G-24

深層学習を用いた質量分析におけるペプチド配列の同定手法の提案

Proposal of identification method of peptide sequence in mass spectrometry using deep learning

橘 勇人* 高橋 篤† 錦織 充広‡ 大星 直樹*
Isato Tachibana Atsushi Takahashi Mitsuhiro Nishigori Naoki Oboshi

1. はじめに

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するプロテオミクス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。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。

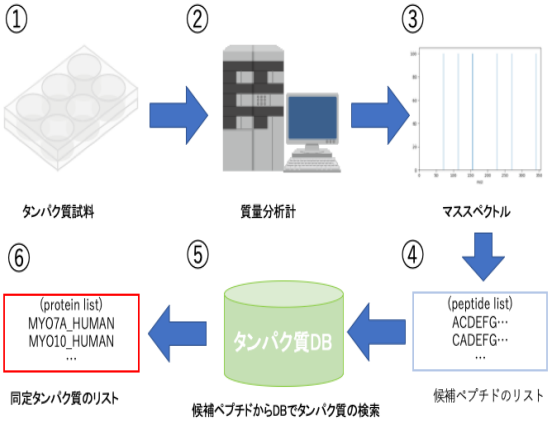

図 1　マススペクトル 取得の一連の流れ

1. タンパク質試料作成
2. タンパク質試料を消化酵素用によって切断したのち、イオン化及び測定
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 と 4.で予測されたペプチド配列を比較
7. 5.で比較したタンパク質と DB の配列が一定長の一致もしくは配列の固有部分が一致するタンパク質を試料に含まれているタンパク質と決定

*近畿大学大学院　総合理工学研究科　Kindai University Graduate School of Science of Engineering
† 国立循環器病研究センター　Faculty of Science and Engineering National Cerebral and Cardiovascular Center
‡ 福岡大学理学部化学科　Fukuoka University Faculty of Science Department of Chemistry

目覚ましい発展を遂げている質量分析技術だが、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、未解決の課題がある。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 解析の目的であるバイオマーカーや創薬指標の発見に役立つであろうと考えられる[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。以下の表 1 のようなアミノ酸と対応する質量を図 2 のようなマススペクトル のピークの 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。

アミノ酸	A(Alanine)	C(Cystein)	D(AsparticAcid)	E(GlutamicAcid)
質量	72.0	103.0	115	129

表 1　アミノ酸配列のその質量の対応表

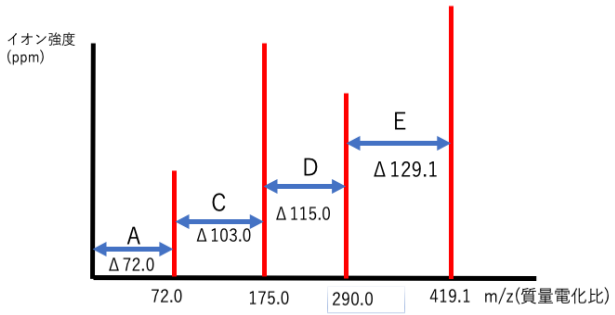

図 2　アミノ酸配列 ACDE を持つマススペクトル の例

3. 目的

近年、この分野においても大きな発展を遂げている深層学習を応用する研究が盛んになっている[2][3]。その一例として、長年の問題である計算が困難であったペプチドの断片化スペクトルに置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おいて、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[4]。

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。

4. 提案手法

本研究では、ProteomeXchange[6]で提供されているある m/z レンジ375~1700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。このデータが抱えている問題として、マススペクトルのピーク情報を表す m/z 及びその強度の組の個数と、ラベルとなるデータのペプチド配列はそれぞれ一定の長さではなく、固定長表現を扱う多くの機械学習アルゴリズムに適応することが出来ない問題を抱えている。したがって、深層学習の中でも RNN[7]と呼ばれる時系列データを扱えるように構築されたニューラルネットワークがある。これを利用し、学習を行う。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。

本研究で使用したネットワークモデルのアーキテクチャを図 3 に示す。

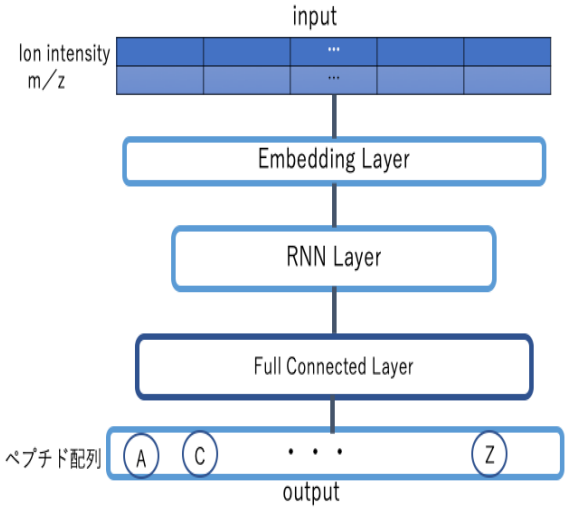

図 3　ネットワークモデルの構造

4.1　埋め込み層(Embedding Layer)

入力の m/z 及びその強度は離散かつ多次元であるのでそのままでは学習させることが非常に困難である。そこで、自然言語処理の分野で主に利用されている埋め込み層を用いて実数値の要素を持つ固定された大きさのベクトルにマッピングする。これによりマススペクトルを表現する特徴を学習し、次元の呪いの影響を抑制する。

4.2　RNN(Recurrent Neural Network)

RNNは一般的なニューラルネットを拡張し時系列データを扱えるようにしたものであり、通常のものとは異なる部分として、ある中間層での出力を次の時系列での中間層に伝える経路を持つように設計されたものである。

本研究で取り扱うマススペクトル、及びラベルであるペプチド配列データは可変長のシーケンスデータである。ペプチドのフラグメントイオンはマススペクトルそれぞれのピーク間m/zの差分に基づいて決定されるため、入力されたピーク情報全体を考慮する必要がある。これらの特徴をふまえ、学習モデルにRNNが適していると考え、このアーキテクチャを選択した。

4.3　全結合層(Full Connected Layer)

今までの層で計算された特徴量を一つのノードに結合し、活性化関数(ReLU)によって計算された値を出力する。最終的な出力結果としてアミノ酸配列のワンホット表現にデコードする役割を持つ。

5．検証

学習用のスペクトルデータ10,000個のうち全体の2割をテスト用データとして無作為に選び、検証を行った結果を表2に示す。

	適合率	再現率	F値
同定精度	71.6%	83.2%	76.9%

表 2　モデルの同定精度

本手法は各アミノ酸配列に分類する多クラス分類であるので適合率と再現率の評価指標はクラス毎の数値を平均したマクロ平均である。

6．まとめ

本研究では、深層学習の時系列データを扱えるように応用した RNN を用いて、質量分析で得られたマススペクトル からペプチド配列を同定する方法を提案した。本手法を翻訳語修飾に対して拡張する場合、学習データに含まれる修飾されたアミノ酸を 21,22 番目のアミノ酸配列として追加することで対応できると考えられる。一方、マススペクトル 計測機器や分析するタンパク質の試料等の要因を入力データに含めておらず、これらをメタデータとして学習モデルに組み込む必要があると考えられる。今後の課題として、本手法によって同定したペプチド配列を用いることによってタンパク質の同定精度に影響するかを検証する必要がある。また、今回学習に利用したデータはほぼ理論値に近いピークを持つデータであったため、実際の観測データに含まれるノイズや分析機器の違い等に対応できるようなモデルを模索する必要がある。

参考文献

[1] Aebersold, Ruedi and Mann, Matthias, “Mass-spectrometric exploration of proteome structure and function”, Nature, Vol.537, 347.EP (2016).

[2] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* 521, 436-444 (2015)

[3] Tiwary, S., Levy, R., Gutenbrunner, P. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat Methods* 16, 519-525 (2019)

[4] Yang, Y., Liu, X., Shen, C. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat Commun* 11, 146 (2020). https://doi.org/10.1038/s41467-019-13866-z

[5] Vizcaíno, J., Deutsch, E., Wang, R. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* (2014).

[6] Röst, H.L. Deep learning adds an extra dimension to peptide fragmentation. *Nat Methods* 16, 469-470 (2019).

[7] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, ″A Novel Connectionist System for Unconstrained Handwriting Recognition,″ in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 5, (2019)