ProtBERT를 활용한 독성 단백질 분류

안성윤O, 이상웅*
*가천대학교 IT융합대학 AI·소프트웨어학부,
O가천대학교 IT융합대학 AI·소프트웨어학부
e-mail: sungyoonahn@gachon.ac.krO, slee@gachon.ac.kr*

Fine-Tuned ProtBERT for Toxic Protein Classification

Sung-Yoon AhnO, Sang-Woong Lee*
*School of Computing, Gachon University,
OSchool of Computing, Gachon University

● 요 약 ●

살아있는 유기체에 의해 분비되는 독소는 대부분의 경우 인간에게 유해하다. 가령 여름철 날것이나 오래된 음식에서 쉽게 식중독에 걸릴 수 있는데, 이는 주로 Clorustidium Botulinum이 만들어낸 보툴리눔 독소가 원인이다. 유기체에 의해 생성된 모든 독소는 단백질이며 이는 아미노산 서열로 나타낼 수 있다. 이를 통해 생물정보학 분야의 많은 연구자들이 많은 머신러닝 기술을 통해 단백질의 독성을 예측할 수 있었다. 최근 몇 년 동안 SVM를 사용하는 BTXpred와 CNN을 사용하는 ToxDL과 같은 모델이 각각 박테리아와 동물 독소의 독성을 예측하기 위해 제안되었다. 시대가 변함에 따라 BERT와 같은 성능이 더욱 뛰어난 모델이 시퀀스 분류를 위해 도입되었다. 본 논문은 독성 단백질을 분류를 위해 ProtBERT를 사용할 경우 이의 성능을 보여주고자 한다.

키워드: 단백질(protein), 독성물질(toxin), BERT

I. Introduction

독은 한 생명체가 스스로를 지키거나 공격하기 위해 생산하는 물질이다. 이렇게 생산된 독은 해당 생명체의 유전 정보에 따라 생산된 단백질인데 이는 단백질 서열로 표기 할 수 있다. 단백질은 아미노산으로 이루어져 있으며 총 20개의 아미노산이 존재한다. 20가지의 아미노산은 각각 다른 알파벳으로 나타낸다. 자연어와 유사한 특징을 가진 덕분에 단백질 서열 데이터에 자연어 처리 기술을 접목하기 용이하다. 본 논문은 독성 단백질을 구분하기 위해 최신 자연어처리 모델인 BERT를 적용하고자 한다.

II. Preliminaries

1. Related works

다양한 이유로 단백질의 독성을 예측하는 모델의 개발에 많은 노력이 기울여졌다. 가령 BTXpred[1]는 박테리아의 독성 단백질 분류를 위해 전통적인 머신러닝 기법인 SVM을 사용하였으며, ToxDL[2]은 동물의 독 단백질을 분류 하기 위해 합성곱 신경망을 사용한다. 단백질 데이터가 문자로 이루어진 시퀀스 데이터 만큼 자연어처리 기술을 접목한 연구도 있으며 대표적으로 ProtBERT[3]는 트랜스포머 모델 기반의 BERT 모델을 단백질 서열 데이터를 활용하여 학습하였다.

III. The Proposed Scheme

1. Benchmark Dataset

본 논문의 성능을 평가하기 위해 두 개의 데이터셋을 사용하였다. 첫 번째 데이터셋은 [2]에서 사용된 동묵 독성 단백질 데이터 이고 두 번째 데이터셋은 [1]에 사용된 박테리아 독성 단백질 데이터셋이다. 두 데이터셋 모두 독성 단백질은 참, 비독성 단백질은 거짓으로 라벨링 되어있다. 아래 표1은 각 데이터셋에 포함된 데이터의 양을 보여준다.

Table 1. Dataset Description

Dataset	Train / Validation	Toxic Protein	non-Toxic Protein
Animal Dataset[2]	Train	4413	5671
	Validation	59	670
Bacteria Dataset[1]	Train	140	402
	Validation	43	92

2. Model Description

1장에서 언급했다시피 독성 단백질과 같은 단백질 데이터는 20개의 알파벳 조합으로 표기가 된다. 표기가 알파벳으로 되어있는 만큼 기존 자연어처리 분야에서 사용되는 모델들을 적용하기가 쉽다. 따라서 본 연구에는 단백질 염기서열을 사용하여 학습된 ProtBERT[3] 모델에 분류 계층을 추가하여 fine-tuning을 진행하였다.

3. Experiment Results

아래 표는 총 20epoch 동안 진행된 실험결과중 가장 좋은 validation 데이터셋의 결과를 보여준다. Validation 데이터셋의 클래스간 불균형으로 정확도 로만 성능을 평가하기에는 객관적이지 않아 F1과 MCC 값으로 모델의 성능을 평가하였다.

Table 2. Experiment Results

Data	Accuracy	F1	MCC
Animal Dataset	0.9726	0.8333	0.8185
Bacteria Dataset	0.9778	0.9639	0.9491

IV. Conclusions

본 논문에서는 ProtBERT[3]를 사용하여 독성 단백질을 분류 하였을 때의 성능을 확인 할 수 있었다. 동물 독 데이터의 성능이 박테리아 데이터셋 보다 떨어지는 것을 볼 수 있으나 이는 박테리아의 단백질서열이 동물의 단백질서열보다 상대적으로 단순하며 데이터셋의 양에 큰 차이가 있어 나타나는 것으로 보인다. 단백질 데이터 특성상 구하기 어려우며 정확한 라벨링이 안되어 있는 만큼 양질의 데이터 생성이 필요할것으로 보인다.

ACKNOWLEDGEMENT

This work was supported by Korea Environment Industry & Technology Institute (KEITI) through Technology Development Project for Biological Hazards Management in Indoor Air Program(or Project), funded by Korea Ministry of Environment(MOE)(2021003380003)

REFERENCES

[1] Saha, S. and Raghava, G.P, “BTXpred: prediction of bacterial toxins”, In silico biology, Vol 7 No.4-5, pp.405-412, Jan 2007.

[2] Pan, X., Zuallaert, J., Wang, X., Shen, H.B., Campos, E.P., Marushchak, D.O. and De Neve, W, “ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity.”, Bioinformatics, Vol.36, No.21, pp.5159-5168, Jan 2021

[3] Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M. and Bhowmik, D. “ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing.”, arXiv preprint arXiv, July 2007