단백질 언어 모델 기반 Titin 단백질 생성에 관한 연구

김형중, 임완수*
성균관대학교
khj3195@g.skku.edu, *wansu.lim@skku.edu

Generation of Titin Protein Sequences using Protein Language Model

Kim Hyeongjung, Lim Wansu*
Sungkyunkwan University

요 약

본 논문은 대규모 단백질 언어 모델(Protein Language Model, PLM)을 활용하여 새로운 Titin 단백질 서열을 생성하는 방법을 제안한다. 이를 위해 대표적인 대규모 PLM인 ProtGPT2와 ProGen2를 적용하고, UniProt에서 제공하는 Titin 단백질 데이터를 이용하여 fine-tuning을 수행하였다. 학습 데이터는 아미노산 서열 길이, 인간 및 비인간 유래 여부, 그리고 도메인(domain) 유형을 기준으로 체계적으로 분할하여 Titin 단백질의 구조적·기능적 특성이 효과적으로 반영되도록 구성하였다. 실험 결과, fine-tuning을 통해 생성 성능이 유의미하게 향상되었다. ProtGPT2의 경우 유효 서열 비율이 기존 42%에서 74%로 증가하여 27.59%의 성능 향상을 보였으며, ProGen2는 38%에서 80%로 증가하여 53.85%의 향상률을 달성하였다. 또한, 생성된 단백질의 구조적 안정성역시 기존 대비 약 10% 향상됨을 확인하였다. 이러한 결과는 Titin 특화 데이터 기반 fine-tuning이 대규모 PLM의 단백질 생성 성능과 구조적 신뢰성을 효과적으로 개선할 수 있음을 보여준다.

Ⅰ. 서 론

최근 단백질 서열 데이터를 대규모로 학습한 단백질 언어 모델(Protein Language Model, PLM)이 인공지능 기반 생명정보학 분야에서 핵심적인 연구 흐름으로 자리 잡고 있다. PLM은 아미노산 서열을 언어적 시퀀스로 해석하여 단백질에 내재된 통계적 규칙성과 진화적 패턴을 학습할 수 있으며, 이를 통해 단백질 서열 생성, 구조 예측, 기능 분석 등 다양한 응용 분야에서 우수한 성능을 보이고 있다 [1].

대규모 서열 데이터로 사전학습된 PLM은 일반적인 단백질 생성에 대해 높은 범용성을 가지지만, 특정 단백질을 목표로 한 생성(task-specific generation)의 경우 해당 단백질의 고유한 서열 분포나 구조적·기능적 특성을 충분히 반영하지 못하는 한계를 가진다. 이러한 문제를 해결하기 위해, 최근에는 특정 단백질 데이터셋을 활용한 fine-tuning기법이 주목받고 있으며, 이를 통해 사전학습된 PLM의 일반화 능력을 유지하면서도 목표 단백질의 특성을 효과적으로 반영할 수 있음이 보고되고 있다.

Titin 단백질은 현재까지 알려진 단백질 중 가장 긴 아미노산 서열을 가지며, 구조적·기능적 특성이 상이한 수백 개의 도메인(domain)으로 구성되어 있다. 각 도메인은 서로 다른 탄성 특성을 가지며, 이러한 조합을 통해 Titin은 근육의 수축 및 이완 과정에서 기계적 안정성과 탄성 조절에 핵심적인 역할을 수행한다. 그러나 현재까지 실험적으로 규명된 Titin 도메인의 구조 정보는 전체 구조의 약 7% 수준에 불과하여 [2], Titin의 구조–기능 관계를 체계적으로 이해하기에는 여전히 큰 한계가 존재한다.
이러한 제한된 구조 정보 문제를 극복하기 위해, 계산적 방법 기반의 단백질 서열 생성 및 구조 예측 기술은 중요한 대안으로 주목받고 있다. 특히, PLM을 활용하여 Titin 특화 서열을 생성하고, 생성된 서열에 대해 구조 안정성 및 타당성을 평가하는 접근은 기존 실험 기반 연구를 보완할 수 있는 새로운 가능성을 제공한다.

따라서 본 연구는 사전학습된 대규모 PLM을 Titin 단백질 데이터로 fine-tuning하여 새로운 Titin 단백질 서열을 생성하고, 생성된 서열에 대해 구조 예측 기반의 성능 평가를 수행함으로써 PLM의 유효성을 검증한다. 이를 통해 본 연구는 Titin 단백질의 구조적 다양성 확장 가능성을 제시함과 동시에, 향후 근육의 역학적 특성 및 구조–기능 관계 연구를 위한 기초적 토대를 마련하고자 한다.

Ⅱ. 본론

2.1 Protein Lannguage Models

본 연구에서는 대규모 단백질 언어 모델을 활용하여 특정 단백질인 Titin 단백질 서열 생성 성능을 비교·분석하기 위해, 대표적인 autoregressive 트랜스포머 기반 모델인 ProtGPT2와 ProGen2를 사용한다. 두 모델은 모두 대규모 단백질 서열 데이터를 통해 사전학습된 기반 모델로, 일반적인 단백질 생성 능력을 보유하고 있으며, 본 연구는 동일한 Titin 데이터셋을 이용해 fine-tuning을 수행함으로써 Titin 특화 생성 성능을 평가한다.

ProtGPT2는 7억 개 이상의 파라미터를 가지는 autoregressive 트랜스포머 모델로, 약 5천만 개의 단백질 서열을 이용해 사전학습되었다. 이 모델은 자연 단백질 서열에 내재된 통계적 특성과 진화적 패턴을 학습하여, 자연 단백질과 유사한 특성을 가지면서도 기존 서열과는 구별되는 새로운 단백질 서열을 생성할 수 있는 능력을 갖는다 [3]. 이러한 특성으로 인해 ProtGPT2는 단백질 서열 생성 연구에서 널리 활용되는 기준 모델로 사용되고 있다.

ProGen2는 ProtGPT2와 동일한 트랜스포머 기반 autoregressive 구조를 따르지만, Rotary Positional Embedding(RoPE)을 적용함으로써 토큰 간 상대적 위치 정보를 보다 효과적으로 모델링할 수 있다는 차별점을 가진다. 이는 긴 아미노산 서열을 가지는 단백질에서 장거리 의존성(long-range dependency)을 학습하는 데 유리한 구조적 특성이다. 본 연구에서는 ProGen2 모델 중 약 10억 개의 단백질 서열로 사전학습된 7억 개 이상의 파라미터를 가지는 medium 모델을 사용하였다 [4]. Titin 단백질 데이터의 규모와 fine-tuning 안정성을 고려할 때, 27억 개 이상의 파라미터를 갖는 large 모델보다 medium 모델이 과적합 위험이 낮고 안정적인 학습 및 서열 생성에 보다 적합하다고 판단하였다.

Model	Data Split	Valid Sequences / Top-50	Valid Ratio (%)
ProGPT2	Random	29/50	58
ProGPT2	Stratified	37/50	74
ProGen2	Random	26/50	52
ProGen2	Stratified	40/50	80

표 1 Stratified 분할 여부에 따른 유효 서열 비율 비교

이와 같이 본 연구는 구조적 특성이 상이한 두 PLM을 동일한 조건에서 fine-tuning하여 비교함으로써, Titin 단백질과 같이 매우 긴 서열과 복잡한 도메인 구조를 가지는 단백질 생성에 적합한 PLM의 특성을 분석하고자 한다.

2.2 실험 설계

2.2.1 데이터셋

본 연구에서 사용한 데이터셋은 UniProt에서 제공하는 인간 Titin 단백질 서열을 기반으로 구성하였다. 데이터의 중복성을 최소화하고 서열 간 유사도를 통제하기 위해, 서열 동일성(sequence identity) 90% 기준을 적용하여 Titin 서열을 선별하였으며, 해당 조건을 만족하는 모든 Titin 서열을 종 구분 없이 사용하였다. 이는 ProtGPT2와 ProGen2 모델이 사전학습 과정에서 UniRef90 데이터셋을 사용한 설정과 동일한 기준을 유지하기 위함이다 [3], [4].

Titin 단백질은 수만 개의 아미노산으로 구성된 초대형 단백질로, 전체 서열을 그대로 학습에 사용하는 경우 모델 학습의 효율성과 안정성이 저하될 수 있다. 이에 따라 본 연구에서는 Titin 서열을 구조적·기능적 특성에 따라 구분되는 도메인(domain) 단위로 분할하여 데이터셋을 구성하였다. 각 도메인 서열은 중복 제거 과정을 거친 후 독립적인 학습 샘플로 사용되었으며, 이를 통해 Titin 단백질의 다양한 국소적 구조 특성이 모델 학습에 효과적으로 반영되도록 하였다.

Fine-tuning에 사용된 최종 데이터셋은 중복이 제거된 약 2,100개의 도메인 단위 서열로 구성되었다. 데이터 입력 형식은 각 모델의 서열 처리 방식에 맞추어 구성하였다. ProtGPT2의 경우, 각 도메인 서열 앞에 종료 토큰(end-of-text token)을 포함하고, 도메인 서열의 길이가 일정 기준(최대 60자)을 초과할 경우 줄바꿈을 적용하여 입력하였다. 반면, ProGen2에서는 Titin 단백질임을 명시하는 특수 토큰을 도메인 서열 앞에 추가하고, 줄바꿈 없이 하나의 연속된 시퀀스로 입력하였다.

이와 같은 입력 형식 차이는 각 모델의 사전학습 구조 및 토큰화 방식에 따른 것으로, 본 연구에서는 모델별 특성을 고려한 입력 설계를 통해 공정하고 안정적인 fine-tuning 환경을 구성하였다.

2.2.2 데이터 분할

본 연구는 fine-tuning된 모델의 일반화 성능을 안정적으로 검증하기 위해 전체 데이터셋을 Train:Test = 8:2 비율로 분할하였다. 단순한 무작위 분할로는 Titin 단백질의 구조적·통계적 특성이 학습 데이터에 편향될 수 있으므로, 본 연구에서는 Titin 단백질의 특성을 반영한 Stratified 분할 전략을 적용하였다.

구체적으로, 데이터 분할은 도메인(domain) 종류, 아미노산 서열 길이, 그리고 인간/비인간 유래 여부를 기준으로 수행하였다. 이러한 분할 기준을 적용한 이유는 다음과 같다. 첫째, Titin 단백질의 도메인별 서열은 길이 분포가 불균등하므로, 단순 분할 시 특정 길이 구간의 서열이 학습 또는 평가 데이터에 과도하게 집중될 수 있다. 이에 따라 서열 길이 기준의 stratification을 통해 각 길이 구간이 학습 및 평가 데이터에 고르게 분포되도록 구성하였다. 둘째, 동일한 도메인은 유사한 구조적·기능적 특성을 공유하므로, 특정 도메인에 대한 학습 편향을 방지하기 위해 도메인 종류를 분할 기준에 포함하였다. 이를 통해 fine-tuning 과정에서 Titin 단백질의 다양한 구조적 특성이 균형 있게 학습되도록 하였다. 마지막으로, 인간 및 비인간 유래 Titin 서열을 분리하여 분할함으로써, 모델이 특정 종에 과적합되는 것을 방지하고 종 간 일반화된 Titin 단백질 서열 생성 능력을 확보하고자 하였다.

이와 같은 stratified 분할 전략의 효과를 검증하기 위해, 본 연구에서는 동일한 Train:Test 비율을 유지한 상태에서 무작위(random) 분할 방식과 성능을 비교·분석하였다. 이를 통해 제안한 데이터 분할 전략이 Titin 단백질 생성 성능 및 일반화 능력에 미치는 영향을 정량적으로 평가하였다.

2.3 실험 방법

본 연구에서는 ProtGPT2와 ProGen2의 구조적 특성을 고려하여, 각 모델에 대해 차별화된 학습 하이퍼파라미터를 설정하였다. 학습률, 배치 크기, 학습 에폭 수 등 주요 하이퍼파라미터는 사전 실험을 통해 각 모델에서 최적의 성능을 보이는 값을 사용하였다.

Fine-tuning은 두 모델 모두 기존에 공개된 공식 라이브러리를 기반으로 수행하였다. ProtGPT2의 경우 HuggingFace 표준 라이브러리를 사용했으며, ProGen2는 해당 모델을 제안한 선행 연구에서 공개한 전용 라이브러리를 활용했다 [5]. 이를 통해 각 모델의 사전학습 설정 및 토큰화 방식이 유지된 상태에서 Titin 데이터에 대한 fine-tuning을 진행했다.

학습 방식으로는 모델의 일부 파라미터만을 업데이트하는 방식이 아닌, 모든 파라미터를 갱신하는 full fine-tuning을 적용했다. 이를 통해 사전학습된 PLM이 보유한 일반적인 단백질 생성 능력을 유지하면서 Titin 단백질의 구조적 특성과 서열 분포를 보다 효과적으로 반영하도록 하였다.

또한, 모델의 과적합 여부와 일반화 성능을 정량적으로 평가하기 위해 전체 데이터셋을 대상으로 5-fold cross validation을 수행하였다. 각 fold에서 학습된 모델의 성능을 비교·분석함으로써, 특정 데이터 분할에 의존하지 않는 안정적인 성능 평가를 수행했다.

2.4 실험 결과 및 성능 평가

본 절에서는 fine-tuning된 PLM이 Titin 단백질의 특성을 반영한 서열을 생성할 수 있는지를 평가하기 위해, ① 서열 유효성(valid sequence)과 ② 구조적 안정성(structural plausibility) 두 관점에서 성능을 분석한다.

2.4.1 서열 유효성 평가

먼저, 언어 모델 기반 생성 품질을 평가하기 위해 Perplexity가 낮은 상위 50개의 생성 서열을 선별하여 분석하였다. Perplexity는 생성 서열이 학습된 분포에 얼마나 잘 부합하는지를 나타내는 지표로, 값이 낮을수록 언어 모델 관점에서 생성 품질이 우수함을 의미한다.

선별된 생성 서열에 대해, 자연 단백질 데이터베이스에 존재하는 Titin 서열과의 서열 일치율(sequence identity)을 계산하였다. 본 연구에서는 서열 일치율이 40–70% 범위에 속하는 경우를, “자연 단백질의 통계적 특성을 충분히 반영하면서도 기존 서열을 단순히 복제하지 않는” 유효한 새로운 Titin 단백질 서열(valid sequence)로 정의하였다. 이는 기존 단백질 생성 연구에서 사용되는 novel-yet-natural기준을 따른 것이다.

표 1은 Stratified 데이터 분할 여부에 따른 유효 서열 비율을 비교한 결과를 보여준다. ProtGPT2의 경우, 무작위 분할(random split)에서는 유효 서열 비율이 58%에 그친 반면, stratified 분할을 적용했을 때 74%로 증가하여 27.59%의 상대적 향상을 보였다. ProGen2에서는 이러한 효과가 더

Model	Data Split	Average TM-Score
ProGPT2	Random	0.525
ProGPT2	Stratified	0.588
ProGen2	Random	0.364
ProGen2	Stratified	0.396

표 2 Stratified 분할 여부에 따른 TM-Score 비교

욱 뚜렷하게 나타나, 유효 서열 비율이 52%에서 80%로 증가하여 53.85%의 향상률을 기록하였다. 이러한 결과는 stratified 분할을 통해 특정 도메인이나 서열 길이에 대한 편향된 학습이 완화되고, Titin 단백질의 다양한 특성이 보다 균형 있게 학습되었음을 의미한다.

2.4.2 구조적 안정성 평가

서열 유효성 평가에 더하여, 생성된 단백질이 물리적으로 실현 가능한 구조를 형성할 수 있는지를 검증하기 위해 구조적 안정성 평가를 수행하였다. 이를 위해 AlphaFold2를 사용하여 생성된 유효 서열의 3차원 구조를 예측하고, 해당 구조를 기존 자연 Titin 단백질 구조와 비교하였다.

구조 비교 지표로는 TM-Score를 사용하였다. TM-Score는 0에서 1 사이의 값을 가지며, 값이 1에 가까울수록 두 구조 간의 유사도가 높음을 의미한다. 일반적으로 0.5 이상의 TM-Score는 동일한 구조적 fold를 가질 가능성이 높음을 시사한다.

표 2는 stratified 분할 여부에 따른 평균 TM-Score를 비교한 결과를 나타낸다. ProtGPT2의 경우, stratified 분할을 적용함으로써 평균 TM-Score가 0.525에서 0.588로 약 10% 향상되었으며, ProGen2 역시 0.364에서 0.396으로 증가하였다. 이는 도메인 단위로 분할된 Titin 서열을 stratified 방식으로 학습할 경우, 구조적으로 더 안정적이고 자연 단백질과 유사한 구조를 생성할 가능성이 높아짐을 의미한다. 종합하면, stratified 데이터 분할은 서열 수준의 유효성뿐만 아니라 구조적 안정성 측면에서도 일관된 성능 향상을 제공함을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 대규모 단백질 언어 모델을 활용하여 Titin 단백질 서열을 생성하고, 데이터셋 분할 전략이 fine-tuning 성능에 미치는 영향을 분석했다. 이를 위해 모델 구조와 학습 환경은 동일하게 유지한 채, 무작위(random) 분할과 Stratified 분할을 적용하여 생성 성능을 비교·평가했다. 실험 결과, Stratified 분할을 적용한 경우 모든 모델에서 유효 서열 비율과 구조적 안정성 지표가 일관되게 향상됨을 확인하였다. 이는 Titin 단백질의 도메인 종류, 서열 길이, 그리고 종 정보를 고려한 데이터 분할이 편향된 학습을 완화하고, 단백질의 구조적·통계적 특성을 보다 균형 있게 학습하는 데 효과적임을 의미한다. 특히, 본 연구의 결과는 PLM이 단순한 서열 생성에 그치지 않고, 구조적으로 타당한 단백질 생성 가능성을 함께 확보할 수 있음을 실험적으로 입증한다. 이러한 결과는 목적 단백질에 특화된 데이터 구성과 분할 전략이 PLM 기반 단백질 생성의 핵심 요소임을 보여주며, Titin과 같이 긴 서열과 복잡한 도메인 구조를 가지는 단백질 연구에 있어 PLM 활용의 실질적인 가능성을 제시한다. 향후 연구에서는 서열 정보뿐만 아니라 3차원 구조, 물리적 특성, 생체역학적 정보를 함께 학습하는 멀티모달(multimodal) 단백질 언어 모델로 확장함으로써, Titin 단백질의 구조–기능 관계를 보다 정밀하게 분석하고, 나아가 근육의 역학적 특성 연구를 위한 새로운 계산적 프레임워크를 제시하고자 한다.

ACKNOWLEDGMENT

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원으로 이루어진 것임 (RS-2025-02223417). 이 성과는 과학기술정보통신부 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2024-00349885).

참 고 문 헌

[1] L. Lv, Z. Lin, H. Li, Y. Liu, J. Cui, C. Chen, L. Yuan, andY. Tian, “ProLLaMa: A Protein Large Language Model for Multi-Task Protein Language Processing,” *IEEE Trans. Artif. Intell.*, Early Access, 2025.

[2] M. Krüger and W. Linke, “The giant Protein titin: a regulatory node that integrates myocyte signaling pathways,” *J. Biol Chem.*, vol. 286, no. 12, pp. 9905-9912, Mar. 2011.

[3] N. Ferruz, S. Schmidt, and B. Höcker, “ProtGPT2 is a deep unsupervised language model for protein design,” *Nat. Commun.*, vol. 13, pp. 1-10, July 2022.

[4] E. Nijkamp, J. Ruffolo, E. Weinstein, N. Naik1, and A. Madani, “ProGen2: Exploring the boundaries of protein language models,” *Cell Syst.*, vol. 14, no. 11, pp. 968-978. Nov. 2023.

[5] H. Hrbáň and D. Hoksza, “Protein Family Sequence Generation through ProGen2 Fine-Tuning,” in *Proc.* BIBM, Lisbon, Portugal, 2024, pp. 7037-7039.