1ZM-05

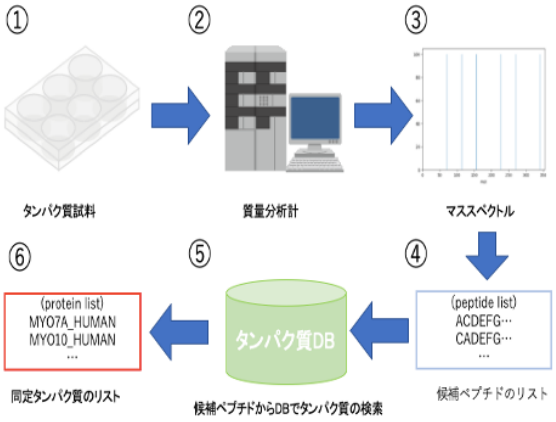
深層学習によるペプチド配列の同定手法の提案

橘　勇人† 　錦織 充広‡ 　高橋　篤* 　大星　直樹§
近畿大学大学院† 福岡大学‡ 国立循環器病センター* 近畿大学§

1. はじめに

タンパク質はアミノ酸と呼ばれるアミノ基とカルボシル基をもつ有機化合物の総称である。こ のタンパク質を同定する手法の一つとして質量分析と呼ばれるものがある。近年、質量分析技術の 進歩により、タンパク質を網羅的に解析し、細胞や組織の状態を明らかにするプロテオーム解析は大きな発展を遂げている。これに伴い、高速かつ大量に高品質なデータを生成できるようになった。正確に整備されたデータベースで人間のものだけで 2 万個、確認できているものだけで 56 万個にも及ぶ。さらに整備されていないものを含めるとタンパク質の数は増大する。この大量の解析データから新たな知見を見つけ出すためには生物学分野の専門知識だけではなく、情報処理の観点からのアプローチも必要である。プロテオーム解析においては質量分析法の中でも液体クロマトグラフィー・タンデム質量分析法(LC/MS/MS)と呼ばれるハイスループットな分析手法が用いられている。

このLC/MS/MS によって試料にどのようなペプチドが含まれているかを同定する。質量分析を用いた同定のフローを図 1 に示す



。図 1 タンパク質同定の一連の流れ
Fig.1: Flow of protein identification

1. タンパク質試料調製
2. タンパク質試料を消化酵素（トリプシン等）よって切断したのち、分離及びイオン化
3. 質量分析計を用いて導入されたイオンの m/z(質量電荷比) とその強度を計測
4. 計測した m/z とその強度によるマススペクトルにおいてピークとなる位置を決定
5. 計測したマススペクトルのピーク間の m/z の距離から当てはまるフラグメントイオンを同定しペプチド配列を決定
6. タンパク質(DB)4. で予測されたペプチド配列を比較
7. 5. で決定したペプチド配列と DB のタンパク質配列の固有部分が一致した場合、そのタン パク質が試料に含まれていると判定

目覚ましい発展を遂げている質量分析技術だが、未解決の課題がある。計測されたスペクトルにおいて全てのフラグメントイオンが検出されるとは 限らないこと、夾雑物の混入、アミノ酸の翻語修飾といった様々な要因により、一般的に質量分析計によって得られる大量のマススペクトルの半数は ペプチド配列の同定まで至らない。ペプチドの同定数を増やすことはタンパク質の同定精度の向上に繋がり、プロテオーム解析の典型的な利用目的 であるバイオマーカーや創薬標的の探索にも役立つと考えられる。[1]

実際にマススペクトルからペプチド配列の推定 するプロセスについて述べる。以下の図 2 のよう にアミノ酸と対応する質量をマススペクトルにおけるピークの m/z 間の差に一致するかどうかを照らし合わせる。これを繰り返すことによりスペクトル全アミノ酸配列の一次構造を決定することができる。

Proposal of peptide sequence identification method by deep learning
† Tachibana Isato,Kinki University Graduate School
‡ Nishigori Mitsuhiro,Fukuoka University
* Takahashi Atsushi, National Cerebral and Cardiovascular Center Hospital
§ Oboshi Naoki,Kinki University

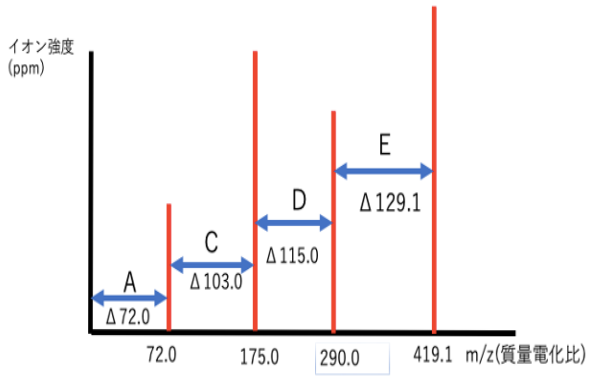


図 2 アミノ酸配列”ACDE”のマススペクトルの例
Fig. 2: Example of mass spectrum of amino acid sequence ”ACDE”

2. 目的

近年、大きな発展を遂げており、この技術を応用した研究が盛んになっている。 その一例としてペプチドの断片化スペクトルにおけるフラグメントイオンの強度予測や検出イオンの保持時間の予測に関して、大きな成果を上げている。これらの研究では、m/z 強度などの情報を利用することでペプチドの同定精度が向上する可能性が示唆されている。[2] そこで本研究では、質量分析によって得られる MS/MS スペクトルのピーク の m/z 値、および多くの既存手法では考慮していないピークの強度情報を入力とする深層学習により、ペプチド配列を同定する手法を提案する。

3. 提案手法

ProteomeXchange で提供されている m/z レンジ 375 から 1700 の質量分析データを学習データとし、マススペクトルのピークからペプチド配列を予測するモデルを作成する。学習のラベルとなる 20 種のアミノ酸によって構成されるペプチド配列のデータは、ワンホットエンコーディングによって 0/1 表現に変換したものを使用した。学習モデルとして、中間ノード数 1024 の全結合 4 層とドロップアウト層を組み合わせたものを利用した。

4. 結果

以下の図 3 に学習結果を示す。この結果は各種アミノ酸に分類する多クラス分類である。したがって適合率と再現率はクラス毎の数値を平均したマクロ平均の値である。
また、以下の図 4 に各ペプチド毎の分類結果をヒートマップにしたものを示す。この図からペプチドのロイシン（表記 L）とイソロイシン（表記 I）の誤分類が多いことが見受けられる。理由として、理論上得られる m/z がロイシンとイソロイシンで同じであり、この二種を分類することは他のペプチドと比べて困難であると考えられる。

精度(accuracy)	0.933
再現率	0.937
適合率	0.938
F1	0.933

図 3 学習結果
Fig. 3: result

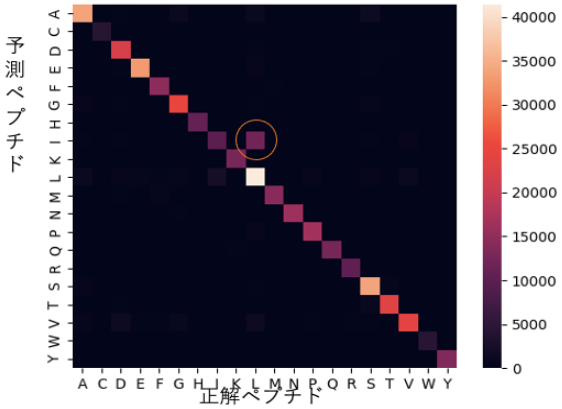


図 4 分類結果のヒートマップ
Fig. 4: Confusion matrix heatmap

5. まとめ

本研究では、深層学習を用いたペプチド配列の同定を行なった。今後の課題として今回学習に利用したデータはほぼ質量分析におけるマススペクトルの理論値に近いデータであり過学習の傾向がみられること、電荷や分析機器の機種などスペクトルのメタデータなども利用して予測を行えるように拡張していく必要があることが挙げられる。

参考文献

[1] Ruedi Aebersold and Matthias Mann. Massspectrometric exploration of proteome structurend function. nature, 2016.
[2] Yang Y. in silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat Commun, Vol. 11, No. 145, 2020.