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

人工知能を用いたアレルゲン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に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、本課題では、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・予測の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規準であり、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規準、タンパク質構造の物理化学的表現に基づく規準、アミノ酸/ジペプチド組成物に基づく規準など、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。課題 1 に関して、2018 年度は食品種目のアレルゲン性、および、非アレルゲン性タンパク質を含むデータベースを構築したが、2019 年度はこれに非食品アレルゲン性タンパク質を追加し、データベースの大規模化、高精度化を行った。課題 2 に関して、2018 年度はアレルゲンタンパク質と非アレルゲンタンパク質それぞれに特異的なパターンを抽出していたが、2019 年度は特定の種や目に限定しないパターンを抽出できるように手法の改良を行った。課題 3 に関して、2018 年度は汎用的な 2 クラス分類モデルを用いていたが、2019 年度はデータベースの特徴を考慮した本研究課題に特化した機械学習法を開発した。本研究におけるアレルゲン性判定・予測システムの概要を図 1 に示す。

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が引き起こされる。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、本課題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で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。2018 年度は、アレルゲン性判定・予測システムのプロトタイプを作成し、その高精度化、高信頼度化、汎用化に向けた問題抽出を行った。2019 年度は 2018 年度のプロトタイプの問題点を列挙し、それぞれを解決するための新たな数理技術、情報技術の開発を行った。

（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。2018 年度には 11 の食品種目のアレルゲンタンパク質と非アレルゲンタンパク質の訓練データベースを作成した。しかしながら、正例数が十分でないため、2019 年度はさらに非食品タンパク質においてアレルゲン性を持つことがわかっているタンパク質を正例として追加する。

（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。2018 年度には食品タンパク質のみを扱っていたため、我々のグループが別の目的で既に確立した方法をそのまま適用することができた。2019 年度は非食品のアレルゲン性タンパク質を正例として追加したため、その対処が必要である。これは、既存の教師あり特徴抽出法を用いると、特定の非食品タンパク質に特化したアミノ酸部分配列がアレルゲン性特異的アミノ酸配列として誤って抽出されてしまうためである。

（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、本研究課題は典型的な教師あり学習問題と解釈できるが、いくつか本研究課題特有の課題を解決する必要がある。まず、本課題の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、アレルゲン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、アレルゲン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、抽出された特徴の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。2018 年度では、訓練データベースが独立同一分布（IID）に従わない点と食品タンパク質における正例と負例の偏りを考慮したモデル作成法を構築した。2019 年度では、さらに非食品タンパク質を訓練データベースに追加した際の対処法を検討した。

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、アレルゲン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。同じくアレルゲン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、アレルゲン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、アレルゲン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ゲン性判定・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、アレルゲン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。2018 年度では、データベースが食品タンパク質のみから構成されていたが、2019 年度には非食品タンパク質も追加した。なお、非食品タンパク質でアレルゲン性のないものを網羅的に収集するのは困難であることが判明したため、本研究では、非食品タンパク質に関しては、アレルゲン性を有する正例のみを扱うこととした。

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、分担者の竹内ら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、アレルゲン性タンパク質に高頻度で含まれ、非アレルゲン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ることは実質的に不可能である。

分担者の竹内らは、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開発した（Sakuma et al., KDD2018）。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、アレルゲン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ゲン性タンパク質と非アレルゲン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(p-value)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。これらの方法の開発と本データベースへの適用は主に 2018 年度に行ったが、2019 年度もアルゴリズムの改良や新たなデータへの適用などを行った。

2019 年度は、主に、非食品タンパク質においてはアレルゲン性を持つ正例のみがデータベースに含まれる点を考慮して特徴抽出を行った。この点を特に考慮せずに通常の機械学習アルゴリズムを適用すると、アレルゲン特異的でなく、非食品タンパク特異的なパターンが誤って検出されてしまう。この問題を回避するため、アレルゲン特異的パターンとして、条件 1） 食品タンパク質に含まれるか、条件 2） 非食品タンパク質のうち複数の目に含まれる、のどちらかの条件を満たすもののみを抽出することとした。2020 年度には、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ゲン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る。

課題３のアレルゲン性判定・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。2018 年度は主にこのプロトタイプモデルに基づく考察を行った。2019 年度は、さらに、パターンが完全に含まれる（exact match）だけでなく、パターンが部分的に類似している場合（non-exact match）も考慮できるような工夫を導入した。20 種のアミノ酸の物理化学的な特徴に基づいてアミノ酸種間の類似度を定義し、タンパク質にパターンが含まれる程度を連続量として定量化した。さらに、2019 年度は、さらに、抽出されたパターンの生物学的な考察として、既存のエピトープとの一致度の確認や、結合性の確認なども行った。

C. 研究結果および考察

2019 年度は 2018 年度に構築したアレルゲン性判定・予測システムのプロトタイプの課題を抽出し、その精度、信頼性、汎用性を向上させるための様々な工夫を行った。

図 2 はアレルゲン性判定・予測システムを作成する際に利用する Leave-Food-Out クロスバリデーションの概要を示したものである。このような工夫をしないと、特定の食物に頻出するアミノ酸部分配列を誤ってアレルゲン特異的パターンとして抽出してしまうリスクが高まる。各食物種をまるごと削除した訓練データを作成して判定・予測システムを作成し、それを削除した食物種のタンパク質のアレルゲン性判定・予測に使うことで、偏りのない判定・予測精度を知ることができる。

図 3 はアレルゲン特異的パターンとして抽出されたパターンを示している。図の各行はアレルゲン性を持つタンパク質のアミノ酸配列を表しており、赤色の部分がアレルゲン特異的パターンとして抽出されたアミノ酸部分配列を表している。図より、アレルゲン性タンパク質が多くのアレルゲン特異的パターンを含んでいることがみてとれる。実際、これらのアレルゲン特異的パターンの生物学的特徴を調べたところ、既知のエピトープと類似していることが確認されている。2020 年度に、さらにこれらの抽出されたパターンの生物学的な分析を行う。

図 4 は 11 種の食物種それぞれに対してアレルゲン性判定・予測を行ったときの ROC 曲線を示している（それぞれのアレルゲン性予測・判定システムは、Leave-Food-Out クロスバリデーションにより、評価対象の食物を一切使わずに作成されて

いることに注意)。従来法を含む複数の判定・予測システムの結果が示されているが、本研究で構築した方法ではおおむねすべての場合において最もよい判定・予測性能を示している。2020 年度はさらにほかのアプローチとの比較も行うことで本システムの有効性の実証を行う予定である。

D. 結論と今後の展望

2019 年度は、2018 年度に構築したアレルゲン性判定・予測システムのプロトタイプにおいて問題点を抽出し、様々な改良を加えた。結果として、訓練データベースの大規模化と高精度化、アレルゲン特異的パターンの信頼性向上、判定・予測システムの精度向上が可能となった。2020 年度は、これまでの取り組みを論文としてまとめるとともに、予測・判定システムの実装を行う。

E. 業績

1. 論文発表

1) Yoshida T., Takeuchi I., Karasuyama M. Safe Triplet Screening for Distance Metric Learning. *Neural Computation*, vol.31, no.12, pp.2432-2491, 2019.

2) Umezu Y., Takeuchi I. Selective inference via marginal screening for high dimensional classification. *Japanese Journal of Statistics and Data Science*, vol.2, pp.2, pages559-589, 2019.

2. 学会発表

1) Ndiaye E, Takeuchi I. Computing Full Conformal Prediction Set with Approximate Homotopy. *Advances in Neural Information Processing Systems* (NeurIPS2019), 2019.

2) Ndiaye E, Le T., Fercoq O., Salmon J., Takeuchi I. Safe Grid Search with Optimal Complexity. *International Conference on Machine Learning* (ICML2019), 2019.

F. 知的財産権の出願・登録状況

該当なし

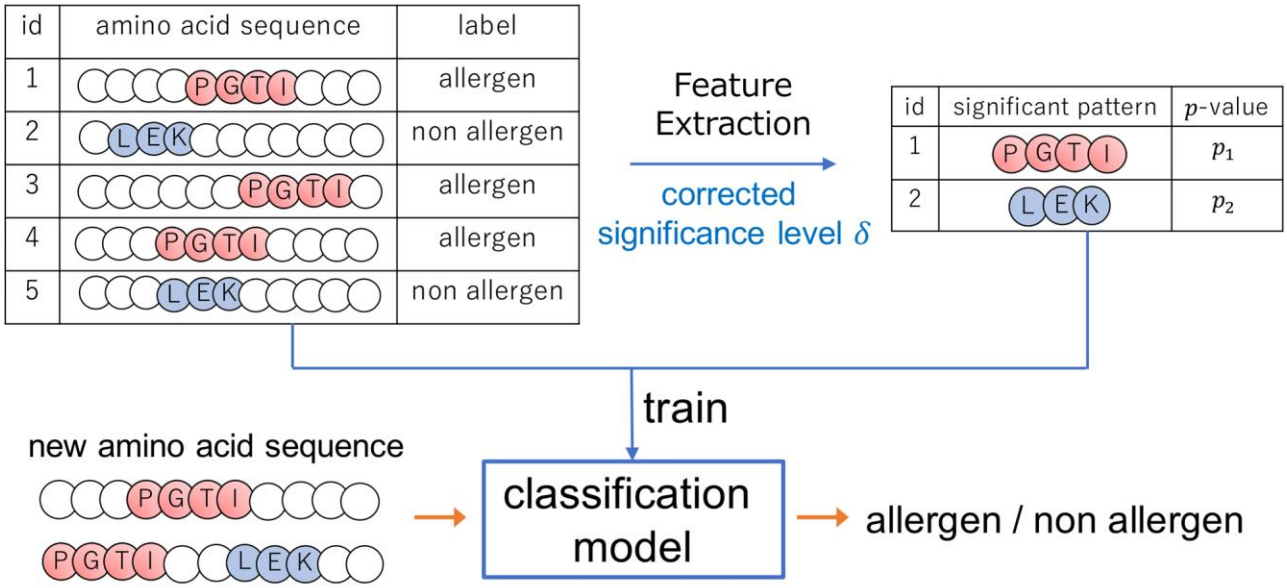

図１　アレルゲン性判定・予測システム構築の全体像

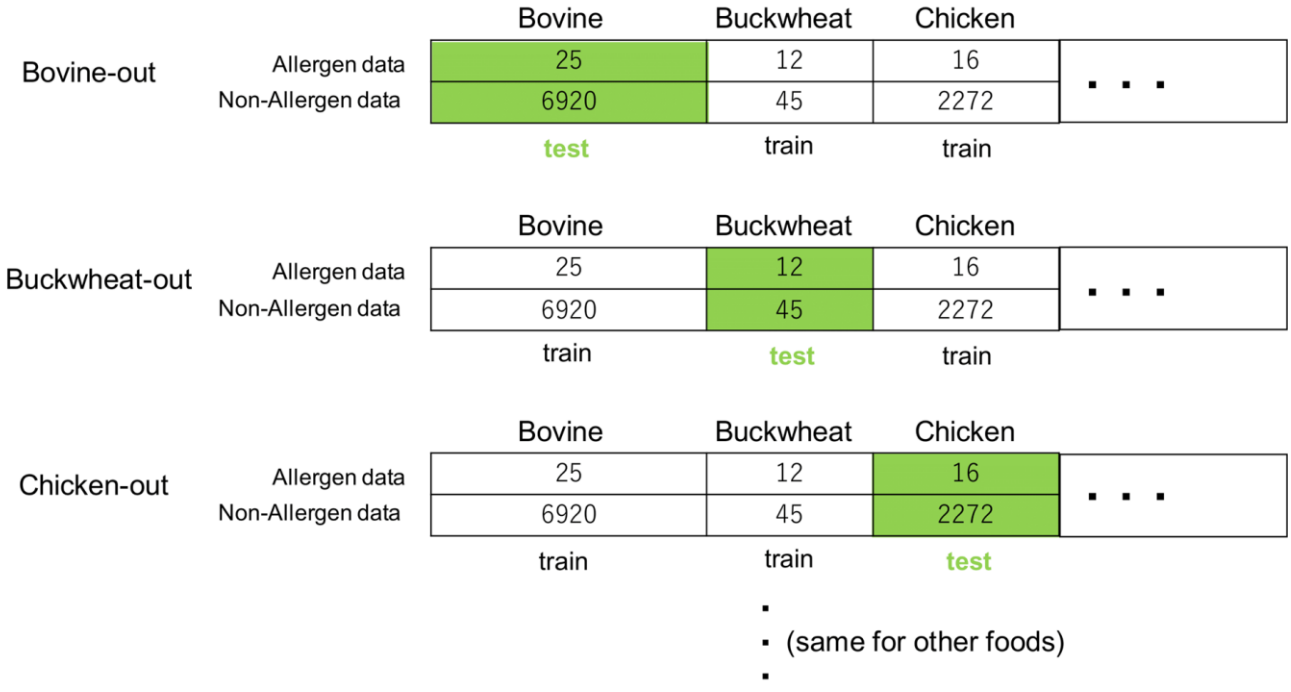

図２　訓練データの非独立同一分布性を考慮した評価方法の概略

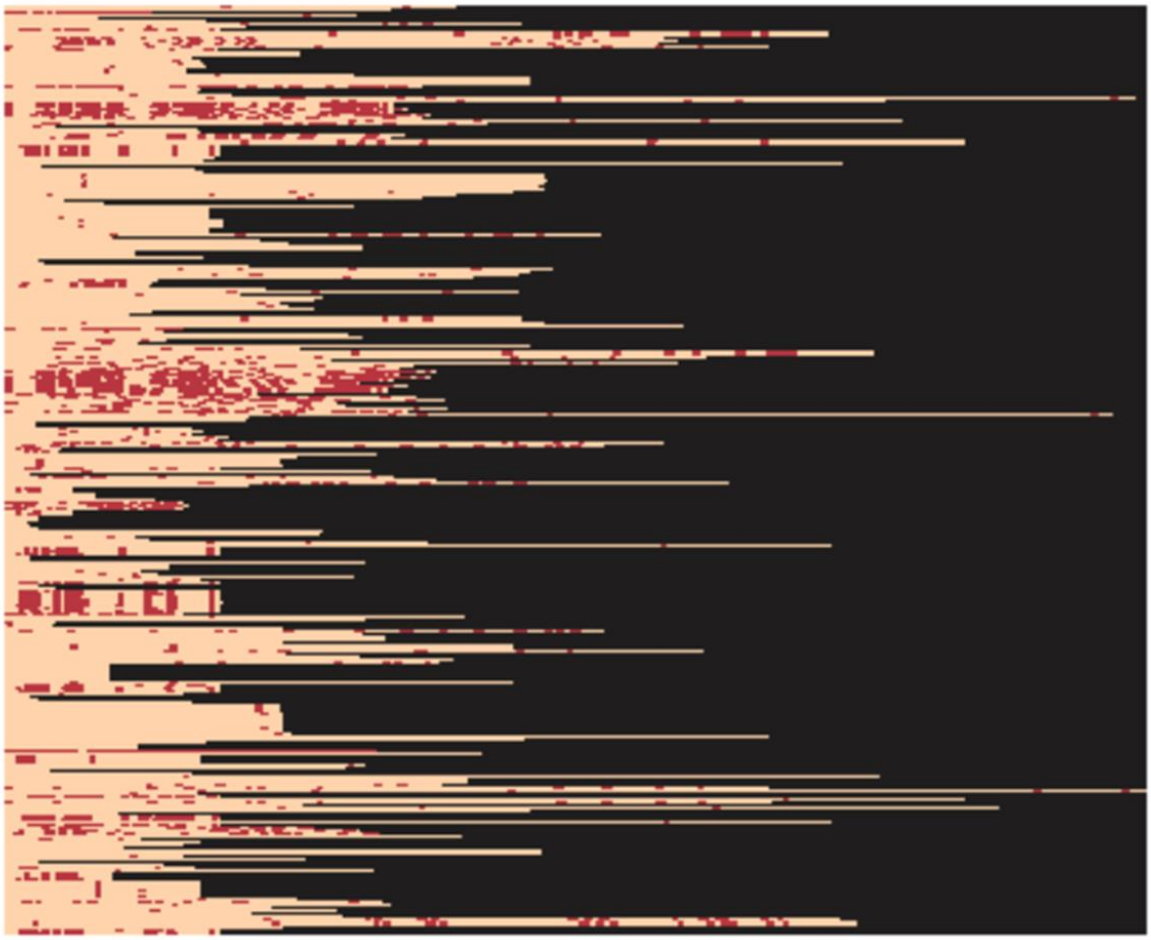

図３　抽出されたアレルゲン特異的パターン（アミノ酸部分配列）の例

各行がアレルゲン性タンパク質を表し、赤くハイライトされている部分がアレルゲン特異的パターンを表している。

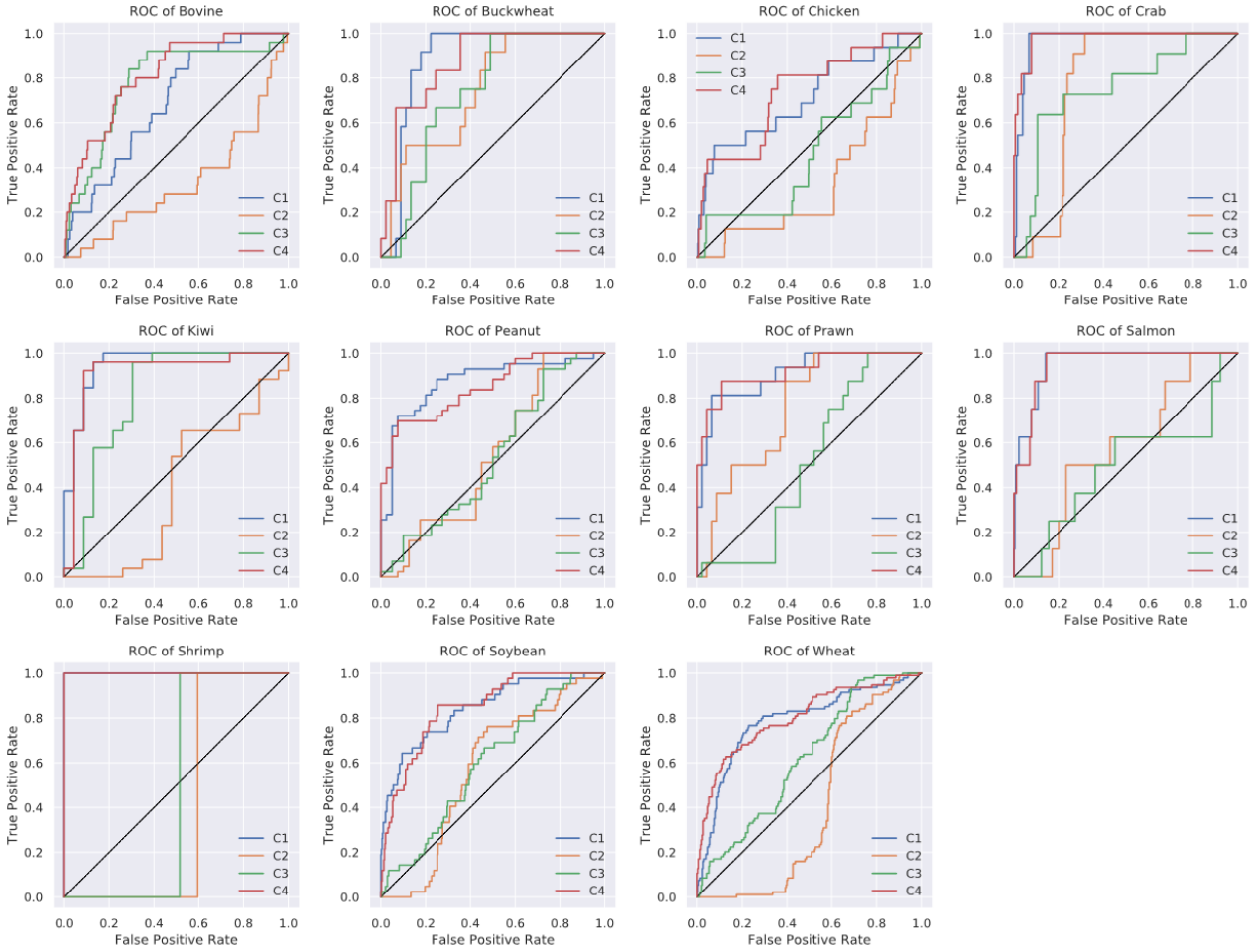

図４　11 種の食品種ごとのアレルゲン性判定・予測結果の AUC 曲線の例
（複数の線は比較した複数の手法に対応）