잠재 표현과 약물 잠재 표현은 연접되어 분류기에 입력되며, 분류기는 약물에 대한 내성 또는 민감 반응을 예측한다. 이러한 구조에서 세포주 잠재 표현과 약물 잠재 표현이 별도로 학습된 후 연접되기 때문에 두 특징 간의 정보 교환이 충분히 이루어지지 않는다는 한계가 있다.

따라서 본 논문에서는 양방향 크로스 어텐션 메커니즘을 도입하여 한계를 해결하였다. 제안 모델은 세포주 오토인코더와 약물 오토인코더를 사용하여 잠재 표현을 생성한 뒤, 두 잠재 표현 간의 양방향 크로스 어텐션을 수행한다. 구체적으로, 세포주 잠재 표현을 C, 약물 잠재 표현을 D라고 할 때, 세포주를 쿼리 (query)로, 약물을 키 (key)와 벨류 (value)로 사용하는 세포주-약물 어텐션은 다음과 같이 정의된다.

$$C' = Attention(Q_C, K_D, V_D) = softmax(\frac{Q_C K_D^T}{\sqrt{d_k}}) V_D \ , \quad (1)$$

여기서 $Q_C = CW_Q^C, K_D = DW_K^D, V_D = DW_V^D$이며, W_Q^C는 세포주 표현에 대한 학습 가능한 가중치 행렬이고, W_K^D, W_V^D는 약물 표현에 대한 학습 가능한 가중치 행렬이다. 또한 d_k는 키 벡터의 차원을 나타낸다. 이를 통해 세포주 표현은 약물의 정보를 반영하여 업데이트된다. 반대로 약물을 쿼리로, 세포주를 키와 벨류로 사용하는 약물-세포주 어텐션은 다음과 같이 계산된다.

$$D' = Attention(Q_D, K_C, V_C) = softmax(\frac{Q_D K_C^T}{\sqrt{d_k}}) V_C \ , \quad (2)$$

여기서 $Q_D = DW_Q^D, K_C = CW_K^C, V_C = CW_V^C$ 이며, W_Q^D는 약물 표현에 대한 학습 가능한 가중치 행렬, W_K^C, W_V^C는 세포주 표현에 대한 학습 가능한 가중치 행렬을 나타낸다. 이 과정에서 약물 표현은 세포주의 정보를 참조하여 업데이트된다.

어텐션 연산 후에는 트랜스포머 인코더 구조 [5]의 잔차 연결 (residual connection), 정규화 (normalization), 피드포워드 (feed forward) 과정이 적용되어 깊고 안정적인 학습이 가능하다. 최종적으로 분류기는 오토인코더를 통해 생성된 세포주 잠재 표현과 약물 잠재 표현, 그리고 양방향 크로스 어텐션을 통해 생성된 표현들을 연접하여 입력으로 받는다. 이를 통해 모델은 원본 특징과 상호작용 특징을 모두 고려하여 더 정확한 약물 반응을 예측할 수 있다. 제안 모델의 구조는 그림 1에 나타나 있다.

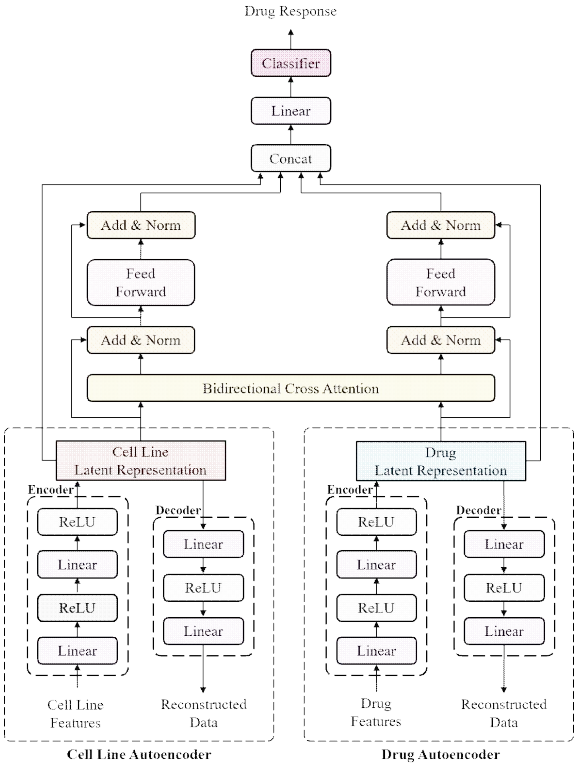

그림 1. 제안하는 약물 반응 예측 모델 구조

오토인코더와 분류기를 동시에 학습하기 위해 손실 함수는 오토인코더의 재구성 손실과 분류기의 예측 손실을 결합하여 정의하였다. 전체 손실 함수 L은 세포주 오토인코더의 손실 L_{C_AE}, 약물 오토인코더의 손실 L_{D_AE}, 분류기 손실 L_{CLS}의 합으로 구성된다. 약물 오토인코더와 세포주 오토인코더의 손실 함수로는 평균 제곱 오차 (mean squared error, MSE)를 사용하였고, 분류기의 손실 함수로는 이진 교차 엔트로피 (binary cross entropy, BCE)를 사용하였다.

$$L = L_{C\_AE} + L_{D\_AE} + L_{CLS} \ . \quad (3)$$

모델의 성능과 일반화 능력을 평가하기 위해 두 가지 실험을 진행하였으며, 성능 평가 지표로는 Accuracy, Precision, Recall, F1-score, AUC, AUPRC를 사용하였다. 첫 번째 실험에서는 CTRP와 GDSC 데이터셋을 결합하여 학습 및 테스트 데이터로 사용하고, 5-fold 교차 검증을 통해 모델의 예측 성능을 평가하였다. 실험 결과는 표 1에 나타나 있다. 제안 모델은 Precision을 제외한 모든 평가 지표(Accuracy, Recall, F1-score, AUC, AUPRC)에서 기존 모델 [1] 대비 성능 향상을 보였다.

표 1. CTRP+GDSC 데이터셋에서의 5-fold 교차 검증 결과

	기존 모델 [1]	제안 모델
Accuracy	0.953	0.956
Precision	0.945	0.939
Recall	0.96	0.969
F1-score	0.952	0.954
AUC	0.992	0.993
AUPRC	0.992	0.993

두 번째 실험은 CTRP와 GDSC를 결합한 데이터셋으로 모델을 학습시키고, 학습에 사용하지 않은 데이터셋인 CCLE로 테스트하는 교차 데이터셋 실험을 수행하였다. 실험 결과는 표 2에 나타나 있다. 제안 모델은 모든 평가 지표에서 개선된 결과를 보였으며, 특히 기존 모델 [1] 대비 AUC가 3.06%, AUPRC가 2.93% 향상되어 우수한 일반화 능력을 보였다.

표 2. CTRP+GDSC 데이터셋으로 학습 후 CCLE 데이터셋으로 테스트한 결과

	기존 모델 [1]	제안 모델
Accuracy	0.804	0.829
Precision	0.832	0.842
Recall	0.823	0.850
F1-score	0.823	0.846
AUC	0.883	0.910
AUPRC	0.887	0.913

Ⅲ. 결론

본 논문에서는 암 세포주와 약물 간의 관계를 효과적으로 학습하기 위해 양방향 크로스 어텐션을 적용한 약물 반응 예측 모델을 제안하였다. 제안 모델은 5-fold 교차 검증 실험에서 기존 모델 [1] 대비 대부분의 평가 지표에서 향상된 성능을 보였으며, 특히 학습에 사용하지 않은 데이터셋을 이용한 교차 데이터셋 실험에서 우수한 성능을 달성하며 강력한 일반화 능력을 입증하였다. 본 연구 결과는 암 세포주에 대한 정확한 약물 반응 예측을 제공함으로써 약물의 새로운 치료 가능성을 탐색하고, 약물 재창출 연구에 크게 기여할 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2025년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2025-RISE-01-019-04)

참 고 문 헌

[1] T. Mohammadzadeh-Vardin, A. Ghareyazi, A. Gharizadeh, et al., "DeepDRA: Drug repurposing using multi-omics data integration with autoencoders," PLOS ONE, vol. 19, no. 7, e0307649, Jul. 2024.

[2] W. Yang, J. Soares, P. Greninger, et al., "Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells," Nucleic Acids Research, vol. 41, no. D1, pp. D955 - D961, Jan. 2013.

[3] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, et al., "Harnessing connectivity in a large-scale small-molecule sensitivity dataset," Cancer Discovery, vol. 5, no. 11, pp. 1210 - 1223, Nov. 2015.

[4] J. Barretina, G. Caponigro, N. Stransky, et al., "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," Nature, vol. 483, pp. 603 - 607, Mar. 2012.

[5] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in Proc. of the 31st Int. Conf. on Neural Information Processing Systems (NeurIPS), pp. 6000 - 6010, Dec. 2017.