CTET Protein 을 사용한 Drug-Drug interaction 예측 Deep Learning Model

서지원 [1], 고윤희 [1]
[1] 한국외국어대학교 바이오메디컬공학부
jiwons7500@hufs.ac.kr, younko@hufs.ac.kr

Drug-Drug interaction predicting deep learning model using CTET protein of drugs

Jiwon Seo[1], Younhee Ko[1]
[1]Div. of Biomedical Engineering, Hankuk University of Foreign Studies

요 약

DDI(Drug-Drug Interaction)는 병원에서 발생하는 약물이상반응의 30%를 유발하는 부작용이지만, 현실적으로 모든 약물쌍의 DDI 를 기존 *in vivo, in vitro* 방식으로 예측하는 것은 불가능하다. 그렇기에, 다양한 *in silico* 방식의 DDI 예측 모델이 연구되고 있다. 본 연구에서는, 단백질 네트워크 상에서 RWR(Random Walk with Restart) 알고리즘을 통해 약물과 직접적으로 상호작용하는 단백질과 간접적으로 상호작용하는 단백질의 정보를 사용하여 DDI 를 예측하는 모델을 개발하였다. 이 모델을 통하여 기존에 발견하지 못한 DDI 를 새롭게 발견하고, 신약 개발 시에도, 신약과 함께 복용 시 문제를 일으킬 수 있는 약물을 예측하여 약물 이상반응을 방지하고자 한다.

1. 서론

약물이상반응은 약물과 관련된 모든 부작용을 의미하며 매년 200 만명의 입원환자에게 영향을 미친다. 이 중 30%는 drug-drug interaction(DDI)에 의해 발생한다.[1] DDI 에 의한 약물이상반응을 예방하는 방법은 미리 DDI 를 유발하는 약물들을 동시에 처방하지 않는 것이다. 그렇기에 DDI를 발견하는 것은 약물 개발 과정에서 매우 중요한 과정이다. 기존에 DDI 를 발견하기 위해 사용된 in vivo, vitro 방식은 시간적으로나, 경제적으로 매우 비효율적인 과정이다.

비효율적인 기존 방식의 대안으로 in silico 방식이 존재한다. 컴퓨터를 통해 진행된 in silico 방식을 사용하면, 임상실험을 하지 않더라도 DDI 에 대한 예측이 가능해지게 되며, 기존에 실험하지 못했던 약물쌍에 대한 DDI 발견도 가능해진다. 추가적으로 신약을 개발하는 과정에서 DDI 를 유발할 가능성이 있는 약물들을 미리 파악할 수 있게 된다. 이러한 이유로 다양한 in silico 방식의 DDI 예측 모델이 개발되었으며, 최근 deep learning 을 사용한 DDI 예측 모델 역시 많아지고 있다.[2][3]

DDI 를 예측하는 딥러닝 모델은 트레이닝에 필요한 데이터로 어떤 데이터를 사용하는지에 따라 달라진다. 임상실험 연구결과 데이터 사용 모델[4], 약물의 화학구조에 관한 데이터 사용모델[5][6][7], 약물의 DDI 데이터 사용모델[8], 약물이 상호작용하는 단백질 데이터를 사용하는 모델[6][9]들이 개발되었으며, 다양한 종류의 데이터들을 동시에 사용하여 DDI 를 예측하는 모델[3][6][9]도 개발되었다.

약물과 상호작용하는 단백질들은 종류에 따라서 carrier, target, enzyme, transporter, 다음과 같이 4 개로 나눌 수 있으며 이를 줄여서 CTET 단백질[9]이라고 칭한다. 하지만 단백질의 전체 개수에 비해 상호작용하는 단백질의 수가 적다. 데이터의 희소성 문제 외에도, 직접적으로 상호작용하는 단백질의 영향만 계산하게 된다.

본 연구에서는 약물과 직접적으로 상호작용하는 단백질의 정보 뿐만 아니라, 추가적으로 RWR(Random Walk with Restart) 알고리즘[10]을 사용하여, 약물과 간접적으로 상호작용하는 단백질의 정보를 사용하여 DDI 를 예측하는 MLP 구조의 딥러닝 모델을 개발하였다.

2. 데이터셋

이번 연구에서 사용된 모든 데이터는 open source

data 를 사용하였으며, drugbank[11]와 STRING[12]에서 제공하는 데이터를 사용하였다.

모델의 input 으로 사용될 약물의 CTET 단백질에 대한 데이터도 drugbank 에서 제공하는 데이터를 사용하였다. 해당 데이터는 6747 개의 약물과 3955 개의 단백질 간의 drug-protein interaction 을 사용하였다.

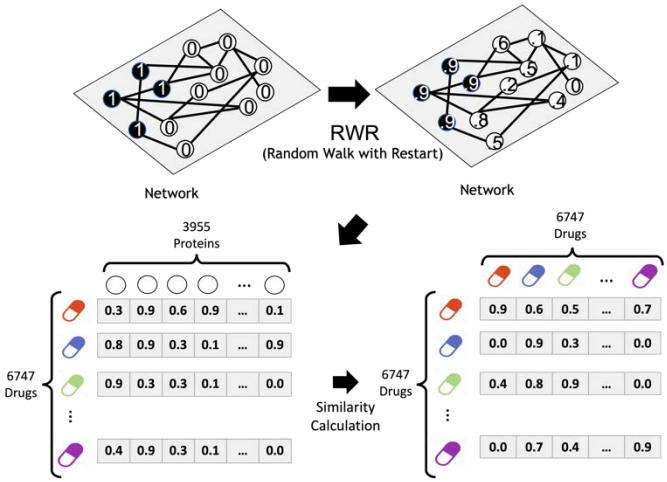
모델 학습을 위한 dataset 으로 drugbank 에서 제공하는 DDI case 데이터를 사용했다. 총 192,284 개의 DDI 중 화학구조에 대한 정보와 CTET 단백질의 정보를 가지고 있는 약물로 이루어진 80 종류의 163,574 개의 DDI 가 최종적으로 training 에 사용되었다.

CTET 단백질에 대해 RWR 을 진행하는 네트워크로는 STRING 을 사용하였다. STRING 에서는 모든 네트워크를 사용하는 것이 아닌, 약물의 CTET 단백질이 포함된 네트워크 부분에서 confidence score 이 상위 40% 인 부분만 사용하였다.

3. 약물의 CTET 단백질 정보를 사용한 Protein Similarity Profile 계산

RWR 알고리즘은 네트워크에서 seed 노드의 값을 기준으로 나머지 노드의 값들을 propagation 하는 알고리즘으로서, seed node 와 나머지 노드들 간의 correlation 의 정도를 파악할 수 있게 된다.

본 연구에서는 STRING 에서 약물의 CTET 단백질에 해당하는 노드를 seed 노드로 설정하여 RWR 을 진행하였다. 임의의 약물 A 의 CTET 단백질을 seed 노드 삼아 RWR 을 진행한 뒤, 다른 약물 B 의 CTET 단백질 노드들이 가지고 있는 값들의 평균치를 약물 A 와 B 의 PSP(Protein Similarity Profile)라 지정하였다. 이런 방식으로 모든 약물들 간의 PSP 를 계산하여 PSP matrix 를 만든 뒤, model 의 input 으로 사용하였다.



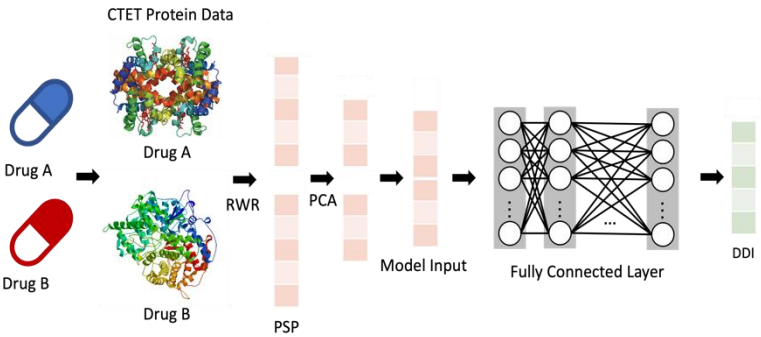
(그림 1) RWR 알고리즘 및 PSP 계산 방식. CTET 단백질을 seed node 로 설정한 뒤 RWR 로 노드의 값들을 propagate 한다. RWR 결과를 사용하여 PSP 를 계산한다.

4. Multi Layer Perceptron 구조의 모델을 사용한 Drug-Drug Interaction Classification

모델은 Multi Layer Perceptron(MLP) 형태의 모델에서 지도학습으로 모델을 학습하여 DDI 를 분류하였다. 하나의 약물의 PSP 는 해당 약물과 나머지 모든 약물의 similarity profile 에 대한 정보를 가지고 있어, PSP 벡터 하나의 차원은 약물의 개수인 6,747 이다. 약물쌍마다 이렇게 긴 차원의 벡터를 넣어주는 것은 차원의 저주에 빠질 수 있는 비효율적인 과정이라 판단하여 Principle of Component Analysis(PCA)를 사용하여 차원을 축소시켰다. PCA 란, 분포된 데이터들의 주성분(principal component)을 찾아주는 방법이다. 데이터의 분포 특성을 가장 잘 표현하는 차원을 새롭게 정의해서, 차원의 수를 줄인다. PCA 를 사용하여 약물 당 PSP 의 차원을 6,747 에서 667 로 축소시켜 입력으로 사용하였다.

Training Set 은 DDI 를 유발하는 두 약물의 PSP 가 concatenate 입력으로 각 약물의 PSP 가 들어가며. 출력으로 해당 입력을 가진 약물쌍이 각 DDI 일 확률이 출력된다. 이 중 가장 높은 출력 값을 가지는 DDI 유형이 해당 약물쌍의 DDI 유형이다. 모델 구조는 1024 의 길이를 가진 hidden layer 4 개가 들어간 형태이다. Overfitting 을 방지하기 위해 dropout 을 0.3 으로 설정하였으며, activation function 으로 ReLU 를 사용하였다. 마지막 layer 에는 Output 을 확률 값으로 얻고 thresholding 을 위해 softmax function 을 사용하였다. Model 의 optimizer 로는 adam 이 사용되었으며, learning rate 는 0.001 로 설정했다.

이렇게 트레이닝을 돌린 결과, CTET 단백질을 전부 다 사용했기에, 기존에 target 단백질만 사용한 모델에 비해서 높은 성능을 보였다. 또한 RWR 을 통해 간접적으로 연관이 있는 단백질 정보까지 활용하였기에, 직접적으로 상호작용하는 CTET 단백질만 사용하여 DDI 를 예측하는 모델에 비해서 좋은 성능을 보였다.



(그림 2) 모델 Training 과정, RWR 을 통해 계산된 PSP 가 PCA 를 통해 차원이 줄어든 뒤, fully connected layer 로 이루어진 MLP 의 input 으로 사용된다. Output 은 해당 약물쌍에 대한 각 DDI 의 확률로 반환된다.

5. 감사의 글

본 연구는 한국연구재단 개인기초연구사업 (2020R1F1A11069672)과 한국외국어대학교 개인 연구 과제의 지원을 받아 수행되었다.

참고문헌

[1] Cresswell, K., Fernando, B., McKinstry, B., and Sheikh, A. (2007) “Adverse drug events in the elderly,” *British Medical Bulletin* 83(1), pp.259-274

[2] Han, K. et al. (2022) “A review of approaches for predicting drug–drug interactions based on machine learning,” *Frontiers in Pharmacology*, 12.

[3] Zhang, C., Lu, Y. and Zang, T. (2022) “CNN-DDI: A learning-based method for predicting drug–drug interactions using convolution neural networks,” *BMC Bioinformatics*, 23(S1).

[4] Zhang, T., Leng, J., & Liu, Y. (2019). “Deep learning for drug–drug interaction extraction from the literature: a review,” *Briefings in Bioinformatics*, 21(5), 1609–1627

[5] Ryu, J.Y., Kim, H.U. and Lee, S.Y. (2018) “Deep learning improves prediction of drug–drug and drug–food interactions,” *Proceedings of the National Academy of Sciences*, 115(18).

[6] Lee, G., Park, C. and Ahn, J. (2019) “Novel deep learning model for more accurate prediction of drug-drug interaction effects,” *BMC Bioinformatics*, 20(1).

[7] Vilar, S., Harpaz, R., Uriarte, E., Santana, L., Rabadan, R., & Friedman, C. (2012). “Drug—drug interaction through molecular structure similarity analysis,” *Journal of the American Medical Informatics Association*, 19(6), 1066

[8] Vilar, S., Uriarte, E., Santana, L., Tatonetti, N. P., & Friedman, C. (2013). “Detection of Drug-Drug Interactions by Modeling Interaction Profile Fingerprints,” *PLoS ONE*, 8(3), e58321.

[9] Dere, S. and Ayvaz, S. (2020) “Prediction of drug–drug interactions by using profile fingerprint vectors and protein similarities,” *Healthcare Informatics Research*, 26(1), p. 42.

[10] Valdeolivas, A. et al. (2018) “Random walk with restart on multiplex and Heterogeneous Biological Networks,” *Bioinformatics*, 35(3), pp. 497–505.

[11] *DrugBank online: Database for Drug and Drug Target Info*, DrugBank Online | Database for Drug and Drug Target Info. Available at: https://go.drugbank.com/

[12] *String website*, STRING. STRING consortium. Available at: https://string-db.org/