통계 정규화와 바이러스 돌파변이 예측 딥러닝 모델 성능 고도화 연구

장효석[1,2,†], 양시형[3,†], 이상철[1,†], 조한얼[1], 정철현[3,*], 김찬수[1,2,*]

1 한국과학기술연구원 인공지능·정보·추론 (AI/R) 연구실

2 과학기술연합대학원대학교 AI-로봇

3 한국과학기술연구원 화학생명융합 연구실

† Equally contributed

*Correspondence should be addressed to che.jeong@kist.re.kr, eau@ust.ac.kr.

Statistical Regularization for Enhancing the Performance of Deep Learning-Based Viral Escape Mutation Prediction Models

Hyoseok Jang[1,2,†], David Shihyung Yang[3,†], Sangchul Lee[1,,†], Haneol Cho[1],

Cherlhyun Jung[3], Chansoo Kim[1,2,*]

1 AI·Information·Reasoning (AI/R) Laboratory, Korea Institute of Science and Tech. (KIST)

2 AI-Robot Department, University of Science and Technology (UST)

3 Chemical & Biological Integrative Research Center, Korea Institute of Science and Tech. (KIST)

요 약

효과적인 항바이러스 치료제와 백신 개발에는 바이러스 돌파 변이의 이해가 필수적이다. 딥러닝 모델의 발전에도 불구하고, 복잡한 생물학적 데이터에서 발생하는 차원의 저주로 인해 변이를 정확하게 예측하는 데 여전히 어려움이 존재한다. 동 문제를 해결하기 위해, 본 연구에서는 뉴클레오타이드 포인트 돌연변이에 대한 통계 분석을 기반으로 하는 정규화 항을 추가한다. 연구 결과, 본 모델에서 가장 높은 예측 점수를 받은 도피 변이체 후보들이 바이러스 감염과 관련된 생물학적으로 중요한 도메인에 집중되어 있으며 일관되게 높은 변이 예측 점수를 받는 것을 확인하였다. 또한, 뉴클레오타이드 치환율과 같은 새로운 변수를 도입함으로써 기존 모델에서 낮은 점수를 받았던 도피 변이체 후보들의 예측 정확도가 크게 향상되었다. 이러한 접근은 잠재적 도피 변이체의 식별 효율성을 개선할 뿐만 아니라 바이러스 도피 메커니즘에 대한 중요한 통찰을 제공하여 표적 항바이러스 전략과 더 효과적인 백신 설계의 개발을 가속화하는 데 기여할 수 있다.

Ⅰ. 서론

COVID-19 팬데믹은 효과적인 바이러스 백신 개발의 중요성을 크게 부각시켰다. 그러나 COVID-19와 같은 빠르게 변이하는 바이러스는 즉각적인 백신 개발을 위한 실험 환경 조성이 물리적 및 재정적 제약으로 인해 어려움을 겪고 있다. 해당 제약사항을 극복하기 위해 모델링 기술과 계산과학 접근법이 점점 더 많이 활용되고 있으며, 특히 인공지능(AI)의 활약이 두드러진다. 최근에는 바이러스 도피 돌연변이를 예측하기 위해 자연어 처리(NLP, Natureal Language Processcing) 딥러닝 모델을 포함한 AI 모델을 활용하려는 노력이 진행되고 있다. 이러한 접근법은 일정 부분 성공을 거두었으나, 생물학적 데이터의 방대한 규모로 인한 차원의 저주로 인해 효과적인 특징 추출에 여전히 한계가 있다. 본 연구에서는 돌연변이 데이터로부터 모델 예측의 정규화에 적용하는 새로운 방법론을 제안한다.

Ⅱ. 본론

자연어처리용 BiLSTM(Bidirectional Long Short Term Memory) 모델을 활용한 기존 연구[1]에서는 바이러스의 아미노산 서열에 자연어와 유사한 패턴의 언어적 의미가 숨어있다고 가정한다. 자연어처리 모델에서 문법의 개념을 바이러스의 구조적 안정성, 변형된 문장의 의미변화(word embedding된 원래 문장과 변형된 문장간의 L1 norm값)를 면역돌파 가능성에 대응시킨다. 이후 모델이 예측하는 특정 바이러스 아미노산서열의 문법점수와 점수의 등수를 합산하는 방식으로 바이러스 돌연변이 후보군들의 돌파변이 가능성을 측정한다.

$$\text{Score} = \text{rank}(f(X')) + \text{rank}(g(X'))$$

여기서 score는 돌파변이 가능성 점수, f는 아미노산 서열을 인자로 받아 문법점수를 도출하는 함수이며 g는 아미노산 서열을 인자로 받아 의미변화점수를 도출하는 함수, rank는 해당점수가 표본데이터 전체대비 등수를 도출하는 함수이다.

초기모델학습

BiLSTM 구조의 인공신경망을 학습시킬 모델로 채택하였다.

- https://www.kaggle.com/therohk/million-headlines 데이터 사용
- 구두점 제거
- 공백으로 구분
- 소문자로 변환

돌파변이바이러스모델구축

초기모델을 기반으로 GISAID(gisaid.org)로부터 다운로드받은 코로나 바이러스 알파변이 및 베타변이 아미노산서열 데이터로부터 총 4172개의 중복없는 데이터를 추출하여 추가적인 모델 학습에 사용하였다.

AI모형개선

실제 돌연변이는 아미노산서열 레벨이 아닌 뉴클레오타이드 서열 레벨에서 발생한다는 생물학적 지식에 착안하여 기존모델의 점수 산출식에 코돈 변이 확률항을 추가하는 방식으로 수정된 산출식을 도출하였다.

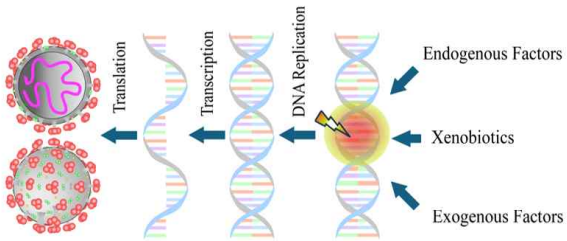

그림 1. 뉴클레오타이드레벨에서 일어나는 바이러스 돌연변이 개형도

아미노산돌연변이확률기반점수의정규화

아미노산 돌연변이 확률 $p(\tilde{x}_i \mid \mathbf{x})$ 을 추정하기 위해 본 연구에서는 word embedding 표현에 기반한 심층신경망 모델 $\hat{p}(\tilde{x}_i \mid \hat{\mathbf{z}}_i)$을 사용한다. 본 논문에서는, 통계적 모델[2]에 기반한 추정 확률을 추가로 고려한 다음의 정규화된 확률을 대신 고려한다.

$$p(\tilde{x}_i \mid \mathbf{x}) \approx \hat{p}(\tilde{x}_i \mid \hat{\mathbf{z}}_i)^{\lambda} p(\tilde{x}_i \mid N_{\mathrm{wt}})^{1-\lambda}$$

여기서 $p(\widetilde{x_i}|N_{wt})^{1-\lambda}$ 는 관찰된 뉴클레오타이드 서열의 통계적 변화율에 기반하여 계산한다.

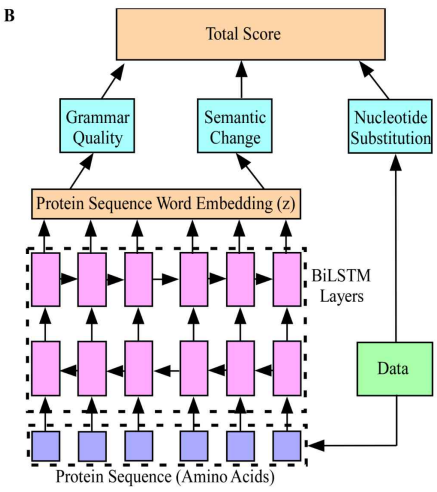

그림 2. 수정된 돌파변이 예측 인공지능 모델 개형도

모델성능

기존 모델에서 escape mutant 검증 데이터셋[3] 점수의 평균이 전체 24,168 중 3,700등으로 상위 15%를 차지한 것에 비교하여 본 연구 모델은 2,200등으로 상위 9%를 차지하여 (15 - 9) / 15 * 100=40% 성능향상이 있었다.

아미노산 위치정보	코로나 바이러스 와일드타입 알파벳	돌연변이 알파벳	기존 모델 랭크	개선된 모델 랭크
417	K	E	7029	3422
445	V	A	439	205
450	N	D	6665	3384
453	Y	F	6508	3713
455	L	F	122	20
485	G	D	949	287
493	Q	K	7816	3655
682	R	Q	1133	332
687	V	G	5187	4011
769	G	E	280	103
779	Q	K	5221	2364
1128	V	A	2262	934
685	R	S	4698	2033
655	H	Y	7237	2251
490	F	L	806	315
490	F	S	2923	1176
486	F	V	3118	2662
484	E	K	5307	1879
444	K	Q	3700	2495

표 1. 기존 모델과 개선된 모델의 돌파변이 예측 성능 결과 테이블
(랭크가 낮을수록 예측성능이 좋음)

Ⅲ.결론

본 논문을 통하여 우리는 기존 딥러닝 기반 모델의 돌파변이 바이러스의 예측원리, 학습방법에 대하여 설명하였고 모델 예측 성능을 향상시키기 위해 정규화 항을 도입하였다. 해당 방법론을 적용하여 구축한 새 모델의 성능이 기존모델의 성능대비 40%가 향상되었음을 확인하였다.

향후연구에서 검증에 활용한 모델외에도 기존에 발표된 다양한 딥러닝 기반 예측 모델에 동 알고리즘을 적용시켜 우리 연구 성과의 범용적 적용 가능성을 증명할 것이다.

ACKNOWLEDGMENT

This research was funded bythe grant Nos. 2021-0-02076 and 2024-00460980 (IITP) funded by the Korea government (the Ministryof Science and ICT).

참 고 문 헌

[1] Brian Hie et al. "Learning the language of viral evolution and escape," Science 371, pp. 284-288, 2021

[2] Yi, K., Mutational spectrum of SARS-CoV-2 during the global pandemic. Exp Mol Med 53, pp. 1229 - 1237, 2021

[3] A. Baum, B. "Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies." Science 369, pp 1014 - 1018 (2020)

[4] Zhiheng Huang, "Bidirectional LSTM-CRF Models for Sequence Tagging", arXiv, https://arxiv.org/abs/1508.01991.