厚生労働科学研究費補助金（食品の安全確保推進研究事業）
「新たなバイオテクノロジーを用いて得られた食品の安全性確保と
リスクコミュニケーションのための研究」
分担研究報告書（令和 4 年度）

新規アレルゲン性評価手法開発の基盤研究と AI のリスク評価への応用

研究分担者　富井　健太郎　　産業技術総合研究所

研究要旨：
HLA クラス II 分子－ペプチド間の結合は、残基間相互作用が本質的に鍵を握っていると考え、既存の深層学習モデルの DeepSeqPanII をベースモデルに、特徴量として（デフォルトの one-hot encoding と BLOSUM62 に加えて）アミノ酸物理化学的インデックス、PSSM、アミノ酸残基間ポテンシャルインデックスを組み合わせてトレーニングを行い、LOAO 交差検証テストによってモデルの予測性能を比較し、性能向上の可能性とどの特徴量が有効であるかを検証した。一次元配列情報データセットに対する性能評価結果では、物理化学的インデックスを特徴量に加えた場合、DeepSeqPanII のデフォルトによる AUC 値よりも平均 2.1%の向上が確認でき、立体構造情報データセットに対する性能評価結果では、特徴量に物理化学的インデックスと PSSM を組み合わせることによって平均 6.9%の向上が得られた。

研究協力者
池田　修己　　産業技術総合研究所
坂無　英徳　　産業技術総合研究所

A. 研究目的

第二次産業革命を契機として、世界の人口増加、温暖化による砂漠化などによる農地面積の減少、作物収穫量の低下、農業従事者の減少などの諸問題によって世界的な食糧不足への懸念が強まっている。こうした問題を背景に、害虫抵抗性や除草剤耐性をもたせることで収量増を見込める遺伝子組み換え作物（GMO）の実用化が進んでいる。日本では遺伝子組み換え作物規制条例で栽培を規制しているが、家畜飼料用がほとんどではあるものの輸入に依存しているトウモロコシ、ダイズ、菜種などは半量が既に遺伝子組換え作物であると推定されている。

GMO に対してはアレルゲン性が遺伝子改変食品の安全性が問われているが、すべての遺伝子組換え食品のアレルゲン性を実験的に評価するのはコストの面から現実的ではない。このため、科学的根拠をもつ信頼性のある評価方法の確立が求められている。適切なリスク管理対策の適用により、遺伝子改変食品のアレルゲン性リスクを低減することが出来るかもしれない。

組換え DNA 技術で導入した新規遺伝子産物（タンパク質）や形質転換による意図しない新規タンパク質のアレルゲン性予測方法としては、FAO（国連食糧農業機関）/WHO（世界保健機関）が提唱しているデータベースに登録済みのアレルゲンタンパク質との相同性比較（[1] 80 個の連続したアミノ酸配列について 35%以上の相同性、[2] 6～8 個の連続したアミノ酸配列の完全一致）が標準的に使用されている。

しかし、配列長が短い既知アレルゲン性ペプチドとの類似性に基づくため偽陽性が高いことが指摘されている。また、進化に基づくアミノ酸置換行列を用いる配列類似性比較は、オフターゲット効果による変異をもつ新規タンパク質に対するア

レルゲン性の判定には十分ではない可能性がある。

機械学習は使用する学習データに強く依存する。アレルギー発症の機序において、HLA 分子は種々のペプチドを結合し、T 細胞に抗原提示を行い、T 細胞受容体との結合によってサイトカインの産生や細胞傷害性活性が起こる。近年、必ずしも HLA 分子とペプチドとのアフィニティ（結合親和性）およびオフレートと T 細胞の活性化が一致していないことが明らかとなっている。したがって、T 細胞活性化/不活性化の実験的根拠をもつ正例／負例データの選定およびデータセットの準備には特段の注意を要する。

そこで本研究では、標的配列と類似した配列のオフターゲット検索しかできない点を克服すべく、人工知能を活用して相同性がないアレルゲン性タンパク質由来ペプチド－HLA クラス II 分子間結合予測法の開発を行うことを目的に取り組んでいる。

研究計画初年度の昨年度では、既存予測手法が使用した機械学習法およびデータセットについて調査を行なった。さらに、HLA クラス II 分子－アレルゲン性ペプチド－T 細胞受容体複合体の立体構造データを用いて、ペプチド結合クレフトにおける HLA クラス II 分子－アレルゲン性ペプチド間の相互作用状態について解析した。

今年度は、まず、既存のデータセット以外に新たにデータセットを 2 種類構築した。次に、これを用いて既存手法の深層学習による予測法に、アミノ酸物理化学的インデックス、位置特異的スコア行列（PSSM）、アミノ酸残基間ポテンシャルインデックスを特徴量として用いたモデルを構築した。LOAO（leave-one-allele-out）交差検証テストによってモデルの予測性能を比較し、性能向上の可能性とどの特徴量が有効であるかを検証した。

B. 研究方法

（1）データセットの構築

前述のとおり、以下の 3 種類のデータセットを用意した。

①既に提案されている予測法 DeepSeqPanII の学習に使用した「BD2016 データセット」を公開サイトからダウンロードし、そのまま使用した。

②オリジナルのデータセットとして「IEDB2022 データセット」を構築した。構築フローを図 2 に示す。HLA クラス II 分子の配列情報およびアレル情報を収めた IPD-IMGT/HLA データベース（release 3.47, 2022-01）からデータファイルをダウンロードした。アレル名に接尾辞「N」、「Q」、「L」、「S」の記載をもつエントリを取り除いた。一方、IEDB（2022-04）から結合ペプチド分子名、HLA 遺伝子名・アレルグループ・アレル型、結合ペプチド配列、アフィニティ、アッセイなどの情報を含むデータをダウンロードした。データの信頼性を確保するために、以下の条件を設けフィルタリング処理によるデータ選別を行った。

- IC50（50%阻害濃度）の値をもち、カラム「Units」の値が「nM」の単位として与えられており、カラム「Quantitative measurement」の値が「null」でないデータ。
- 対象とするアレル名に「/」が含まれるデータ（α 鎖タイプ/β 鎖タイプ）、および「HLA-DRB」であるデータ。
- 同じアレルかつ同じペプチドが存在する場合、その重複のペプチドが正例と負例の両方に存在する場合、それら全てを除外する。正例または負例に重複のペプチドが存在する場合は、その中で IC50 が最小のものを採用し、それ以外は削除する。
- アレル名に接尾辞「N」、「Q」、「L」、「S」の記載をもたないデータ。

③高分子立体構造データベース PDB に登録されているペプチド結合状態の HLA クラス II 分子を、ウェブツール IMGT を用いて検索・抽出した。さらに、以下の条件でデータの選別を行った。

- CD4、TCR エントリを除去。

・ HLA-DM タイプのエントリを除去。
・ HLA のペプチド結合領域、結合ペプチドに非標準アミノ酸／unknown を含むエントリを除去。
・ HLA のペプチド結合領域に欠失／置換のあるエントリを除去

同じく IMGT を用いて、HLA クラス II 分子と結合ペプチド間において水素結合を形成する残基情報、位置、結合数の情報を抽出し、「PDB-HB データセット」として構築した。構築フローを図 3 に示した。

（2）特徴量

予測法に組み込む特徴量は以下の 5 種類を用いた（ただし、One-hot encoding と BLOSUM62 に関しては、DeepSeqPanII において採用されている特徴量である）。

・ One-hot encoding：20 種類のアミノ酸を 20 次元の bit ベクトルで表現。
・ BLOSUM62：アミノ酸置換行列の値を 23 次元のベクトルで表現。
・ AAindex 物理化学的インデックス：20 種類のアミノ酸に KEGG AAindex データベースに登録されている 566 種類のインデックスの値を割り当てる。566 種類の全てを次元圧縮した場合と、類似したインデックスを相関係数に基づいて非冗長な 62 種類に削減したうえで次元圧縮した場合とを別々に特徴量として追加。
・ PSSM：マルチプルアラインメント処理によって得られたカラム重複部分の重み付きアミノ酸出現頻度として、20 次元のベクトルで表現。
・ AAindex 残基間ポテンシャルインデックス：KEGG AAindex データベースに登録されている 20 種類のアミノ酸残基間におけるポテンシャルのスコア行列（47 種類）を用いて、立体構造既知の HLA クラス II 分子とペプチド間で 3.5 オングストローム以内に近接する残基ペアに対してスコアを割当。

（3）深層学習ベースモデル

既に提案されている予測法の DeepSeqPanII をベースモデルとして、前節で述べた 5 種類の特徴量を組み合わせて、3 種類のデータセットごとに LOAO 交差検証テストを行った。

ここで、DeepSeqPanII をベースモデルに用いた利点は、第一に、DeepSeqPanII が再帰的ニューラルネットワーク（Recurrent Neural Network；RNN）の一種である LSTM（long short-term memory）を用いていることである。LSTM の強みは、時系列データの学習や予測（回帰・分類）にあり、HLA クラス II の α 鎖、β 鎖、ペプチドで構成される 3 組の一次元配列間における線形での結合状態を考慮する上で、モデルが複雑にならないことが期待できる（図 1）。第二に、注意機構（attention mechanism）を採用しており、RNN が記憶しきれない過去の情報を記憶にキャッシュすることによって、ニューラルネットワークの内部を可視化することができる長所をもつ。第三に、一連のプログラムが GitHub にて公開されており、MIT License として再利用が認められていることである。したがって、新たに別の特徴量を組み込んで利用することは問題にならない。

（4）LOAO 交差検証テストと AUC の算出

LOAO（leave-one-allele-out）交差検証では、データセットに含まれるアレル分子群から 1 つのエントリだけを抜き出してテスト事例とし、残りをトレーニングとする。この処理を全てのエントリが一回ずつテスト事例となるよう検証を繰り返す。

LOAO 交差検証テストの結果から、各 HLA クラス II 分子に対し、

$$FPR = FP / (FP + TN)$$

$$TPR = TP / (TP + FN)$$

を算出できる。横軸に FPR（偽陽性率）、縦軸に TPR（真陽性率）をとった時に描かれる ROC 曲

線下側の面積（AUC：area under the curve）の値に基づいて、予測性能を評価する。

C. 研究結果および考察

（1）構築したデータセット

LOAO 交差検定テストに使用する「IEDB2022 データセット」および「PDB-HB データセット」の構築フローを、それぞれ図 2 と図 3 に示した。信頼性の高い学習効果を得ることを目的に、データ選別あるいはクリーニングの条件を設けた。

3 種類のデータセットにおけるアレルタイプ別の配列数を表 1 にまとめた。また、「BD2016 データセット」と「IEDB2022 データセット」における正例と負例ペプチドデータ数を表 2 にまとめた（「PDB-HB データセット」は立体構造既知データに由来しているため、結合ペプチドデータは全て正例として扱う）。

「IEDB2022 データセット」は「BD2016 データセット」に比べて新しいデータが含まれており、また、曖昧な実験情報あるいは矛盾した実験結果が記載されているエントリを含めないデータ選別処理を行ったため、「BD2016 データセット」よりもやや少ない構成となっている。「PDB-HB データセット」に関しては立体構造データそのものが限られており、「IEDB2022 データセット」と同様に HLA-DP タイプが含まれていない。

（2）LOAO 交差検証テスト結果

①多くの予測法が学習用データセットとして利用している「BD2016 データセット」に対する AUC による LOAO 交差検証テスト結果を図 4 に示す。One-hot encoding と BLOSUM62 を特徴量に用いているオリジナルの予測法 DeepSeqPanII よりも、AAindex を特徴量に加えた場合が全体的に高い AUC 値を示した。特に、AAindex（次元圧縮）が 54 アレル中 50 アレルで DeepSeqPanII よりも高く、2 アレルは同じ AUC 値であった。AAindex（次元圧縮）のほうが AAindex（非冗長 62 インデックス＋次元圧縮）よりも、2 アレルを除いて、優れていることが分かった。

②本研究で構築したオリジナルデータセットである「IEDB2022 データセット」に対する LOAO 交差検証結果を図 5 に示す。「BD2016 データセット」と同様に AAindex（次元圧縮）は DeepSeqPanII よりも AUC 値が高い傾向にあり（15/19 アレル）、AAindex（非冗長 62 インデックス＋次元圧縮）よりも全アレルで優れていることが分かった。正例／負例の閾値（500 [nM]、1000 [nM]）の違いは AUC 値に顕著な影響を与えなかった。図には示していないが、AAindex を次元圧縮せずに 566 インデックスを特徴量に加えても AAindex（次元圧縮）の AU 値を全アレルで超えられていなかった。

③HLA クラス II 分子－ペプチドの立体構造データから選別した「PDB-HB データセット」に対する LOAO 交差検証結果を図 6 に示す。AAindex（次元圧縮）＋PSSM による AUC 値が 8 アレル中 5 アレルにおいて最も高い AUC 値を示した。1 アレル（DRA*01:01－DRB1*01:01）のみ、DeepSeqPanII のデフォルトと同値であるが、それ以外の 7 アレルでは顕著な差（AUC：6.9±3.2）があった。この特徴量の単独と比較すると、組み合わせによる効果が 5/8 アレルで有意であることが分かった。アミノ酸残基間ポテンシャルを加えた結果は、AAindex（次元圧縮）＋PSSM よりも高い AUC が得られた一方で、1 アレル（DQA1*03:01－DQB1*03:02）では他の特徴量に比べて顕著に低い値となった。

D．結論

深層学習を用いた DeepSeqPanII をベースモデルに、特徴量（デフォルトの one-hot encoding と BLOSUM62 に加えて、アミノ酸物理化学的インデックス、PSSM、アミノ酸残基間ポテンシャルインデックス）を組み合わせてトレーニングさせた。

2 種類の一次元配列情報データセットに対するLOAO 交差検定による予測性能評価結果では、物理化学的インデックスを用いた場合、DeepSeqPanII の AUC 値よりも平均 2.1%向上した。ペプチド結合残基位置が明らかな立体構造情報データセットに対する性能評価結果からは、特徴量に PSSM と残基間ポテンシャルインデックスを組み合わせることによって平均 6.9%の予測精度を向上できることを明らかにできた。

計画 3 カ年中の 2 年目において概ね予定通りに進捗している。最終年度となる次年度では、実際に食物アレルギーの発症が確認されているアレルゲンペプチドデータへの予測と評価を行う。また、他の既存の予測法との性能比較を行う予定である。論文誌上ならびに学会や研究会などでの成果発表を行い、研究に使用したデータセットや開発した一連のプログラムをコミュニティ向けに限定公開し、改良を重ねつつ利用促進活動を行う。

E. 研究発表・業績

1. 論文発表

無し

2. 学会発表

無し

F．健康危険情報

該当なし

G. 知的財産権の出願・登録状況

該当なし