RibonanzaNet과 IPA를 활용한 RNA 3D 구조 예측 모델

박솔, 임지섭, 조성우, 이종태*, 김재현
아주대학교 전자공학과, 아주대학교 AI융합네트워크학과*
{solbak2, tjqwldla, thdtjsgk, jtlee830*, jkim}@ajou.ac.kr

RNA 3D structure prediction model using Ribonanzanet and IPA

Sol Park, Jisub Lim, Seongwoo Cho, Jongtae Lee, Jae-Hyun Kim
Department of Electrical Engineering, Ajou University
Department of Artificial Intelligence Convergence Network, Ajou University

요 약

Ribonucleic acid(RNA)의 3D 구조 예측은 질환이나 신약 개발 등에 널리 활용되지만, 큰 복잡성으로 인해 도출하는 데 어려움이 존재한다. 본 논문은 기존 RNA 3D 구조 예측 모델인 RibonanzaNet 구조에 기하학적 관계를 고려 가능한 Invariant Point Attention(IPA) 모듈을 추가한 예측 모델을 제안한다. RNA 서열 데이터 기반의 Root Mean Absolute Error(RMAE) 평가에서 제안하는 모델은 기존 모델보다 14.6% 향상된 성능을 확인했다.

Ⅰ. 서론

RNA 구조는 암을 비롯한 다양한 질환과 밀접한 관련이 있으며, RNA를 표적으로 하는 신약 개발에서도 구조 정보는 필수적이나 삼차원 구조를 규명하는 일에는 여전히 많은 제약이 따른다 [1], [2]. 기존 단백질 구조 예측 모델인 Alphafold[3]의 구조와 유사한 Foundation model인 RibonanzaNet[4]이 제시되었으나, 해당 모델은 주로 서열 정보에 의존해 실제 모양 및 구조를 잘 반영하지 못하기 때문에 예측 성능이 저하되는 한계가 있다. 따라서 본 연구에서는 기존 RibonanzaNet 기반 3D RNA 구조 예측 모델에 Invariant Point Attention(IPA) 모듈을 추가하여 예측 정확도 향상을 위한 Ri-PA 모델을 제안한다.

Ⅱ. 본론

1. RibonanzaNet

RibonanzaNet 모델은 RNA 염기 서열 정보만을 입력으로 사용하여 입체 구조적 특성을 예측하는 딥러닝 모델로, Transformer Encodoer Block만을 이용하고, Base Pairing Probability와 같이 사전에 계산된 물리적 정보에 의존하지 않는 특징이 있다[4]. 특히 1차원 서열 표현과 2차원 쌍 표현 간에 양방향으로 정보를 교환하여 표현을 점진적으로 정교화하는 구조를 통해, 복잡한 RNA 구조 패턴을 효과적으로 포착이 가능하다. 그러나 RibonanzaNet 모델의 Transformer 기반 아키텍처는 서열 내의 1차원적 관계나 문맥을 파악하는 데에는 뛰어나지만, 뉴클레오타이드 간의 3차원적 공간 배치나 기하학적 관계를 직접적으로 고려하여 모델링하는 데에는 잠재적인 한계가 있다.

2. IPA 모듈

IPA 모듈은 3차원 구조를 반복적으로 정제하고 정확하게 예측하기 위해 사용된 핵심적인 신경망 모듈로, 이는 각 잔기의 국소 좌표계와 전체 구조의 전역 좌표계를 오가며 정보를 처리하는 방식으로 작동한다. 먼저

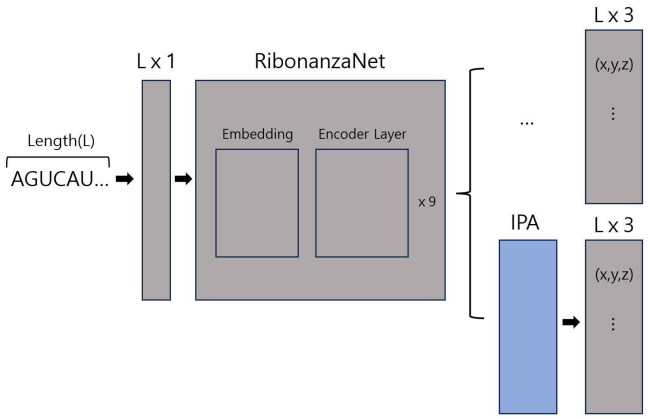

그림 1. 제안하는 Ri-PA 시스템 모델 구조

각 잔기의 국소 좌표계 내에서 어텐션 계산에 사용될 Query, Key, 그리고 Value에 해당하는 3차원 점들을 생성한다. 다음으로, 이 점들을 잔기의 골격 위치와 방향 정보를 이용해 전역 좌표계로 변환하고, 전역 좌표계 상에서 점들 사이의 거리를 기반으로 어텐션 가중치를 계산한다. 이때 가까운 잔기일 수 록 더 높은 가중치를 부여받는다. 계산된 가중치를 반영하여 Value matrix를 업데이트한 후, 최종 결과는 다시 각 잔기의 국소 좌표계로 변환된다. 이러한 좌표계 변환 과정을 통해 구조 전체가 회전하거나 이동하는 강체 변환이 발생하더라도 잔기 간의 상대적인 거리와 방향은 보존되므로, 최종 계산 결과는 항상 동일하게 유지되는 특징을 갖는다[5].

3. 제안하는 Ri-PA 모델

본 논문에서 제안하는 Ri-PA 모델은 기존 RibonanzaNet 모델의 기하학적 측면에서의 한계점을 IPA 모듈을 추가하여 문제를 개선한다. RibonanzaNet 모델의 블록에서는 RNA 서열로부터 Embedding과 행렬

표 1. 실험 Parameter

Optimizer	Adam	
Epoch	50	
Multiheadblock	8	
Pairwisedimension	256	
EncoderBlock	9	
Datasize	Set1	Set2
	100	564

곱 연산을 통해 sequence feature와 pairwise feature를 생성한다. 생성된 pairwise feature와 잔기 위치정보를 Self-Attention을 통해 IPA 모듈이 처리할 수 있도록 2차원 행렬을 생성된다. 이 행렬은 상대적 위치로 계산된 구조 정보가 함축된 고차원 정보로 단순한 MLP로 처리하지 않고, IPA의 기하학적 계산을 추가하여 단일 서열 정보만으로는 포착하기 어려운 긴 거리 상호작용이나 고차 구조적 접힘 패턴까지도 모델이 학습할 수 있으며 실제 공간상의 이해도를 향상시킨다.

Ⅲ. 모의실험 결과

1. 모의실험 환경

본 논문에서는 기존의 RibonanzaNet 모델과 제안한 Ri-PA 모델을 각각 두 개의 훈련 데이터세트로 학습하고, 하나의 검증 데이터세트를 통해 비교한다. RNA 실측 데이터는 'Protien Data Bank(PDB)[6]'의 PDF file를 활용했으며, 이는 RNA의 이름, 잔기 서열, 각 잔기의 위치가 포함된다. 잔기의 위치는 잔기가 리보스와 결합하는 C1' 원자의 위치로 설정하여 뉴클레오티드 분자 중 한 점을 특정하였다. 그 외, 추가적인 시스템 파라미터는 상기 표 1과 같다. Set 1, Set 2는 RNA 실측 데이터를 랜덤으로 뽑은 훈련 데이터의 표본 집합을 의미한다. 검증 데이터는 각 Set에 동일하게 적용하여 성능을 비교하였다. 성능 분석 지표로는 Root Mean Absolute Error(RMAE)를 활용한다. 점수는 RMAE 값의 역수로 정규화하여 예측 좌표와 실측 좌표 간의 오차를 0-1 범위로 변환한다. 이 점수는 1에 가까울수록 모델의 예측 성능이 우수함을 직관적으로 보여준다.

$$Score = 1/(1+E_{MAE}), \tag{1}$$

$$E_{MAE} = \frac{1}{Z}\left(\frac{1}{3L}\sum_{i=1}^{L}\|p_{aligned,i} - q_i\|_1\right) \ (Z=10), \tag{2}$$

여기서 L은 서열의 길이, $P_{aligned}$는 회전 및 이동 변환을 거친 후의 3차원 예측 좌표의 집합, Q는 실측 좌표를 의미한다.

2. 모의실험 결과

모의실험을 통한 결과는 그림 2에서 나타낸다. 모의실험 Set 1에서는 기존 RibonanzaNet 모델은 0.643의 성능 점수를 보였으나, 제안하는 Ri-PA 모델에서는 0.651의 성능 점수로 0.008 만큼의 성능 향상을 확인했다. 모의실험 Set 2에서는 기존 RibonanzaNet 모델에서 0.582, Ri-PA 모델이 0.735로 0.135 만큼의 성능 점수 향상을 확인했다. 모의실험 결과를 통해 제안하는 Ri-PA가 두 개의 모의실험 세트에서 성능 점수가 더 높은 것을 확인할 수 있다. 기존 RibonanzaNet 모델에서는 데이터의 크기가 증가할 때, 오히려 성능이 감소하는 Overfitting 현상이 발생하였으나, 반면 기하학적 계산이 고려된 Ri-PA 모델은 이러한 Overfitting 없이 데이터 크기가 늘어남에도 높은 성능을 유지할 수 있음을 확인할 수 있다. 이는,

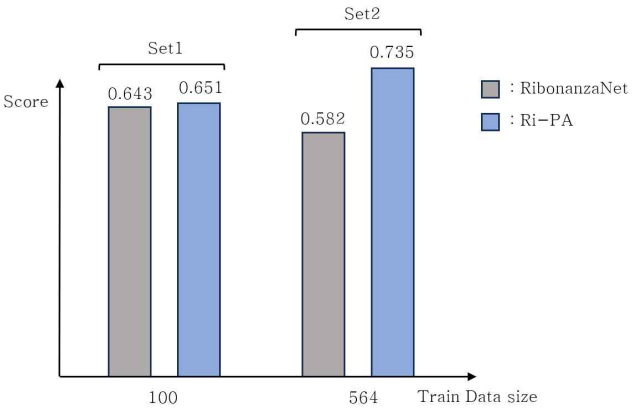

그림 2. 모의실험을 통한 성능 분석 결과

의도한 바와 같이 Ri-PA 모델이 공간 이해도가 증가하였다는 것을 확인할 수 있으며 기존 모델에서 서열 간의 관계로 구조를 예측하는 것 이외의 새로운 축으로의 구조 예측 알고리즘에 대한 가능성을 열었다.

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 RNA 서열 정보 기반의 RibonanzaNet 모델에 3차원 기하구조를 직접 학습하는 IPA 모듈을 결합한 Ri-PA를 제시했다. 모의실험 결과, 제안한 Ri-PA 모델은 기존 RibonanzaNet 모델보다 예측 정확도가 향상된 성능을 보였으며, 특히나 학습 데이터가 많을수록 성능 향상 폭이 크게 향상되는 효과를 확인했다. 이는 RNA 구조 예측에 있어 3차원 공간 정보를 직접적으로 활용하는 IPA 모듈을 같이 결합하는 방안이 효과적임을 입증한다. 향후 연구로는 계산 효율성을 높이기 위한 모델 최적화와 더 다양한 데이터 세트를 활용한 일반화 성능 검증을 진행할 예정이다.

참 고 문 헌

[1] Y. Shao, and Q. C. Zhang, "Targeting RNA structures in diseases with small molecules," Essays in Biochemistry, vol. 64, no. 6, pp.955-968, Dec. 2020.

[2] A. Chari. and H. Stark, "Prospects and Limitations of High-Resolution Single-Particle Cryo-Electron Microscopy," International Journal of Molecular Sciences, vol. 24, no.9, Art. no. 8312, May 2023.

[3] J. Jumper, et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, pp. 583-589, Aug. 2021.

[4] S. He, et al., "Ribonanza: deep learning of RNA structure through dual crowdsourcing," bioRxiv, preprint, doi:10.1101/2024.02.05.578931, Feb. 2024.

[5] A. Liu, et al., "Flash Invariant Point Attention," arXiv, preprint, arXiv:2505.11580, May. 2025.

[6] [Online]. Available: Protein Data Bank: https://www.rcsb.org/