食品衛生基準科学研究費補助金（食品安全科学研究事業）
「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推進のための研究」
令和 6 年度　分担研究報告書

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　爲廣紀正　国立医薬品食品衛生研究所 生化学部 第三室長

研究要旨
本研究では、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 5 年 6 月から令和 6 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 10 報のピアレビューを行い、4 種のアレルゲンについて、総数 14 のエピトープ情報を ADFS に追加し、データベースの更新を行った。これらの情報更新により ADFS のアレルゲンおよびイソアレルゲンのアミノ酸配列情報は 2,447、エピトープ既知のアレルゲン数は 315 となった。加えて、データベースへの不正なアクセス等を回避するため ADFS のシステムを大幅に改修し、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となるよう充実化を進めた。
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指し、液体クロマトグラフィー質量分析法（LC-MS/MS）を活用した網羅的アレルゲン性評価システムの開発を試みた。アレルゲンを含む食品としてカシューナッツに着目し、ADFS などで集積した知見を基にエピトープ配列を含む酵素消化ペプチド断片を標的ペプチドとして分析したところ、2 種類のペプチドが分析対象候補として有用であることを確認した。また、先行研究により開発が進められてきた AI を活用した新規高精度アレルゲン性予測手法に関する検証等を進めた結果、LLM システムを採用することで予測性能の向上が期待できることが示された。

研究協力者
田口千恵　国立医薬品食品衛生研究所生化学部

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性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、意図しない変化、およ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（国立衛研）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、遺伝子改変技術応用食品等のリスク評価に活用できるよう、過去一年間に報告された新規アレルゲンおよび新規エピトープに関する情報を整理し、内容の充実化を図る。また、システムやデータに損壊を与える可能性がある情報セキュリ

ティ上のリスクを回避するため、データベース自体の脆弱性に関する見直しを進める。

近年、アレルゲンが食品に含まれているかを確認する分析手法として、高速液体クロマトグラフ-タンデム四重極型質量分析計（LC-MS/MS）による網羅的な定性・定量分析が国際的に検討されている。こうした国内外での現状を踏まえ、本研究では、バイオテクノロジー技術を用いて開発された食品に含まれるアレルゲンを予測・検出する方法として、LC-MS/MS を活用し、ADFS で集積したアレルゲンに関する知見に基づく新たな網羅的アレルゲン性評価システムを開発する。また、先行研究に引き続き、最新の AI 技術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。

B. 研究方法

ADFS エピトープ情報の追加

新たなバイオテクノロジーを用いて得られた食品に含まれるタンパク質についてアレルゲン性等を予測・評価するため、これまで国立衛研では、アレルゲンデータベース（ADFS）を管理・公開している。最新の情報が利用できるデータベースとして継続的に運用するため、令和 5 年 6 月から令和 6 年 5 月までの過去 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、エピトープ配列決定に関するものを抽出・収集した。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、確度の高いエピトープ情報が収載され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 データベースに追加した。

ADFS システムの改修

Web データベースのセキュリティ上に生じた弱点が利用された場合、データが不正に読み取られるほか、改ざんなどによって利用者が深刻な被害を受ける可能性がある。ADFS では、管理システムの脆弱性に関する対応を公開以降適時に実施しているが、前回の大幅な改修から時間経過によって不具合が発生している可能性が心配された。そこで、ADFS のシステムの安定性や、セキュリティの向上そして、将来的な拡張性や保守性を考慮し、全面的なシステムの改修を実施した。手順として、オペレーティングシステム（OS）バージョンアップに伴い、フレームワークの改修し、関連ツールのアップデートの後、公開する際の環境において動作確認を実施した。最後に、本改修によって新たに導入された脆弱性が存在しないかを確認するため、外部からの攻撃を想定したペネトレーションテストを実施した。

LC-MS/MS を活用した新たな網羅的アレルゲン性評価システムの開発

分析対象として、食物アレルギー表示における特定原材料への移行が検討されているカシューナッツを選定した。アレルゲンの抽出バッファー（I:デオキシコール酸ナトリウム、ラウロイルサルコシン酸ナトリウム、炭酸水素トリエチルアンモニウムを含む溶液、II:ELISA 用検体抽出液）に食品中のカシューナッツ含有量が 10 μg/g となるよう一次標準粉末を添加し、アレルゲンタンパク質を抽出した。次に、遠心分離およびろ過により不溶物を除去し、上清を採取した。I で得られた上清は、ヨードアセトアミド溶液および炭酸水素トリエチルアンモニウムを添加し遊離チオール基の還元アルキル化反応を行い、遠心エバポレーターで濃縮した。得られた抽出タンパク質はトリプシンあるいはキモトリプシンにより酵素消化し、ペプチドに分解した。安定同位体標識した内部標準ペプチドは後述の固相抽出前（I）あるいは酵素消化前（II）に添加した。消化後のペプチド混合物は、固

相抽出カートリッジ(I: C18 カラム、II:陽イオン交換係カラムおよび陰イオン交換系カラム)で精製した。I については、加えて陰イオン交換レジンによる固相抽出を実施した。得られた固相抽出液は、ロータリーエバポレータ等の濃縮装置により乾固し、分析用再溶解液(0.1%ギ酸―アセトニトリル水溶液)で溶解後、分析用検体とした。分析用の液体クロマトグラフィー(LC)分離条件を表 1 に示す。LC により分離されたペプチドは、質量分析計(MS/MS)に導入され表 2 の条件で解析した。なお、カシューナッツ特異的なペプチド(Ana o 2、Ana o 3 由来)を多重反応モニタリング(MRM)モードで選択的に検出し、同位体標識ペプチドを内部標準として用い分析条件を確定した。

AI を活用した新規高精度アレルゲン性予測手法の性能検証

自然言語処理技術を応用してタンパク質配列からアレルゲン性を予測するモデルを構築し、複数の代表的な深層学習モデルによる予測性能を比較・評価した。特に、タンパク質配列を文字列として扱い、それを言語モデルで処理する手法のアレルゲン性評価における有効性を明らかにすることを目的として検証を実施した。

今年度は、以下の 4 つのタンパク質自然言語モデルを用い、タンパク質配列からアレルゲン性を予測した。

- ProtGPT2:GPT アーキテクチャに基づくタンパク質生成モデル
- ProLLaMA:軽量化された LLM(Large Language Model)ベースのタンパク質モデル
- ProtBERT:トランスフォーマーに基づく事前学習済みモデル(BERT)
- LSTM:配列処理に特化した再帰型ニューラルネットワーク

各モデルには、同一のタンパク質配列データセットを用い、機能ラベル(アレルゲン性に関する二値データ)を教師信号として付与した。データは種目によってカテゴリーごとに分割し、1 カテゴリーをテスト・検証用、残りのカテゴリーを訓練用として Leave-Category-Out Cross-Validation を実施した。予測性能の評価指標は、ROC(Receiver Operatorating Characteristic)曲線の AUC(Area Under Curve)、PR-AUC、F1 スコア、accuracy(正解率)、precision 等を使用した。

モデルのトークナイゼーション、ハイパーパラメータ調整、および学習率等は統一された設定下で行い、モデル間での公平な比較を担保した。

C. 研究結果

1. ADFS エピトープ情報の追加

令和 5 年 6 月から令和 6 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 10 報を選択し、ピアレビューを行った。その結果、4 報の論文から 4 種のアレルゲンについて、総数 14 のエピトープ情報を新たに追加した。(表 3)

上記のアレルゲンおよびエピトープ情報更新作業により、最終的に、ADFS のアレルゲンおよびイソアレルゲンのアミノ酸配列情報は 2,447、エピトープ既知のアレルゲン数は 315、構造既知のアレルゲン数は 210、糖鎖付加アレルゲン数は 127 となった。

2. ADFS システムの改修

ADFS のセキュリティ脆弱性への対応、および長期的なデータベースの運用におけるサポート体制を考慮し、ソフトウエアの推奨バージョンへの更新等に伴う改修作業を実施した。

OS は、インプレースアップグレードではなく、新規インストールにて環境を構築した。これにより、現行のバージョンより長期的なサポートが提供されるため、安定したシステム運用が見込まれる。加えて、関連ツールやパッケージのアップグレードによって、フレームワークとしての性能が向上し、クエリ処理が最適化され解析精度が向上した。ま

た、動作検証により既存ワークフローに影響がないことを確認した。最後に、改修作業によって新たに生じた脆弱性を調査するため、ペネトレーションテストを実施し、情報セキュリティの更なる強化を図った。

3. LC-MS/MS を活用した新たな網羅的アレルゲン性評価システムの開発

LC-MS/MS を用いた新たなアレルゲン性分析評価法を開発するため、実施可能なカシューナッツの分析法に関する標的配列を整理した。いずれの分析法においても、標的ペプチド配列はアレルゲンタンパクである Ana o 2（レグミン様タンパク質）あるいは Ana o 3（2S アルブミン）配列に由来していた。トリプシンあるいはキモトリプシンによる酵素消化により生成されたペプチドの 9 種類を分析対象とした。

これらの 9 種の標的配列のうち ADFS データベース情報に基づきエピトープ配列が含まれるものを相同性検索したところ、標的ペプチドの一つが Ana o 2 で報告されているエピトープ配列 HSLDRTPRKFHLAGNPK の C 末端側7アミノ酸と一致すること。さらに他の標的ペプチドは Ana o 2 で報告されているエピトープ配列 VFQQQQQHOSRGRNL の N 末端側11アミノ酸と一致することが明らかとなった。そこで、各分析法に基づき、カシューナッツ一次標準粉末が 10 µg/g 相当となるように添加したアレルゲン抽出液から、分析試料を調製し、LC-MS/MS による測定を実施した。なお、各標的ペプチドの測定条件については、事前に合成ペプチドを用いて最適化を実施した。測定結果を図 1 に示す。それぞれのペプチドの検出感度は、安定同位体標識ペプチドのカラム保持時間を参照して確認を行い、カシューナッツに含まれるアレルゲンタンパクとしては国内のアレルギー表示の基準値に準ずる感度で判定が実施できるものと示唆された。このことから、エピトープ配列を標的ペプチドとする LC-MS/MS を用いた新たなアレルゲン性分析評価法の開発はフィジビリティが高いと考えられる。

そこで、より高感度でかつ特異的な標的ペプチドを選定するため、対象ペプチドの範囲をさらに広げた検討に取り組んだ。検討対象として、カシューナッツのアレルゲンタンパクである Ana o 1（ビシリン様タンパク質）を追加し、Ana o 2 や Ana o 3 についても新たな標的候補を検討した。その結果、それぞれ 9、12、3 種のペプチドを検証候補として追加し、そのうち 8 ペプチドが ADFS のエピトープ情報と相同性を有する配列であることが分かった。特に Ana o 3 由来の標的候補配列は、同アレルゲンのエピトープとして報告されている QRQFEEQQR の内部配列であったため、分子量が小さいペプチドではあるが、バイオテクノロジー技術を用いて開発された食品に対するアレルゲン性評価の対象として期待が持たれる。

4. AI を活用した新規高精度アレルゲン性予測手法の性能検証

先行研究である機械学習を活用したアレルゲン性予測手法 allerStat の開発時に整理した学習データセットを用い、自然言語モデルのアレルゲン性予測における性能比較を行った。各モデルの最終的な評価指標のスコアを表 4 に示す。protBERT は、最も高い F1 スコア（0.611）を示し、次いで protGPT2 がほぼ同等の性能（0.606）を示し、proLLaMA と LSTM は少し低い性能（0.561、0.559）であったが、依然として実用に耐える性能を保っていた。また、機械学習モデルを活用した allerStat のスコアは 0.517 であったことから、自然言語処理技術の応用によるアレルゲン性予測性能の向上が認められた。他の評価指標である Accuracy、Precision、ROC-AUC、PR-AUC についてもおおよそ同等の結果が得られている。

D. 考察

本研究では、令和 5 年 6 月から令和 6 年 5 月までの 1 年間に報告された 4 種のアレルゲンに

ついて、総数 14 のエピトープ情報を ADFS に追加した。ADFS は食品安全委員会　遺伝子組換え食品等専門調査会の遺伝子組換え食品の食品健康影響評価に関する技術的文書においてアレルゲン性の予測に用いる解析手法の一つとして、脚注に収載されている。このため、国内外や国際機関等から公表されたアレルゲンに関する情報を今後も継続的に収集し、当該データベースに反映させることは、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評価において重要であると考えられ、引き続きデータベースの充実化に努める必要がある。

また、ADFS システムの脆弱性を放置すると、不正アクセスやデータ漏洩のリスクが高まり、公開停止といった深刻な状況を招く可能性がある。そこで今年度は、ADFS で使用しているアプリケーション等の脆弱性への対応を目的とし、システム改修を実施した。本作業により適切なセキュリティ対策が講じられ、利用者に対する社会的責務が果たされたと考えられる。脆弱性への対応は、信頼性の確保とコンプライアンス遵守が可能となることから、公開データベースとして継続していくためには今後も定期的な見直しが必要と考えられる。

アレルゲンを検査する方法のうち、LC-MS/MS を用いたアレルゲン測定法は、以下のような利点が挙げられる。まず、特定のタンパク質やペプチドを高感度かつ高精度で検出できるため、微量なアレルゲンの存在が正確に確認できる。次に、抗体を用いる ELISA 法と異なり、交差反応のリスクが低く、類似タンパク質との識別が可能である。また、糖鎖等で修飾されたアレルゲンも検出することができる。さらに、複数のアレルゲンを同時に測定できるマルチプレックス解析が可能で、効率的かつ包括的な分析が行える。これらのことから、LC-MS/MS はアレルゲン測定において、信頼性・再現性・網羅性の面で優れた分析手法として期待されている。本事業で開発する網羅的アレルゲン性評価システムは、アレルゲンとしての抗原性獲得に重要な IgE エピトープを標的の基準として検出する分析法であり、今年度の解析結果により、我が国での食物アレルギー表示制度における閾値に準ずる感度において、判定を実施できることが確認できた。このため、新たなバイオテクノロジーを用いて得られた食品を想定し、さらに適用範囲を広げた標的ペプチドを選定することにより、より包括的で高分解能の評価システムが構築できるものと考えられる。

本事業と関連する令和 5 年度までの先行研究班では、アレルゲンおよび非アレルゲンタンパク質から機械学習により抽出した特徴的なアミノ酸配列パターンを利用してアレルゲン性を予測する手法（アレルゲン性予測手法：allerStat）を開発した。一方、最近では医療分野での深層学習の活用や規制の取り組みが多く認められており、食品分野やリスク評価分野においても深層学習の活用が今後進んでいくものと考えられる。そこで本研究班では、自然言語処理技術を応用することでアレルゲン性予測システムの性能の向上を試み、規制用途での実用化の可能性を探った。本検討の結果において、ProtBERT が最も高い性能を示したことは、周囲の文脈を同時に考慮する Masked Language Modeling を採用した自然言語処理が、事前学習によってタンパク質配列中のアレルゲンパターンやモチーフをうまく捉えることができることを示唆している。特に、長距離依存関係の処理に優れた BERT 構造が、アレルゲン性予測において有利に働いたと考えられる。一方、ProtGPT2 は、主に生成タスク向けに最適化されているため、分類タスクにおいては若干性能が劣る結果となったと考えられた。ただし、タンパク質の新機能探索や生成応用には今後も有望であると考えられる。また、ProLLaMA は性能面では BERT や GPT2 より劣っていたが、コンシューマー GPU でも QLoRA の活用により効率よくファインチューニングできるため、推論時間やメモリ効率

の面で有利であり、エッジ用途への展開が期待される。LSTM については、入力長が長くなると勾配消失や文脈保持の限界があり、タンパク質配列のような長大な系列データに対しては不利であったことが予想された。今後は、BERT をベースとした更なるファインチューニングや、他の事前学習済みバイオインフォマティクスモデル（ESM など）との比較検討も行うことで、より高精度なタンパク質機能予測システムの開発が期待される。また今回検討した BERT 以外のモデルにおいても、状況に応じた利点があるため、目的や環境等に応じてモデルを選定する柔軟性が求められると考えられた。

E. 結論

本研究では、令和 5 年 6 月から令和 6 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 10 報のピアレビューを行い、4 種のアレルゲンについて、総数 14 のエピトープ情報を ADFS に追加した。さらに、データベースへの不正なアクセス等を回避するため、システムを大幅に改修し、ADFRS の運用環境を最新の技術スタックに近づけた。脆弱性に対応したデータベースとして公開することにより、利用者の信頼性確保とコンプライアンスを遵守した運用に繋がると考えられる。

また、バイオテクノロジー技術を用いて開発された食品のアレルゲン性の評価において、より高い精度での判定を可能にすることを目的として、LC-MS/MS を活用した新たな網羅的アレルゲン性評価システムの開発と AI を活用した新規高精度アレルゲン性予測手法の性能検証を実施し、最先端の技術導入により高性能のアレルゲン性予測ツールが開発可能であることを示した。

F. 研究発表

1. 論文発表

なし

2. 学会発表

Tamehiro N: ML/AL Based Allergenicity Prediction of Novel Food. 14th Global Summit on Regulatory Science（2024 年 9 月 19 日、米国）

G. 知的財産権の出願・登録状況

特許取得

該当なし

実用新案登録

該当なし

その他

爲廣紀正：レギュラトリーサイエンス教材ポイントシリーズ. PHARM TECH JAPAN. 2024, 40.

表1　高速液体クロマトグラフ（LC）測定条件

	条件
ポンプ	Nexera X3
移動相 A	0.2％(v/v)酢酸水溶液
移動相 B	0.2％(v/v)酢酸含有アセトニトリル
流速	0.3 mL/min
カラム	C-18
カラムオーブン	50℃

表2　質量分析計（MS）測定条件

	条件
分析計	LCMS-8060NX
インターフェイス電圧	1.0 kV
インターフェイス温度	250℃
DL 温度	150℃
ヒートブロック温度	200℃
コンバージョンダイノード電圧	10.0 kV

表3　新たに登録されたエピトープ情報

	Name	start	end	Sequence	Method	CTYPE	Reference	UniProt acc.No
001	Der f 40	79	106	VLIKNEQKVHSFSGASEPKLREAIQQYS	Western Blot/Dot blot/ELISA /Cell based assay	L	PMID 37577028	A0A922I5V1
002	Bet v 1.0101				ELISA/Cell based assay	C	PMID 37701941	P15494
003	Ara h 2.0201	56	71	DEDSYGRDPYSPSQDP	ELISA/Peptide array/Dot blot	L	PMID 37706599	Q6PSU2-1
004	Scy p 1	28	43	EGLHELHVKYNAEHVQ	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	151	165	GEGRKRNQISVGSQS	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	187	205	APSGLEEPCFLKKLPNGHL	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	257	273	EGQTHKENQFTIDTRDA	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	321	335	NHVPGSPFTVKVTGE	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	615	632	PFRLRIGKDEADPAAVSV	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	646	660	TDFIVDTCNAGAGTL	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	726	741	ESSSVVVETVEKTKSG	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	743	761	KGHHGTIIPKFHSDANKVT	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1	817	828	SYKVKERGNHIL	ELISA/Dot blot	L	PMID 37931089	A0A5J6X3F8
	Scy p 1			E216, T270, Y699, V704	ELISA/Dot blot	C	PMID 37931089	A0A5J6X3F8

表4　AI を活用したアレルゲン性予測システムの性能評価

Model	accuracy	precision	F1	ROC_AUC	PR_AUC
AllerSTAT			0.517	0.873	
LSTM	0.898	**0.578**	0.559	0.928	0.667
ProtBERT	**0.901**	0.577	**0.611**	**0.949**	**0.734**
ProtGPT2	0.894	0.569	0.606	0.946	0.729
ProLLaMA	0.882	0.553	0.561	0.898	0.660

図1　カシューナッツ由来酵素消化ペプチドの LC-MS/MS による分析結果