大規模言語モデルの Fine-Tuning による B 細胞エピトープ予測		
黒田研究室	学籍番号:22261010	氏名　伊藤紫苑

[背景]

機械学習モデルによる B 細胞エピトープ（抗体結合部位）の特定は、中和抗体の誘導を目的とした抗原設計において極めて重要である。従来、立体構造や配列情報に基づく予測手法が数多く提案されてきたが、その利用には限界がある。そのため、膨大な学術論文などのテキストデータの活用が期待されている。近年、大規模言語モデル（LLM）はテキストデータ処理において高い性能を示すことで注目されている。本研究では、LLM に対して免疫学の学術知識と B 細胞エピトープデータを段階的に学習させることで、B 細胞エピトープ予測能の改良を検証した。

[研究手法]

ベースモデルとして Mistral-7B-v0.3（以下、オリジナルモデル）を用い、以下の 2 段階の学習を実施した。一つ目は、新しい知識の獲得を目的とした継続事前学習（CPT）である。Immune Epitope Database（IEDB）より得られた論文 1035 報を学習データに用い、モデルに専門用語の文脈および科学的知見を学習させた。二つ目は、定着させた知識を元に、応答を調節する教師あり学習（SFT）である。Structural antibody database（SabDab）よりフィルタリングした、抗原タンパク質と B 細胞エピトープ部位の対応データ 793 件を用い、配列情報から B 細胞エピトープ残基をリストアップするよう微調整した。

これらの学習を通し、オリジナルモデル、CPT モデル、SFT モデル、CPT-SFT モデルの 4 種のモデルを作成した。作成した各モデルに対し、評価用データとして未学習の抗原タンパク質を用い、予測精度の検証を行った。検証条件として、タンパク質のアミノ酸全長配列を入力とする場合と、タンパク質の表面露出残基のみを入力とする場合の 2 群を設定した。

[結果]

Fine-Tuning を行う前の LLM のエピトープ残基を予測する能力を有していなかった。入力に全長配列を用いた場合、AUC=0.5382 であった。また、Fine-Tuning を行ったモデルにおいても、予測性能は向上しなかった。

一方、入力配列にタンパク質の表面露出残基を用いた場合では、CPT-SFT 学習を行ったモデルが AUC=0.6060 と最も高いスコアを示したが、オリジナルモデルのスコアは AUC=0.5739 であり、顕著な変化は見られなかった。

	Original model	CPT model	SFT model	CPT-SFT model
Input: Full sequence	0.5382	0.5617	0.5199	0.5617
Input: Surface residues	0.5739	0.5903	0.6021	0.6060

考察として、エピトープ予測にはタンパク質の表面露出残基といった三次構造情報が不可欠であることが考えられる。したがって、今後はテキストデータを学習させた LLM に加え、タンパク質の三次構造情報をベースとするモデルを組み合わせたフレームワークを開発することで、構造的制約を考慮した高精度なエピトープ予測能の獲得を目指す。