

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2年度）

人工知能を用いたアレルギー性評価のためのアルゴリズム開発

研究分担者 竹内 一郎（名古屋工業大学）

研究要旨：

ゲノム編集技術などを用いて人工的に生成した食品のアレルゲン性を確認する方法は明らかになっていない。アレルゲン性の主要な識別子とされる単一の因子は知られておらず、複数の因子が複雑に関連することでアレルゲン性を持つことが示唆されている。また、人工的に生成された食品のアレルゲン性を都度実験的に検証するのは様々なコストがかかり、現実的でない。そこで、本研究では、人工知能やデータ科学のアプローチを用い、食品のアレルゲン性を高精度、高信頼度で汎用的に判定・予測できるシステムを開発することを目指す。これまでに、様々なアレルゲン性を判定・予測するための分析ツールが開発されてきた。残念ながら、これら既存の方法には様々な問題点が存在する。国際連合食糧農業機関/世界保健機関によるガイドラインはアミノ酸配列の類似性に基づく基準であり、精度が低く、大規模データの分析には適していない。また、既知のIgE エピトープに基づく基準、タンパク質構造の物理化学的表現に基づく基準、アミノ酸/ジペプチド組成物に基づく基準など、タンパク質に関する生物科学的・物理化学的な知見に基づく単一、もしくは少数の因子を採用したツールが提案されているが、これらはアレルゲン物質の多様性を十分に反映できるものとなっていない。本研究では、以下の3つの課題に取り組む：(課題1) 既存のデータベースを拡張し、アレルゲンタンパク質と非アレルゲンタンパク質のデータベースを作成する。(課題2) 課題1で作成したデータベースをもとに、アレルゲン特異的なパターン(アミノ酸配列)を統計学的手法により抽出する。(課題3) 課題1で作成したデータベースと課題2で抽出したパターンをもとにアレルゲン性判定モデルを人工知能・機械学習手法により作成する。本年度は、課題1に関して、年度新たなデータを追加してデータベースを更新した。課題2に関して、更新後のデータベースをもとにアレルゲンパターンの再抽出を行った。課題3に関して、更新後のデータベースをもとにアレルゲン性予測システムを再学習し、既存の方法との比較検証のうえで、予測精度の観点で提案するシステムの性能がよいことを確認した。

A. 研究目的

ゲノム編集技術を用いた人工的な農産物の合成が行えるようになり、これまでにない食用タンパク質製品が登場する可能性がある。新たに合成された食用タンパク質は未知の特性を有しており、特定の人が摂取するとアレルギー反応が起こってしまうリスクがある。免疫反応においてタンパク質抗原のアミノ酸配列のうち、抗体が結合する部位をエピトープと呼び、エピトープを認識する抗体を人が持っている場合にアレルギー反応を引き起こされる。これまでの様々な研究から、いくつかのアレルゲンタンパク質において共通のエピト

ープ配列が見出されているが、アレルゲン性の単一因子は知られておらず、複数の因子が複雑に関連することでアレルゲン性を持つことが示唆されている。既存のアレルゲン性判定・予測ツールのうちもっとも基本的なアプローチはアレルゲン性を持つタンパク質とのアミノ酸配列の類似性(アミノ酸配列相同性)に基づくものである。しかし、このようなアプローチは偽陽性が高いことが指摘されており、ゲノム編集技術によって合成される新規タンパク質のアレルギー性判定には十分でない。また、別のアプローチとしては、タンパク質に含まれるアミノ酸の物理化学的な特徴の統計量

に基づいてアレルゲン性を判定する試みもなされている。このようなアプローチではアミノ酸の順序や位置関係を適切に考慮できないため、十分な精度ではないことが確認されている。アミノ酸配列パターンを用いたアプローチとして、Alledictorと呼ばれる方法が提案されたが、この方法では一定の長さのアミノ酸配列のみを抽出するものであり、すべてのエピトープを網羅できるようなものではない。このような背景のもと、本研究では人工知能や機械学習のアプローチを用い、食品のアレルゲン性を高精度かつ高信頼度で汎用性のあるアレルゲン性判定・予測が行えるシステムを開発することを目指す。本研究では、まず、食品タンパク質の大規模データベースを整備し、アレルゲン特異的な様々な長さのアミノ酸配列を抽出し、これらに基づいてアレルゲン性判定・予測システムを構築する。さまざまな数理技術、情報技術を活用することで、高精度で信頼性が高く汎用性のあるアレルゲン性判定・予測システムを開発することを目的とする。令和2年度は、平成30年度に作成したプロトタイプおよび令和元年度のシミュレーションで明らかになった課題を解決し、AIによるアレルゲン性判定・予測システムを完成させた。

**(課題1)** 人工知能や機械学習で判定・予測システムを構築するには訓練データベースが必要である。既存のアレルゲン性判定・予測システムで使われていたデータベースはアレルゲンタンパク質のみを用いたものであった。人工知能や機械学習では正例 (positive example) だけでなく、負例 (negative example) もあると有効なため、後者をデータベースに追加する必要がある。負例の追加では、アレルゲン性とは無関係のタンパク質データベースを取得し、そこからアレルゲン性のあるものを取り除く作業により行った。本データベースにおいて注意すべき問題は、アレルゲン性タンパク質数 (正例数) と非アレルゲン性タンパク質数 (負例数) に偏りがあることである。正例は生物学的な実験によって判定されたものであるため数が少なく、負例は通常のタンパク質データベースから大量に取得できる。一方、通常のタンパク質データベースから大量に取得した負例には誤陰

性 (False Negative) が多く含まれてしまうため、なんらかの対処が必要である。また、正例と負例の数が食物種目ごとにバラつきがある場合、特定の食物種目に特化したアミノ酸配列がアレルゲン性特異的なアミノ酸配列と誤って発見されてしまうリスクが生じる。本年度は、新規にアレルゲン性が判明した食品タンパク質をデータベースに増加するとともに、非アレルゲンタンパク質として登録されていたタンパク質のうち、不確実性が高いものをデータベースから取り除いた。

**(課題2)** 人工知能や機械学習でタンパク質の物性を判定・予測するにはタンパク質の特徴を機械学習が使える数値データとして抽出しなくてはならない。生物情報学で採用されているアプローチとして主に2通りのものがある。1つ目のアプローチは、タンパク質を構成するアミノ酸の物理化学的な特徴 (疎水性、分子量など) を求め、その平均、分散、相関などを特徴として抽出することである。2つ目のアプローチは、アミノ酸の部分配列のうち、特定の物性を有するタンパク質に特化して頻出する部分配列を特徴として抽出することである。アプローチ1ではアミノ酸の順序や位置を考慮できないため、本研究ではアプローチ2を採用する。また、一般に、機械学習における特徴抽出は、教師なし学習と教師あり学習の2つのアプローチが存在する。本研究においては、前者はアレルゲン性タンパク質の情報のみから特徴抽出を行うことに相当し、既存のアレルゲン性判定・予測システムの多くではこのアプローチを採用されている。本研究では、より判定・予測に有用な特徴を抽出するため、教師あり特徴抽出のアプローチを採用する。本年度は、課題1で更新されたデータベースをもとにパターンの再抽出を行った。さらに抽出したパターンと既知のエピトープとの一致性を検証した。

**(課題3)** 正例と負例を含む訓練データベースを用いて、正負が未知の事例を判定・予測する問題は教師あり学習 (supervised learning) と呼ばれている。アレルゲン性タンパク質を正例、非アレルゲン性タンパク質を負例とみなせば、典型的な教師あり学習問題と解釈できるが、いくつか本研究

特有の課題を解決する必要がある。まず、課題の1つ目の特徴は訓練データベースに含まれるタンパク質が独立同一分布 (i.i.d.; independently, identically distributed) に従わない点である。この場合、通常の教師あり学習で多用されるクロスバリデーションなどのリサンプリング法をそのまま利用することができず様々な工夫が必要となる。また、正例数と負例数に偏りが生じてしまう点も本課題の特徴であり、注意深く対処する必要がある。本研究で用いるデータベースにおいて、食品タンパク質に関しては正例が負例に比べて極端に少なくなってしまうっており、非食品タンパク質に関しては正例のみが存在する状況になってしまっている。また、アレルギー性の原因となるエピトープはさまざまな長さであることが知られているため、さまざまな長さのアミノ酸部分系列特徴を抽出できるような工夫が必要である。さらに、アレルギー性の判定は統計的信頼性が担保されたものである必要があるため、抽出された特徴的信頼性定量化を行う必要がある。加えて、特定の食品種目に特化したものでなく、一般的な特徴を抽出するための工夫が必要である。本年度は、課題2で更新されたパターンをもとに予測モデルの再学習を行った。既存のアレルギー性予測システムとの比較を行い、提案法の予測精度が高いことを確認した。

## B. 研究方法

課題1の訓練データベースの構築においては、アレルギー性を持つ食品タンパク質の正例としてCOMPAREデータベースのものを利用した。同じくアレルギー性のない食品タンパク質の負例としてUniProtデータベースより取得した。UniProtデータベースは汎用的なタンパク質データベースであるため、アレルギー性を持つものも含まれている。そのため、既存のエピトープを含むもの、アレルギーに関連するキーワードが付記されているものなどを削除した。またプロトタイプとして作成したアレルギー性判定・予測システムにおいて偽陽性であったタンパク質に関して個別にデータベースを精査し、アレルギー性を持つ可能性があるものは削除するなどの措置をとった。後述のように、課題2、3においては食品種目の情報を活

用するため、食品種目分類の精査を行い、あいまい性のあるタンパク質はデータベースから削除するプロセスを行った。その他にもプロトタイプシステムや諸々のタンパク質データベースを活用することで訓練データベースの大規模化と高精度化を実現した。上述のように、本データベースに含まれる事例(タンパク質)は独立同一分布(IID)に従わないので、食品種目ごとにデータ分割を行うLeave-Food-Outクロスバリデーションと呼ぶ方法に基づいてデータ分析を実施した。2020年度は、本研究開始後に新たにアレルギー性として登録されたアレルギー性食品タンパク質をリストアップし、これをデータベースに追加した。一般に、タンパク質が非アレルギー性であることを完全に確定するのは困難であるため、非アレルギー性タンパク質としてデータベースに登録されているもののうち、False Negativeが存在する可能性が排除できない。本年度は、非アレルギー性タンパク質を精査し、不確実性の高いものを削除してデータベース的信頼性を向上させた。

課題2の特徴抽出においては、本研究に特化したさまざまな工夫を行った。まず、異なる長さのアミノ酸部分配列を抽出できるようにするため、我々が開発したデータマイニング分野の技術を利用した。系列データから特定の性質を持つ部分系列を抽出する技術は系列マイニングと呼ばれ、さまざまな方法が提案されている。系列マイニングでは、系列を木構造と呼ばれるデータ構造で表現し、枝刈りと呼ばれる手順を導入することにより、膨大な部分系列から、特定の性質を満たすものを探索することができる。本研究の基本的な方針は、アレルギー性タンパク質に高頻度で含まれ、非アレルギー性タンパク質には低頻度でしか含まれない(あるいはまったく含まれない)ような部分配列を探索することである。頻度の違いを定量化する指標には様々なものがあるが、本研究ではフィッシャーの正確検定(Fisher Exact Test)に基づく指標を利用した。例えば、20種類のアミノ酸において長さ10までのアミノ酸の種類は10の20乗となり、その頻度を数えたデータテーブルを作成することは実質的に不可能である。

我々は、系列マイニングにおける木構造の枝刈りをフィッシャーの正確検定と統合する方法を開

発した。詳細は割愛するが、この方法では、統計的に有意となり得ない部分配列を木構造の枝刈りによって排除できるため、膨大な数の候補から予測に最適な部分配列を選択することができる。また、アレルギー性予測モデルの信頼性を高めるため、統計的な有意性を持つ部分配列のみを用いることが望ましい。ある部分配列の出現頻度がアレルギー性タンパク質と非アレルギー性タンパク質で異なるかどうかの統計的検定を行う場合、フィッシャーの正確検定の  $p$  値を利用することができる。しかしながら、膨大な部分系列の候補のなかから特に頻度の違いの大きなものを抽出してきた場合、選択バイアスが生じてしまい、所望の誤検出率を制御できなくなる。この選択バイアスの問題は多重検定問題 (multiple hypothesis testing) と呼ばれており、その補正を行うためにはフィッシャーの正確検定によって得られた  $p$  値を適切に補正しなくてはならない。もっともよく使われている多重検定補正にボンフェローニ補正 (Bonferroni correction) と呼ばれるものがあるが、選択における候補数が多い場合、補正が保守的になってしまう問題点が指摘されている。本研究ではこの問題に対処するため、Westfall Young 法と呼ばれるランダム化に基づく方法を採用した本年度は、諸々のタンパク質データベースを活用し、非食品タンパク質でアレルギー性を持たないものをデータベースに加えることができないか検討を進めた。また、昨年度に引き続き、パターン同定アルゴリズムの高速化を行った。さらに同定されたパターンと既存のエピトープとの一致度の検証を行った。

課題 3 のアレルギー性判定・予測システムの構築は、上述の Leave-Food-Out クロスバリデーションを利用した教師あり学習によって行った。アミノ酸部分配列パターンを特徴として抽出したため、テスト対象のタンパク質がパターンを含むか否かをバイナリ表現した線形分類器をベース手法として採用した。パターン数が多いと解釈性が低く過学習のリスクがあるため、スパース正則化や二次正則化 (Ridge Regression) を導入した。本年度は、アレルギー性予測システムの性能を定量評価するため、既存の方法である Alledictor 法、MEME 法、Allertop 法と比較し、本研究で構築し

た方法の予測精度が高いことを確認した。

### C. 研究結果および考察

本年度は、課題 1 に対して、データベースの更新を行った。新たにアレルギー性タンパク質であることが報告されたものをデータベースに加えるとともに、データベースの精査を行った。特に、タンパク質に完全にアレルギー性がないことを確定させるのは困難であるため、非アレルギー性タンパク質として登録されているもののなかには一定の False Negative が含まれている可能性が否めない。

課題 2 に対しては、アレルギー特異的パターンの定義を整理し、その解釈を明確化した。近年の AI 研究では、データ駆動型 AI システムの説明性・信頼性が重要であることが要請されるが、本研究で作成したシステムは、統計的信頼性が担保されたパターンのみを用いており、これらの要請を満たしている。

課題 3 に対しては、既存のアプローチである Alledictor 法、MEME 法、Allertop 法と比較を行い、本研究で作成した方法の予測精度がよりよいことが確認できた。

### D. 結論と今後の展望

令和 2 年度は、平成 30 年度に作成したプロトタイプと令和元年度に行ったシミュレーション検証の結果をもとに、タンパク質のアレルギー性判定・予測システムを完成させた。課題 1 に関しては、データベースの更新と精査を継続的に行っていくためのプロトコルの確立が必要である。課題 2 に関しては、タンパク質の立体構造を考慮したアレルギー性パターンの同定が望まれる。課題 3 に関しては、現状においても予測精度は十分に高いとは言えず、誤検出や未検出が起こってしまう。問題の性質上、誤検出や未検出を完全になくすることはできないため、予測結果を踏まえてリスク考慮型の意思決定を行う手順の整備などが必要である。

### E. 研究発表・業績

#### 1) 論文発表

Inatsu Y., Karasuyama M., Inoue K., Takeuchi

I. Active learning for level set estimation under input uncertainty and its extensions. *Neural Computation*: **32**, 2486-2531. 2020

2) 学会発表

1. Tanizaki K., Hashimoto N., Inatsu Y., Hontani H., Takeuchi I. Computing Valid P-values for Image Segmentation by Selective Inference. IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (CVPR2020), 2020.
2. Duy V.N.L., Toda H., Sugiyama R., Takeuchi I. Computing Valid p-value for Optimal Changepoint by Selective Inference using Dynamic Programming. The 34th Conference on Neural Information Processing Systems (NeurIPS2020). 2020