タンパク質言語モデルを用いた有効変異の効率的スクリーニング手法

竹　村　俊　晃[*1]
内　田　雅　人[*1]

Efficient Screening Method of Desired Variants Using Protein Language Model

Toshiaki TAKEMURA
Masahito UCHIDA

Improving protein function using directed evolution is widely used in the fields of industry and medicine. In recent years, there has been a lot of research on predicting desired variants with machine learning. In particular, protein language models that assume the amino acid sequence to be a string of characters have been successful. These models, adapted from natural language processing are used to predict the effects of mutations on protein stability and enzyme activity. However, previous research has not proposed a desired variant screening method that is compatible with the experimental operations of directed evolution. To solve this problem, we present an effective model that focuses on the mutate position of a protein, and then propose an efficient screening method for desired variants that is compatible with the experimental operations of directed evolution. In previous models, experimental values were predicted using an embedding learned from the field of natural language processing such as classification tokens. In this model, the prediction accuracy is improved by using an embedding of mutate position. Furthermore, we show a method that can efficiently search for variants by using active sampling, which takes advantage of the characteristic that only one mutation can be introduced per site in a protein. This method allows for more efficient searches than conventional directed evolution methods. This search method can be widely used to modify protein functions and has the potential to accelerate the development of new effective proteins.

進化分子工学手法によるタンパク質の機能改良は産業・医療の分野を中心に広く用いられる。近年では進化分子工学手法に機械学習を組み合わせた有効変異の予測が盛んに研究されている。特にアミノ酸配列を文字列とみなし、自然言語処理分野で用いられるモデルを適用したタンパク質言語モデルは、タンパク質の安定性や酵素活性などの変異効果予測に用いられる。しかし、これまでの研究では進化分子工学手法の実験操作に即した有効変異探索手法が提案されてない。この問題を解決するために、タンパク質の変異位置に着目したモデルの有効性を示した後に、進化分子工学手法の実験操作に対応した有効変異の効率的スクリーニング手法を提案する。従来のモデルでは自然言語処理分野で用いられる Embedding を特徴量として実験値を予測していた。本モデルでは変異位置に着目した Embedding を特徴量に用いることで予測精度が向上した。さらに、タンパク質は１部位につき１変異しか導入できない特性を活かしてアクティブサンプリングすることで効率よく探索できる手法を示す。本手法により従来の進化分子工学手法よりも効率的なスクリーニングが可能になった。このようなスクリーニング手法は、タンパク質の機能改変に広く用いられ、新規有効タンパク質の開発を加速させる可能性がある。

＊１　MI センター　MI2 グループ

1. 序論

私たちの身の回りの製品の一部には、進化分子工学手法によって改良されたタンパク質が広く用いられている。進化分子工学手法は、ランダム変異導入と選択を繰り返すことで、特定の目的を持つタンパク質や酵素の機能を改良・最適化する手法である。まず、Error-Prone PCR[1] などを用いて様々な部位のアミノ酸が別種のアミノ酸に変わった変異体を大量に作成し（ライブラリ）、変異導入された遺伝子を宿主細胞に導入してタンパク質を発現させ、その後、発現したタンパク質や酵素の性能を基に測定することで、望ましい特性を持つ変異体（有効変異）を選び出すことが可能となる。また、選ばれた変異体に対してさらに変異を導入するといった、これら一連のプロセスを繰り返すことで、目的とするタンパク質の特性が徐々に向上することも知られている。こうした進化分子工学手法の応用例として、例えば洗剤に使用される酵素開発が知られている。洗剤用酵素は高温の洗濯条件でも安定して機能する必要があるため [2]、進化分子工学手法を用いることでこれらの酵素の耐熱性を向上させることにより、より効果的な洗浄力を持つ製品開発が可能となる [3]。また、医薬品開発の分野では、抗体の特異性や親和性を高めるために進化分子工学手法が利用されており、例えば、がん細胞を特異的に認識する抗体を最適化することで、高い治療効果を持つ抗体医薬品が開発されている [4]。

一方で、タンパク質の変異効果を予測するために、機械学習などの計算科学的手法を用いるアプローチが提案されている [5-8]。このアプローチでは既存のデータからタンパク質の変異とその機能の関係性を学習するモデルを構築することで、未知の変異体の機能を予測することが可能となる。例えば、Tian らはタンパク質の熱安定性指標の一つである自由エネルギー変化をサポートベクターマシンやランダムフォレストで予測した [5]。さらに、2010年代に台頭したニューラルネットワークによって様々なタンパク質の変異効果が予測されている [6, 7, 8]。このようなモデルの予測値から最も有望な変異を選び出して実験的に検証することで、効率的に有効変異を同定できると考えられる。

近年、自然言語処理（NLP）で用いられる深層学習モデルの技術を応用した、タンパク質のアミノ酸配列を解析・予測するタンパク質言語モデル（pLM）を用いた研究が増えている [9-13]。pLM は、多数のタンパク質のアミノ酸配列を「言語」として扱う。すなわち、各アミノ酸を「単語」と見なし、1つのアミノ酸配列を「1文」として解析する手法である。まず大規模なタンパク質データベースのアミノ酸配列データを用いてアミノ酸の連なりやその文脈情報を pLM に学習させる。その後、目的に合わせて pLM を転移学習または Fine-Tuning することで、未知のタンパク質の構造や機能 [10]、相互作用 [11] の予測が可能になるとともに、変異効果の評価 [12] や新規タンパク質の設計 [13] にも応用可能である。pLM は、従来の物理化学的手法や統計的解析に比べて、より迅速かつ高精度に解析できることから注目を浴びている。

しかしながら、pLM を用いた有効変異探索の先行研究では、進化分子工学手法の実験サイクルに即した探索手法は少なく、既に存在する大規模なデータセットを用いることが多いが、こうした大規模なデータセットを最初に構築することは、民間の研究機関や大学の一研究室では実務的に難しい場合が多い。

そこで我々は、上記問題を解決するため進化分子工学手法の実験サイクルに pLM を組み込むことで、開発プロセスを加速可能なアクティブサンプリング手法を開発した。本稿では、まず変異位置に着目したモデル構造が従来法よりも高精度になることを示す。従来の pLM を用いた変異効果予測では、NLP 分野で用いられる分類トークンの Embedding[14] や全配列の Embedding を pooling したもの [15] を回帰タスクの特徴量に用いていたが、今回、変異位置とその両隣の Embedding を特徴量に用いることで予測精度が向上することを示す。次に進化分子工学手法の実験サイクルに合わせた有効変異の効率的スクリーニング手法を示す。タンパク質には一つの部位に一つの変異しか導入できないが、この制約条件を利用したアクティブサンプリング手法がランダム変異手法を模したランダムサンプリング手法と比較して効率的に改良できることを示す。

2. 方法

[1] タンパク質の変異効果予測に適切なモデル構造の検討

タンパク質の変異効果を予測するために、pLM に回帰 Head を組み合わせたモデル構造を採用した。学習済みの pLM には Evolutionary Scale Modeling（ESM）を用いた [10]。ESM は UniProt が提供する UniRef データセットのアミノ酸配列を事前学習した BERT モデル [16] である。また、回帰 Head には2層の全結合ニューラルネットワークを使用した。

本手法では変異位置とその両隣のトークンの

Embedding に対して最大値を返す Max pooling 処理を行い回帰 Head に入力するモデル（mut-around モデル、**Fig.1** (a)）を提案する。アミノ酸配列を BERT に入力すると各アミノ酸に対応した Embedding が得られることから、本手法では、変異位置とその周辺のアミノ酸に対応した Embedding には変異情報が含まれると仮定した。検証用のベースラインモデルとしては、変異位置のトークンの Embedding のみを入力するモデル（mut モデル、**Fig.1** (b)）と、分類トークンの Embedding を入力するモデル（CLS モデル、**Fig.1** (c)）[12] と、全トークンの Embedding を Max Pooling して入力するモデル（Max-Pool モデル、**Fig. 1** (d)）[13] を採用した。

各モデルの精度検証のために **Table 1** に示す 15 個のデータセットを用いた [17]。これらは Deep Mutation Scan などで測定されたデータセットであり、野生型の実験値が 0 となる対数尺度でスケールされている。

転移学習・Fine-Tuning では各データセットの 80% を学習データセットに、20%をテストデータセットに用いた。オプティマイザには Adam を、損失関数は MSE を使用し、適切な学習率で学習させた。各モデル構造について、ESM のパラメータを固定した場合（Transfer-Learning）としない場合（Fine-Tuning）で検証し、テストデータの評価関数にはスピアマンの順位相関係数を用いた。

［2］変異体の効率的スクリーニング

進化分子工学手法の実験サイクルに即した有効変異

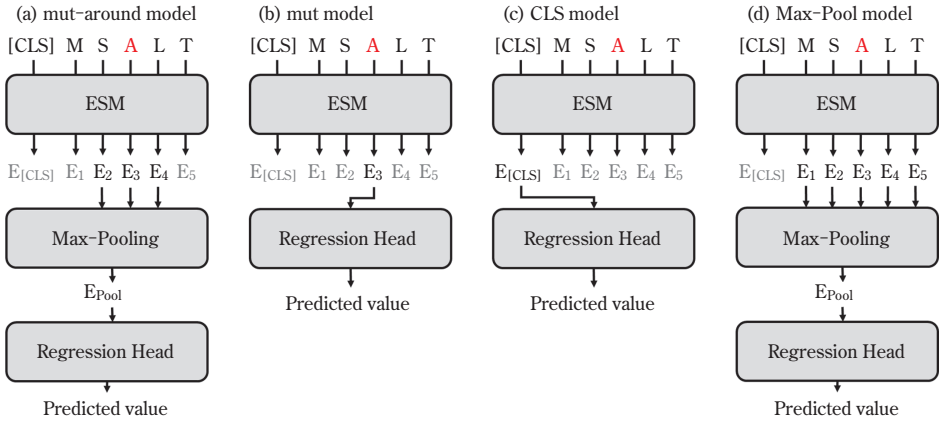

Fig. 1　Model architectures for predicting experimental values

MSALT is the amino acid sequence of the protein. A refers to the mutation position. Each model inputs the following Embeddings into the regression head. (a) mut-around model: Max-pooled embedding corresponding to the mutation position and the amino acids on both sides of the mutation (b) mut model: Embedding corresponding to the mutate position (E_3) (c) CLS model: Embedding corresponding to the classification token ($E_{[CLS]}$) (d) Max-Pool model: Max-pooled Embedding of all sequences embedding (E_{pool})

Table. 1　Datasets for model architectures with different embeddings

Dataset	Number of sample	Sequence length
POLG_HCVJF_Sun2014-fitness	1632	114
DLG4_RAT_Ranganathan2012-CRIPT	1577	101
RL401_YEAST_Bolon2013-selection_coefficient	1161	76
BG_STRSQ_Abate2015-enrichment	2634	501
BLAT_ECOLX_Palzkill2012-ddG_stat	4808	263
BLAT_ECOLX_Ostermeier2014-linear	4610	263
BRCA1_HUMAN_Fields2015-y2h	1335	110
BLAT_ECOLX_Tenaillon2013-singles_MIC_score	951	263
HSP82_YEAST_Bolon2016-selection_coefficient	4104	240
BLAT_ECOLX_Ranganathan2015-2500	4807	263
GAL4_YEAST_Shendure2015-SEL_C_40h	1123	75
BRCA1_HUMAN_Fields2015-e3	1382	110
RL401_YEAST_Bolon2014-react_rel	1295	76
KKA2_KLEPN_Millenlsen2014-Kan18_avg	4384	264
MTH3_HAEAESTABILIZED_Tawfilk2015-Wrel_G17_filtered	1634	330

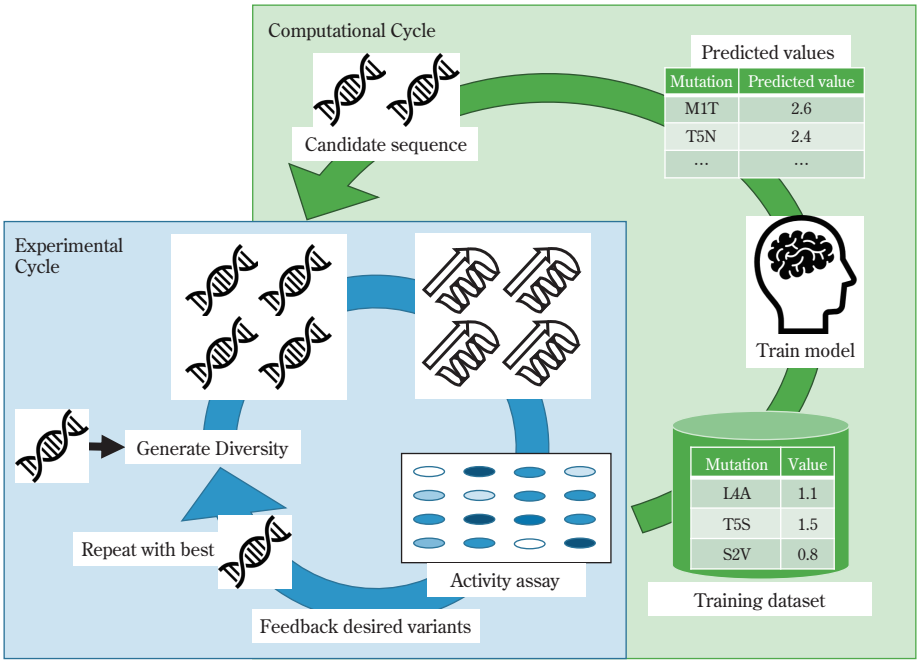

Fig. 2　Overview of efficient desired variants screening method

の効率的スクリーニング手法の概要を示す（**Fig.2**）。進化分子工学手法では「ライブラリ作成→スクリーニング→有効変異の集積」のサイクルを繰り返す（**Fig.2**, experimental cycle）。そこで、この実験サイクルと協奏しながら、計算上有効と思われる変異をサンプリングする手法を提案する（**Fig.2**, computational cycle）。本手法では最初の実験サイクルのスクリーニングで得られた 200 個の教師データでモデルを構築し、未探索の単変異体の実験値を予測する。その中から候補変異を 20 個選択し、次の実験サイクルで性能を評価する。次の計算サイクルでは元の教師データに評価した 20 個の候補変異の実験データを追加する。mut-around モデルを用いて 2.1 と同様の計算条件で計算し、4 サイクル探索したときの 80 個（20 サンプル×4 サイクル）の変異体を評価した。

20 個の候補変異を選択するためのアクティブサンプリング手法を示す。単変異に限定しても、探索空間は配列長×19 アミノ酸と膨大であるため、全空間を実験的に検証することは難しい場合が多い。しかしながら、タンパク質は各部位ごとに 1 変異しか導入できないため、1 部位で多数の変異を検証するよりも、各部位ごとに 1 つの変異を検証した方が効率的である。そこで、各部位ごとに予測値が最も良い変異を抽出し、その中の上位 20 変異を候補配列とした。これにより配列長×19 変異の変異空間の中から効率よく高活性体をサンプリングできると仮定した。比較対象として、未探索の変異体の中から 20 変異体をランダムに選択した場合（ランダムサンプリング）についても検証を行った。

検証用に ProteinGym[18] で公開されている 4 つのデータセットを用いた（**Table 2**）。これらのデータセットは単変異体の探索空間の 95%以上を網羅している。

今回、我々は、進化分子工学手法に合わせた評価指標を定義した。進化分子工学手法では一般に元のタンパク質よりも高活性な変異を導入するが、これは有効変異を集積するほど活性が向上する「加法性」を仮定している。そこで、4 回のサイクルでサンプリングした 80 個の変異体の内、野生型よりも高活性（値が 0 以上）な実験値の「合計」を評価指標とした（サンプリングスコア）。なお、データセットが野生型を 0 と

Table. 2　Datasets for comparing sampling methods

Dataset	Number of sample	Sequence length
KCNE1_HUMAN_Muhammad_2023_expression	2339	129
DYR_ECOLI_Nguyen_2023	2916	159
RNC_ECOLI_Weeks_2023	4277	226
OXDA_RHOTO_Vanella_2023_expression	6769	364

Table. 3　Average of Spearman's rank correlation coefficient

		mut-around model	baseline model		
			mut model	CLS model	Max-Pool model
ESM	Fine-Tuning	0.756	0.748	0.666	0.713
	Transfer-Learning	0.729	0.718	0.631	0.679

Table. 4　Sampling scores for each dataset

Dataset	active sampling	random sampling
KCNE1_HUMAN_Muhammad_2023_expression	47.436	19.662
DYR_ECOLI_Nguyen_2023	35.12	12.62
RNC_ECOLI_Weeks_2023	1.011	0.438
OXDA_RHOTO_Vanella_2023_expression	2.409	1.2

したログスケールであることから、サンプリングスコアは同一データセットのサンプリング方法の違いを比較するための指標であり、データセット間で比較できないことに注意する必要がある。

3. 結果と考察

［1］Embedding の変異位置依存性

15 個のデータセットを用いて mut-around モデルと 3 つのベースラインモデルの予測精度を比較した。**Table 3** に各条件における 15 データセットの順位相関係数の平均を示す。Fine-Tuning 時に ESM のパラメータを固定しない場合（Fine-Tuning）、mut-around モデルは 0.756, mut モデルは 0.748, CLS モデルは 0.666, Max-Pool モデルは 0.713 であった。固定した場合（Transfrer-Learning）はそれぞれ 0.729, 0.718, 0.631, 0.679 であった。ESM のパラメータ固定の有無に関わらず、mut-around モデルがいずれのベースラインモデルよりも高精度であった。また ESM のパラメータを固定しない方が高精度であった。全条件の中で最も高精度なモデルは Fine-Tuning した mut-around モデルであった。

今回、提案手法である mut-around モデルはベースラインモデルと比較して変異効果を正確に予測できることが示された。今回の結果からは、変異位置とその周辺の Embedding を用いて予測し、変異していない部位の情報を用いないことで高精度になると考えられる。ESM が採用する BERT モデルには Attention 機構があり、一つのトークンに配列全体の情報が取り込まれることから、NLP 分野では配列先頭の分類トークンの Embedding を配列全体の特徴量として用いる。また、Max-pool モデルでは、配列全体の Embedding を Max Pooling して回帰 Head に入力するため、配列全体の情報で予測していると見なせる。このことから、配列全体を代表する Embedding では変異効果の予測が困難であるのに対し、変異位置の Embedding はその他の野生型と同じアミノ酸配列の情報の影響を受けにくいため予測精度が高くなったことが示唆される。

［2］変異体の効率的スクリーニング手法

次に探索・集積のサイクルを繰り返す進化分子工学手法に合わせた有効変異のスクリーニング手法の結果を示す。4 回のサイクルのサンプリングスコアを **Table 4** に示す。各データセット間を比較すると、ランダムサンプリングと比較してアクティブサンプリングは 2 倍以上のサンプリングスコアをマークした。

このように、アクティブサンプリングはランダムサンプリングよりも有効変異を効率的に探索できることが判る。今回、200 データ程度の比較的少ない実験データでモデルを構築し候補配列を決定することで、ランダムサンプリングよりも効率的に変異をサンプリングすることができた。探索空間の数パーセントのデータ数にも関わらずランダムサンプリングより有効であり、**Table 2** のような 1000 オーダーのデータセットを検討初期に作製しなくとも効果的に予測ができ、開発効率を高められることが示唆された。

4. まとめ

今回、我々は変異位置に対応する Embedding が分類トークンや配列全体の Embedding を Max Pooling したものよりも変異効果を高精度で予測できることを見出した。すなわち、pLM における変異情報の配列位置依存性が示唆される結果であり、これまでに提案されている様々な変異効果予測モデルについても、配列全体の Embedding ではなく、変異位置の

Embeddingを使用することで更に精度が向上する可能性がある。

また、pLMを用いた効率的スクリーニング手法はランダムサンプリングよりも効率的に探索が可能であることを示した。有効変異は各部位に一つずつしか導入できないため、各部位ごとに最も高活性と思われる変異を検証することは有用であり、さらに探索空間の数パーセントのデータ数のみでモデルを構築してもランダムサンプリングより効率が高いことが示された。

一方で私たちが提案した効率的スクリーニング手法には二つの課題がある。一つ目は1サイクル目の学習データセットの取得である。従来の1000オーダーのデータセットと比較して取得コストは低減してはいるものの、200変異体の配列データ取得が本手法の実務的なボトルネックとなり得る。このボトルネックを解決するためには実験的な配列データ取得のスループット向上が必要である。二つ目は多重変異体を考慮していない点である。一般的に多重変異体は単変異体よりも実験値を予測することが難しいことが知られている。さらに今回検証に用いたデータセット内の多重変異体のデータ不足から、今回は多重変異体については考慮しなかった。実務的には多重変異の変異効果を予測可能とすることで集積過程の開発期間も短縮できるため、さらに効率的な高機能タンパク質の探索手法となることが期待される。

本稿で提案したpLMを用いた効率的スクリーニング手法により、進化分子工学手法による開発フェーズを短縮できる可能性が示された。本手法によりライフサイエンス関連製品の開発が加速され、人の健康と福祉に貢献することを期待している。

5. 参考文献

［1］D. W. Leung, E. Chen, and D. V. Goeddel, *Technique*, **1**, 11 (1989).

［2］L. Vojcic, C. Pitzler, G. Körfer, F. Jakob, R. Martinez, K. H. Maurer, and N. Schwaneberg, *Nat. Biotechnol*., **32**, 629 (2015).

［3］K.-H. Maurer, *Curr. Opin. Ciotech*., **15**, 330 (2004).

［4］R. Amon, R. Rosenfeld, S. Perlmutter, O. C. Grant, S. Yehuda, A. Borenstein-Katz, R. Alcalay, T. Marshanski, H. Yu, R. Diskin, R. J. Woods, X. Chen, and V. Padler-Karavani, *cancers*, **12**, 2824 (2020).

［5］J. Tian, N. Wu, X. Chu, and Y. Fan, *BMC Bioinfo*., **11**, 370 (2010).

［6］H. Cao, J. Wang, L. He, Y. Qi, and J. Z. Zhang, *J. Chem. Inf. Model*., **59**, 1508 (2019).

［7］Y. B. L. Samaga, S. Raghunathan, and U. D. Priyakumar, *J. Phys. Chem. B*, **125**, 10657 (2021).

［8］S. Wang, H. Tang, P. Shan, Z. Wu, and L. Zuo, *Comput. Bio. And Chem*., **17**, 107952 (2023).

［9］山口秀輝、齋藤裕、*JSBi Bioinformatics Review*、**4**、52 (2023).

［10］Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shimueli, A. D. S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, *Science*, **379**, 1123 (2023).

［11］K. Zhou, C. Lei, J. Zheng, Y. Huang, and Z. Zhang, *Plant Methods*, **19**, 141 (2023).

［12］E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, *Nat. Methods*, **16**, 1315 (2019).

［13］N. Ferruz, S. Schmidt, and B. Hocker, *Nat. Commun*., **13**, 4348 (2022).

［14］D. Umerenkov, F. Nikolaev, T. I. Shashkova, P. Strashnov, M. Sindeeva, A. Shevtsov, N. Ivanisenko, and O. L. Kardymon, *Bioinfo*., **39** (2023).

［15］R. Schmirler, M. Heinzinger, and B. Rost, *bioRxiv*, doi.org/10.1101/2023.12.13.571462 (2023).

［16］J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *arXiv*, 1810.04805 (2019).

［17］T. A. Hopf, J. B. Ingraham, F. L. Poelwijik, C. P. I. Scharfe, M. Springer, C. Sander, and D. S. Marks, *Nat. Biotech*., **35**, 128 (2017).

［18］P. Notin, A. Kokkasch, D. Ritter, L. van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, J. Frazer, M. Dias, D. Franceschi, Y. Gal, and D. Marks, in *NeurIPS* (2023).