Vol.2020-BIO-64 No.1
2020/12/7

大規模タンパク質データベースに基づくBERTを用いたペプチド結合予測

玉木 竜二[1,a)]　農見 俊明[1]　佐藤 尭彰[1]　井上 誠一[1]　貞光 九月[1]
坂口 誠[2]　天満 昭子[2]　中神 啓徳[3]

概要：ワクチン開発において，B 細胞エピトープ予測と，MHCII に対するペプチドの結合予測はいずれも重要な予測タスクである．B 細胞エピトープを予測することは，抗原に特異的な抗体産生を誘導するワクチンの設計・開発のために有益である．一方，感染の重症度を低減する T 細胞を活性化するワクチン開発に対しても，MHCII に対するペプチドの結合を予測する必要がある．これら予測タスクに対する機械学習を用いた従来手法には，以下の二つの課題がある．一点目は離れたアミノ酸間の複雑な依存関係を捉えていない課題，二点目は学習データが不十分な場合に精度が低いという課題である．これらの課題に対処するために，本稿では大規模タンパク質データベースにより事前学習した，自己注意機構を持つ BERT モデルを用いた手法を提案する．実験の結果，提案手法は B 細胞エピトープ予測，MHCII に対するペプチドの結合予測の実験で従来よりも高い性能を達成した．

キーワード：B 細胞, T 細胞, MHCII, ペプチド結合予測, タンパク質 , アミノ酸配列 , BERT, 事前学習

Prediction of peptide binding using BERT based on a large scale protein database

1. はじめに

ワクチンは感染症に対して免疫を獲得するために人為的に投与される，弱毒化した病原体などの物質である．歴史的には弱毒化した生の病原体を直接摂取することで体内の免疫応答を生じさせてきたが，後に死んだ病原体やその抗原，毒素のみでもワクチンとして利用されている．感染症に対する免疫獲得にはリンパ球，特に B 細胞と T 細胞が関与する．生体内において抗原特異的な免疫応答を誘導する B 細胞はそれ自身の細胞膜に B 細胞受容体 (BCRs) を発現し，抗原タンパク質と直接結合してそのエピトープ領域を認識する．これにより，抗原特異的な抗体を大量に産生することができる．一方 T 細胞が病原体タンパク質を認識するためには，病原体タンパクが免疫原として抗原提示細胞に貪食処理されてできた抗原ペプチドが，主要組織適合遺伝子複合体 (major histocompatibility complex; MHC) 分子と結合することで細胞表面に提示される必要がある．その後 T 細胞は 提示されたペプチドを認識して応答し，抗体を産生する B 細胞を活性化したり，感染細胞を直接破壊したりする．

これらの性質に基づき，最近では B 細胞や T 細胞が受容する構造を持ったペプチド小片を投与することで免疫獲得を試みる，ペプチドワクチンが研究されている [1], [2]. 特に特定の抗原エピトープ領域の構造を模したワクチンを投与することで，その抗原特異的なモノクローナル抗体産生を誘導できる可能性がある．

B 細胞，T 細胞に対して働くペプチドワクチンの候補を探索するためには，B 細胞受容体に結合するペプチドや MHC 分子に結合するペプチドをそれぞれ予測することが必要である．この予測には従来タンパク質同士の立体結合を精密に明らかにする必要があるとされてきた [2], [3]．抗体抗原複合体の立体構造を実験的に決定するには非常に時間や労力がかかり，ワクチン開発のように大量の候補ペ

[1] フューチャー株式会社 , Future Corporation
[2] 株式会社ファンペップ, FunPep Co., Ltd.
[3] 大阪大学大学院, Osaka University Graduate School of Medicine.
[a)] r.tamaki.k3@future.co.jp

プチドをアッセイするには不適である．計算機シミュレーションによる結合予測では，タンパク質を剛体とみなした場合の結合しか調査することができない課題 [4], [5] や，分子動力学法ベースの手法を用いた場合でも時間と計算資源を膨大に要する課題が残る [6].

一方，近年ではタンパク質同士の結合を陽に扱わない，機械学習 [7], [8], [9], [10] を用いた結合予測の研究が進んでいる．しかし，機械学習を用いた B 細胞エピトープ予測と MHC 分子，特に MHC クラス II 分子 (MHCII) に対するペプチドの結合予測には以下 2 点の課題があるため未だ高い精度を達成できていない.

１点目は長距離の依存関係を学習するのが困難という課題である．近年ではアミノ酸配列の系列データを対象として，深層学習の一種である LSTM(Long Short-Term Memory)[11] を用いることで長距離依存関係の学習が試みられている [8], [9], [12]．LSTM はセルと呼ばれる記憶部分により長距離依存関係の問題に対処するモデルであるが，依然として長距離のアミノ酸間のネットワークを経由する必要があり，情報が失われる可能性がある [13]．更にタンパク質に適用する際には，タンパク質の高次構造を原因とした依存関係の把握も課題として残る．特に MHCII に結合するペプチド予測のタスクでは，MHCII，ペプチド双方のアミノ酸の複雑な相互作用を捉える必要がある．LSTM ではこのような相互作用を捉えることが難しい．なお，MHC 分子の主要な 2 つのクラスのうちの 1 つである MHC クラス I 分子（MHCI）に結合するペプチド予測では，MHCII と比べアミノ酸配列が短いため，長距離の依存関係の学習の問題が少なく，機械学習を用いた予測により高い精度を得られることが先行研究 [14] で報告されていることから，本稿の扱うタスクの対象外とした.

２点目の課題は教師付きの学習データが少ない場合に十分な汎化性能が得られないという課題である [9], [10], [15]. 汎化性能を向上させるための最も単純なアプローチはより多くの教師付き学習データを準備することであるが，学習データを増やすための生物学実験には多大なコストが必要なため容易ではない.

これらの課題を解消するために，我々は大規模タンパク質データベースに基づいた BERT(Bidirectional Encoder Representations from Transformers)[16] を用いることを提案する．BERT は，LSTM のように系列情報を中間ベクトルとして保持することは行わず，注意機構のみを用いることで可変長かつ長い配列をモデリングできる．LSTM ではアミノ酸間の距離が遠いほど多くのネットワークを経由する必要があり，その結果遠く離れたアミノ酸の情報を失うリスクがある．この距離が短ければ短いほど長距離の依存関係を学習しやすくなるが [13]，注意機構は配列内のアミノ酸の位置に関わらず，アミノ酸間の関係を直接モデル化できる利点がある．この注意機構により，2 つの系列データ間の複雑な相互作用を捉えることもできる [17], [18].

BERT のもう一つの優れた特徴は、大量の教師無しデータを事前学習として利用することが可能な点である。本タスクの場合、大量のタンパク質データを元に BERT を事前学習することで、学習データが少ない場合でも汎化性能を向上できる可能性がある。

本研究では B 細胞のエピトープ予測と MHCII に対するペプチドの結合予測の 2 つのタスクに対して，3100 万のタンパク質ドメインのデータベースである Pfam[19] により事前学習された BERT モデルを用いることで長距離依存関係の問題と，学習データが少量の場合における精度の低下の課題に同時に対応する．実験の結果，2 つのタスクにおいて先行研究よりも高い精度を達成し，提案手法がこれらのタスクの双方に有効であることを示す.

2. 関連研究

B 細胞エピトープ予測では，研究初期はタンパク質を構成するアミノ酸の物理化学的性質のみを特徴量として用いた予測であった [20]．その後，アミノ酸配列自体の情報を組み入れた機械学習に基づく手法が，比較的高い精度を達成しており，サポートベクターマシンを用いた手法 (公開ツール Lbtope[21])，Ramdom Forest を用いた手法 (公開ツール BepiPred-2.0[10])，順伝播型ニューラルネットワークを用いた手法 (公開ツール DLBepitope[15])，リカレントニューラルネットワークを使用した手法 (公開ツール ABCpred[22])，注意機構付き LSTM を用いた手法 [8] 等多くの手法が提案されている．Lbtope と BepiPred-2.0 はジペプチド組成の特徴量を用いているものの，抗原タンパク質における長距離特徴量はモデルに取り込めていない．DLBepitope[15] では単純な順伝播型ニューラルネットワークを用いているため，系列情報や注意機構を持つモデルに比べて複雑なアミノ酸間の依存関係を学習するのが難しい．ABCpred は RNN を用いて系列情報を学習しようとしているが，RNN は長距離の依存関係を捉えることができない．注意機構付き LSTM を用いた手法 [8] では，そのような長距離の依存関係を考慮した特徴量と，抗原タンパク質全体の構造的・化学的特徴量を併用することで，BepiPred-2.0 より高い精度を達成している．しかし，LSTM を用いているため，前節で述べた通り，遠く離れたアミノ酸間の情報が失われる可能性がある.

MHCII に対するペプチドの結合予測でも多くの機械学習法が提案されているが，以下の 2 つの課題に対し適切に対処する必要がある．1 つ目の課題はアミノ酸配列のような長い系列データに対するアルゴリズムにおいて，勾配消失問題や長距離依存性の問題が発生する問題である [13]. DeepSeqPanII[9] は注意機構付き LSTM に更に CNN を組み合わせたモデルを使用することで対処しているが，モデルに LSTM を用いているためこの問題は残される．2 つ

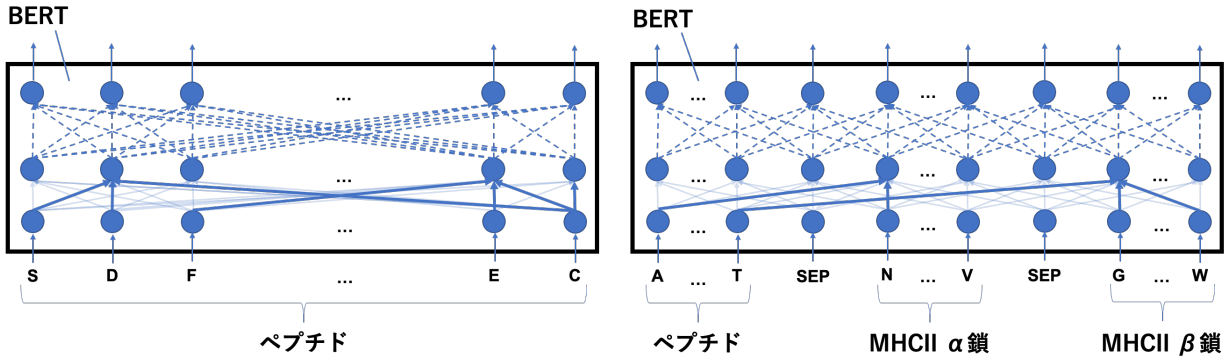

図 1 左：B 細胞エピトープ予測の入力の表現．右：MHCII に対するペプチドの結合予測の入力の表現．図中下部の矢印の線の太さは注意の重みを表す．図中上部の点線は層の省略を表す．SEP は separater を表す．

目の課題はペプチドと MHCII の相互作用を考慮する必要がある点である．先行研究 [23] では，化合物のグラフ表現とタンパク質のアミノ酸配列を入力とし，化合物のグラフ表現を Graph neural network(GNN)[24]，アミノ酸配列を Convolutional neural network(CNN)[25] で低次元のベクトルに射影し，それらのベクトル間の相互作用を注意機構を用いることで捉えており，既存手法よりも高い性能を達成した．しかし GNN は構造の情報が入力に必要なため，後述する大規模タンパク質データベースで事前学習を行うことができず，さらに CNN では長距離の依存関係を捉えることができない．本研究ではモデルに注意機構のみを用いることにより，大規模タンパク質データベースで事前学習を行い汎化性能を向上させつつ，長距離依存性の問題とペプチドと MHCII の相互作用を考慮する問題を同時に解決することに取り組む．

さらに機械学習一般に，学習データが少ない場合に学習データから特徴を十分に学習できないという課題がある．単純に学習データを増やすことは多大なコストが必要なため難しく，近年ではデータ拡張，事前学習といった手法が用いられる．データ拡張とは，類似したデータに対し機械学習モデルで予測をし，仮のラベルをつけることで学習データを拡張する手法である．先行研究 [26] では，MHCII に対するペプチドの結合予測において，質量分析から得られるリガンドデータを用いてペプチドの結合予測の学習データを拡張することでモデルの性能を上げている．事前学習とは，解きたい対象のタスクの前に違うタスクで学習を行うことにより，対象タスクに有効な特徴を獲得する手法である [16]．本研究ではこの事前学習の手法に着目し，学習データが少ない場合にも頑健な予測を可能にするため，大規模タンパク質データベースで BERT で事前学習することにより，本課題の解決に取り組む．

3. 提案手法

本稿で扱うタスクの課題として，離れたアミノ酸間の複雑な依存関係を学習することが難しい，学習データが不十分な場合に精度が低いという 2 つの課題がある．本稿ではこれらの課題を解決するために，BERT を応用したペプチド結合予測の手法を提案する．BERT の利点として以下の点が挙げられる．

- 注意機構により長距離の依存関係や、MHCII とペプチド結合におけるアミノ酸間の相互作用を学習できる
- 教師なし大規模データベースを事前学習に用いることで、モデルの汎化性能を向上できる

3.1 BERT

BERT の主な特徴の 1 つは双方向の Transformer[27] をアーキテクチャに採用したことで，LSTM よりも長距離の依存関係を学習できるようになった点である．Transformer は注意機構のみからなるニューラルネットワークであり，Transformer は LSTM より長距離の依存関係を捉えられることが知られている [28]．自然言語処理の様々なタスクで LSTM よりも単方向の Transformer を多層にしたモデルの方が性能が良いことが知られている [29]．BERT は更に Transformer を双方向にすることにより，逆向きからの系列情報も併用し，性能を上げている．今回のタスクでは各アミノ酸の長距離の依存関係を捉えることが重要になる．特に MHCII に対するペプチドの結合予測では，MHCII，ペプチド間の複雑な相互作用を学習することも重要である．BERT はこのような長距離の依存関係を捉えることができ，更に注意機構を多層にすることにより複雑な相互作用を学習することが期待できる．

BERT のもう 1 つの主な特徴は，大規模なデータセットに対して Masked Language Modeling で事前学習することにより，汎化性能を向上できる点である．自然言語処理における Masked Language Modeling とは，入力の文章中から一部の単語を欠損させ，その欠損させた単語を周りの文脈から予測するタスクであり，このタスクにより文脈を考慮した単語自体の特徴を学習することができる．Masked

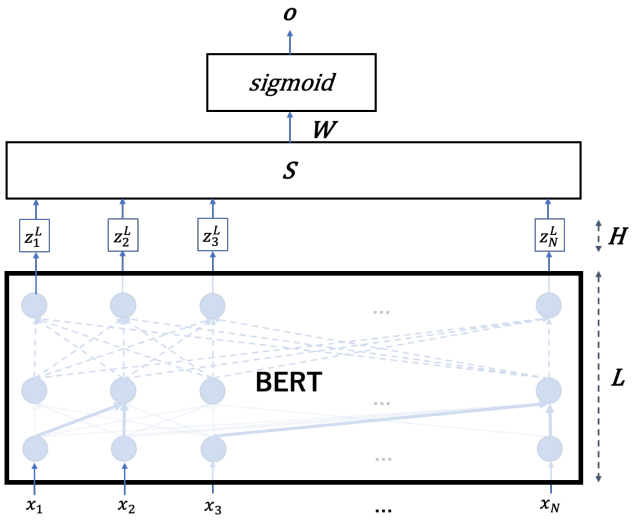

図 2 BERT モデルのネットワーク構造

Language Modeling は単語自体が一種の教師有りデータとして機能するため，大量の教師なしテキストデータを利用することができる．BERT はこの事前学習により，自然言語処理の様々なタスクで更に性能を向上させた．また事前学習は，データが少ない場合においても他の手法に比べ効率よく学習できることが報告されている [29], [30].

3.2 BERT のペプチド結合予測への適用

本稿では，長距離のアミノ酸間の依存関係を捉えるため BERT をペプチド結合予測に適用する．B 細胞エピトープ予測では単独のペプチド配列のみを入力とするため，図 1 のようにペプチド配列をそのまま BERT の入力とすることで，BERT の直接利用が可能である．MHCII 分子のペプチド結合予測では，ペプチド配列，MHCII 分子の α 鎖のアミノ酸配列，MHCII 分子の β 鎖のアミノ酸配列の 3 つ組を入力として扱うため，BERT の適用方法は自明ではない．従来，BERT 以前のモデルでは複数の入力を独立にベクトルに射影してから双方向の注意機構を適用していた [17], [18]．本研究では入力を separate token と呼ばれる特別なトークンを間に入れて連結し，双方向の自己注意機構を適用することで効果的に複数の入力間の関係をモデル化する（図 1）．これにより，入力が単一のアミノ酸配列か複数のアミノ酸配列かに関わらず，同様にモデル化できることが期待できる．

しかし，4 節で後述する通り，教師あり学習データのみに対し BERT の学習を行った場合，即ち事前学習を行わなかった場合において，BERT がうまく学習できないことが判明した．これは B 細胞のエピトープ予測，MHCII に対するペプチドの結合予測で使用する学習データが少ないために，BERT が持つ巨大なパラメータを調整するのが困難であるためと考えられる．そこで，大規模タンパク質データベースの Pfam[19] を用いて BERT の事前学習を実施することとした．B 細胞のエピトープ予測で利用する教師あり学習データが約 20 万，MHCII に対するペプチドの結合予測で利用する教師あり学習データが約 5 万であるのに対し，Pfam は約 3100 万のタンパク質データベースである．そのため，Pfam に対し Masked Language Modeling で事前学習することで，BERT が持つ巨大なパラメータを調整することが可能であると考える．また，Pfam で事前学習した BERT はタンパク質の構造や機能の学習 [31] や，二次構造予測，接触予測といったタスクの汎化性能を向上させられることが先行研究によって報告されている [32]．本稿が扱うタスクにおいても，事前学習により性能が向上することが期待できる．

モデルのアーキテクチャを図 2 に示す．層数 L は 12, 隠れ層のサイズ H は 768 である．このモデルは入力として N 文字のアミノ酸のシーケンス $x = (x_1 \dots x_N)$ を受け取った後，出力では，l 層において H 次元の embedding vector のシーケンス $Z^l = (Z_1^l \dots Z_N^l)$ を出力する．本稿では BERT の最終層の出力 $Z^L = (Z_1^L \dots Z_N^L)$ の平均ベクトル S を入力シーケンスの集約表現として使用する．ベクトル $S \in \mathbb{R}^H$ は分類層の重み $W \in \mathbb{R}^H$ と内積を取ることにより，1 次元に射影される．最終的な出力 o は sigmoid 関数により活性化され，最適化には交差エントロピー誤差を用いた．なお本稿では事前学習ありの BERT として，大規模タンパク質データベース Pfam に基づき事前学習された BERT[*1] [32] を実験に用いた．

$$S = \frac{1}{n}\sum_{i}^{n} Z_i^L$$

$$o = sigmoid(SW^T)$$

4. 実験

本節では提案手法の有効性を確認するために，B 細胞のエピトープ予測と，MHCII に対するペプチドの結合予測における実験結果を示す．

4.1 B 細胞エピトープ予測

4.1.1 実験設定

B 細胞エピトープ予測における提案手法の有効性を確認するために，以下の手法との比較を行う．

- 注意機構付き LSTM[8]
- bepipred-2.0[10]
- DLBepitope[15]

先行研究 [8] では抗原タンパク質全体の構造的・化学的特徴量及びペプチド前後の配列を用いているが，本研究で使用した注意機構付き LSTM は提案手法と条件を合わせるためにそれらの特徴量を使用しなかった．既存研究の報告

[*1] https://github.com/songlab-cal/tape

表 1 B 細胞エピトープ予測の実験結果

	テスト		検証
	Lbtope_Fixed	ABCpred16	
bebipred-2.0 [10]	0.532	0.553	-
DLBepitope [15]	0.738	0.655	0.908
LSTM w/ attention [8]	0.750	0.684	0.878
BERT w/o pre-train	0.586	0.602	0.788
BERT w/ pre-train	**0.774**	**0.734**	**0.911**

の他に，事前学習の効果を確かめるために，事前学習なしの BERT モデルでも実験をした．事前学習ありの BERT と同じパラメータをもつ BERT を事前学習なしで実験した結果，学習が進まないことが実験よりわかった．そのため，事前学習ありと比べパラメータ数を 2%まで小さくして実験を行った．モデルの評価指標として，Area Under Curve(AUC) を用いた．

4.1.2 データセット

教師あり学習用データには DLBepitope[15] で使用されたデータセットの 1 つである DLBepitope20[*2]を使用した．DLBepitope20 には 225210 個のペプチドを含む DLBepitope20_train と，6454 個のペプチドを含む DLBepitope20_test という 2 つのデータがあり，学習 (train) 用データとして*"train"*，検証 (validation) 用データとして*"test"*を使用した．このデータセットは，IEDB[33] から取得した 10〜50 までのペプチドを 20 の長さに揃えた Positive 208437 個，Negative 23227 個のものである．揃え方は，National Center for Biotechnology Information (NCBI) database からダウンロードした抗原タンパク質を元にペプチドの中心位置から前後 10 アミノ酸を取得するという方法で行われた．また，テスト用データセットとして，Lbtope_Fixed dataset[21]，ABCpred16 dataset[22] を使用した．これは Lbtope と ABCpred で使用していたデータセットから DLBepitope dataset の共通部分を削除したものであり，DLBepitope で他のモデルと比較するためのデータセットとして使用されたものである．Lbtope_Fixed dataset は，長さ 20 の Positive 8661 個，Negative 16492 個のデータセットで，ABCpred16 は，長さ 16 の Positive 107 個と Negative 196 個のデータセットである．DLBepitope との正確な比較を行うため，LBtope Fixed dataset の評価は DLBepitope と同じ実験設定で行った．ABCpred16 の評価に関しては，DLBepitope では長さを揃えるため DLBepitope16 という長さが 16 のデータを使用して学習して評価しているが，本論文では汎用性のあるモデルの評価を行うため DLBepitope20 で学習した結果を用いて ABCpred16 の予測を行った．全てのデータは DLBeptiope のサイト[*2]に掲載のデータを用いた．

4.1.3 実験結果

表 1 に B 細胞のエピトープ予測の実験結果を示す．同条件での比較となる DLBepitope，注意機構付き LSTM と比較すると，両方のデータセットで事前学習ありの BERT は精度が大きく向上していることが確認できる．また事前学習あり BERT と事前学習なし BERT を比較すると事前学習のない BERT は検証データ，テストデータ両方で事前学習あり BERT より大きく精度が劣ることが確認できた．これは DLBepitope20 の学習データが少なかったため，巨大なパラメータをうまく学習させることができなかったためと考えられる．また事前学習あり BERT は DLBepitope20 という長さ 20 のみのデータセットのみで学習しても，長さ 16 のデータセットである ABCpred16 も高い精度で予測することが出来ている．これは事前学習で様々な長さのペプチドを学習させているため，長さの変化に対応できたと考えられる．

[*2] http://ccb1.bmi.ac.cn:81/dlbepitope/

表 2 MHCII 分子に結合するペプチド予測の実験結果

	テスト		検証	
	AUC	SRCC	AUC	SRCC
Single Model				
DeepSeqPanII [9]	0.73	0.41	-	-
LSTM w/ attention [8]	0.72	0.39	0.88	0.71
BERT w/o pre-train	0.70	0.37	0.88	0.74
BERT w/ pre-train	**0.76**	**0.45**	0.87	0.70
Ensemble model				
NetMHCIIpan-3.1[34]	0.78	0.50	-	-

4.2 MHCII 分子に結合するペプチド予測

4.2.1 実験設定

MHCII 分子に結合するペプチド予測における提案手法の有効性を確認するために，以下の手法との比較を行う．

- DeepSeqPanII[9]
- 注意機構付き LSTM[8]
- NetMHCIIpan-3.1[34]

DeepSeqPanII は，weekly benchmark dataset において単一のモデルで最高精度が報告されている手法であるため，提案手法の比較対象とした．注意機構付き LSTM は B 細胞エピトープ予測のタスクで高い性能を示した手法であり，MHCII 分子に結合するペプチド予測でも高い性能が期待できるため，提案手法の比較対象に加えた．注意機構付き LSTM の入力形式は，提案手法と同じく separate token を用いて入力を連結する方法を用いた．教師あり学習データの分割の手法，モデルの数が異なるため同条件の比較にはならないが，同じテストデータにおいて高い性能を示している NetMHCIIpan-3.1[34] も比較対象に加えた．事前学習なしの BERT のパラメータ数は，4.1.1 の実験設定と同じく，2%とした．

MHCII 分子に結合するペプチド予測では，MHCII 分子とペプチドの結合親和性を予測する．このタスクでは，結合親和性が 500 nM 以下のデータを陽性，それより大きい

表 3 weekly benchmark dataset での各 MHCII 分子における性能比較
太字は単一モデルで最高のスコアであることを示す

		Single Model						Ensemble Model	
		BERT w/ pre-train		DeepSeqPanII[9]		LSTM w/ attention[8]		NetMHCIIpan-3.1[34]	
Allele	測定方法	AUC	SRCC	AUC	SRCC	AUC	SRCC	AUC	SRCC
HLA-DQA1*01:02/DQB1*05:01	ic50	0.62	0.26	0.6	0.21	**0.69**	**0.38**	0.60	0.21
HLA-DQA1*01:02/DQB1*06:02	ic50	0.69	**0.21**	0.56	0.12	**1.00**	0.10	0.81	0.22
HLA-DQA1*01:03/DQB1*06:03	ic50	0.75	**0.35**	0.56	0.10	**0.79**	0.34	0.81	0.42
HLA-DQA1*02:01/DQB1*03:01	ic50	**0.79**	**0.55**	0.75	0.48	0.73	0.44	0.81	0.59
HLA-DQA1*02:01/DQB1*03:03	ic50	0.73	0.47	**0.76**	**0.50**	0.73	0.42	0.76	0.54
HLA-DQA1*02:01/DQB1*04:02	ic50	0.55	0.11	0.56	0.11	**0.61**	**0.18**	0.52	0.04
HLA-DQA1*03:01/DQB1*03:02	ic50	**0.80**	0.38	0.36	0.02	0.63	**0.46**	0.97	0.55
HLA-DQA1*03:03/DQB1*04:02	ic50	0.53	0.00	**0.57**	0.06	0.56	**0.09**	0.48	-0.08
HLA-DQA1*05:01/DQB1*03:02	ic50	**0.78**	**0.52**	0.74	0.48	0.75	0.45	0.77	0.57
HLA-DQA1*05:01/DQB1*03:03	ic50	0.75	0.45	**0.76**	**0.47**	0.69	0.36	0.81	0.61
HLA-DQA1*05:01/DQB1*04:02	ic50	0.55	0.10	0.56	0.10	**0.59**	**0.17**	0.58	0.14
HLA-DQA1*06:01/DQB1*04:02	ic50	0.54	0.02	0.52	0	**0.58**	**0.07**	0.50	-0.06
HLA-DRA*01:01/DRB1*01:01	binary	0.84	0.53	**0.85**	**0.54**	0.73	0.35	0.84	0.52
HLA-DRA*01:01/DRB1*01:01	ic50	**0.82**	**0.65**	0.81	0.63	0.79	0.59	0.80	0.63
HLA-DRA*01:01/DRB1*03:01	binary	**0.61**	**0.18**	0.49	-0.01	0.56	0.09	0.62	0.21
HLA-DRA*01:01/DRB1*03:01	ic50	**0.78**	**0.55**	0.76	0.5	0.77	0.53	0.85	0.71
HLA-DRA*01:01/DRB1*04:01	binary	**0.80**	**0.52**	0.73	0.39	0.68	0.31	0.76	0.46
HLA-DRA*01:01/DRB1*04:01	ic50	0.76	**0.42**	**0.77**	**0.42**	0.65	0.29	0.84	0.58
HLA-DRA*01:01/DRB1*04:04	ic50	0.84	0.66	**0.85**	**0.67**	0.82	0.64	0.86	0.71
HLA-DRA*01:01/DRB1*07:01	binary	0.79	0.50	**0.8**	**0.52**	0.71	0.36	0.88	0.65
HLA-DRA*01:01/DRB1*07:01	ic50	**0.86**	**0.70**	**0.86**	**0.70**	0.81	0.60	0.88	0.76
HLA-DRA*01:01/DRB1*08:01	ic50	**0.83**	**0.67**	0.82	0.64	0.81	0.62	0.86	0.72
HLA-DRA*01:01/DRB1*08:02	ic50	0.73	0.23	**0.74**	**0.31**	0.73	0.30	0.74	0.44
HLA-DRA*01:01/DRB1*09:01	binary	**0.72**	**0.38**	0.66	0.28	0.66	0.28	0.84	0.58
HLA-DRA*01:01/DRB1*09:01	ic50	**0.84**	**0.61**	**0.84**	0.59	0.81	0.55	0.87	0.70
HLA-DRA*01:01/DRB1*11:01	binary	**0.66**	**0.27**	0.64	0.23	**0.66**	**0.27**	0.75	0.42
HLA-DRA*01:01/DRB1*11:01	ic50	**0.86**	**0.72**	0.84	0.68	0.84	0.67	0.89	0.78
HLA-DRA*01:01/DRB1*12:02	binary	**0.86**	**0.61**	0.84	0.59	0.74	0.41	0.80	0.51
HLA-DRA*01:01/DRB1*13:01	binary	**0.99**	**0.85**	0.91	0.72	0.93	0.74	0.86	0.63
HLA-DRA*01:01/DRB1*13:01	ic50	**0.86**	**0.69**	0.82	0.59	0.80	0.57	0.77	0.53
HLA-DRA*01:01/DRB1*13:02	ic50	**0.76**	**0.21**	0.71	0.14	0.47	0.03	0.90	0.62
HLA-DRA*01:01/DRB1*14:54	ic50	**0.86**	**0.66**	0.85	0.66	0.83	0.61	0.89	0.71
HLA-DRA*01:01/DRB1*15:01	binary	0.56	0.10	0.52	0.03	**0.57**	**0.11**	0.58	0.12
HLA-DRA*01:01/DRB1*15:01	ic50	**0.86**	0.64	0.85	**0.65**	0.79	0.53	0.76	0.50
HLA-DRA*01:01/DRB1*15:02	binary	**0.85**	**0.51**	**0.85**	**0.51**	0.63	0.20	1.00	0.74
HLA-DRA*01:01/DRB1*15:02	ic50	0.87	0.52	**0.93**	**0.69**	0.80	0.56	0.67	0.41
HLA-DRA*01:01/DRB3*01:01	ic50	**0.76**	**0.48**	0.66	0.32	0.65	0.28	0.84	0.60
HLA-DRA*01:01/DRB3*02:02	ic50	**0.72**	**0.40**	0.7	0.38	0.69	0.35	0.74	0.43
HLA-DRA*01:01/DRB3*03:01	ic50	0.72	0.43	**0.73**	**0.45**	0.69	0.39	0.78	0.56
HLA-DRA*01:01/DRB4*01:01	binary	0.66	0.25	**0.75**	**0.39**	0.71	0.32	0.72	0.35
HLA-DRA*01:01/DRB4*01:01	ic50	**0.70**	0.35	0.63	**0.38**	0.68	0.37	0.80	0.52
HLA-DRA*01:01/DRB4*01:03	ic50	**0.80**	**0.58**	**0.80**	**0.58**	0.77	0.53	0.79	0.54
HLA-DRA*01:01/DRB5*01:01	binary	**1.00**	**0.82**	0.97	0.77	0.96	0.75	0.96	0.75
HLA-DRA*01:01/DRB5*01:01	ic50	**0.82**	**0.69**	0.79	0.63	0.80	0.64	0.84	0.74
Mean		**0.76**	**0.45**	0.73	0.41	0.72	0.39	0.78	0.50

データを陰性と定義したときの 2 値分類での評価と，結合親和性の連続値を直接予測する回帰予測での評価を行う．モデルの評価指標として，2 値分類の性能を評価するため Area Under Curve(AUC) と，回帰予測の性能を評価するためスピアマンの順序相関係数 (SRCC) を用いる．提案手法と従来手法の精度の全体傾向比較のため，各分子の AUC と SRCC の全体平均を表 2 に示し，分子毎の詳細傾向を比較するため，各分子個別の AUC と SRCC を表 3 に示す．DeepSeqPanII と NetMHCIIpan-3.1 は，学習時に使用された検証データが不明であるため，テストデータの評価のみを示す．

4.2.2 データセット

IEDB[33] より作られたベンチマークデータセット [35] を用いる．このデータセットのうち, BD2013 と呼ばれる 2013 年までに報告されたデータで学習し，weekly benchmark data と呼ばれる 2016〜2017 年に報告されたデータで評価する．同じ分子でも，測定方法が $IC50$ と *binary* の 2 種類あるものもあり, これらは別々に評価する．測定方法が *binary* のデータは，結合度を 0, 1 の二値と表現したデータである．測定方法が $IC50$ のデータは, $1-\log(IC50)/\log(50000)$ の計算により，0 から 1 までの連続値に対数変換されている．学習データは 51023 件，テストデータは 20640 件である．すべてのデータ，評価スクリプトは DeepSeqPanII[9] のレポジトリ[*3]から利用した．また，エポック数を決めるための検証 (validation) 用データは学習データから 10%を用い，MHCII 分子の種類により分割する．これは DeepSeqPanII の実験設定と同じである．

4.2.3 実験結果

実験 1 の結果を表 2 に示す．テストデータの各分子における AUC の平均は，提案手法が比較手法の DeepSeqPanII 及び注意機構付き LSTM を大きく上回っていることが確認できた．これは，MHCII 分子に結合するペプチド予測においては，長距離の依存関係を捉えることが重要であり，DeepSeqPanII 及び注意機構付き LSTM では，長距離の依存関係の学習が十分ではなかったためと考えられる．また，事前学習あり BERT と事前学習なし BERT のテストデータでの AUC，SRCC の比較から，事前学習が汎化性能の向上に寄与していることがわかる．これは B 細胞エピトープ予測の実験と同じく，大規模データベースによる事前学習を行わない場合，自由度の高い BERT のパラメータを十分に学習できないためと考えられる．

次に実験 2 の結果を表 3 に示す．提案手法である事前学習あり BERT，DeepSeqPanII 及び注意機構付き LSTM の単一モデル同士で比較すると，提案手法は 44 種類のテストデータのうち，AUC では 26 種類の分子で他の手法と同等以上の性能を達成しており[*4]，SRCC では 25 種類の

[*3] https://github.com/pcpLiu/DeepSeqPanII

[*4] DeepSeqPanII[9] の論文の実験の値は小数点第 2 位までの記述のため，小数点第 2 位のスコアが同じ値の場合は，性能が同等とみなした．

分子で提案手法は他の手法と同等以上の性能を達成している[*4]．各分子毎の評価指標の比較からも，提案手法は高い性能を示していることがわかる．アンサンブルモデルで最高精度を示したと報告されている NetMHCIIpan-3.1 と比較すると，44 種類のテストデータのうち，AUC では 13 種類の分子で NetMHCIIpan-3.1 を上回り，28 種類の分子で劣り，他 3 種類の分子で同等であった[*4]．SRCC では 14 種類の分子で NetMHCIIpan-3.1 を上回り，28 種類の分子で劣り，他 2 種類の分子で同等であった[*4]．提案手法は NetMHCIIpan-3.1 に多くの分子で性能が劣るものの，NetMHCIIpan-3.1 のような結合部位周辺構造の情報を明示的に与えておらず，モデルのアンサンブルも行っていないという違いがある．これらの工夫は提案手法にも導入可能であるため，今後性能が改善されることが期待できる．

5. まとめと今後の発展

本論文では B 細胞のエピトープ予測と，MHCII に対するペプチドの結合予測における，離れたアミノ酸間の複雑な依存関係を捉えていない課題と，学習データが不十分な場合に精度が低いという課題に対処するため，教師なしの大規模タンパク質データベースから事前学習した BERT を用いた手法を提案した．提案手法は，B 細胞エピトープ予測，MHCII 分子に結合するペプチド予測の両方で高い性能を達成できることを示した．今後の発展として，NetMHCIIpan-3.1 のように結合部位周辺構造の情報を明示的を与えたり，事前学習とは別にデータ拡張を行ったりすることで，さらなるモデルの性能の向上に取り組むことが挙げられる．また BERT は注意を可視化することで，学習したモデルがどこに注意して予測を行っているかがわかる．先行研究では BERT が 3 次元構造において近い位置にあるアミノ酸を注目したり，結合部位を注目していることがわかっている [31]．本研究では注目度の解釈は行わなかったが，今後そのような可視化を試みることで，メカニズムを明らかにし，今後の更なる改善に活かしたい．

参考文献

[1] Bhattacharya, M., Sharma, A. R., Patra, P., Ghosh, P., Sharma, G., Patra, B. C., Lee, S. S. and Chakraborty, C.: Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): Immunoinformatics approach, *Journal of medical virology*, Vol. 92, No. 6, pp. 618–631, (2020).

[2] Li, W., Joshi, M. D., Singhania, S., Ramsey, K. H. and Murthy, A. K.: Peptide Vaccine: Progress and Challenges, *Vaccines*, Vol. 2, No. 3, pp. 515–536, (2014).

[3] Droppa-Almeida, D., Franceschi, E. and Padilha, F. F.: Immune-Informatic Analysis and Design of Peptide Vaccine From Multi-epitopes Against Corynebacterium pseudotuberculosis, *Bioinformatics and Biology Insights*, Vol. 12, 1177932218755337, (2018).

[4] Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T. and Weng, Z.: ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers, *Bioinformatics*, Vol. 30, No. 12, pp. 1771–1773, (2014).

[5] Pierce, B. G., Hourai, Y. and Weng, Z.: Accelerating protein docking in ZDOCK using an advanced 3D convolution library, *PLoS One*, Vol. 6, No. 9, e24657, (2011).

[6] Hou, T., Wang, J., Li, Y. and Wang, W.: Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations, *Journal of Chemical Information and Modeling*, Vol. 51, No. 1, pp. 69–82, (2011).

[7] Fast, E., Altman, R. B. and Chen, B.: Potential T-cell and B-cell Epitopes of 2019-nCoV, *bioRxiv*, DOI: 10.1101/2020.02.19.955484, (2020).

[8] 農見俊明，藤田春佳，貞光九月，坂口誠，天満昭子，中神啓徳：注意機構付き LSTM を用いた抗原タンパク質のエピトープ領域予測，第 60 回バイオ情報学研究会, (2019).

[9] Liu, Z., Jin, J., Cui, Y., Xiong, Z., Nasiri, A., Zhao, Y. and Hu, J.: DeepSeqPanII: an interpretable recurrent neural network model with attention mechanism for peptide-HLA class II binding prediction, *bioRxiv*, DOI: 10.1101/817502, (2019).

[10] Jespersen, M. C., Peters, B., Nielsen, M. and Marcatili, P.: BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes, *Nucleic Acids Research*, Vol. 45, No. W1, pp. W24–W29 , (2017).

[11] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp.1735–1780, (1997).

[12] Jurtz, V. I., Johansen, A. R., Nielsen, M., Almagro Armenteros, J. J., Nielsen, H., Sønderby, C. K., Winther, O. and Sønderby, S. K.: An introduction to deep learning on biological sequence data: examples and solutions, *Bioinformatics*, Vol. 33, No. 22, pp. 3685–3690, (2017).

[13] Kolen, J. F. and Kremer, S. C.: Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies, *A Field Guide to Dynamical Recurrent Networks, IEEE*, pp. 237–243, (2001).

[14] Liu, Z., Cui, Y., Xiong, Z., Nasiri, A., Zhang, A. and Hu, J.: DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction, *Scientific Reports*, Vol. 9, Article number 794, (2019).

[15] Liu, T., Shi, K. and Li, W.: Deep learning methods improve linear B-cell epitope prediction, *BioData Min*, Vol. 13, Article number 1, (2020).

[16] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *NAACL*, Vol. 1, pp. 4171–4186, (2019).

[17] Parikh, A., Täckström, O., Das, D. and Uszkoreit, J.: A Decomposable Attention Model for Natural Language Inference, *EMNLP*, pp. 2249–2255, (2016).

[18] Seo, M., Kembhavi, A., Farhadi, A. and Hajishirzi, H.: Bidirectional Attention Flow for Machine Comprehension, *ICLR*, (2017).

[19] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E.

L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. and Finn, R. D.: The Pfam protein families database in 2019, *Nucleic Acids Research*, Vol. 47, No. D1, pp. D427–D432, (2019).

[20] Selvan, S. R., Sanchez-Trincado, J. L., Gomez-Perosanz, M. and Reche, P. A.: Fundamentals and Methods for T- and B-Cell Epitope Prediction, *Journal of Immunology Research*, Vol. 2017, Article ID 2680160, (2017).

[21] Singh, H., Ansari, H. R. and Raghava, G. P.: Improved method for linear B-cell epitope prediction using antigen's primary sequence, *PLoS One*, Vol. 8, No. 5, e62216, (2013).

[22] Saha, S. and Raghava, G.: Prediction of Continuous B-cell Epitopes in an Antigen Using Recurrent Neural Network, *Proteins*, Vol. 65, No. 1, pp. 40–48, (2006).

[23] Tsubaki, M., Tomii, K. and Sese, J.: Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, *Bioinformatics*, Vol. 35, No. 2, pp. 309–318, (2019).

[24] Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. and Sun, M.: Graph Neural Networks: A Review of Methods and Applications, *CoRR*, abs/1812.08434, (2019).

[25] LeCun, Y. and Bengio, Y.: Convolutional Networks for Images, Speech, and Time Series, *The Handbook of Brain Theory and Neural Networks*, pp. 255–258, (1998).

[26] Reynisson, B., Alvarez, B., Paul, S., Peters, B. and Nielsen, M.: NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data, *Nucleic Acids Research*, Vol. 48, No. W1, pp. W449–W454, (2020).

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., L ukaszKaiser, Polosukhin, I. : Attention is All you Need, *NIPS*, Vol. 30, pp. 5998–6008, (2017).

[28] Al-Rfou, R., Choe, D., Constant, N., Guo, M. and Jones, L.: Character-Level Language Modeling with Deeper Self-Attention, *AAAI*, (2019).

[29] Radford, A., Narasimhan, K., Salimans, T. and Sutskever., I.: Improving Language Understanding by Generative Pre-Training, Technical report, OpenAI, (2018).

[30] Howard, J. and Ruder, S.: Universal Language Model Fine-tuning for Text Classification, *ACL*, pp. 328–339, (2018).

[31] Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R. and Rajani, N. F.: BERTology Meets Biology: Interpreting Attention in Protein Language Models, *bioRxiv*, DOI: 10.1101/2020.06.26.174417, (2020).

[32] Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P. and Song, Y. S.: Evaluating Protein Transfer Learning with TAPE, *NeurIPS*, Vol. 32, pp. 9689–9701 (2019).

[33] Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A. and Peters, B.: The immune epitope database (IEDB) 3.0, *Nucleic Acids Research*, Vol. 43, No. D1, pp. D405–D412, (2015).

[34] Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., Buus, S. and Nielsen, M.: Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification, *Immunogenetics*, Vol. 67, No. 11–12, pp. 641–650, (2015).

[35] Andreatta, M., Trolle, T., Yan, Z., Greenbaum, J. A., Peters, B. and Nielsen, M.: An automated benchmarking platform for MHC class II binding prediction methods, *Bioinformatics*, Vol. 34, No. 9, pp. 1522–1528, (2018).