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　５ 年　 ６ 月 １９ 日現在

機関番号：３３２０２
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０３２４５

研究課題名（和文）質量分析によるアミノ酸配列de novo決定のための新規手法開発

研究課題名（英文）Development of a novel method for de novo sequencing of amino acid sequences by mass spectrometry

研究代表者
河野　信（Kawano, Shin）

富山国際大学・現代社会学部・教授

研究者番号：４０４７００７５

交付決定額（研究期間全体）：（直接経費）　　13,600,000円

研究成果の概要（和文）：本研究では、アミノ酸配列を既知配列を用いずに質量分析データのみによって決定する新規手法を開発した。一般に既知配列情報を用いない同定法では、探索空間が膨大なため最適解が得にくい。そこでアミノ酸配列の物性情報も活用することで大幅に探索空間を限定し、現実的な計算時間で計算が完了するようにした。このための物性情報の調査を行い、アミノ酸組成と高速液体クロマトグラフィーの保持時間の活用を決定した。そして、ベンチマークの結果から、結果の候補配列を構成するときにペプチド配列の部分アミノ酸配列情報をまず決めるという戦略が有効であると結論付け、これらの要素を用いた新規手法を開発した。

研究成果の学術的意義や社会的意義
現在、疾患研究を含む医学研究ではプロテオーム解析の重要性が高まっている。しかしプロテオーム・データの解析ではゲノム情報に基づく既知配列しか探知できず、測定データの3割程度しか同定ができていない。癌のように変異が多い場合には既知配列が利用できないが、このような場合に用いる特殊な同定方法では、既存の代表的なソフトウェアを用いても正解率は4割程度である。本研究では、今まで用いられてこなかった情報を補助的に用いることで正解率を上げる手法を開発しており、今後の疾患タンパク質研究や抗体医薬などの応用研究への進展が期待される。

研究成果の概要（英文）：In this study, we developed a novel method to determine amino acid sequences using only mass spectrometry data without known sequences. In general, the identification method that does not use known sequence information has a huge search space, making it difficult to obtain an optimal solution. Thus, by utilizing the physical property information of the amino acid sequence, the search space is greatly limited, and the calculation can be completed in a realistic calculation time. We investigated the physical property information of amino acid sequence and decided to utilize the amino acid composition and the retention time of high-performance liquid chromatography. Based on the benchmark results, we concluded that the strategy of first determining the partial amino acid sequence information of the peptide sequence when constructing the resulting candidate sequence is effective. We developed a novel method using these elements.

研究分野：バイオインフォマティクス

キーワード：タンパク質　質量分析　アミノ酸配列決定　アミノ酸組成分析　de novo sequencing

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

質量分析はタンパク質を対象とした分析化学の強力な手段であり、近年では ”Cancer Moonshot” と総称される一連の癌化機構研究など、医学研究でのプロテオーム解析の劇的増加に伴って、プロテオームデータ解析の需要が急速に高まっている。しかし、現在の配列同定手法は、実験的に得られたマススペクトルから抽出したピークと既知配列データベースから作成した理論ピークを照合してペプチド配列を推定するデータベース検索法が主流であり、既知のペプチド配列しか同定できない。したがって、疾患細胞研究などで重要な変異プロテオームの解析にはゲノムの決定が必須である。また現状では、測定されたマススペクトルのうち、ペプチド同定に至る割合は約 3 割程度であり、多くのマススペクトルはペプチドに帰属できていないという問題がある（Nat Methods, 2016, Gris J. et al.）。さらに近年は、癌治療などのために抗体医薬がホットな開発テーマとなっているが、抗体可変領域の配列情報はゲノムにコードされていない。このため通常は、非常に高コストなファージディスプレイ法を用いた大規模スクリーニングが必要になっている。

このような問題を解決するには、既知配列データベースを用いず、質量分析データのみからアミノ酸配列を決定する「de novo sequencing（de novo 決定）」を行えばよい。しかし現状の de novo 決定には以下のような大きな問題がある。(a) MS/MS 測定によって「対象ペプチドのいずれかの末端を含む部分配列」が得られるが、de novo 決定にはこれら部分配列の全種類に対応するマスピークを取得する必要がある。しかしアミノ酸配列によってイオン化効率は一定ではなく、すべての部分配列のマスピークは測定できない可能性が高い。またノイズピークの存在も計算量を増加させる。(b) 計算によって最適な配列を推定するためには、可能性のあるアミノ酸配列の全て（全長が N 残基の場合 20^N: アミノ酸 20 種類の残基数(N)乗）を計算する必要がある。しかしこのような検証は場合の数が指数関数的に増加するため、実行が困難である。配列の類似性検索を行う Smith-Waterman アルゴリズムなどでは動的計画法を用いることで場合の数の“枝刈り”を行っているが、その場合の置換行列に相当する適切な評価関数が質量分析データの場合には存在しない。このため動的計画法を利用すると不完全な評価関数を用いることになり、部分問題を解いている間に最適解が捨てられてしまうことが多い。de novo 決定ソフトウェアは複数存在しているが、前述の問題 (a), (b) のうち、主に (b) 評価関数の問題のみに対応しようとしているものが多い。しかしこのような MS/MS スペクトル情報（マスピークの間隔）のみに基づいた処理の場合、代表的な商用ソフトウェアであるカナダ Bioinformatics Solutions 社の PEAKS ソフトウェアの場合でも、予備的研究でのベンチマークでは配列既知ペプチドの正解率は 40%程度であり、正解率を改善するような手法の開発が求められている。

２．研究の目的

本研究では、アミノ酸配列を質量分析によって決定する新規 de novo 決定法を開発する。de novo 決定法は一般的なデータベース検索法と異なりゲノム情報が不要で、アミノ酸を一残基ずつ確定するため精度が高いが、探索空間が膨大なため最適解が得にくい。そこで、ペプチドのマススペクトルに加えて、アミノ酸組成などの物性情報も実験的に取得し併用することで大幅に探索空間を限定し、現実的な計算時間で de novo 決定を行う手法を開発する。「計算すべき場合の数の全数を削減する」という目標のために追加の実験系を開発し、今まで de novo 決定で用いられていない種類の情報を取得し酵素消化ペプチドの質量分析測定と組み合わせることによって、動的計画法を使わずにすべての組み合わせを数え上げる。例えばアミノ酸 7 個の配列を従来法で決定すると、検討すべき場合の数の総数は $20^7 = 12.8$ 億になるが、7 個のアミノ酸組成が判明していれば、場合の数は最大で $7! = 5040$ に過ぎない。上記の PEAKS のベンチマークと同時に行った予備的研究ではアミノ酸組成情報のみを追加したが、同一試料に対する正解率は約 40%から約 80%に上昇した。このアプローチは新規であり、情報解析の内容を念頭に置いて実験系を開発することによって、実験系と情報処理系が相互に弱点を補完し合うことを可能にしている。

３．研究の方法

・　ペプチド物性情報の取得手法確立

ペプチド試料の物性情報で、de novo 決定に利用するのに適切なもの及びそれを取得する最も簡便な方法について検討を行った。予備的研究において有効性を確認済みであるアミノ酸組成情報に加えて、既存の手法で取り入れられていない情報を中心に、等電点・酵素切断部位に当たるアミノ酸の個数・液体クロマトグラフィーの保持時間などの活用を検討した。また、MS/MS 測定で確認できるイミニウム(インモニウム)イオンをアミノ酸組成決定に利用可能か検討した。まず、可能なアミノ酸配列の一時的データベースを用いたデータベース検索法を利用することで作業を進めた。また、従来法によるアミノ酸組成情報を利用した方法についても検討を行った。

・ ペプチド配列からタンパク質配列を再構成する手法の開発

上記のような de novo 決定が誤りを含む可能性に対応するため、複数の手法で酵素消化を実施することで、切断部位が異なるより短いペプチド断片を生成した。それぞれについて de novo 決定を行い、その結果から元のタンパク質配列を再構成した。また再構成では多数決原理を用いてより信頼性の高い配列を採用するプロセスを実装した。

４．研究成果

1. まず、アミノ酸配列の絞り込み（枝刈り）を行うのに有効と思われる物性情報の調査を行い、

・ de novo sequencing を行うための先行ソフトウェアである PEAKS(商用ソフト)を用いたベンチマーク実験
・ 仮想配列のデータベースを用いた de novo 決定法のシミュレーション
・ 複数の合成ペプチドの複数酵素による消化物の LC/MS による測定

を実施した。これらの結果から、アミノ酸配列の絞り込み（枝刈り）を行うのに有効と思われる物性情報として、アミノ酸組成に加え、高速液体クロマトグラフィーの保持時間を利用できる、と結論した。

2. また先行ソフトウェアである DeepRT(+)のベンチマークの結果、保持時間予測が利用可能な充分な精度を持つと結論した。したがって物性情報としては、候補配列を構成する段階でのアミノ酸の組成に加えて、本手法によって得られた候補配列を絞り込むという目的で予測保持時間を利用する。
3. 次に PEAKS を用いたベンチマークの結果から、候補配列の構成においてアミノ酸配列の seed をまず決める、という戦略が有効であると結論した。そこでこの seed の決定用に、長さ 3 個のアミノ酸配列（配列タグ）を MS/MS スペクトルから機械学習で同定する手法を開発した。この段階ではトレーニングデータとして、プロテオーム統合データベース jPOST において再解析済みのデータから 3,291,902 個の MS/MS スペクトルを取り出し、そこから抽出された 624,357 種類の配列タグ情報を用いた。
4. さらに、ペプチド配列からタンパク質配列を再構成する手法の開発については、塩基配列の de novo アセンブルを参考に、同定できたタグ・ペプチドを重ね合わせ接続するためのプログラムを試作した。
5. このほか、本研究グループ外との共同開発によって、本手法実装公開用のオープンソース・ソフトウェア・プラットフォーム（Mass++ ver.4）の開発を行い、プラットフォーム部分のベータ版を公開した。

上記のように方法論の骨格を為す要素技術の開発は完了したが、実際の試料に対して本方法論を試行する段階は本基盤研究期間中には終了しなかったため、各要素技術の更なる改善を行いつつ、今後の発展研究に展開していく。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　4件／うちオープンアクセス　2件）

１．著者名 Deutsch Eric W.、Vizcaino Juan Antonio、Jones Andrew R.、Binz Pierre-Alain、Lam Henry、Klein Joshua、Bittremieux Wout、Perez-Riverol Yasset、Tabb David L.、Walzer Mathias、Ricard-Blum Sylvie、Hermjakob Henning、Neumann Steffen、Mak Tytus D.、Kawano Shin、et al.	４．巻 22
２．論文標題 Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work	５．発行年 2023年
３．雑誌名 Journal of Proteome Research	６．最初と最後の頁 287～301
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acs.jproteome.2c00637	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Deutsch Eric W、Bandeira Nuno、Perez-Riverol Yasset、Sharma Vagisha、Carver Jeremy?J、Mendoza Luis、Kundu Deepti J、Wang Shengbo、Bandla Chakradhar、Kamatchinathan Selvakumar、Hewapathirana Suresh、Pullman Benjamin、Wertz Julie、Sun Zhi、Kawano Shin、et al.	４．巻 51
２．論文標題 The ProteomeXchange consortium at 10 years: 2023 update	５．発行年 2022年
３．雑誌名 Nucleic Acids Research	６．最初と最後の頁 D1539～D1548
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nar/gkac1040	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 LeDuc Richard D.、Deutsch Eric W.、Binz Pierre-Alain、Fellers Ryan T.、Cesnik Anthony J.、Klein Joshua A.、Van Den Bossche Tim、Gabriels Ralf、Yalavarthi Arshika、Perez-Riverol Yasset、Carver Jeremy、Bittremieux Wout、Kawano Shin、Pullman Benjamin、Bandeira Nuno、Kelleher Neil L.、Thomas Paul M.、Vizca?no Juan Antonio	４．巻 21
２．論文標題 Proteomics Standards Initiative’s ProForma 2.0: Unifying the Encoding of Proteoforms and Peptidoforms	５．発行年 2022年
３．雑誌名 Journal of Proteome Research	６．最初と最後の頁 1189～1195
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acs.jproteome.1c00771	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Deutsch Eric W.、Perez-Riverol Yasset、Carver Jeremy、Kawano Shin、Mendoza Luis、Van Den Bossche Tim、Gabriels Ralf、Binz Pierre-Alain、Pullman Benjamin、Sun Zhi、Shofstahl Jim、Bittremieux Wout、Mak Tytus D.、Klein Joshua、Zhu Yunping、Lam Henry、Vizca?no Juan Antonio、Bandeira Nuno	4．巻 18
2．論文標題 Universal Spectrum Identifier for mass spectra	5．発行年 2021年
3．雑誌名 Nature Methods	6．最初と最後の頁 768～770
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41592-021-01184-6	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計11件（うち招待講演　4件／うち国際学会　5件）

1．発表者名 Satoshi Tanaka, Masaki Murase, Masaki Kato, Hiroyuki Yamamoto, Tsuyoshi Tabata, Maiko Kusano, Shin Kawano, Susumu Goto, Yasushi Ishihama, Akiyasu C. Yoshizawa
2．発表標題 Mass++ ver.4 -An open-source MS data viewer with enhanced basic functions and easy implementation of external software-
3．学会等名 71st ASMS Conference on Mass Spectrometry and Allied Topics（国際学会）
4．発表年 2023年

1．発表者名 Henry Lam1, Tytus D. Mak, Joshua A. Klein, Wout Bittremieux, Ralf Gabriels, Yasset Perez-Riverol, Tim Van Den Bossche, Andrew R Jones, Pierre-Alain Binz, Shin Kawano, et al.
2．発表標題 Proteomics Standards Initiative (PSI) proposed peak annotation format (mzPAF) and spectral library format (mzSpecLib) standards
3．学会等名 71st ASMS Conference on Mass Spectrometry and Allied Topics（国際学会）
4．発表年 2023年

1．発表者名 河野信
2．発表標題 プロテオームデータの収集と標準化
3．学会等名 疾患プロテオミクス研究会（招待講演）
4．発表年 2023年

１．発表者名
Satoshi Tanaka, Masaki Murase, Masaki Kato, Hiroyuki Yamamoto, Tsuyoshi Tabata, Maiko Kusano, Shin Kawano, Susumu Goto, Yasushi Ishihama, Akiyasu C. Yoshizawa

２．発表標題
Mass++ ver.4 -MS data viewer meets online databases-

３．学会等名
The HUPO 2022 Congress（国際学会）

４．発表年
2022年

１．発表者名
河野信

２．発表標題
jPOSTの現状とプロテミクスデータの標準化に向けた国際連携

３．学会等名
日本プロテオーム学会2021年大会（招待講演）

４．発表年
2021年

１．発表者名
Tim Van Den Bossche, Eric W. Deutsch, Yasset Perez-Riverol, Jeremy Carver, Shin Kawano, Luis Mendoza, Ralf Gabriels, Pierre-Alain Binz, Benjamin Pullman, Zhi Sun, Jim Shofstahl, Wout Bittremieux, Tytus D. Mak, Joshua Klein, Yunping Zhu, Henry Lam, Juan Antonio Vizcaino, and Nuno Bandeira

２．発表標題
The HUPO-PSI Universal Spectrum Identifier (USI) for mass spectra

３．学会等名
HUPO reconnect 2021（国際学会）

４．発表年
2021年

１．発表者名
吉沢明康

２．発表標題
マススペクトルに於けるペプチドイオンピークの深層学習による検出法の開発

３．学会等名
第17回日本臨床プロテオゲノミクス研究会（招待講演）

４．発表年
2021年

１．発表者名
吉沢明康，守屋勇樹，小林大樹，張智翔，奥田修二郎，田畑剛，河野信，幡野敦，高見知代，松本雅記，山ノ内祥訓，荒木令江，岩崎未央，杉山直幸，福島敦史，田中聡，五斗進，石濱 泰

２．発表標題
jPOSTdb: COVID-19データベースの構築

３．学会等名
トーゴーの日シンポジウム2021

４．発表年
2021年

１．発表者名
有馬佳奈美、岡本瑠璃、小林大樹、吉沢明康、河野信

２．発表標題
jPOSTrepoメタデータのSDRF化

３．学会等名
トーゴーの日シンポジウム2021

４．発表年
2021年

１．発表者名
Tanaka, S., Murase, M., Kato, M., Tabata, T., Kusano, M., Kawano, S., Goto, S., Ishihama, Y., Yoshizawa, A.C.

２．発表標題
An extension of Mass++ ver.4, a data viewer, for proteome analysis

３．学会等名
ASMS 2020 Reboot（国際学会）

４．発表年
2020年

１．発表者名
吉沢 明康

２．発表標題
質量分析インフォマティクスの世界：どこから来て、何者で、どこに向かうのか

３．学会等名
第9回生命医薬情報学連合大会（招待講演）

４．発表年
2020年

〔図書〕　計1件

1．著者名 日本遺伝学会	4．発行年 2022年
2．出版社 丸善出版	5．総ページ数 690
3．書名 遺伝学の百科事典	

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	岩崎　未央 (Iwasaki Mio) (10722811)	京都大学・iPS細胞研究所・講師 (14301)	
研究分担者	小林　大樹 (Kobayashi Daiki) (20448517)	新潟大学・医歯学系・助教 (13101)	
研究分担者	吉沢　明康 (Yoshizawa Akiyasu) (70551159)	京都大学・薬学研究科・特定助教 (14301)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
英国	European Bioinformatics Institute			
米国	Institute for Systems Biology	University of California San Diego		