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

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：
食物アレルギー発症の細胞分子生物学的メカニズム、アレルギー物質と関与生体分子を登録するデータベース（IPD-IMGT/HLA、IEDB）、HLA 分子－ペプチド結合予測法の先行研究（NNAlign_MA、NetMHCIIpan-4.0、DeepSeqPanII、MHCAttnNet、MHCII3D など）が用いた学習用データセットに関する調査を行なった。これらには、HLA 遺伝子名・アレルグループ・アレル型、アレルゲンペプチド配列、アフィニティー、アッセイなどの情報を含んでおり、機械学習に要する正例／負例のデータセットとして使用する。次に、ヒト由来 MHC クラス II 分子配列情報（UniProtKB/Swiss-Prot：15 配列）と構造情報（PDB：α 鎖=216、β 鎖=193）とを関連付けた。現在、アレルゲンペプチド－MHC クラス II 分子複合体の構造データを IMGT のデータを用いて解析し、MHC クラス II 分子とペプチドとの結合部位における両者の残期間相互作用の関係性を抽出する作業を重点的に行なっている。これらのリソースは、新規アレルゲン性予測法の開発に役立つことが期待できる。

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、MHC 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 MHC 分子とペプチドとのアフィニティー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 positive/negative 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。

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の予測や非天然型アミノ酸を含むペプチド－MHC クラス II 分子間結合予測法の開発を行う。

研究計画初年度の今年度は、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した。次に、MHC 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 MHC 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。

B. 研究方法

（1）公的データベース

機械学習に使用するための MHC クラス II 分子の配列情報およびアレル情報を収めたデータベース（UniProt/Swiss-ProtKB、IPD-IMGT/HLA）からデータファイルをダウンロードした。UniProt エントリに記載されている生体高分子立体構造データベース PDB との相互参照情報を利用し、両データベース間のエントリ情報を関連づけた。

さらに、IEDB からアレルゲン性ペプチドとエピトープ、結合分子名、HLA 遺伝子名・アレルグループ・アレル型、アレルゲンペプチド配列、アフィニティー、アッセイなどの情報を含むデータをダウンロードした。これらの多くは、既存の予測法が学習/テスト用データセットとして使用したエントリを含んでいる。

（2）先行研究予測法

先行研究として、既に提案されている主な予測法（ProPred、RANKPEP、MHCPred、MHC2Pred、SVRMHC 、 MixMHC2pred 、 NNAlign_MA 、NetMHCIIpan-4.0、MHCII3D、DeepSeqPanII）について調査した。さらに、これらのうち、NNAlign_MA、NetMHCIIpan-4.0、MHCAttnNet、MHCII3D、DeepSeqPanII 予測法が使用したデータセットについても調査した。

（3）抗原ペプチドと MHC クラス II 分子間相互作用解析

ペプチドと MHC クラス II 分子間相互作用の関係性を調べるために、IMGT において抗原ペプチドと MHC クラス II 分子と T 細胞受容体の複合体の立体構造データを検索した。さらに、PDBe PISA（Proteins, Interfaces, Structures and Assemblies）を用いて、IMGT への検索で得られた複合体構造における抗原ペプチドと MHC クラス II 分子間の相互作用状態について調査した。

（4）人材育成（統計学、バイオインフォマティックス、AI 分野）

分担研究者および協力研究者と共同で行うことで、インフォマティクス関連技術の取得に努める。

C. 研究結果および考察

（1）公的データベース

UniProt/Swiss-ProtKB からヒト MHC クラス IIα 鎖（各 1 配列、計 6 配列：DRA、DQA1、

DQA2、DPA1、DMA、DOA)、β 鎖（各 1 配列、計 9 配列：DRB1、DRB3、DRB4、DRB5、DQB1、DQB2、DPB1、DMB、DOB）を抽出し、構造情報（PDB：α 鎖=216、β 鎖=193）と関連付けた。さらに IPD-IMGT/HLA からアレル遺伝子由来タンパク質配列情報（合計 8,331 配列）をダウンロードした（表 1)。IPD-IMGT/HLA データベースにおける HLA 遺伝子データ登録数は年々増加しているが、HLA 抗原型ごとによって配列数に大きな開きがある。

（2）先行研究予測法

提案されている 11 種類の MHC クラス II－ペプチド結合予測法（ProPred、RANKPEP、MHCPred、MHC2Pred、SVRMHC、MixMHC2pred、NNAlign_MA、NetMHCIIpan-4.0、MHCAttnNet、MHCII3D、DeepSeqPanII）に関して、採用しているアルゴリズム、サーバ URL の一覧を表 2 にまとめた。これらのうち、5 種類の予測法が開発の際に使用した学習用データセットについても調査した（表 3)。多くの予測法は IEDB に登録されているデータをリソースとして使用し、データセットを継承あるいは統合するなどの工夫を行ない、量と質の確保に努めている。MixMHC2pred では著者らが LC-MS/MS を用いた実験によって新規ヒト MHA クラス II 結合ペプチド（77,189 リガンドペプチド）を同定し、これを評価用データセットとして使用している。

ただし、これらのすべてのペプチドデータが MHC クラス II 分子との結合が実験によって検証されたデータだとしても、学習用データとして利用する場合は、T 細胞受容体が結合し T 細胞応答が確認できているデータかどうかを IEDB の登録情報を基に精査する必要がある。

（3）抗原ペプチドと MHC クラス II 分子間相互作用解析

IMGT のウェブサイトにてペプチド－MHC クラス II 分子－T 細胞受容体の複合体を形成している立体構造データを検索した結果、29 の複合体が得られた。この 29 複合体中には、DRA：24、DQA1：25、DRB1：18、DRB3：2、DRB5：4、DQB1：24 チェインが含まれていた。現在、PDBe PISA を用いて、得られた MHC クラス II 分子とアレルゲン性ペプチドとの結合部位における両者の残期間相互作用の関係性を抽出する作業を重点的に行なっている。これらのデータは、新規アレルゲン性予測法の開発に役立つことが期待できる。

D．結論

計画 3 カ年中の 1 年目において概ね予定通りに進捗し、先行研究の調査やデータセットの整備などの基盤構築を行えることができた。次年度では、グラフ表現学習法あるいは他の機械学習法との組み合わせに適用することによって、食物アレルギー発症に重要な 3 次元構造の特徴を抽出する学習モデルの構築、モデルと予測精度の検証・評価を行う。

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