단백질 다중 서열 정렬을 활용한 약물-표적 상호작용 딥러닝 모델 연구

방준일, 이세규, 홍성은, 김선욱, 김화종*
강원대학교, 바이오에이아이, *강원대학교

tkfka965@gmail.com, dltprb655@chol.com
sungkenh@gmail.com, king950411@gmail.com, *hjkim3@gmail.com

Drug-target interaction deep learning model study using protein multi-sequence alignment

Junil Bang, Segyu Lee, Sungen Hong, Seon Uk Kim, Hwajong Kim*
Kangwon National Univ., BioAI, *Kangwon National Univ.

요 약

본 논문은 약물-표적 상호작용을 위한 딥러닝 모델의 성능 증대를 위한 연구를 소개한다. 기존 연구된 딥러닝 모델들에서 확인 할 수 있었던 데이터 손실적인 측면에서 접근하여 모든 시퀀스 데이터를 활용할 수 있는 방법으로서 단백질 시퀀스 데이터의 다중서열 정렬을 활용한 딥러닝 프로세스를 연구하였으며, 현재 진행된 다중서열정렬 실험까지의 과정을 소개하였다.

Ⅰ. 서 론

딥러닝 기법을 통해 비용 및 시간적 측면에서 인간이 직접 하기 힘든 업무를 인간과 비슷하거나 더 뛰어난 결과를 보여주게 되면서, 여러 분야에서 연구 및 사용되기 시작했다. 특히, 화학이나 생물학 분야에서 사용되는 방법으로, 새로운 선도물질 발굴이나, 약물 용도 예측을 위한 약물-표적 상호작용 예측 등의 여러 방안으로 연구되고 있으며, 실제 해당 방법들을 통해 컨설팅 및 서비스를 하는 인공지능 신약개발 기업들이 늘어나고 있다. 그 중, 약물-표적의 상호작용을 예측하기 위한 방안으로 인공지능 모델을 사용한 스코어 예측이 사용되기 시작하면서, 예측 성능을 높이기 위한 여러가지 딥러닝 모델 구조들이 사용되기 시작했다. [1,2]

현재까지 좋은 성능을 보여준 타 논문에 게재된 연구의 경우, 약물의 정보를 얼마나 잘 표현하고 반영하는지가 성능에 높은 영향을 주었다는 결과를 보여주었다.[3] 해당 방법을 위해 어텐션 기법을 활용하여 약물의 정보를 더욱 반영할 수 있는 딤러닝 모델이 소개되었으며, 단백질의 경우 단순 CNN 을 활용한 기본적인 특징 추출로 동작하게 된다. 약물정보의 초기 임베딩을 MBERT 라고 명명된 Transformer 를 활용한 초기 임베딩 작업 후, FeedForward 와 Self-Attention 레이어를 추가하여 DTI 를 위한 Molecule 파트의 벡터로서 임베딩하는 방법을 사용하였다.[4, 5]

이러한 연구와는 다르게 추가적으로 단백질의 정보를 보다 잘 표현 및 반영하는 모델의 추가적인 활용을 통해, 보다 높은 성능의 예측 모델을 만들 수 있을것이라 가정하고 연구하였다.

Ⅱ. 본론

본 논문에서는 약물-표적 상호작용 예측을 위해 단백질의 정보를 보다 잘 표현할 수 있는 임베딩 방법을 연구하였다. 기존 연구되는 딥러닝 모델은 단백질을 1 차원으로 표현한 Fasta 포맷 형태의 서열로 표현된다.[6,7]이러한 1 차원 형태의 서열 즉, 단백질 시퀀스를 자연어처리의 최신 모델인 Transformer 를 통해 임베딩 하거나, CNN 모델을 통해 임베딩 하는 방식으로 벡터화 시킨다.[8]

단, 각 단백질들은 대부분 서열의 길이가 다르기 때문에, 딥러닝 방법을 사용하여 임베딩하기 위하여 단백질 서열의 길이를 맞추어주는 전처리 과정을 진행하게 된다. 이러한 인풋 사이즈를 맞추어주는 작업을 위해, 임의의 길이를 정해놓고 그에 맞추어주는 처리과정을 거치게 된다. 예를 들어, 시퀀스의 길이가 각, 10, 50, 80 인 세 단백질이 있을 경우 최대 길이를 70 으로 설정한다면, 첫번째와 두번째 단백질의 뒤에 각 60 개, 20 개의 X 패딩을 하여 길이를 늘려주고, 세번째 단백질의 경우 후반 10 개의 단백질서열을 잘라 길이를 줄여 전부 70 의 길이로 맞춰주게 된다.

이러한 전처리 과정을 거칠 경우 단백질이 가진 전체 정보를 활용하지 못하고 버리는 데이터가 생길 수 밖에 없으며, 이는 반드시 정보의 손실 이후 딥러닝 모델을 학습할 수 밖에 없게 된다. 이러한 정보의 손실을 막고, 가진 데이터를 모두 활용하기 위해 다중서열정렬 방법을 연구했다. 본래 모든 정보를 활용하기 위해 단백질 3 차 및 4 차구조 즉, 3 차원구조를 통해 분석하고자 하였다. 이를 위해, 구글 딥마인드가 CASP 를 위해 개발한 알파폴드를 참고하여 전처리 작업을 하려 했으나, 공개된

코드가 완전 구현이 되어있지 않으며 모델이 너무 무거워질 것을 우려하여, 향후 연구에서 진행하기로 하였다.[9]

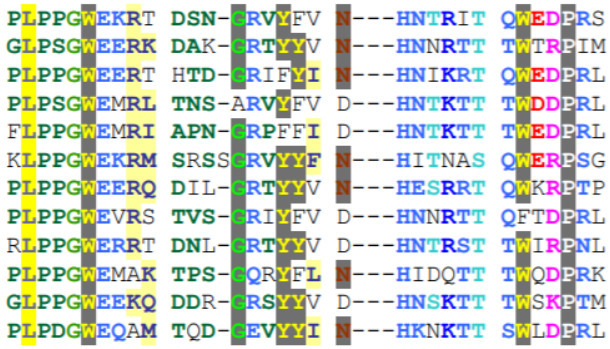

그림 1. 다중서열정렬(MSA) 예시

단백질의 서열 정보를 활용한 단백질 데이터 전처리를 위하여, 오픈데이터베이스인 BindingDB 의 모든 단백질 데이터의 유니크 값을 추출하였다. 약 5,500 개의 유니크한 데이터를 추출할 수 있었고, 다중서열정렬을 위하여 각 데이터의 길이를 체크하였다. 짧은 데이터는 두 자리 숫자의 길이를 가지고 있었지만, 긴 데이터는 네 자리 숫자의 길이를 지니고 있어, 이를 한번에 다중서열정렬 형태의 데이터로 변환하는 작업은 쉽지 않았다.

파이썬으로 구현된 Muscle 모듈을 사용하여 전체 데이터를 변환하려 하였으나, 단순 CPU 로 동작하기 때문에 모든 단백질 데이터를 활용한 연산을 감당할 수 없었기 때문에, GPU 를 활용한 연산이 가능한 프로그램을 찾게 되었다.[10] GitHub 에 공개되어 있는 코드들 중 C 언어로 짜여진 GPU 를 활용한 다중서열정렬 코드를 찾아 이를 참조하여 단백질 데이터를 정렬하였으며, 모든 데이터를 활용한 동작 연산이 가능함을 확인하였다.

Ⅲ. 결론

본 논문에서는 진행하고 있는 약물-표적 상호작용 예측 딥러닝 모델의 전처리 과정까지의 연구를 소개하였으며, 이를 활용하고자 하는 방향까지를 서술하였다. 다만, 해당 다중 서열 정렬 방법을 사용할 경우 전체 약 5,500 개의 단백질 데이터의 모든 시퀀스를 포함한 다중서열로서 데이터가 정렬되므로 각 데이터의 길이가 모두 길어진다는 점이 있다. 이를 보다 잘 활용하기 위하여, 각 데이터를 보다 잘 필터링 하여 활용할 수 있는 CNN 레이어를 파인튜닝하여 사용하거나, Transformer 를 활용하여 각 시퀀스의 어텐션을 정밀히 계산하여 활용할 수 있는 딥러닝 모델이 필요하다.

향후 데이터 전처리를 위해 HH-suite 를 활용한 유사 단백질 서열의 클러스터를 가져와 데이터 어그멘테이션 작업과, 단백질 3 차구조 정보와 그래프 구조를 활용한 입력데이터 및 딥러닝 모델 개발까지 진행 할 예정이다[11,12].

ACKNOWLEDGMENT

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019007059)

참 고 문 헌

[1] Ali Oskooei, Jannis Born, Matteo Manica, Vigneshwari Subramanian, Julio Sáez-Rodríguez, María Rodríguez Martínez, “PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks”, Nucleic Acids Research, Volume 48, Issue W1, 02 July 2020, Pages W502- W508

[2] Hakime Öztürk, Arzucan Özgür, Elif Ozkirimli, “DeepDTA: deep drug- target binding affinity prediction”, Bioinformatics, Volume 34, Issue 17, 01 September 2018, Pages i821- i829

[3] Bonggun Shin, Sungsoo Park, Keunsoo Kang, Joyce C. Ho, “Self-Attention Based Molecule Representation for Predicting Drug-Target Interaction”, Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR 106:230-248, 2019

[4] Ashish Vaswani Noam Shazeer Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin, “Attention is All You Need”, Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1

[6] Lipman, David J. Pearson, William R., “Rapid and Sensitive Protein Similarity Searches”, Science, Volume 227, Issue 4693, pp. 1435-1441

[7] W R Pearson, D J Lipman, “Improved tools for biological sequence comparison.”, Proc Natl Acad Sci U S A. 1988 Apr; 85(8): 2444- 2448.

[8] Ahmed Elnaggar, el al, “ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High Performance Computing”, COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

[9] Andrew W. Senior. et al, “Improved protein structure prediction using potentials from deep learning”, Nature volume 577, pages706- 710 (2020)

[10] Robert C Edgar, “MUSCLE: a multiple sequence alignment method with reduced time and space complexity”, BMC Bioinformatics volume 5, Article number: 113 (2004)

[11] Johannes Söding, “Protein homology detection by HMM- HMM comparison”, Bioinformatics, Volume 21, Issue 7, 1 April 2005, Pages 951- 960

[12] Michael Remmert, Andreas Biegert, Andreas Hauser & Johannes Söding, “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment”, Nature Methods volume 9, pages173- 175 (2012)