Annual Conference of KIPS

CT-50 (Cell Type IC50 Regression): 딥러닝 모델을 이용한 세포 특이적 약물 반응성 예측

이준혁[1,2], 고윤희[2]

[1]한국외국어대학교 국제통상학과, [2]한국외국어대학교 바이오메디컬공학부

Background

■ NGS 기술의 발전으로 맞춤형 치료의 중요성이 부상하면서, 개개인의 유전체 정보에 기반해 환자 특이적인 약물 반응성을 찾아내는 연구가 활발해졌다. 특히 암 치료에 있어서 질병의 진단이나 예후 예측, 치료제 선택 및 약효 예측을 하는 것이 매우 중요한 문제로 떠오르게 되었다. 유전자들이 특정 약물이나 화합물들에 대한 반응이 다양한 실험을 통해 이루어지고 있으며, 이는 정교한 약물의 반응성을 예측하는데 매우 큰 기여를 하고 있다. 더 나아가 암환자들의 유전체 정보를 고려한 약물 선택은 실제 임상적으로 매우 효과가 있음을 보여주고 있다. 하지만 환자마다 모든 약물에 대한 반응성을 측정하는 것은, 비용적,시간적 한계로 인해 불가능하다. 따라서 암 치료약물에 대한 대규모 in vitro 실험 데이터를 이용해 약물의 반응성을 예측하고자 하는 연구가 진행되고 있다.

■ 그에 따라, 다양한 머신러닝(Machine Learning) 기술들을 활용한 약물 반응성 예측 모델이 만들어지고 있다. 예를 들어, Elastic net, regularized matrix factorization, kernel methods, linear regression 등과 DNN이 활용되고 있다. 하지만 현재까지 진행된 대다수의 연구들은 cell line level에서 각 약물에 반응하는 전체 유전자들의 발현 패턴을 기반으로 약물의 반응성을 예측하였다.

■ 따라서 본 연구에서는 'Cell line 특이적 유전자'를 고려하여 그에 따른 가중치를 설정하고, 이를 통해 Cell line level로 약물 반응성을 예측 하는 딥러닝 모델을 개발하였다. 본 연구에서는 세포 특이적 유전자를 선별하고, 이들이 각 약물에 반응하는 발현 패턴 및 이러한 유전자들이 가지는 돌연변이 발생 패턴 및 약물의 화학적 구조와 함께 고려함으로써, 세포 특이적인 약물의 반응성을 효과적으로 예측할수 있었다. 본 연구는 추후 개인의 유전체 데이터에 기반한, 약물의 반응성 예측이 가능하고,이를 통해 임상적으로 치료제를 선택하고, 약물의 복용량을 결정하는데 활용될 수 있을 것이다.

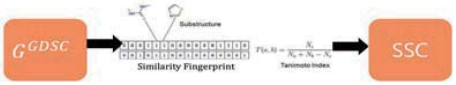
Method

■ 학습 및 시험 데이터(training & test data) :

- 입력데이터는 유전자 정보와 약물의 구조적 정보를 담고 있다. Genomics of Drug Sensitivity in Cancer (GDSC)와 The Cancer Cell Line Encyclopedia (CCLE)는 암 연구에 관련된 공개 데이터베이스이다. GDSC의 G^{GDSC} 는 조직별 약물의 민감성 수치를 담고 있으며, CCLE의 E^{CCLE}, M^{CCLE}는 cell line별 종양 관련 유전자의 발현정보와 변이정보를 담고있다.
- 민감성 수치로는 반수 최대 억제 농도를 나타내는 IC50값을 활용하였다. 유전자 발현수치는 각 cell line에서의 유전자 발현개수의 로그값으로 표현되었으며, 변이 정보는 nonsynonymous 변이를 1로, 이진법으로 표현되었다.

$$E^{CCLE} = log_2(tpm_{g,c}^{CCLE} + 1), where\ tpm\ is\ number\ of\ transcripts\ per\ million\ of\ gene\ g\ in\ cell\ line\ c$$
$$M^{CCLE} = m_{g,c}^{CCLE}, where\ m: mutation\ state$$

- 약물의 구조적 정보는 GDSC에서 제공하는 PubChem 정보를 이용하여 얻었다. The Simplified Molecular Input Line Entry System (SMILES)는 RDKit을 활용해 Morgan Fingerprint로 전환하여, 약물들의 구조적 유사성 정보를 담은 대칭 매트릭스를 입력 데이터로 활용하였다. 이는 Tanimoto distance를 사용하여 Structure Similarity of Chemicals (SSC)로 변환하여 사용하였다.



■ 모델 설계 :

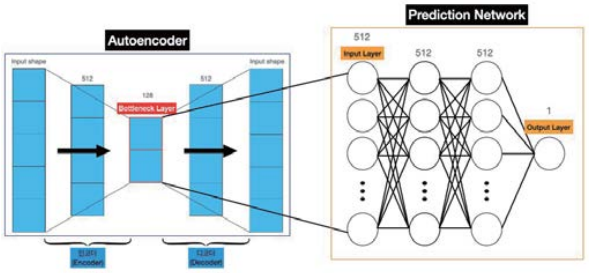
- 본 연구에서는 고차원 유전자 발현 데이터의 차원 축소를 위해 오토인코더(Autoencoder) 모델을 활용하였다. 축소된 유전정보 데이터는 약물정보 데이터와 합쳐져 MLP(Multi-layer Perceptron)의 입력데이터로 활용되어 각 약물의 IC50값을 예측하기 위해 사용되었다.

CT-50 Model

■ 모델 설계 : 오토인코더는 비지도 학습을 이용한 딥러닝 구조로, 대칭적인 인코더와 디코더로 구성되어 있다. 가장 큰 특징으로는 입력 데이터와 모델이 학습을 통해 재구성한 데이터의 차이를 최소화시키는 작업을 한다. 이는 병목 계층(Bottleneck layer)을 통한 원본 데이터의 중요한 특징을 추출하고 차원 축소에도 활용된다. 각 세부 네트워크는 세개의 층으로 이루어져 있으며 본 연구에서는 병목계층이 128차원으로, 입력데이터의 차원을 줄이는 데 활용되었다.

■예측 모델 : 오토인코더를 사용해 입력 데이터의 차원을 줄인 후, 딥러닝 모델을 사용해 IC50 값을 예측한다. Multi-Layer Perceptron 딥러닝은 세개의 은닉층으로 각 은닉층은 512개의 노드로 이루어졌다. 각 은닉층의 활성화함수로는 ELU가 사용되었으며, 드롭아웃(Dropout)은 0.3으로 설정하였다.

$$ELU(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha * (\exp(x) - 1), & \text{if } x \le 0 \end{cases}$$



Result

■해당 모델의 성능 평가지표로는 RMSE(Root mean squared error)와 R^2가 사용되었다.이는 주로 회귀분석에서 쓰이는 지표로 모델이 예측한 IC50값이 얼마나 실제 정답과 일치하는 지를 비교하는 평가방식이다. 또한 다양한 모델들의 성능을 평가하기 위해 다음의 방법들이 사용되었다. 1)입력 데이터의 차원을 줄이는 방식을 상이하게 하는 방식, 2)입력데이터 중 유전자 데이터를 하나씩 제거하면서 효과적인 입력 데이터를 찾아내는 방식, 그리고 3) 기존의 다른 모델과 비교하는 방식을 통해, 우리 모델의 성능을 평가하는 방식들이 사용되었다.

■ 데이터 차원 축소 여부 및 방법을 결정하기 위해, 원래 고차원 데이터, 오토인코더 및 PCA (Principal Component Analysis) 를 이용해 차원 축소를 한 모델을 비교하였다. 본 연구에서는 오토인코더를 사용한 모델이 가장 좋은 성능을 나타냄을 확인하였다. 또한, 유전자 발현 데이터가 실제 약물의 반응성을 예측하는데 매우 중요한 역할을 함을 확인하였다. 마지막으로 선형회귀(Linear Regression), 랜덤 포레스트(Random Forest) 모델 등과의 비교를 통해 본 논문에서 제안한 모델이 가장 효과적으로 IC50값을 예측하는 것을 확인할 수 있다.

Acknowledgement

■ 본 연구는 한국연구재단 개인기초연구사업 (2020R1F1A1069672) 과 한국외국어 대학교 개인 연구 과제의 지원을 받아 수행되었다.

Reference

■ Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol, 2014;32(12):1202-12

■ Zhaorui Zuo, Penglei Wang et al. SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. BMC Bioinfo, 2021;22(434)