KEGG 데이터베이스와 유전자 발현 데이터를 통합한 머신러닝 기반 약물반응성 예측 모델

최승연° [1], 김지태[2], 송하원[2], 최병권[2], 허예림[2], 박치현[3]
강원대학교 동물응용과학과[1]
강원대학교 의생명융합학부[2]
강원대학교 컴퓨터공학과[3]
tmddus1553@kangwon.ac.kr, nuafam@kangwon.ac.kr, gkdnjs1626@kangwon.ac.kr,
rnjsdlso@kangwon.ac.kr, yerim62345@kangwon.ac.kr, chihyun@kangwon.ac.kr

Machine learning-based drug response prediction model using KEGG database and gene expression data

Seungyeon Choi° [1], Jitae Kim[2], Hawon Song[2], Byeongkwon Choi[2], Yerim Heo[2], Chihyun Park[3]
Department of Applied Animal Science, Kangwon National University [1]
Division of Biomedical Convergence, Kangwon National University[2]
Department of Computer Science and Engineering, Kangwon National University[3]

요 약

같은 약물에 대해서도 암 환자들은 서로 다른 유전자 발현량으로 인해 다른 약물 반응성을 보일 수 있어 개인별로 적절한 항암제를 찾는 것이 중요하다. 기존의 연구에서는 이를 위해서 여러 딥러닝 방식을 통해 유전자들의 발현량만을 가지고 약물반응성을 예측한다. 하지만 본 연구에서는 환자에게 적절한 항암제를 선정할 수 있도록 pathway와 expression value를 통해 학습할 수 있는 파라미터를 가지는 모델을 제안한다. Pathway 데이터를 통해 유전자의 발현량 뿐만 아니라 유전자 간에 관계까지 학습할 수 있도록 pathway별로 독립적인 GCN(Graph Convolution Network)연산을 수행하여 IC50를 예측하는 연구가 진행됐다. 기존의 GCN 모델에서 더 나아가 여러 advanced GCN을 통한 IC50 예측은 기존의 GCN모델보다 더 높은 성능을 보였고 제안된 모델을 통해 약물반응성 예측에 중요한 biomarker 추출의 가능성까지 제시하였다.

1. 서론

암세포는 비정상적인 속도와 주기로 세포분열을 하거나, 새로운 혈관을 신생시키는 등 세포 생장과 생존에 특화되어 빠르게 증식한다. 이러한 암세포를 억제하기 위해서 사용하는 약물들은 안타깝게도 강한 항암 효과를 가진 약물일수록 강한 부작용이 따른다. 항암 화학요법에서 사용하는 항암제의 경우 정확하게 암세포만을 표적으로 작용하기 쉽지 않기 때문에, 주변 세포 혹은 정상 세포의 손상을 일으킨다. 하지만 환자들은 강력한 항암제라도 사용할 수밖에 없고, 부작용을 고려한다면, 올바른 항암제를 사용하는 것이 환자의 예후를 결정하는 중요한 요인이 된다.[1]

환자에게 투여하는 항암제를 선정하는 과정은 머리가 아플 때 두통약을 먹는 것처럼 간단한 문제가 아니다. 같은 약물이 투여된 경우에도, 개인마다 여러 가지 생화학적 요인들에 의해 모두 다른 약물반응성을 보일 수 있기 때문에 환자에게 항암제가 적절한지 검사하는 절차가 필요하다. 지금까지 많은 연구에서는 개개인의 유전자의 발현량에 초점을 두어 tree 기반 머신러닝을 통해 약물반응성을 예측하거나 각 유전자들을 node로 하는 MLP(Multi-Layer Perceptron)를 통해 예측하는 방식으로 연구를 해왔다[2]. 하지만 본 연구는 앞서 말한 연구에서 추가로 유전자의 발현량 뿐만 아니라 유전자들의 관계까지 고려하여 약물 반응성을 예측하도록 제안한다.

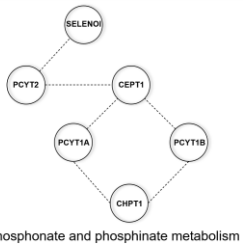
본 연구는 pathway의 네트워크가 graph와 같은 형태임에 착안하여 그래프 기반 딥러닝 기법인 GCN(Graph Convolutional network) 모델을 활용한다. 각 pathway를 하나의 graph로 설정하여 adjacency matrix를 만들고 pathway에 속한 유전자들의 발현량을 각 노드의 feature로 설정하여 입력값으로 사용했다. Pathway별로 graph convolution 연산에 의한 결과를 모아 최종적으로 FC(Fully Connected) layer의 입력값으로 사용되어 약물에 대한 세포의 생존률인 IC50 value를 예측하도록 했다. 하나의 cell line은 하나의 sample로서, 마치 한 명의 환자에 대한 약물 반응성을 예상하는 것처럼 모델을 구상하였다.

2. 데이터 수집

항암제에 대한 반응성을 평가하는 지표로 GDSC(Genomics of Drug Sensitivity in Cancer) Database[3]에서 제공하는 IC50 value를 사용하였다.

그 IC50 value는 모델의 이진 분류 학습 및 예측을 위해 sensitive 그룹과 resistance 그룹으로 구분하였다. 또한 GDSC Database에서 5-Fluorouracil에 대한 IC50 value가 구해진 801개 cell line들의 gene expression data를 얻었다. GDSC에서 얻은 각 cell line 별 RNA seq FPKM value를 TPM으로 normalization하여 사용하였다. KEGG(Kyoto Encyclopedia of Genes and Genomes) database[4]에서 각각의 KEGG pathway별로 function을 가지는 gene들을 확인하여 KEGG pathway 내에 명시된 역할을 가지는 gene이 없는 경우를 제외했다. 위 과정을 거쳐 Pyrimidine metabolism, One carbon pool by folate, Phosphonate and phosphinate metabolisms을 포함한 총 308개의 KEGG pathway를 선택하여 분석에 활용했다. 선택된 KEGG pathway의 그래프 정보를 사용함과 동시에 pathway에 속해 있는 gene과 동일한 GDSC RNAseq data의 gene의 expression값을 GCN연산에 사용되었다. GDSC에서 얻은 각 cell line 별 RNA seq FPKM value를 TPM으로 normalization하여 사용하였다.

(a). 그래프



(b). 인접행렬

	PCYT2	CEPT1	SELENOI	PCYT1A	PCYT1B	CHPT1
PCYT2	1	1	1	0	0	0
CEPT1	1	1	0	1	1	0
SELENOI	1	0	1	0	0	0
PCYT1A	0	1	0	1	0	1
PCYT1B	0	1	0	0	1	1
CHPT1	0	0	0	1	1	1

그림. 1 Pathway 그래프, 노드 간 인접행렬

그림.1과 같이 그래프는 G=(V,X,A)로 표기할 수 있다. V는 노드의 집합을 나타내며 $A \in \mathbb{R}^{n\times n}$는 간선의 집합을 나타낸다. X= $H_k^{(0)} \in \mathbb{R}^{n\times F}$ 는 노드의 특성 집합이며 $W^{(0)} \in \mathbb{R}^{F\times F'}$ 는 노드의 정보를 학습이 진행됨에 따라 업데이트 해줄 수 있는 가중치의 집합이다. 여기서 n은 그래프의 노드를 의미하고 F는 Feature Matrix의 feature를 의미하며 F' 는 Weight Matrix의 output dimension을 의미한다.

3. GCN Model 개요

Graph convolution network에서 사용된 convolution 연산은 CNN에서 사용된 convolution 연산과 동일한 개념이다. CNN에서는 동일한 weight를 가지는 filter를 통해 이미지의 local feature를 계산한다. Graph 구조에서도 같은 과정을 수행하기 위해 feature matrix에 weight matrix를 곱하여 나온 matrix에 인접행렬을 행렬곱을 하는 것이다. 그 결과로 해당 노드는 인접한 노드들의 정보가 추가되어 업데이트된다. 층이 거듭될수록 연결이 많이 된 노드들은 많은 양의 정보가 추가되는 것처럼 해당 pathway의 구조에 따라 속해 있는 노드(유전자)의 hidden state도 영향을 받게 된다.[5]

앞서 제안된 모델에서의 GCN연산은 기존에 연구되었던 GCN모델처럼 해당 네트워크의 label을 예측하는 것이 아닌 supervised learning이 이루어지는 MLP의 input 데이터를 생성하는데 목적을 둔다. 즉, 위 모델은 두 가지 정보(유전자 간의 관계와 그 유전자들의 발현량)를 포함하는 input 데이터를 만드는데 의미가 있는 것이다.

본 연구에서 제안하는 GCN Model의 전체적인 개요는 1) GCN 연산 수행, 2) Concatenation, 3) MLP의 과정에 따라 진행된다. GCN 학습에 사용되는 그래프의 vertex(node)는 유전자, edge는 KEGG Database에서 얻은 유전자 간 상호관계 여부로 설정되었다. 위의 과정은 특정 Cell line의 gene expression과 gene과의 관계를 통해 sensitive, resistance로 이진 분류하는 목적을 가진다.

3.1 GCN 연산

$$f(H_k^{(l)}, A) = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H_k^{(l)}W^{(l)}) \quad (1)$$

식(1)에서 $H_k^{(l)}$ 는 l번째의 은닉층, $\tilde{A}$는 Adjacency Matrix에 Self-Loop를 추가한 Matrix, $W^{(l)}$는 layer 별로 학습이 가능한 Weight Matrix, $\tilde{D}$는 라플라시안 정규화(Laplacian normalization)를 적용한 Degree Matrix를 의미한다. k는 308개의 pathway 중 해당 pathway를 의미한다.[6]

Pathway 별 GCN 연산 수행 시 Gene expression data에서 KEGG pathway의 노드로 구성되어 있는 유전자들만 선별하여 그 유전자들의 발현량을 $H^{(0)}$의 Feature Matrix로 설정하였다. 그림.1에서 [1.GCN]처럼 각 pathway 별 식(1)의 결과값은 마지막 과정에서 readout layer를 통해 해당 pathway의 노드 수의 크기를 가지는 벡터로 나오게 된다. 이 때 pathway별로 서로 독립적인 GCN연산이 수행된다.

이 때 layer가 깊어질수록 기울기 소실(Gradient Vanishing)로 인해 각 layer의 학습이 제대로 이루어지지 않아 학습이 제대로 되지 않을 수 있다. 이를 해결하기 위해 식(2)처럼 Residual Architecture을 추가한 Skip Connection GCN 모델을 제안한다.[7]

$$H_k^{(l,sc)} = f(H_k^{(l)}, A) + H_k^{(l)} \quad (2)$$

또한 gated recurrent unit(GRU), long-short term memory(LSTM)와 같이 recurrent cell units에 사용되는 gate 방식을 추가한 Gated Skip Connection GCN도 제안되었다. LSTM에서의 cell state update과정과 마찬가지로 이 모델에서 gate는 이전 정보에서 어느 부분을 전달할지 계산해주는 구조로 되어있다. σ 는 LSTM에서와 같이 시그모이드 함수를

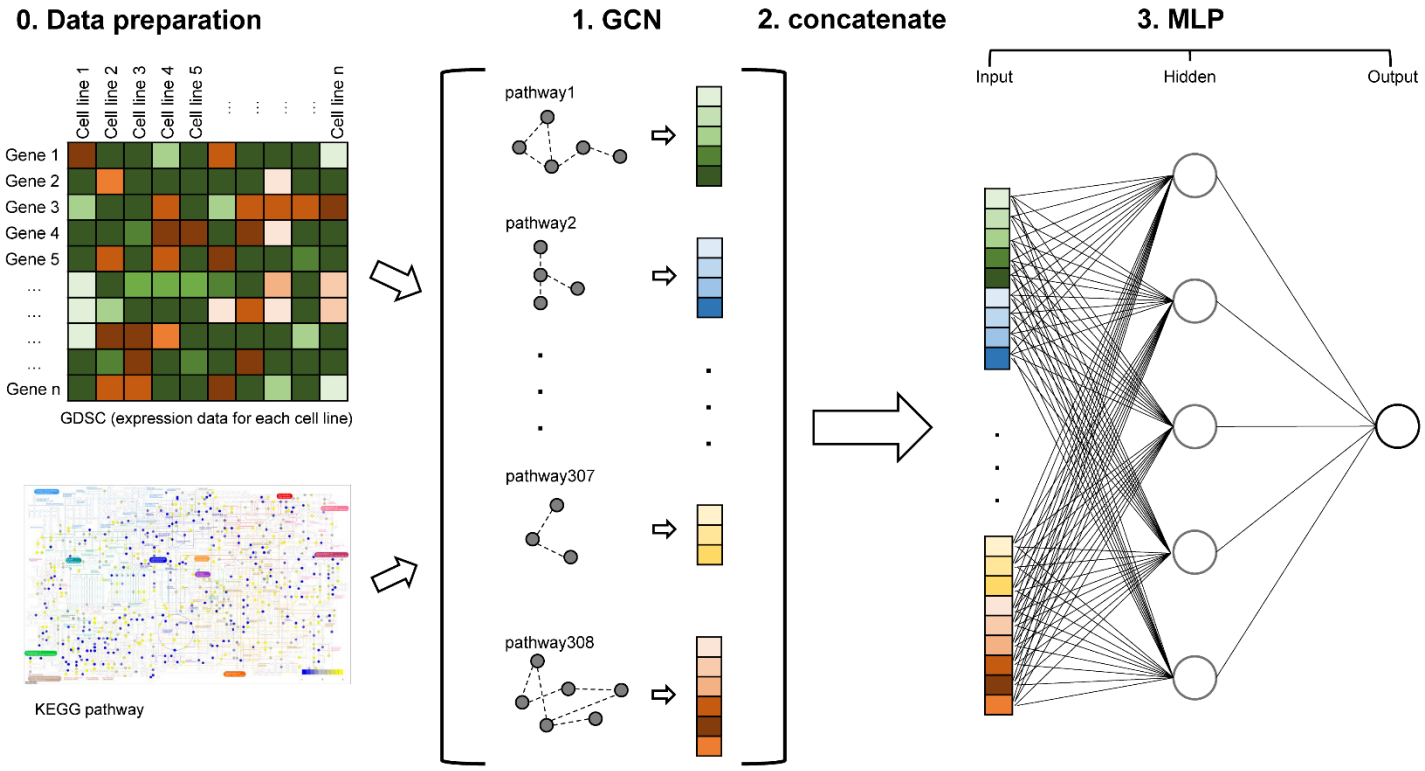


그림 2. 제안된 GCN 모델

활성함수로 사용하였고 그 범위는 0~1로 지정되었다.

$$H_k^{(l,sc)} = z_i \odot f(H_k^{(l)}, A) + (1 - z_i) \odot H_k^{(l)} \quad (3)$$

$$z_i = \sigma(W[f(H_k^{(l)}, A), H_k^{(l)}] + b_z) \quad (4)$$

기존의 GCN 연산에서는 노드의 중요도는 고려하지 않고 인접한 노드 간의 정보를 업데이트하는 방식으로 진행됐다. 이에 spatial 관점에서의 convolution GNN과 같이 인접한 노드 간의 관계에 초점을 두어 $\tilde{A}$에 노드 간에 상관관계에 대한 정보를 추가하여 유전자 간에 관계를 더 깊게 파악할 수 있는 Graph Attention 알고리즘이 제안되었다. [7]

attention coefficient는 수식(6)과 같이 구할 수 있으며 학습 시 특정 layer에서 인접 노드 간에 중요도를 계산할 수 있다. Layer가 진행되면서 i번째 노드와 j번째 노드의 관계 및 중요성이 계속 업데이트 되어 어떤 유전자가 IC50에 영향을 미치는지 측정할 수 있다.

$$f(H_k^{(l)}, A) = \sigma(a_{ij}^{(l)} \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H_k^{(l)} W^{(l)}) \quad (5)$$

$$a_{ij}^{(l)} = \tanh((H_i W)^T (H_j W)) \quad (6)$$

이와 같이 본 연구에서는 앞서 언급했던 Vanilla GCN 뿐만 아니라 여러가지 알고리즘이 추가된 Advanced GCN을 폭넓게 사용하여 IC50를 예측하고자 했다.

3.2 Concatenation

308개의 KEGG pathway 별로 GCN연산이 수행되어 나온 308개의 벡터($H_1^{(l)}, H_2^{(l)} \cdots H_{308}^{(l)}$)는 그림. 1의 [2. concatenate]처럼 25116의 크기를 가지는 하나의 벡터로 통합된다. 여기서 통합된 하나의 벡터는 MLP 연산이 수행되는 과정의 input layer로 정의된다.

3.3 MLP

최종 약물 반응(IC50)을 예측하기 위해 GCN의 결과값을 concatenation한 벡터를 다층 퍼셉트론(Multi-Layer Perceptron, MLP)으로 연결하여 IC50의 최종 분류 결과를 출력한다.

4. 실험 및 결과

본 연구에서는 Cell line 별 biological pathway를 통해 유전자 간의 관계와 그 유전자의 발현량을 통해 IC50를 예측하는 실험을 수행하였다. Training set, validation set, test set을 8:1:1 비율로 분리하고 배치사이즈를 64로 두어 총 30 epochs를 통해 제안된 모델을 학습하였고 AdamW 알고리즘을 사용하여 최적화를 진행했다. GCN의 hidden layer는 6개를 사용하였고 MLP의 4개의 hidden layer를 사용하였다. GCN, MLP의 가중치 초기화는 He Initialization 알고리즘을 사용하였다.

수행한 실험결과는 표.1과 같이 8 가지 모델의

성능을 비교하였다. 측정 기준으로는 테스트 데이터에서의 loss와 AUC가 있다.

본 연구에서는 여러가지 Graph Convolution Network를 통해 약물반응성을 가장 잘 예측할 수 있는 최적의 모델을 설계하는데 목적을 둔다. 기본적인 GCN 모델에서부터 layer가 쌓일수록 생길 수 있는 기울기 소실(Gradient Vanishing)문제를 해결해고자 residual learning을 추가했으며 더 나아가 residual learning을 통해 유지되는 gradient의 비율을 학습하는 모델까지 제안되었다. GCN의 학습과정에서 생길 수 있는 문제를 해결하는데에 그치지 않고 유전자 상호관계 여부의 정보만을 사용하지 않고 유전자 간에 관계성에 대해 학습할 수 있는 알고리즘을 추가하여 향후 biomarker 추출에 도움이 될 수 있는 방향을 제시하였다.

표.1 모델 별 테스트 데이터에 대한 성능 비교

Model		AUC
Attention	Vanilla GCN	0.62
	scGCN	0.65
	gscGCN	0.88
Not Attention	Vanilla GCN	0.61
	scGCN	0.71
	gscGCN	0.74
RandomForest		0.72
XGBoost		0.69

표.1에서 볼 수 있듯이 Attention 메커니즘과 gated skip connection 알고리즘을 혼합한 GCN 모델이 가장 높은 성능임을 알 수 있다. 그 모델을 제외한 다른 모델들은 검증 모델로서 설계된 tree기반 ensemble모델의 성능과 비슷한 것으로 나왔다. 하지만 attention 알고리즘을 사용하지 않았을 때에 gated skip connection 알고리즘은 attention을 사용했을 때에 비해 좋은 성능을 보이지 못했다. 결론적으로 개별로 attention 알고리즘과 gated skip connection 알고리즘을 사용하는데에 성능면에 큰 의미를 볼 수는 없지만 동시에 활용했을 때 의미 있는 결과를 볼 수 있었다.

학습된 모델에서 pathway 별 가중치를 확인해본 결과 그 값이 가장 높게 나온 pathway는 p53 signaling pathway로 확인되었으며 그 뒤로 Thyroid cancer pathway, Melanoma pathway 가 약물반응성을 예측하는데 중요하게 사용되었음을 알 수 있었다. p53 signaling pathway에서 TP53gene으로 암호화된 p53은 실제로 암 억제 유전자로 알려져 있으며 항암치료에 중요한 영향을 미치는 단백질이다. 이를 통해 본 연구에서 제안된 모델을 통해 잠재적 biomarker를 추출할 수 있다는 가능성을 제시한다.

5. 결론 및 향후 연구

본 연구에서는 KEGG pathway를 통한 유전자들의 관계와 그 유전자의 발현량을 통한 cell line별 약물 반응성을 예측하였다. 제안된 모델에서는 각기 다른 pathway를 개별로 GCN 연산을 수행하여 마지막에 통합하여 FC연산을 수행하는 과정을 거친다는 점에서 기존 일반적인 GCN연구와 차별점을 가진다. 해당 pathway에 대한 파라미터를 독립적으로 학습하면서 다양한 biological process에 대한 정보를 가지고 약물반응성을 예측할 수 있었다.

차후 연구에서는 attention 알고리즘과 gated skip connection 알고리즘을 동시에 사용했을 때 좋은 성능을 거둘 수 있었던 이유에 대해서 더 깊이 있게 연구하고자 한다. 또한 bioinformatics 분야의 고질적인 문제인 데이터셋의 부족으로 인한 overfitting 문제 또한 GAN(Generative Adversarial Network) 알고리즘을 통해 해결하는 주제에 대해서도 연구할 계획이다.

6. 사사

본 논문은 미래창조과학부의 재원으로 한국연구재단의 지원(2019R1I1A1A01058458)과 2020년도 강원대학교 대학회계 학술연구조성비의 지원을 받아 수행된 연구임.

참고논문

[1] Kim Ik-Yong(김익용). 대장암 항암요법의 향상된 예후와 최신 지견. 대대항지. 2006;22(2):137-48.

[2] Robert J, Vekris A, Pourquier P, Bonnet J. Predicting drug response based on gene expression. Critical Reviews in Oncology/Hematology. 2004 September 1,;51(3):205-27.

[3] Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research. 2013 January 1,;41(D1):D955-61.

[4] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 - 01-01;28(1):27-30.

[5] Sun M, Zhao S, Gilvary C, Elemento O, Zhou J, Wang F. Graph convolutional networks for computational drug development and discovery. Briefings in Bioinformatics. 2019;21(3):919-35.

[6] Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. IEEE transactions on neural networks and learning systems. 2021 January;32(1):4—24.

[7] Ryu S, Lim J, Kim WY. Deeply learning molecular structure–property relationships using graph attention neural network. CoRR. 2018;abs/1805.10988.