Vol.2019-BIO-60 No.3
2019/12/2

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

農見俊明[†1†2]　藤田春佳[†1]　貞光九月[†1]
坂口誠[†3]　天満昭子[†3]　中神啓徳[†4]

概要：生体内において抗原に特異的な免疫応答を誘導する B 細胞は，抗原タンパク質の部分領域（エピトープ領域）を認識することで抗原特異的な抗体を大量に産生し，抗体を抗原タンパク質に結合させることでその機能を阻害することができる．エピトープ領域を予測することは，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り，予測精度の向上が求められている．従来のエピトープ領域予測手法では，抗原タンパク質全体のアミノ酸配列のうち予測対象配列のみに注目しており，抗原タンパク質全体の配列や特徴量を十分に考慮できていなかった．本研究では，予測対象配列に加え抗原タンパク質全体の特徴を考慮するために深層学習の一種である注意機構付き LSTM の適用法を提案する．提案法は IEDB データを用いた実験において，従来法によるエピトープ領域予測精度を上回る精度を示した．

キーワード： B 細胞エピトープ予測，タンパク質，アミノ酸配列，エピトープ, LSTM，注意機構

Epitope Prediction of Antigen Protein using Attention-Based LSTM Network

TOSHIAKI NOUMI[†1†2]　HARUKA FUJITA[†1]　KUGATSU SADAMITSU[†1]
MAKOTO SAKAGUCHI[†3]　AKIKO TENMA[†3]　HIRONORI NAKAGAMI[†4]

1. はじめに

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，抗原タンパク質の部分領域（エピトープ領域）を認識することで，抗原特異的な抗体を大量に産生する．抗体は抗原のエピトープ領域に結合することで，抗原の機能を阻害する[1]．エピトープの構造・機能を真似た物質を「ワクチン」として生体に投与し生体内で特定の抗体を誘導することができる．副作用のない安全で効果的なワクチンの設計に役立てることを目的に，これまで多くのエピトープに関する研究が行われてきた．B 細胞に認識されるエピトープを明らかにする確実な方法は，X 線[2]や NMR[3]による抗体・抗原複合体の立体構造解析であるが，時間，費用，労力の点でコストが高い．そのためコンピュータによるエピトープ予測が取り組まれてきた．近年は機械学習を適用した手法が多く提案され[4][5], 性能が改善されているものの，機械学習に用いる特徴量が予測対象のアミノ酸配列内に限定されていたり，モデルの表現能力が不十分であるという課題があった．

本研究では，B 細胞エピトープ予測において，予測対象のアミノ酸配列のみでなく，タンパク質全体における長距離の特徴を予測に取り込むことを可能とするため，深層学習の一種である注意機構付き LSTM (Attention-based Long Short Term Memory Network) [6]を用いた手法を提案する．LSTM によってタンパク質の長距離の特徴を表現するとともに，注意機構[7]によって，予測対象エピトープ候補内外の各アミノ酸に対して，予測にとっての注目すべき箇所を自動的に推定することが可能となる．更に，データスパースネスの問題に対応するため，抗原タンパク質全体の構造的・化学的特徴を深層学習ネットワーク内で同時に考慮可能な手法として拡張した．

提案法を用いて，免疫エピトープの公開データベースである IEDB (Immune Epitope DataBase, [8])を対象とし実験を行った結果，提案法によって，既存手法 BepiPred2.0 [4]の予測精度を上回る性能を達成したことを示す．

2. 本研究で取り組む問題設定

2.1 問題設定

本稿で対象とする B 細胞エピトープの予測は，抗原タンパク質を構成する長鎖アミノ酸配列から，B 細胞が認識するエピトープ領域を予測するものである．過去の免疫エピトープに関する研究結果を蓄積した公開データベースである IEDB に登録されているエピトープ領域は，5-20 アミノ酸程度の範囲である[9]．我々の提案手法は，エピトープ領域の長さに依存せず適用可能であるが，本稿の問題設定で

†1 フューチャー株式会社, Future Corporation
†2 東京大学大学院, The University of Tokyo
†3 株式会社ファンペップ, FunPep Co., Ltd.
†4 大阪大学大学院, Osaka University Graduate School of Medicine
第一著者，第二著者は筆頭著者として同等に貢献した

は，エピトープ候補となるペプチド（短いアミノ酸配列）の長さを 8 アミノ酸に限定した．これは，生体内にて B 細胞誘導とは異なる細胞性の免疫応答を引き起こすことを避けるためである[10]．これらペプチドが抗体誘導活性をもつ（陽性），もたない（陰性）の 2 分類を予測する問題に取り組む．

2.2 本研究で用いるデータセット

8 アミノ酸ペプチドが抗体誘導活性をもつか否か（活性ラベル）の情報は， 多くの先行研究でも用いられる IEDB[8]から取得可能なため，これらを教師データとした．一方，エピトープ候補アミノ酸配列（ペプチド）と活性ラベルデータは，B 細胞エピトープデータ[11]より取得した．抗体タンパク質は IEDB 中で最も収録数の多い IgG のみに限定した．同じペプチドに対して異なる抗体活性(Qualitative Measure)を示すレコードは，今回実験対象から便宜的に除外した．最終的に 2472 ペプチド，86 タンパク質，活性ラベル陽性：陰性＝1：4.5 のデータセットが得られた．

この母集団に対し，学習データと評価データの比率を 95:5 程度に，各セット内で重複しないよう分割し，これを 3 セット分抽出した．なお，陽性：陰性の比率は 1：3.4～4.6 となった．本データを，以降用いることとする．

3. 関連研究とその課題

コンピュータによる初期の B 細胞エピトープ予測手法は，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指標にした予測であった[10]．このような人手で設計された特定の指標に基づく予測に対し，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，比較的高い性能を達成している[4][5]. 多くの機械学習アルゴリズムを用いた手法が提案されており，Random Forest を用いた手法（公開ツール BepiPred-2.0 [4]）や，SVM (Support Vector Machine) と k 近傍法を用いた手法（公開ツール LBtope）[5]が比較的新しい手法として知られている．これら手法の性能は，実験で用いられているデータセットが異なるため，一概に比較できないものの，単純な値としては BepiPred-2.0 が最も高い値を示している[4]．これら機械学習を用いた先行研究において, 対象とするペプチド前後の短いアミノ酸配列（1～3 アミノ酸）を特徴量に加えた実験も行われているものの，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．

抗原タンパク質におけるペプチド外の長距離情報を取り扱うための手法として，アミノ酸配列そのものを系列データとみなすアプローチが考えられる．具体的な先行研究として，深層学習の一種である RNN (Recurrent Neural Network: 再帰ニューラルネットワーク) [12]を用いた研究

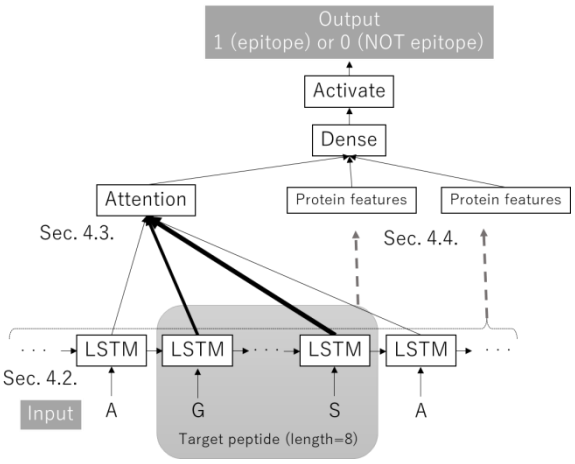

図 1. 本稿モデルのネットットワーク構造. 図中の Sec 番号は各部分に対応する節番号を示す．図中の矢印上の数字は伝播するベクトルの次元数を表す.

[13]が挙げられるが，エピトープ外のアミノ酸配列情報は取り込んでいない．これは RNN の学習過程において勾配消失問題を生じてしまうため，長距離情報をうまく扱えないことが課題の一つと考えられる． 本研究では，ペプチド自身，およびペプチド前後のアミノ酸配列を系列データとして扱いつつ，勾配消失問題に頑健な LSTM[6] を適用することで本課題の解決に取り組む.

さらに機械学習の抱える本質的な問題点として，学習データが少ない場合に予測がうまく行えないという課題も残存する．先行研究[5]では SVM の特徴量にペプチド内の，アミノ酸の種類や組成，疎水性，極性，安定性などの物理化学的性質を用いる．提案法ではさらに，アミノ酸系列情報に加え，抗原タンパク質全体の構造的・化学的特徴量をニューラルネットワークの内部で併用するようにモデルを拡張することで本課題の解決に取り組む.

4. 提案手法

4.1 手法概要

エピトープを予測する際，エピトープ内の特徴量やアミノ酸配列のみでは予測のための情報が不足することが考えられる．本研究ではタンパク質の長距離情報を扱うため，エピトープ内外のアミノ酸配列を系列データとみなした注意機構付き LSTM を用いた手法を提案する．

図 1 に提案手法の全体図を示す. 以下の節では, LSTM, 注意機構について概説した後，提案法について詳しく述べる．

4.2 LSTM

系列データを扱う深層学習として RNN[12]があるが，モデル学習時の誤差逆伝播の際に値が小さくなる勾配消失の問題により，長距離の系列情報を扱うことが難しかった．

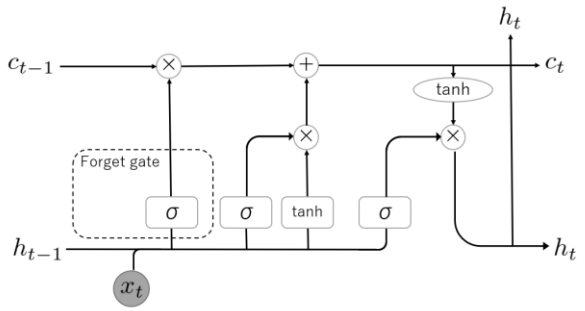

図 2. ある系列時点 t における LSTM ブロックの構造．x_tは時点 t における入力を表し，本研究の場合 t 番目のアミノ酸に相当する．c_t は長距離情報を保持するメモリセル，h_tは隠れ状態を表す埋め込みベクトル，σ はシグモイド関数，tanh は双曲線正接関数，×は行列のアダマール積，+は単純加算を表す．

LSTM [6] は RNN に対し，メモリセルとゲートを採用することにより，長距離の系列情報を扱うことを可能としたモデルである．図 2 に LSTM の各系列時点でのブロック(図 1 中”LSTM”と記述した四角形)の模式図を示す．メモリセル(図 2 中 c_t) の逆伝播の流れ (図 2 中 c_{t-1} への流れ) を追うと，単純加算とアダマール積のみを通ることが分かる．単純加算の逆伝播は偏微分すると 1 となるため勾配の変化は生じず，一方アダマール積の偏微分は，ゲートの 1 種である忘却ゲート (図 2 中 ”Forget gate”) の出力のみに依存する．つまりゲートが忘れるべきと判断した要素の勾配は小さくなり，逆に忘れるべきでないと判断した要素の勾配は維持したまま過去方向へと伝わる．そのため，メモリセルは長期記憶しておくべき情報を消失させることなく伝播することが可能となる．本研究では，系列と同じ方向の順方向 LSTM に加え，逆方向 LSTM も併用する双方向 LSTM (Bi-directional LSTM) [14] を用いることで，逆向きの系列から得られる情報も併用することとした．

4.3 注意機構

LSTM は RNN に比べ長距離情報を保持する能力を有するが，入力系列の情報は圧縮した 1 つのベクトル (h_T) として表せないため，入力系列中の細かな情報は失われやすい．そこで入力系列の情報を直接参照可能とした仕組みが注意機構である．注意機構を用いることで，各系列時点の LSTM ブロックから出力されるベクトルを記憶した上で，各系列時点に対する重みを掛け，文脈中のどの要素に着目すべきかを考慮したベクトル (context vector) を得ることが可能となる（図 1 中”Attention”部）．注意機構付き LSTM は，自然言語処理等をはじめ多くの成果を挙げており[7]，本研究でもタンパク質のどの部分を重視すべきかをモデル内で考慮すべく，注意機構を導入した．

4.4 アミノ酸配列の構造的・化学的特徴の利用

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．我々は学習データが少ない場合でも頑健な予測を可能とするため，ペプチド内および抗原タンパク質全体の構造的・化学的特徴量を用いる．

ペプチドの構造的・化学的特徴としては，β ターン[15]，表面到達性[16]，抗原性[17]，親水性[18]を用い，IEDB にて提供するエピトープ予測 API[19]を用いて取得した．抗原タンパク質全体の特徴量は,タンパク質とエピトープの結合し易さに影響すると期待されるため採用し，等電点，芳香環を持つアミノ酸の割合，疎水性，安定性の 4 種について，Biopython ライブラリ[20]を用いて取得した．これら合計 8 種の構造的・化学的特徴について，次節において提案手法のネットワーク内に統合して用いる．

4.5 注意機構付き LSTM を用いたエピトープ領域予測

4.2 節の LSTM によって長距離の系列情報を得て，4.3 節の注意機構によって注目すべき箇所の予測機構を得て，4.4 節の構造的・化学的特徴量によって学習データが少ない場合でも頑健な推定を可能とする特徴を得た．本節では，本稿で対象とするエピトープ予測に対し，これら 3 つの特長を 1 つのモデルの中で自然に利用できることを示す．

図 1 において，入力となる系列情報それぞれは，ペプチド内外の各アミノ酸(A,G,…,S)であり，各系列時点毎の LSTM ブロックがアミノ酸情報を受け取る．LSTM は当該系列時点のアミノ酸情報に加え，前後のアミノ酸情報を埋め込みベクトル(h_t)を伝搬させていくことで，長距離のアミノ酸配列情報を捉えることを可能とする．

次に注意機構を用いて，どのアミノ酸が特にエピトープ推定において注目すべき対象であるかを推定する. 例えば，ペプチド外の情報よりも，ペプチド内の情報の方が重要度は高いと考えられるため，ペプチド内に高い重みを付与して情報を伝播させていくことが可能となる．

注意機構付き LSTM から得られた最終的な埋め込みベクターと 4.4 節で得たアミノ酸配列の構造的・化学的特徴（図 1 で”peptide/protein features”と表記）を全結合層(“dense”層)によって結合し，最後にシグモイド関数（”activation”層）を用いて，エピトープであるか否かを二値として推定する．

なお事前のモデル学習フェーズでは，各学習ステップ時点のモデルパラメータを用いて上記推定を行った後，エピトープか否かの判定に対する損失を逆伝播させ，各層のモデルパラメータを更新することで学習を行う．

5. 評価実験

5.1 実験条件

提案手法の評価のため，同じデータセットを用いてベ

表 1. 各手法の特徴量の比較

	範囲	BepiPred2.0	LightGBM	Proposed
埋め込みベクトル	ペプチド	–	■ペプチド内の各アミノ酸ごとに 20 次元の埋め込みベクトル	■ペプチド内の各アミノ酸及び前後各 16 アミノ酸に関し，それぞれ 20 次元のベクトルを算出．以下のペプチド，タンパク質に関する特徴量を全結合層で結合したのち，20 次元で出力
	ペプチド前後情報	–	■ペプチドの前後各 15 アミノ酸ずつに関し，それぞれ 20 次元の埋め込みベクトルを算出した後，Truncated SVD により 50 次元に削減	
特徴量	ペプチド内の構造的・化学的特徴量	アミノ酸単位での分子量，疎水性，極性	βターン，表面到達性，抗原性，親水性	βターン，表面到達性，抗原性，親水性
	タンパク質全体の構造的・化学的特徴量	表面到達性，タンパク質の予測二次構造	–	等電点，芳香環を持つアミノ酸の割合，疎水性，安定性

ースラインの精度と比較する．本実験では，ベースラインとして， BepiPred2.0[4] と， 機械学習の一種である LightGBM[15]を用いた．LightGBM は近年多くのデータ分析タスクで優れた結果を示す機械学習法であり，SVM を用いた先行研究[5]の発展的実験として位置付けた．

LightGBM は複数の決定木アルゴリズムを勾配ブースティング（Gradient Boosting）により組合せたアンサンブル法であり，データ分析における多くの分類問題において高い精度が報告されている[21]．一方で LightGBM は，LSTM のように系列データを明示的に取り扱う機構は有していない．

表 1 に BepiPred2.0，LightGBM と提案法で用いた特徴量を比較提示する．LightGBM の特徴量は，ペプチド内外の埋め込みベクトル及びペプチド内アミノ酸配列の構造的・化学的 4 特徴量を用いた．この際の埋め込みベクトル算出のため，事前に全タンパク質のデータを用いた word2vec[22]を学習した後，各アミノ酸に対する埋め込みベクトル算出を行った．

各モデルは最終的にエピトープらしさを示す 0.0-1.0 の値を示すが，この値が 0.5 を超える場合をエピトープ，0.5 以下を非エピトープとみなした．なお，BepiPred2.0 についてはアミノ酸単位で予測値が算出されるため，ペプチドを構成する 8 アミノ酸のスコアを平均したものを用いた．

評価指標として， 陽性ラベルの予測に注目するため，陽性ラベルに関する 3 指標，

・適合率(Precision)＝真陽性／(真陽性＋偽陽性),
・再現率（Recall）＝真陽性／(真陽性＋偽陰性),
・F1 値＝2×適合率×再現率／(適合率＋再現率),

さらに，陽性・陰性を通した正解率（accuracy）を用いた．なお本実験のチャンスレートは，すべて陽性と判定した場合，再現率は 1.0 となり，Precision は 0.21， F 値は 0.35 となる．一方すべて陰性と判定した場合の正解率は 0.79 である．正解率が高めなのは，データ中に陰性ラベルが多く含まれる偏りによるものである．

LightGBM および提案法については，学習データ，テストデータともに 2.2 節で述べたデータセットを用いた．これら機械学習では学習過程にランダム性が存在し，試行毎に算出スコアに差が生じるため，4 回の試行の平均を代表値とした．BepiPred2.0 は学習済みのモデルが API[19]として公開されており，本 API を用いて 2.2 節の評価データに適用した．

5.2 実験結果・考察

BepiPred2.0, LightGBM, 提案法それぞれの実験結果を表 2 に示す．最下行には参考として，注意機構をモデルから取り除いた提案法の結果 (表中，”w/o Attention”) も付記した．

提案法は，比較手法である BepiPred2.0，LightGBM に比べ， 全ての精度指標において最も高い値を示した．BepiPred2.0 と比べ，F1 値で 8 ポイント，正解率 (accuracy) で 25 ポイント，と大幅な向上を達成した．次に提案法と注意機構を提案法から取り除いた結果と比較すると，適合率 (precision) ではほとんど差がないものの，再現率 (recall) において精度を大きく落とす結果となり，注意機構によるペプチド内外の各アミノ酸への重要度の付与がエピトープ予測において重要な役割を果たす可能性が示唆された．一方 LightGBM は正解率および再現率では BepiPred2.0 を上回ったが，適合率は BepiPred2.0 を下回り，結果的に F1 値では BepiPred2.0 と同等程度となっている．

次に，図 3 に提案法及び LightGBM に関する 4 回の試行結果の全ての値をプロットして示す．比較的分散の大きい LightGBM に比べ，注意機構付き LSTM が全体的に高い精度かつ分散を抑えられていることが確認でき，提案法がより頑健に推定できていることが分かる．

表 2. 各手法による予測検証結果．下線部は比較手法中，最良の精度を示す．

	precision	recall	F1-value	accuracy
BepiPred2.0	0.75	0.57	0.65	0.64
LightGBM	0.68	0.63	0.64	0.84
Proposed	0.77	0.70	0.73	0.89
Proposed w/o Attention	0.78	0.48	0.57	0.85

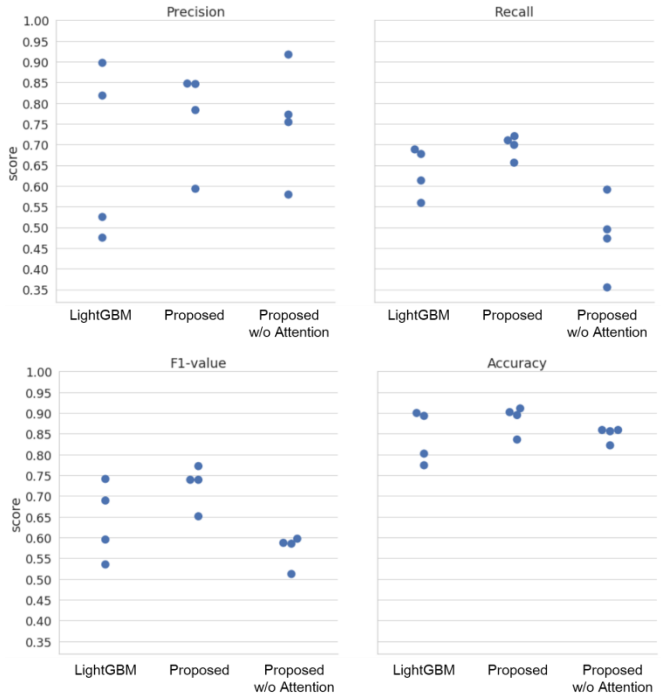

図 3. 各手法における全 4 回の試行それぞれの精度の分布

5.3 提案手法による検出例

前節の実験で得られた提案法と BepiPred2.0 の予測事例を表 3 に示す． 表中，タンパク質 Q39967 に含まれるペプチドは非エピトープが多いが，BepiPred2.0 では 0.5 以上のスコアが付与され，エピトープと判定されたのに対し，提案法では正しく非エピトープと認識できている．一方タンパク質 Q5EFD8 に含まれるペプチドはエピトープが多く，ここでも提案法は正しくエピトープとして検出できている．タンパク質 71037364 では提案法，BepiPred2.0 共に多くのペプチドで正しく推定できているものの，ペプチド FLGMINTI に対しては推定を誤っている．このような推定を誤る事例の原因究明と改善が今後の課題である．

6. おわりに

本研究では，B 細胞エピトープ予測を行う新しい手法として以下の特徴を持つモデルを提案し，提案手法が既存の手法（BepiPred2.0）に比べ優れることを示した．

表 3. エピトープ予測事例．Ans.は正解ラベルで 1 がエピトープであることを示し，背景色を色付けしている．提案法 (Proposed) と BepiPred2.0 の列の各値は推定結果を表し，推定結果が 0.5 より大きい場合エピトープと判定されるものについて，背景色を色付けしている．即ち背景色が一致しているペアが正解となる．

Parent Protein ID	Peptide	Ans.	Proposed	BepiPred2.0
Q39967	EQETADAT	0	0.371	0.536
Q39967	ESAATALP	0	0.303	0.521
Q39967	ETADATPE	0	0.385	0.532
Q39967	ETATTEVP	0	0.327	0.525
Q39967	EVESAATA	0	0.307	0.523
Q39967	EVTKAEET	0	0.323	0.511
Q39967	ITEAAETA	0	0.279	0.508
Q39967	KAEETKTE	0	0.346	0.508
Q5EFD8	DDLYTTNP	1	0.806	0.531
Q5EFD8	DKARYGGK	1	0.826	0.534
Q5EFD8	ERIQKYTR	1	0.847	0.463
Q5EFD8	FYDEEKKL	1	0.813	0.496
Q5EFD8	GGLLVKKY	1	0.694	0.431
Q5EFD8	LAARSSAP	1	0.719	0.569
Q5EFD8	LDYENWTK	1	0.839	0.503
Q5EFD8	LKEHDMLA	1	0.667	0.534
Q5EFD8	NARLGQRV	1	0.753	0.471
Q5EFD8	NPITIKKG	1	0.812	0.521
Q5EFD8	TVLKKNIG	1	0.734	0.541
Q5EFD8	YLGRVTLA	1	0.569	0.488
71037364	DVRDLQNK	1	0.726	0.556
71037364	ESIDHQTK	1	0.714	0.544
71037364	EVMPHILT	1	0.665	0.538
71037364	FLGMINTI	0	0.728	0.519
71037364	IKDDEANW	1	0.796	0.537

1）ペプチドだけでなく，タンパク質をアミノ酸の系列データとして扱い注意機構付き LSTM を用いてモデル化した．
2）ペプチドおよびタンパク質全体の構造的・化学的特徴量を組み合わせることで，学習データ量が少ない場合でも頑健な予測を可能とした．本手法は，タンパク質部分配列およびタンパク質全体配列の相互作用を予測する際にも応用できる可能性がある．

今後の課題として，実験における IgG 以外の抗体タンパク質への適用による提案法の有効性の確認や，アミノ酸配列の構造的・化学的特徴量を比較手法と統一した場合の提案法の有効性の確認が挙げられる．また，非線形エピトープの予測に用いられるタンパク質立体構造のモデル化[23]も有望な情報源と考えており，今後はタンパク質の立体構造の活用によって，さらなる予測精度向上に取り組みたい．

参考文献

[1] Van Regenmortel Van Regenmortel, M. H. The concept and operational definition of protein epitopes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* vol.323 pp.451–466 (1989).
[2] Rux, J. J. & Burnett, R. M. Type-Specific Epitope Locations Revealed by X-Ray Crystallographic Study of Adenovirus Type 5 Hexon. *Molecular Therapy* vol.1, pp.18–30 (2000).
[3] Mayer, M. & Meyer, B. Group Epitope Mapping by Saturation Transfer Difference NMR To Identify Segments of a Ligand in Direct Contact with a Protein Receptor. *Journal of the American Chemical Society* vol.123, pp.6108–6117 (2001).
[4] Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research* vol.45, W24–W29 (2017).
[5] Singh, H., Ansari, H. R. & Raghava, G. P. S. Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence. *PLoS ONE* vol.8, pp.1-8 (2013).
[6] Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* vol.9 pp.1735–1780 (1997).
[7] Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* 1409.0473 (2014).
[8] Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Research* vol.43 D405–D412 (2015).
[9] http://www.thinkpeptides.com/bcell.html
[10] Sanchez-Trincado, J. L., Gomez-Perosanz, M. & Reche, P. A. Fundamentals and Methods for T- and B-Cell Epitope Prediction. *Journal of Immunology Research* 2017 (2017).
[11] http://www.iedb.org/ database_export_v3.php
[12] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* vol.323 pp.533–536 (1986).
[13] Saha, S. & Raghava, G. P. S. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins: Structure, Function and Genetics* vol.65, pp.40–48 (2006).
[14] Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. in *Neural Networks* vol.18 pp.602–610 (2005).
[15] Chou, P. Y. & Fasman, G. D. Prediction of the Secondary Structure of Proteins From Their Amino Acid Sequence. in *Advances in Enzymology and Related Areas of Molecular Biology* vol. 47 pp.45–148 (2006).
[16] Emini, E. A., Hughes, J. V, Perlow, D. S. & Boger, J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of virology* vol.55 p.836–839 (1985).
[17] Kolaskar, A. S. & Tongaonkar, P. C. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Letters* vol.276 pp.172–174 (1990).
[18] Parker, J. M. R., Guo, D. & Hodges, R. S. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry* vol.25 pp.5425–5432 (1986).
[19] http://tools.iedb.org/main/tools-api/
[20] https://biopython.org/
[21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: a highly efficient gradient boosting decision tree. In Proc. *Advances in Neural Information Processing Systems*, pp. 3149–3157 (2017).
[22] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In Proc. *Advances in Neural Information Processing Systems* vol.26 pp.3111–3119 (2013).
[23] Sun, J. *et al.* SEPPA: A computational server for spatial epitope prediction of protein antigens. *Nucleic Acids Research* vol.37 W612-616 (2009).