様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 １０ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００３９０

研究課題名（和文）深層学習による質量ピーク探知法の開発

研究課題名（英文）Development of mass peak picking method by deep learning

研究代表者
吉沢　明康（Yoshizawa, Akiyasu）

京都大学・薬学研究科・特定助教

研究者番号：７０５５１１５９

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：　質量分析のデータ解析で必須のプロセスである「ピーク検出」を、ニューラルネットの深層学習に基づいて行う方法を開発した。教師用データとしては、ペプチド実測データから既存の手法で検出されたピークのうち、従来法によって高いスコアでペプチドが同定できたピークを採用した。我々の作成した判定器は現在までに、従来法と同等の検出性能を持ち、従来法では検出できなかったピークを少数ながら検出できるようになっている。

研究成果の学術的意義や社会的意義
質量分析はイオン化した試料の質量を測定する方法であるが、得られたマススペクトルから正確な質量の値を求めるには、「ピーク検出」が必須である。しかしこの過程には経験的な手法が用いられており、現状では全スペクトルの1/3程度しか同定できない。
そこで、最近大きなブレークスルーのあった、ニューラルネットの深層学習（deep learning）に基づいて、ピークを検出する新しい方法を開発した。機械学習研究での深層学習の応用例として、また今まで同定できなかったピークを効率的に検出するツールとしての発展と、それを利用した効率的な生命科学研究や医療技術への応用が期待できる。

研究成果の概要（英文）：We developed a novel method of “peak picking,” an indispensable step in mass spectrometry data analysis, by deep learning in neural network. Mass peaks used for training data were those that were picked and identified as peptides with high scores by conventional methods. The current implementation of our classifier has almost equal detection performance as the conventional methods, and can also detect mass peaks that could not be detected by conventional methods.

研究分野：プロテオミクスに於けるバイオインフォマティクス方法論

キーワード：バイオインフォマティクス　質量分析　機械学習　深層学習　プロテオミクス

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

質量分析(MS)を用いたプロテオーム解析(LC/MS/MS 解析)では、高速液体クロマトグラフィー(HPLC)及び質量分析計を用い、それぞれ、（通常）疎水性と質量という二つの基準によって対象分子を分離している。特にプロテオーム試料の場合、試料であるタンパク質（混合物）を消化酵素によってペプチドに切断、これを HPLC で分離して質量分析計に導入するが、試料ペプチドをイオン化した後、マスアナライザーによって 2 度の分離が行われる。

1 回目の分離 (MS^1) では特定の *m/z*（ペプチドの質量をイオンの電荷数で割った値に等しい）のペプチドが分離される。このペプチドが断片化（1 カ所で切断）され、2 回目の分離 (MS^2 または MS/MS）でペプチド断片（開裂部位が異なる、種々のペプチド断片）の *m/z* 情報が得られる。

最終的に、MS^1 から得られるペプチド全体の質量と電荷数、及び MS^2 から得られるペプチド断片の *m/z* のリスト（ピークリスト）と、タンパク質配列データベースから得られる理論 *m/z* を突き合わせて、最も可能性の高いペプチドを推定する（データベース検索法）。

しかしながら現時点では、この手法を用いても一般に MS^1 で選択したピークの 1/3 程度しかペプチドが同定されていないと言われている。この原因の一つとして考えられることが、ペプチドの精密質量と電荷数推定の精度である。取得されたマススペクトルには、タンパク質由来のペプチドイオンの他、夾雑物（ゴミ）由来のピークやノイズなどが混ざるが、天然物であるペプチドのピークには、ペプチド内の原子が同位体（主に天然存在比 1.07%の 13C と 0.36%の 15N）であった場合、それに由来するピーク（同位体ピーク）が付随して現れる。

同位体は中性子の個数が異なるため、ピークは中性子の質量をイオンの電荷数で割った値に相当する間隔ごとに出現し、従ってこの同位体ピークの間隔からペプチドイオンの電荷数を推定することが可能である。また重い同位体を含まない、最も質量の小さいピーク（モノアイソトピック・イオン・ピーク）から、精密質量を指定することが可能である。しかし実際には、ほぼ同一の質量・同一の電荷を持ったペプチドイオンの重複や、ノイズ、イオン化効率が低いペプチドのシグナル強度がノイズ以下になることなどの影響を受け、正しい電荷数と精密質量を推定することは難しい。

２．研究の目的

一般に用いられている conventional な手法では、同位体ピークの出現規則から組み立てられたアルゴリズムに基づいて、ペプチドイオンの精密質量と電荷数を予測している。しかし MS^1 の出力であるマススペクトルは、 *m/z* と retention time の 2 次元から成るベクトル（行列）の形で、signal intensity を要素の値として表現することが可能である。そこで本研究では、このベクトルを入力データとして用いることによって、深層学習を用いたペプチドイオンの精密質量と電荷数を予測する手法を開発した。

３．研究の方法

(1) 教師 (training) データ

教師データとして、正しい精密質量と電荷数の判明しているスペクトルのデータが必要であるが、これを大量に集めることは困難である。しかし、質量分析を用いたプロテオーム解析の実データの中で、ペプチド同定のスコアが高いものを教師データとして利用できることが確認できたため、この手法によって教師データを収集した。即ち、高分解能の質量分析計から得られた生データから、既存ソフトを用いてピークの精密質量と電荷数を推定し、同じく既存データベース検索ソフトを用いてタンパク質配列データベースを検索した結果、ペプチド同定の信頼性の指標の一つである peptide expectation value が 0.001 以下となる、極めて信頼性の高いものから、約 10 万ピークを選択し教師データとした。この際、後述する教師データの分類ラベルの頻度が大きく偏らないように収集を行った。

(2) 入力データ

MS^1 スペクトルはペプチド混合物全体のスペクトルであり、複数のペプチドに由来するピークが混在していて、スペクトル全体をそのまま training に用いるのは適切ではない。

そこで、マススペクトル及びクロマトグラム (XIC) のデータから、測定時に質量分析計が「MS/MS を測定するピーク」と自動判定したピークの前後数 *m/z* を含むマススペクトル、及び Retention time 方向に前後数 Scan 相当の範囲を抜き出し、最大 Intensity で正規化した 2 次元ベクトルを、学習及び予測の入力データとした。

(3) 分類ラベル

電荷数は整数値であるため、そのまま教師データの分類ラベルとして用いた。また、精密質量は連続値であり分類ラベルとして適切でないため、精密質量ピークの、基準となるピーク（質量分析計が検出したピーク）からの、同位体ピーク上での位置に変換し、整数値に置き換えることで分類ラベルとした。

(4) ネットワークモデル

畳み込み層 2 層、全結合層 2 層からなる畳み込みニューラルネットワークを、深層学習のネットワークモデルとして採用した。深層学習のフレームワークには Chainer（ver. 5.4.0）を用いた。精密質量と電荷数を分類するための学習モデルをそれぞれ作成し、電荷数の予測と精密質量の予測を 2 段階で行った。

４．研究成果

(1)概要

作成した教師データを用い学習を行った結果、予測精度（training:test=8:2）は電荷数の予測で 99%以上、精密質量の予測では 92%以上であった。

(2)性能評価

教師データ作成時に用いた既存のピークピッキング・ソフトウェア 2 つとの比較を行った。学習に用いたものとは別の 4 つの MS サンプルに対して、本手法と既存ソフトウェア 2 つを用いて電荷数と精密質量の予測をそれぞれ行い、ペプチド同定を行った。同定結果中で、peptide expectation value が 0.01 以下になる PSM(Peptide Spectrum Match)の数を、ペプチド同定に至ったピークの数とみなして比較した。

表１：ペプチド同定に至ったピークの数（PSM 数）

	サンプル a	サンプル b	サンプル c	サンプル d
本手法	1,990	4,065	5,576	3,174
既存手法 A	2,005	4,099	5,699	3,217
既存手法 B	1,758	3,620	5,479	3,093

ペプチド同定に至ったピークの数を表１に示した。本手法の同定数は既存手法 B より多く、既存手法 A より少ないという結果になった。既存手法 B はプロテオーム解析でよく用いられる手法であるが、本手法はこれより１割以上多くペプチドを同定できた。また、既存手法 A との差は僅差であった。学習に用いた教師データとしては実際のプロテオームデータを用いたため、分類ラベルの頻度が一定ではなく、頻度の少ない分類ラベルを学習しきれていない可能性があり、このことが既存手法 A に僅かに性能が劣った要因の一つと考えられる。

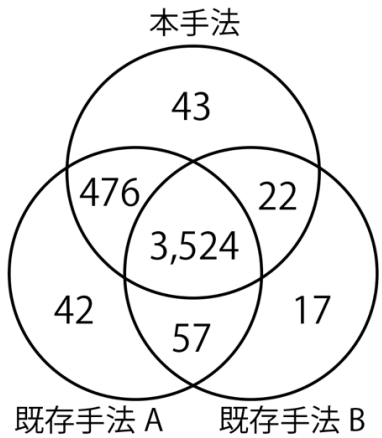

図１：サンプル b における PSM 数の内訳

図 1 に示したのは、サンプル b の場合の PSM 数の内訳である。既存手法 2 つは同様の同位体ピークの出現規則に基づいた予測を行っており、また本手法はこれら 2 つの手法を教師として学習を行っているため、大部分のピークは同一のペプチド同定に至っている。

次いで、本手法と既存手法 A のみで共通して同定されているペプチドが多い。これは、既存手法 B には存在せず既存手法 A のみに存在している予測アルゴリズムの特徴を、本手法で学習できていることが示唆される。また、少数ながらも本手法のみで同定に至ったピークも見られ、本手法をプロテオーム解析に用いることで、これまでの手法を用いた解析では取りこぼしていたピークも拾うことが可能となったといえる。

なお、本研究成果は既に複数の学会で発表しているが、本年度中を目処に論文発表と、一般的に利用可能なツールとしての実装を計画している。

５．主な発表論文等

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　2　件）
①守屋勇樹，田畑剛，岩崎未央，河野信，五斗進，石濱泰，瀧川一学，吉沢明康，深層学習に基づくペプチド由来イオンピークの新規検出手法，第 67 回質量分析総合討論会（2019 年 5 月 17 日，つくば），[3D-O1] オープンデータからデータサイエンスへ，3D-O1-1030
②守屋勇樹，田畑剛，岩崎未央，河野信，五斗進，石濱泰，瀧川一学，吉沢明康，機械学習に基づくペプチド由来イオンピークの新規検出手法，第 21 回情報論的学習理論ワークショップ (IBIS2018)，（2018 年 11 月 5-7 日，札幌），D1-53

〔図書〕（計　1　件）
①吉沢明康，第 10 章　プロテオーム解析，講談社，よくわかるバイオインフォマティクス入門（藤博幸、岩部直之、川端猛、浜田道昭、門田幸二、須山幹太、光山統泰、黒川顕、森宙史、東光一、吉沢明康、片山俊明　著），2018 年，p.139-155

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕
ホームページ等　（現時点ではなし）

６．研究組織

(1)研究分担者
研究分担者氏名：守屋　勇樹
ローマ字氏名：MORIYA, Yuki
所属研究機関名：情報・システム研究機構
部局名：データサイエンス共同利用基盤施設
職名：特任助教
研究者番号（8 桁）：40773841

(2)研究協力者
研究協力者氏名：田畑　剛
ローマ字氏名：TABATA, Tsuyoshi

研究協力者氏名：岩崎　未央
ローマ字氏名：IWASAKI, Mio

研究協力者氏名：河野　信
ローマ字氏名：KAWANO, Shin

研究協力者氏名：五斗　進
ローマ字氏名：GOTO, Susumu

研究協力者氏名：石濱　泰
ローマ字氏名：ISHIHAMA, Yasushi

研究協力者氏名：瀧川　一学
ローマ字氏名：TAKIGAWA, Ichigaku