트랜스포머 인코더-BiLSTM 기반 Lysine Crotonylation 위치 예측 모델

이종태, 김정현

세종대학교

22012153@sju.ac.kr, j.kim@sejong.ac.kr

A Transformer Encoder and BiLSTM-Based Model for Lysine Crotonylation Site Prediction

Jongtae Lee, Junghyun Kim

Sejong Univ.

요 약

본 논문에서는 Kcr 과정이 발생하는 라이신 아미노산의 정확한 위치를 찾기 위해 One-hot 인코딩과 트랜스포머 인코더, BiLSTM 기반의 단순화된 예측 모델을 제안한다. 기존 예측 모델은 다중 인코딩과 복잡한 모델 구조로 인해 파라미터 수가 많고 계산 비용이 크다는 한계가 있다. 제안 모델은 One-hot 인코딩을 사용한 단일 인코딩과 단순화된 모델 구조를 통해 기존 모델의 한계점을 보완하는 동시에 예측 성능을 향상시킨다. 실험 결과, 제안 모델은 정확도, 민감도, 특이도, 매튜 상관계수, AUC 5가지 평가 지표 모두에서 기존 모델보다 향상된 성능을 보였다. 또한 제안 모델의 파라미터 수와 테스트 시간은 기존 모델 대비 각각 39%, 22% 감소하였다.

Ⅰ. 서 론

Lysine crotonylation (Kcr)은 크로토닐-CoA를 기질로 하여 crotonyltransferase에 의해 히스톤 단백질의 특정한 라이신 아미노산 (K)에 크로토닐기를 전달하는 과정이다[1]. Kcr은 유전자 발현, 크로마틴 재구성, 생식 과정 조절 등 다양한 생명 활동에서 중요한 역할을 한다[2]. 다양한 생명 활동에 영향을 주는 Kcr 과정이 발생하는 라이신 아미노산의 위치를 찾기 위한 전통적인 방법으로 안정동위원소 표지, 고성능 액체 크로마토그래피, 면역 친화 크로마토그래피 등이 있다[3].

하지만 기존의 전통적인 방법들[3]은 많은 시간과 비용을 요구하기 때문에 최근에는 머신러닝 기반 예측 방법[4]과 딥러닝 기반 예측 방법[5]이 주를 이루고 있다. 특히 딥러닝의 발전으로 대규모 데이터셋에서 Kcr 위치를 정확하고 빠르게 예측할 수 있게 되었다. 딥러닝을 통한 Kcr 위치 예측은 Kcr에 대한 이해를 도울 뿐 아니라 다양한 세포 환경에서 단백질의 동적 조절 방식을 연구하는 데 유용한 도구를 제공한다[6]. 이러한 딥러닝의 발전에도 불구하고, Kcr 위치 예측을 위한 기존 모델[6]은 다중 인코딩과 복잡한 모델 구조로 구성되기 때문에 파라미터 수가 많고 계산 비용이 크다는 한계점이 있다.

본 논문에서는 One-hot 인코딩만을 사용하여 전처리 과정을 간소화하며 트랜스포머 인코더와 Bidirectional Long Short Term Memory (BiLSTM) 기반의 단순화된 예측 모델을 제안한다. 제안 모델은 트랜스포머 인코더를 사용해 전역적 특징을 추출한 후, BiLSTM을 통해 시퀀스의 양방향 정보를 효과적으로 학습한다. 실험 결과, 제안 모델은 기존 모델[6]보다 우수한 예측 성능을 보이는 동시에 단순화된 모델 구조를 통해 39% 감소한 파라미터 수와 22% 감소한 테스트 시간을 보여준다.

Ⅱ. 본론

본 논문은 Kcr 위치 예측을 위한 기존 모델[6]과의 원활한 비교를 위해 동일한 데이터셋을 사용한다. 해당 데이터셋은 [3]에서 HeLa 세포 유래 데이터를 기반으로 처음 구축하였으며, CD-HIT[7]을 통해 14,311개의 Kcr 위치 데이터 중 서열 유사도가 30% 이상인 데이터를 제거하여 완성하였다. 모든 데이터는 라이신 아미노산을 중심으로 양쪽에 15개의 아미노산을 포함하여 총 31개의 아미노산으로 구성된다. 학습 데이터셋은 Kcr 과정이 발생한 양성 샘플과 Kcr 과정이 발생하지 않은 음성 샘플 각각 6,975개씩, 총 13,950개의 샘플로 구성된다. 모델 성능 평가를 위한 테스트 데이터셋은 양성 샘플과 음성 샘플 각각 2,989개씩, 총 5,978개의 샘플로 구성된다.

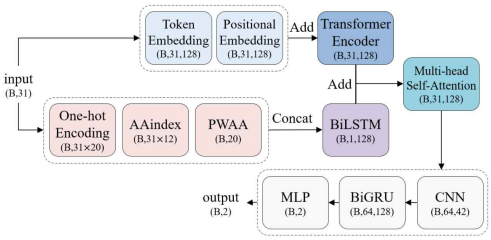

그림 1. Kcr 위치 예측을 위한 기존 모델 구조

그림 1은 Kcr 위치 예측을 위한 기존 모델의 구조를 나타낸다. 기존 모델은 학습 가능한 임베딩과 수작업으로 구성된 세 개의 인코딩을 사용하여 입력 데이터를 전처리하며, 트랜스포머 인코더와 BiLSTM을 사용하여 입력 데이터의 특징을 추출한 후 Multi-head Self-Attention으로 통합한다. 이후 모델은 통합된 특징을 Convolutional Neural Network (CNN), Bidirectional Gated Recurrent Unit (BiGRU), 다층 퍼셉트론 (Multi-Layer Perceptron, MLP)에 순차적으로 통과시켜 최종 예측을 수행한다. 이처럼 기존 모델은 다중 인코딩과 복잡한 모델 구조로 인해 파라미터 수가 많고 계산 비용이 크다는 한계가 있다.

본 논문은 One-hot 인코딩만을 사용하여 전처리 과정을 간소화하며, 트랜스포머 인코더와 BiLSTM 기반의 단순화된 Kcr 위치 예측 모델을 제안한다.

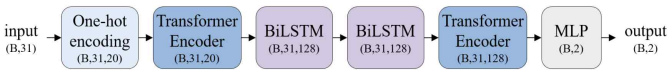

그림 2. Kcr 위치 예측을 위한 제안 모델 구조

그림 2의 제안 모델은 입력 데이터를 One-hot 인코딩 방식으로 전처리한다. One-hot 인코딩은 다음과 같이 표현한다.

$$X_{i,j} = \begin{cases} 1 & \text{if } s_i = a_j \\ 0 & otherwise \end{cases} \quad (1)$$

여기서 s_i는 입력 데이터 i번째 위치에 해당하는 아미노산이며, a_j는 20종의 표준 아미노산 집합에 포함된 j번째 아미노산을 의미한다. 입력 데이터 길이가 31, 아미노산의 종류가 20종이므로 One-hot 인코딩으로 변환된 입력 데이터는 31×20 크기의 행렬로 표현된다. 제안 모델은 변환된 입력 데이터로부터 전역적 특징을 추출하기 위해 트랜스포머 인코더를 사용하며, 이후 두 개의 BiLSTM 레이어를 통해 양방향 정보를 학습한다. 최종 예측 단계에서는 트랜스포머 인코더와 다층 퍼셉트론을 사용한다. 제안 모델은 트랜스포머 인코더를 사용함으로써 최종 예측 단계에서 전역 정보를 효과적으로 반영할 수 있다. 반면, 기존 모델은 CNN, BiGRU, 다층 퍼셉트론을 사용하여 최종 예측을 수행하는데, CNN의 제한된 수용 영역과 BiGRU의 순차적 학습 방법으로 인해 전역 정보를 충분히 반영하기 어렵다.

제안 모델은 총 에포크 45, 배치 크기 64, 학습률 0.0009의 Adam 옵티마이저를 사용하여 학습한다. 모델 성능 평가에는 정확도, 민감도, 특이도, 매튜 상관계수 (Matthews Correlation Coefficient, MCC), Area Under the receiver operating characteristic Curve (AUC) 5가지 지표를 사용한다. 5가지 지표는 다음의 수식으로 계산된다.

$$\text{정확도} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2)$$

$$\text{민감도} = \frac{TP}{TP + FN}, \quad (3)$$

$$\text{특이도} = \frac{TN}{TN + FP}, \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}, \quad (5)$$

$$AUC = \frac{\sum_{i}^{n_{pos}} rank_i - \frac{n_{pos}(n_{pos}+1)}{2}}{n_{pos} n_{neg}}, \quad (6)$$

여기서 TP, TN, FP, FN은 각각 진양성, 진음성, 위양성, 위음성의 수를 의미하며 $n_{pos}, n_{neg}, rank$는 각각 양성 샘플의 수, 음성 샘플의 수, 양성 샘플에 대해 모델이 부여한 예측 점수의 순위를 나타낸다.

표 1. 제안 모델과 기존 모델 성능 비교

	정확도	민감도	특이도	MCC	AUC
기존 모델[4]	0.856	0.876	0.835	0.712	0.931
제안 모델	0.873	0.886	0.860	0.747	0.938

표 1은 제안 모델과 기존 모델의 예측 성능을 보여준다. 제안 모델은 5가지 평가 지표 모두에서 기존 모델보다 우수한 예측 성능을 보인다.

표 2. 전처리 방식에 따른 정확도 및 입력 차원 크기 비교

	학습 가능한 임베딩	HF	One-hot	AAinedx	PWAA
정확도	0.785	0.773	0.770	0.650	0.615
차원 크기	128	71	20	20	31

표 2는 기존 모델이 사용한 전처리 방식에 따른 정확도 및 차원 크기를 비교한 결과이다. 여기서 Hand-crafted Feature (HF)는 One-hot, Amino Acid index (AAindex), Position-Weighted Amino Acid composition (PWAA)를 연결한 방식이다. 평가를 위해 제안 모델의 다층 퍼셉트론 층만을 추출하여 간단한 분류 모델로 이용했다. 비교 실험 결과 One-hot 인코딩은 학습 가능한 임베딩, HF 인코딩 방식과 비교하여 차원 크기는 각각 4배, 3배 작음에도 불구하고 정확도는 0.015, 0.003의 작은 수치만큼 하락하였다. 따라서 학습 가능한 임베딩과 HF를 모두 사용한 기존 모델과 다르게, 제안 모델은 One-hot 인코딩만을 사용하여 계산 비용을 줄이는 동시에 성능을 유지함을 확인할 수 있다.

표 3. 제안 모델과 기존 모델의 파라미터 수 및 테스트 시간 비교

	기존 모델[4]	제안 모델
파라미터 수	6,363,714	3,870,428
테스트 시간	0.137s	0.107s

표 3은 제안 모델과 기존 모델의 파라미터 수와 테스트 시간을 보여준다. 제안 모델은 기존 모델 대비 39% 감소한 파라미터 수와 22% 감소한 테스트 시간을 달성하여 계산 비용이 감소했음을 확인할 수 있다.

Ⅲ. 결론

본 논문은 Kcr 과정이 발생하는 히스톤 단백질 내 라이신 아미노산의 정확한 위치를 찾기 위해 One-hot 인코딩만을 사용하는 간소화된 전처리 과정과 트랜스포머 인코더, BiLSTM 기반의 단순화된 예측 모델을 제안한다. 실험 결과, 정확도, 민감도, 특이도, 매튜 상관계수, AUC에서 각각 0.017, 0.010, 0.025, 0.035, 0.007 만큼 향상된 성능을 보였다. 또한 제안 모델은 기존 모델 대비 39% 감소한 파라미터 수와 22% 감소한 테스트 시간을 통해 감소한 계산 비용을 확인하였다.

ACKNOWLEDGMENT

Put sponsor acknowledgments.

참 고 문 헌

[1] Tan, M., Luo, H., Lee, S., Jin, F., Yang, J. S., et al, “dentification of 67 histone marks and histone lysine crotonylation as a new type of histone modification,” Cell, vol. 146, no. 6, pp. 1016-1028, Sep, 2011.

[2] Soffer, R L., “Post-translational modification of proteins catalyzed by aminoacyl-tRNA-protein transferases,” Mol Cell Biochem, vol. 2, no. 1, pp. 3-14, Nov. 1973.

[3] Yu, H., Bu, C., Liu, Y., Gong, T., Liu, X., et al., “Global crotonylome reveals CDYL-regulated RPA1 crotonylation in homologous recombination-mediated DNA repair,” Science Advances, vol. 6, no. 11, pp.1-16, Mar. 2020.

[4] Ahmed, S., Rahman, A., Hasan, M. A., Rahman, J., Islam, M. K. B., et al., “predML-Site: Predicting Multiple Lysine PTM Sites With Optimal Feature Representation and Data Imbalance Minimization,” IEEE/ACM transactions on computational biology and bioinformatics, vol. 19, no. 6, pp. 3624-3634, Apr. 2023

[5] Qiao, Y., Zhu, X., and Gong, H., “BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models,” Bioinformatics, vol 38, no. 3, pp. 648-654, Feb, 2022.

[6] Liang, Y., and Li, M., “A deep learning model for prediction of lysine crotonylation sites by fusing multi-features based on multi-head self-attention mechanism,” Scientific Reports, vol. 15, no. 1, pp. 1-12, May. 2025.

[7] Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W., “CD-HIT Suite: a web server for clustering and comparing biological sequences,” Bioinformatics, vol. 26, no. 5, pp. 680-682, Mar. 2010.