염기 서열 인코딩을 위한 Conformer 기반 최소 자유 에너지 예측 스크리닝 모델

황우진, 박호성

전남대학교

ruddy_1632@jnu.ac.kr, hpark1@jnu.ac.kr

Conformer-based Screening Model for Minimum Free Energy Prediction in DNA Encoding Sequences

Woojin Hwang, Hosung Park

Chonnam National University

요 약

본 논문은 DNA 염기 서열 인코딩을 위한 Conformer 기반 모델을 제안한다. 제안하는 방법은 self-attention과 Convolution 모듈을 통합한 Conformer 구조를 MFE(Minimum Free Energy) 예측의 핵심 백본으로 도입하여, 염기 서열 내 장거리 상호작용과 연속적인 염기 패턴을 단일 구조에서 동시에 모델링한다. 특히 Conformer 내부의 convolution 모듈을 다중 커널을 병렬로 적용하는 inception 형태로 확장하여, 서로 다른 길이의 연속적인 염기 패턴을 동시에 학습할 수 있도록 설계하였다. 이를 통해 순차 의존성을 제거하고 병렬 처리가 가능한 MFE 예측 모델을 구성하여 기존 BiLSTM-Transformer 기반 구조 대비 예측 정확도를 크게 향상시킨다.

Ⅰ. 서 론

DNA 기반 데이터 저장 시스템에서는 디지털 정보가 다수의 염기 서열로 인코딩되며, 각 서열들은 상보적인 염기 결합으로 인해 이차 구조를 형성하여 원치 않은 반응이 일어날 수 있다. 이차 구조의 형성 정도는 최소 자유에너지 값으로 정량화될 수 있으며, MFE는 염기 서열의 구조적 안정성을 평가하는 핵심 지표로 활용된다. 따라서 인코딩 단계에서 염기 서열의 MFE를 예측하고 이를 기반으로 서열을 선별하는 스크리닝 과정은 중요한 절차이다. 염기 서열의 MFE는 일반적으로 NUPACK[1]과 같은 동적 프로그래밍 기반 소프트웨어를 통해 계산되며, 이러한 계산 방식은 서열 길이가 증가할수록 계산 복잡도가 급격히 증가하는 한계를 가진다. 특히 대규모 인코딩 환경에서 다수의 염기 서열을 반복적으로 평가하는 경우, 계산 비용이 실질적인 병목으로 작용할 수 있다. 이로 인해 MFE 계산을 효율적으로 근사할 수 있는 딥러닝 기반 예측 모델에 대한 필요성이 제기된다. 기존 연구[2]에서는 BiLSTM(Bidirectional Long Short Term Memory)과 Transformer를 결합한 순차적 모델을 통해 염기 서열의 MFE를 예측하였으나, 순차 처리 구조로 인한 병렬성 및 확장성의 한계가 존재한다. 본 논문에서는 self-attention과 convolution 모듈을 통합한 Conformer[3] 구조를 MFE 예측의 핵심 백본으로 도입하고, 서로 다른 길이의 연속적인 염기 패턴을 효과적으로 모델링 하기 위해 convolution 모듈을 Inception[4] 형태의 다중 커널 병렬 구조로 확장한다.

Ⅱ. 모델 구조

2.1 입력 표현

입력 데이터는 길이 50-150nt 의 DNA 염기 서열로 구성되며, 각 염기는 원-핫 인코딩 이후 임베딩 벡터로 변환된다. 모델이 염기 간 상대적 위치 정보를 학습할 수 있도록 Relative positional encoding[5]을 적용하며, 이를 통해 장거리 상호작용을 고려한 특징 학습이 가능하도록 한다. 이러한 입력 표현은 이후 Conformer 기반 백본의 self-attention 및 convolution 연산에 공통적으로 사용된다.

2.2 Conformer 기반 백본 구조

제안하는 모델의 전체 구조는 그림1 과 같이, 입력된 DNA 염기 서열을 Conformer 기반 백본을 통해 처리 후 완전연결층을 통해 자유에너지 값을 출력하는 회귀 모델로 구성된다. Conformer 기반 백본은 다수의 Conformer 블록을 적층한 구조로, 각 블록은 Feed-forward network, multi-head self-attention, convolution 모듈을 포함하며, 잔차 연결과 층 정규화를 통해 안정적인 학습을 유도한다. 본 연구에서는 Conformer 논문에서 제안된 구조를 따르며, 이를 아래의 수식과 같이 정의한다.

$$\tilde{x}_i = x_i + \frac{1}{2}FFN(x_i)$$
$$x'_i = \tilde{x}_i + MHSA(\tilde{x}_i)$$
$$x''_i = x'_i + Conv(x'_i)$$
$$y_i = LayerNorm(x''_i + \frac{1}{2}FFN(x''_i))$$

수식의 MHSA는 염기 서열 전반에 걸친 장거리 상호작용과 상보적 관계를 모델링 하는 역할을 수행하며, Conv는 연속적인 염기 패턴으로부터 국소적인 구조적 특징을 추출한다. 이러한 두 연산을 단일 블록 내에서 결합함으로써 전역적 및 국소적 특징을 동시에 학습할 수 있다. 기존의 BiLSTM 기반 모델과 달리, 제안하는 구조는 순차 처리에 대한 의존성을 제거하고 병렬 처리가 가능하다는 장점을 가진다. 이로 인해 서열 길이가 증가하더라도 효율적인 학습과 추론이 가능하다.

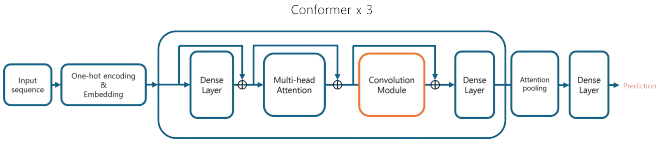

그림 1. Conformer 기반 자유에너지 예측 스크리닝 모델 구조

2.3 Inception 기반 convolution 구조

Conformer 구조의 convolution 연산은 염기 서열 내 국소적인 연속패턴

을 모델링 하는 핵심 요소이다. 그러나 [3]논문의 Conformer 구조에서는 단일 커널을 사용한 convolution 연산을 통해 국소적 특징을 추출하므로, 서로 다른 길이와 형태를 갖는 연속적인 염기 패턴을 동시에 표현하는 데에는 한계가 존재한다. NUPACK의 MFE 계산은 특정 패턴에 의해 단독으로 결정되는 것이 아니라 다양한 크기의 국소적 상호작용이 복합적으로 작용하여 계산된다. 따라서 MFE 예측을 위해서는 단일 수용 영역에 기반한 특징 추출보다, 서로 다른 수용 영역을 갖는 국소 패턴을 동시에 고려하는 다중 스케일 모델링이 필요하다. 이를 위해 본 논문에서는 Conformer 블록 내부의 convolution 연산을 Inception 형태의 구조로 확장한다. 제안하는 convolution 모듈의 구조는 그림2 와 같다.

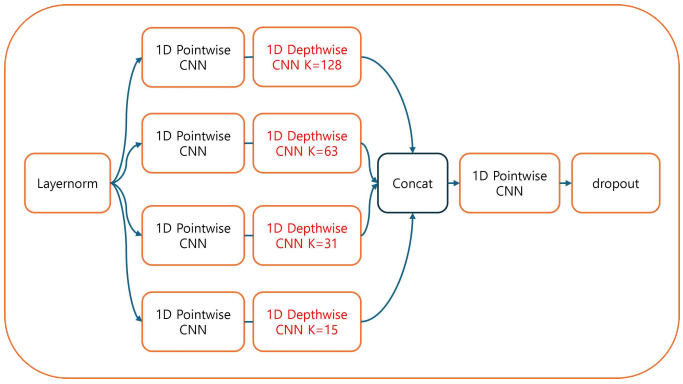

그림 2. Convolution module 내부 구조

동일한 특징 맵에 대해 서로 다른 커널 크기를 갖는 다수의 convolution filter를 병렬로 적용한다. 각 병렬 경로는 서로 다른 수용 영역을 가지므로, 다양한 길이의 염기 패턴을 동시에 포착함으로써 기존의 단일 커널 기반 Conformer 대비 복합적인 연속 패턴을 효율적으로 모델링 할 수 있다.

Ⅲ. 모의 실험

학습 데이터셋은 길이가 50~150nt인 총 1.500.000개의 DNA 염기서열로 구성되며, GC-content(45~55%)와 homopolymer run(≤3)제약을 만족하도록 무작위로 생성하였다. 각 서열의 MFE는 NUPACK을 통해 계산하여 사용하였다. 테스트 데이터셋은 Microsoft 연구 그룹에서 공개한 실제 DNA 시퀀싱 데이터셋을 사용하였다. 해당 데이터셋은 길이 110nt의 DNA 염기 서열 10,000개로 구성되어 있고 MFE는 NUPACK을 통해 계산되었다.

그림 3은 Epoch 별 모델의 결정계수 비교를 보여준다. Conformer는 기존 BiLSTM-Transformer 모델 대비 약 0.1의 예측 성능 향상을 달성하였다. Conformer와 Conformer-inception을 비교하였을 때, 수렴 속도는 Conformer가 더 빠르지만 Conformer-inception이 더 높은 예측 성능을 보인다. 표1 은 각 모델의 최고 성능과 실행 시간을 보여준다. 제안하는 모델은 BiLSTM-Transformer 대비 파라미터 수가 다소 증가하였으나, Conformer 구조는 순차적인 시간 의존성을 갖는 BiLSTM과 달리 병렬 연산이 가능한 구조를 갖기 때문에 서열 길이가 길어져도 병목 현상이 발생하지 않는다.

Ⅲ. 결론

본 논문은 MFE 예측을 위한 Conformer 기반 딥러닝 모델을 제안하였다. 제안하는 Conformer-inception 모델은 NUPACK 기반 MFE 값을 효과적으로 근사하며 기존 순차 모델 대비 향상된 예측 성능을 보였다. 또한 병렬 처리가 가능한 구조적 특성으로 GPU 환경에서의 학습 및 추론 과정에서 효율적인 병렬 처리를 가능하게 하여 향후 대규모 DNA 서열 스크리닝과 같은 확장된 문제에 쉽게 적용이 가능하다. 향후에는 보다 더 높은 예측 성능을 위해 모델 구조 및 파라미터에 대해 추가적인 개선을 진행할 예정이다.

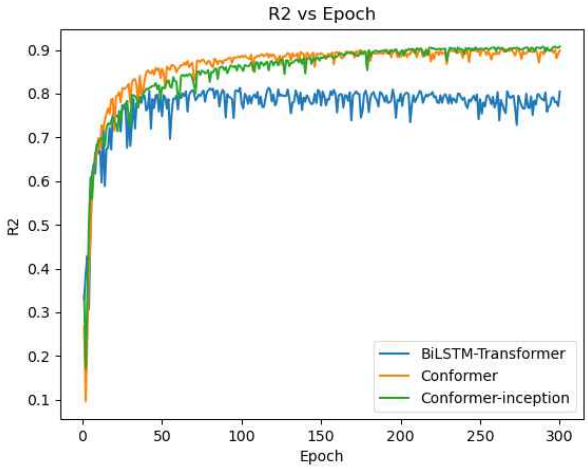

그림 3. Epoch별 모델 테스트 결과

모델 종류	결정계수	평균상대오차	파라미터 수	실행 시간(s)
BiLSTM-Transformer	0.8134	0.1073	862,837	0.0006089
Conformer	0.9033	0.0759	971,377	0.0006929
Conformer-inception	0.9075	0.0736	1,027,633	0.0007012
NUPACK 계산				0.0024308

표 1. 최고 성능 비교 결과.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00410005). 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (과제번호 2022M3C1A3090857).

참 고 문 헌

[1] Zadeh, Joseph N., et al. “NUPACK: Analysis and design of nucleic acid systems.” Journal of computational chemistry 32.1(2011): 170-173.

[2] Lin, Wanmin, et al. “Predict the degree of secondary structures of the encoding sequences in DNA storage by deep learning model.” Scientific Reports 15.1 (2025): 20920.

[3] Gulati, Anmol, et al. “Conformer: Convolution-augmented transformer for speech recognition.” arXiv preprint arXiv:2005.08100(2020).

[4] Chollet, Francois. “Xception: Deep learning with depthwise separable convolutions.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[5] Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. “Self-attention with relative position representations.” arXiv preprint arXiv:1803.02155(2018).