厚生労働科学研究費補助金（食品の安全性確保推進研究事業）
「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）

新規アレルゲン性予測手法開発のための基盤的研究

研究分担者　安達　玲子　　国立医薬品食品衛生研究所生化学部　室長

研究要旨：
本研究では、遺伝子改変技術応用食品のアレルゲン性について、より高い精度での評価・予測を可能とすることを目的として、アレルゲン性予測手法（allerStat）の機能拡充を行った。分類器として、これまで使用してきた線形 SVM に加えて、非線形 SVM、ロジスティック回帰、lightGBM、LSTM、proteinBERT を実装し、その性能を検証した。今後、バイオテクノロジー技術応用食品のリスク評価手法としての実用化を目指す。また、国立医薬品食品衛生研究所にて運用・公開しているアレルゲンデータベース（Allergen Database for Food Safety, ADFS）に関して、令和 3 年 6 月から令和 4 年 5 月までの 1 年間に NCBI PubMed に収載された論文から、エピトープ配列決定に関する 18 報のピアレビューを行い、14 種のアレルゲンについて、総数 52 のエピトープ情報を ADFS に追加した。また HESI が運営するアレルゲンデータベース COMPARE の登録アレルゲンに関するアップデートを ADFS に反映させた。これらの情報更新により、ADFS のアレルゲン及びイソアレルゲンのアミノ酸配列情報は 2,403、エピトープ既知のアレルゲン数は 274 となり、遺伝子改変技術応用食品のアレルゲン性評価に有用なデータベースである ADFS を充実させることができた。また、医療・食品分野やリスク評価分野等における AI の活用実態の調査を実施した。現状では医療分野への活用や規制の取り組みが多いことが示され、食品分野やリスク評価分野への活用は今後進んでいくものと考えられた。

研究協力者
爲廣　紀正　国立医薬品食品衛生研究所生化学部
協力研究員

A. 研究目的

遺伝子改変技術を応用した食品開発は、技術的には、外来遺伝子導入による遺伝子組換え食品から、内在性遺伝子の改変を行うゲノム編集技術応用食品へ、また、酵母等に多数の外来遺伝子を導入し新規食品機能成分を産生させる合成生物学の利用へと変化している。現在、ゲノム編集技術では多様な手法が生み出されており、これらの手法による意図しない塩基変化も一様ではないことが明らかになりつつある。従って、このような意図しない変化、及びそこから生じる代謝成分の変化を検出または予測し、その変化が与える影響を正確に評価することは、食品の安全性確保において急務の課題である。

バイオテクノロジー技術を用いて開発された食品のリスクの 1 つに、アレルゲン性増大の可能性がある。本研究では、国立医薬品食品衛生研究所生化学部にて管理・公開している、アレルゲン性予測機能（FAO/WHO 法等）を装備したアレルゲン・エピトープ情報データベース（Allergen Database for Food Safety, ADFS）について、新規アレルゲン及びエピトープ情報の収集・解析等によりアレルゲン性評価に関する検討を行い、遺伝子改変技術応用食品のリスク評価に資するデータベースとなるよう、情報を更新し内容を充実させる。

また、人工知能（AI）を活用した新規高精度アレルゲン性予測手法の開発を進める。令和 2 年度

までの先行研究班では、アレルゲン及び非アレルゲンタンパク質から抽出した特徴的なアミノ酸配列パターンを利用して機械学習によりアレルゲン性を予測する手法（allerStat）を開発してきた。本研究班では、この予測システムの機能を拡充し、高精度アレルゲン性予測法としての実用化を進める。

また、医療・食品分野やリスク評価分野等における AI の活用実態を調査し、課題を整理する。

B．研究方法

アレルゲン性予測手法-1（allerStat）の機能拡充

現システムの分類器である線形SVM（Support Vector Machine)に加えて、非線形 SVM、ロジスティック回帰、LightGBM（Gradient Boosting Machine)、LSTM（Long short-term memory)、proteinBERT（Bidirectional Encoder Representations from Transformers）を分類器として追加実装し、それぞれの分類器を用いた場合の予測性について検証した。

ADFS 登録アレルゲンのアップデート

昨年度までは米国ネブラスカ大学リンカーン校が運営しているアレルゲンデータベース（AllergenOnline）における登録アレルゲンのアップデート内容を ADFS に反映させていたが、2021 年 2 月を最後にデータ更新が行われなくなっている。そのため今年度は、HESI（the Health and Environmental Sciences Institute、国際生命科学協会 ILSI のグローバルブランチ、非営利組織）が運営しているアレルゲンデータベースである COMPARE （Comprehensive Protein Allergen Resource）における登録アレルゲンのアップデート内容を、ADFS に反映させた。

ADFS エピトープ情報の追加

令和 3 年 6 月から令和 4 年 5 月までの 1 年間に NCBI PubMed に収載された論文から、キーワード検索により、エピトープ決定に関するものを抽出した。キーワードとしては、IgE、epitope、linear、conformational、sequence、recognition 等々のワードを使用し、これらを複数組み合わせて 6 通りの検索式を作成して検索を行った。この検索により抽出されてきた論文についてピアレビューを行った。その結果エピトープ情報を報告していると判断された論文について、そのエピトープ情報を整理し、ADFS のデータに追加した。

AI のリスク評価分野等への応用に関する調査

Web of Science にて、AI（artificial intelligence)、machine learning、deep learning、neural network のキーワードを用いて、2020 年以降に出版された総説を検索し、ヒットした 14,626 報の中から、引用数及び内容を考慮して食品・農業分野、医療分野、生物学分野の合計 168 報を入手した。入手した総説からさらに 48 報（食品分野 18 報、医療分野 20 報、生物学分野 10 報）を選択し、その内容を精査した。

続いて、精査した総説において出現頻度及び重要度の高い機械学習・深層学習関連の 15 種の用語（Neural network、Machine learning、artificial intelligence、Random forest、Support Vector Machine、Deep Learning、Logistic regression、Ensemble、Principal component analysis、Boosting、k-nearest neighbor、Decision tree、Naive Bayes、Long Short Term Memory、Natural language processing）のいずれかを含み、かつ“risk assessment”または“safety assessment”をキーワードとして含む 2020 年以降に出版された論文を、Web of Science にて検索したところ、5,181 報がヒットした。その中から、引用数及び内容を考慮して食品・農業分野、医療分野の合計 474 報について検討し、さらに 65 報（食品分野 7 報、医療分野 58 報）を選択して、その内容を精査した。

また、国際機関や各国規制当局（世界保健機関（WHO)・国際連合食糧農業機関（FAO)・コーデックス委員会・医薬品規制調和国際会議（ICH)・薬事規制当局国際連携組織（ICMRA)・国際医療機

器規制当局フォーラム（IMDRF）・米国食品医薬品局（FDA）・米国国農務省（USDA）・欧州食品安全機関（EFSA）・欧州医薬品庁（EMA）・英国食品基準庁（FSA）・英国医薬品・医療製品規制庁（MHRA）・中国国家衛生健康委員会（NHC））における AI 関連の取り組みについても調査を行った。

C. 研究結果

アレルゲン性予測手法（allerStat）の機能拡充

allerStat について、線形 SVM に加えて新たに非線形 SVM、ロジスティック回帰、LightGBM、LSTM、proteinBERT を分類器として追加実装し、それぞれの予測性について比較検討した。20 品目の食品のアレルゲン及び非アレルゲンデータを用いて leave-category-out cross-validation を行い、予測性の指標となる ROC（Receiver Operatorating Characteristic）曲線の AUC（Area Under Curve）を算出した。各分類器における AUC を表 1 及び図 1 に示す。ロジスティック回帰、LightGBM、及び LSTM では、線形 SVM とほぼ同程度の AUC が得られた。非線形 SVM では線形 SVM と比較して AUC が小さい品目が多く、また品目ごとの AUC の変動が大きかった。一方 proteinBERT では線形 SVM と比較して AUC が大きい品目が多く、品目ごとの AUC の変動が小さかった。

これまで分類器として使用してきた線形 SVM、及び、上記クロスバリデーションにおいて大きな AUC を示した proteinBERT について、allerStat の学習データ（アレルゲン及び非アレルゲンデータ）の分布図を図 2 に示す。線形 SVM では、アレルゲンデータのほとんどは高値の領域に存在していたが、非アレルゲンデータは比較的広い領域に分布していた。proteinBERT では、アレルゲンデータのほとんどは高値の領域に、非アレルゲンデータのほとんどは低値の領域に分布しており、線形 SVM と比較して分類性能が改善されていることが示された。このように、線形 SVM 以外の分類器を使用することにより予測性能を向上させることが可能であることが示された。

ADFS 登録アレルゲン（アミノ酸配列情報）のアップデート

米国ネブラスカ大学リンカーン校が運営しているアレルゲンデータベースである AllergenOnline は、登録アレルゲンの全てが国際的なアレルギーの専門家チームによるピアレビューを経ており、登録データの信頼性が非常に高いデータベース（但しエピトープ情報は含まない）である。ADFS における登録アレルゲンは平成 20 年度に AllergenOnline の登録アレルゲンと統合し、その後も、昨年度まで AllergenOnline のアップデートに伴って ADFS 登録アレルゲンのアップデートを行ってきた。しかし、2021 年 2 月を最後に AllergenOnline のデータ更新が行われなくなっている。そのため今年度は、AllergenOnline と同様に専門家によるピアレビューにより登録データの信頼性が非常に高いデータベースである、COMPARE（ILSI のグローバルブランチである HESI が運営している）における登録アレルゲンのアップデート内容を、ADFS に反映させた。

ADFS エピトープ情報の追加

令和 3 年 6 月から令和 4 年 5 月までの 1 年間で、キーワード検索により抽出された論文は 23 報であった。その中からエピトープ情報が記載されていると思われる 18 報を選択し（表 2）、ピアレビューを行った。その結果、11 報の論文から 14 種のアレルゲンについて、総数 52 種のエピトープ情報を新たに追加した（表 3）。

上記のアレルゲン及びエピトープ情報更新作業により、最終的に、ADFS のアレルゲン及びイソアレルゲンのアミノ酸配列情報は 2,403、エピトープ既知のアレルゲン数は 274、構造既知のアレルゲン数は 194、糖鎖付加アレルゲン数は 127 となった。

AI のリスク評価分野等への応用に関する調査

各分野において AI 技術の活用が進められる中で、特に医療分野においては、画像解析やプログラム医療機器（Software as Medical Device: SaMD）をはじめとする様々な方面で AI 技術の導入及び検討が進められている。各国規制当局は、この革新的ツールの重要性を認識しており、多くの国が、これらの手法や技術の研究、開発、採用を促進するために、国家的な AI 戦略や政策を策定している（後述）。食品（農業）分野における AI の活用は医療分野ほどには進んでいないが、これまでに、土壌特性や気象パターンの予測、作物収量予測、病気や雑草の検出、スマート灌漑、家畜の生産・管理、インテリジェント収穫、生産・流通管理、品質管理、需要・購買行動予測、食品表示からの成分・栄養価予測等に関する報告がある。生物学分野においては、マイクロバイオーム研究、バイオマーカー選択、タンパク質コード配列同定、メタゲノム解析データの活用等に関する報告がある。一方、AI 活用における課題としては、データセットの入手・内容・質、サンプルサイズ、モデルの性能評価・実用化・説明可能性等が挙げられている。

また、医療分野では、様々な疾病関連のリスク予測への AI 活用事例の論文報告がいくつかあった。食品分野では、不確実状況下でのサプライチェーンにおけるリスク予測、家畜の病気に関するリスク予測、乳製品の品質に関するリスク評価、病原菌・重金属・化学物質による汚染に関するリスク予測等の論文報告があった。

国際機関や各国規制当局の取り組みやリスク評価への活用については、次のような調査結果が得られた。

医療・健康分野に関して、WHO では、デジタルヘルスとして AI 技術の活用について多くの検討が進められている。ITU/WHO Focus Group on artificial intelligence for health（FG-AI4H）では ガイダンス文書“Ethics and governance of artificial intelligence for health”が取りまとめられており、健康分野における AI 活用状況、AI に適用される法律・政策及び原則、AI の倫理原則や倫理的課題等がまとめられている。

ICH では医薬品承認申請におけるデジタル関連技術活用のための取り組みが進められている。

ICMRA では、Informal Innovation Network が AI に関するホライズン・スキャンニングとして、2 件のケーススタディ（中枢神経系アプリ、ファーマコビジランスにおけるシグナル管理）を実施し、その結果を踏まえて提言を行っている。

IMDRF では、AI 関連の医療機器に関するワーキンググループが 2013 年に結成され、イノベーションを支援するガイダンスの開発と、安全で効果的な SaMD へのアクセスを世界規模でタイムリーに行うことを目指して活動が行われている。

FDA では、デジタルヘルスとして AI 技術の規制の検討が進められている。デジタルヘルスイノベーション行動計画、デジタルヘルスソフトウェア事前認証プログラム（Pre-Cert プログラム）の試験運用、AI/ML ベースの SaMD に関する行動計画の公表、デジタルヘルス専門組織（DHCoE）の設立、医療機器開発のための指針原則、リアルワールドエビデンス（RWE）の活用ガイダンス、画期的なデバイスプログラム（BDP）関連の取り組み等が行われている。

EMA では、AI と新しいデジタル技術を活用して医薬品規制プロセスの効率を改善し、データに対する洞察を深める目的でプロジェクト・勧告を運営・発表している。

MHRA では、ソフトウェアと AI に対する規制要件を明確にし、患者を保護することを目的とした医療機器規制の改革プログラムを 2021 年 9 月に公表している。

食品・農業分野に関して、FAO では、農業情報・知識へのグローバルアクセスの強化、リモートセンシングと GEO-AI を用いた農作物季節学と農事暦の生成、衛星リモートセンシングデータを用いた水ストレス（乾燥期や干ばつ）の検出、FAO データラボによるツールの実装、写真データからの害虫（ツマジロクサヨトウ）の検出、FAO デジタルポートフォリオ、ヒレの写真からのサメの種類

の識別、水生産性（作物収量を生産に使われた水量で除した値）の向上を目的としたポータルサイトの公開等、AI 技術を活用した多くの取り組みが進められている。

Codex では、AI を食品不正に対抗する革新的な手段の 1 つであるととらえ、2022 年 9 月に、食品偽装を検出するための AI に関する国際会議を開催している。

USDA では、進行中または計画されている AI の使用事例の目録を作成し公表している。2022 年 12 月時点で、USDA の組織内外における AI の使用事例 26 件が登録されている。

EFSA では、リスク評価プロセスのエビデンス管理段階において AI 手法を実装するためのアプローチを開発するため、「リスク評価におけるエビデンス管理のための AI 関連活動のためのロードマップ」プロジェクトを 2021 年 5 月に発足させ、その報告書を 2022 年 5 月に公開している。2027 年までに人間の専門知識と密接に共存する人間中心の AI を適用することによってエビデンスのアクセス性と範囲を拡大し、リスク評価プロセスの信頼性を高めることを目標に掲げ、ロードマップを作成している。

FSA では、2016 年にデータ戦略アクションプランを公表しデータの活用の検討等を進めている。AI を食品衛生評価制度（Food Hygiene Rating Scheme：FHRS）に基づく地方自治体による食品施設の衛生検査を効率化するためのスキームに活用している。

なお、NHC に関しては、2022 年 12 月時点で AI 関連の取り組みに関する英語で記された情報は確認できなかった。

一方、課題としては、デジタルデバイド（情報通信技術を利用して恩恵を受ける者と、利用できずに恩恵を受けられない者との間に生ずる知識・機会・貧富などの格差、情報格差）が AI の導入に影響を与える可能性、デジタル化が想定されていない書類様式等のデジタル化への対応、法令や規制体制の整備（枠組み作成、規制側の人材確保等）の必要性、患者や公衆の安全の確保（AI 導入におけるリスクの考慮、要件の明確化、市販後調査等）、サイバーセキュリティの確保、人間の解釈可能性の程度／人間の解釈が安全性や有効性に影響する可能性、AI を導入したシステムの継続的な学習及び性能確保、公的機関が AI を利用する場合の透明性・公正性の確保、データ利用に関する責任（合法的、安全、公正、倫理、持続性、説明可能な方法でのデータ利用）等が挙げられている。

D. 考察

本研究では、AI を活用した新規アレルゲン性予測手法開発に向けて、アレルゲン性予測手法（allerStat）の機能拡充を行った。分類器として、これまで使用してきた線形 SVM に加えて、非線形 SVM、ロジスティック回帰、lightGBM、LSTM、proteinBERT を実装し、その性能を検証した。現在細かな点の修正等プログラムの整備を進めており、今後、バイオテクノロジー技術応用食品のリスク評価手法としての実用化を目指す。また、ADFS に、アレルゲン及びイソアレルゲンのアミノ酸配列情報を 24 種追加、また、14 種のアレルゲンについて総数 52 のエピトープ情報を追加し、アレルゲンデータベースとしての継続的な充実を進めた。AI の活用状況の調査では、現状では医療分野への活用や規制の取り組みが多いこと、食品・農業分野やリスク評価分野への活用は今後進んでいくと考えられることが示された。また AI 導入の問題点としては、各分野に共通のものとして、データセットの入手・内容・質、サンプルサイズ、モデルの性能評価・実用化・説明可能性等が挙げられる他、規制当局側の問題点としては、法令や規制体制の整備、患者や公衆の安全確保、透明性・公正性の確保、デジタルデバイドへの対応等が挙げられる。

E．結論

本研究では、遺伝子改変技術応用食品のアレル

ゲン性について、より高い精度での評価・予測を可能とすることを目的として、アレルゲン性予測手法（allerStat）の機能拡充を行った。また、令和3年6月から令和4年5月までの1年間にNCBI PubMed に収載された論文から、エピトープ配列決定に関する18報のピアレビューを行い、14種のアレルゲンについて、総数52のエピトープ情報を ADFS に追加した。また、COMPARE の登録アレルゲン（アミノ酸配列情報）に関するアップデートを ADFS に反映させた。これらの情報更新により、遺伝子改変技術応用食品のアレルゲン性評価に有用なデータベースである ADFS を充実させることができた。AI の活用状況の調査では、現状では医療分野への活用や規制の取り組みが多いことが示され、食品分野やリスク評価分野への活用は今後より進むものと考えられた。

F. 研究発表

1. 論文発表

1）Goto K, Tamehiro N, Yoshida T, Hanada H, Sakuma T, Adachi R, Kondo K, Takeuchi I. AllerStat: Finding Statistically Significant Allergen-Specific Patterns in Protein Sequences by Machine Learning. J Biol Chem, submitted.

2. 学会発表

1）安達玲子．医療分野等における AI 技術の利用状況とリスク評価分野への適用性．日本薬学会第143年会（2023年3月26日、札幌）

2）爲廣紀正．機械・深層学習を利用した新たなアレルゲン性予測手法の開発．日本薬学会第143年会（2023年3月26日、札幌）

H. 知的財産権の出願・登録状況

1. 特許取得

なし

2. 実用新案登録

なし

3. その他

なし