様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５月１０日現在

機関番号：１５４０１
研究種目：基盤研究(C)
研究期間：２００８　～　２０１０
課題番号：２０５００２７１
研究課題名（和文）質量分析プロテオミクスにおける網羅的スプライシング部位解析法の開発

研究課題名（英文）Development of a comprehensive splice-site analysis using mass spectrometry-based proteomics

研究代表者
石野　洋子（ISHINO YOKO）
広島大学・産学・地域連携センター・研究員
研究者番号：９０３７３２６６

研究成果の概要（和文）：
発現しているアミノ酸の部分配列を質量分析プロテオームデータから直接同定する方法（*de novo* 法）に、新たにペプチドごとの物理化学的特性の情報を取り入れた探索を組み合わせることで、網羅的に選択的スプライシング部位を検出する新規な方法を考案することを目指した。しかし、実データによる検証では、あまり良い結果が得られなかったため、原因のひとつと考えられる質量分析データの精度を改善することを検討した。その結果、測定後に計算によってデータを較正する方法を新たに開発し、分裂酵母のプロテオームデータで有効性を確認した。

研究成果の概要（英文）：
This study aimed to establish a method of accurately and comprehensively detecting the splice donor and acceptor sites of alternative splicing. The method is based on a *de novo* sequencing method, which directly identifies partial amino acid sequences of expressed proteins from the mass spectrometry-based proteome data. In addition, predicting the ease of peptide detection in proteomics using theoretical physicochemical properties of peptides is newly incorporated into the method in order to improve the accuracy of the splice site detection. Although the proposed method was evaluated with proteomics experimental data, it produced poor results. However, this led to the study of improving the mass accuracy of the mass spectrometry-based proteome data. A computational *a posteriori* calibration method was newly proposed, and based on the proteome data of fission yeast it was confirmed that the developed method increased mass accuracy.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,300,000	390,000	1,690,000
２００９年度	1,400,000	420,000	1,820,000
２０１０年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：　　バイオインフォマティクス
科研費の分科・細目：情報学・生体生命情報学
キーワード：プロテオーム情報処理、機械学習、バイオインフォマティクス

１．研究開始当初の背景

近年、ソフトイオン化質量分析技術と情報技術の一分野であるデータベース検索技術とを組み合わせることで、特定の細胞や組織で発現している数千種類ものタンパク質を網羅的に同定することが可能となった。一般的なプロテオーム解析では、測定されたペプチド断片の質量の組み合わせをもとに、データベースにあらかじめ格納されている遺伝子群のアミノ酸配列に対して確率統計的な計算を行い、発現タンパク質を同定する。このとき、遺伝子のアミノ酸配列はあらかじめ一意に決定されている。この方法では、遺伝子注釈に誤りが含まれている場合は同定精度が下がるという欠点を持ち、また、注釈されていない遺伝子を同定することもできない。そのため，最近では、ゲノムの DNA 配列上に存在する可能な全ての読み枠（Open Reading Frame）を抽出してアミノ酸配列に翻訳し、そのデータに対して検索を行うという試みがなされている（Mann M., Pandey A. Trends Biochem Sci. 2001, 26, 54-61.）。しかし、この方法だけでは、選択的スプライシング検出の核となるスプライス部位の正確な予測・検出はできない。

質量分析を用いたプロテオミクスのデータからスプライス部位の予測を行うことが理論的に可能かどうかを試みた先行研究としては、Chen らの研究（Chen T. Proc. of the Fifth Annual International Conference on Computational Biology; ACM., 2001, 87-94.）や Colinge らの研究（Colinge J., Cusin I., et al. J Proteome Res. 2005, 4, 167-174.）がある。両者の研究はともに、既存の“発現ペプチド同定検索法”と“スプライス部位予測法（マルコフモデルに代表される）”とをタンデムにつなげたものである。この方法では、アミノ酸配列をその物理化学的な性質は無視して単なる文字列として扱い、すでに明らかになっているスプライス部位配列の特徴を学習してスプライス部位を予測する。そのため、結果は学習データに大きく依存してしまい、出現頻度の低い特徴は正しく検出できないという欠点がある。

また、そもそもプロテオミクスのデータの精度が発現ペプチドの同定精度に大きな影響を及ぼすということも無視できない。タンパク質の質量分析を行う方法や機器には多種類のものが存在するが、特にハイスループットなプロテオーム解析にふさわしいといわれているものに、低速の液体クロマトグラフィーを質量分析の前段に加えたエレクトロスプレーイオン化・飛行時間型タンデム質量分析（LC/ESI-TOF MS/MS）がある。LC/ESI-TOF MS/MS では、通常、タンパク試料を測定する前に Glu1-Fibrinopeptide B などの外部標準物質を用いてキャリブレーションを行うが、ハイスループットなプロテオーム測定には、それだけでは不十分である。なぜなら飛行時間型の質量分析器では、測定中の微妙な温度変化が飛行チューブ長に影響し測定誤差を生むことが知られており、ハイスループットなプロテオミクスでは測定が何十時間も連続して行われるのでその影響が無視できないほど大きくなるからである。

２．研究の目的

選択的スプライシングとは、一つの遺伝子から性質の異なる複数のタンパク質（アイソフォーム）を生成する合理的かつ効率的な遺伝子発現調節機構であり、高等生物ではこの選択的スプライシングが高頻度で起きている。発生の段階に応じて、また細胞組織の環境の違いに応じて、状況に適したアイソフォームを発現させることで高等生物は多様な生命活動を実現している。生体が正常に機能していくためには、選択的スプライシングが適切に行われる必要があり、その異常は時として疾患を引き起こす。このように選択的スプライシングは生体の機能維持に大変重要であることから、その網羅的解析は益々重要視されている。

しかしながら、現在のプロテオミクスで一般的に用いられる検索アルゴリズムでは、発現遺伝子の同定は出来ても、選択的スプライシングの正確な把握までは不可能である。なぜなら、現行の検索アルゴリズムは、データベース中の“静的”遺伝子配列のみを検索する目的で構築されており、選択的スプライシングが起きている“動的”遺伝子配列を検索しようとするとスプライス部位の可能な組み合わせを全てデータベースに格納してから検索しなければならず、計算量が膨大になり解析が困難となるからである。

そこで本研究では、選択的スプライシング部位の精度の良い予測・検出が可能な新規アルゴリズムを考案することを目的とする。また、そのためには、質量分析プロテオームデータの精度が高いことが必須であるため、測定後に計算で誤差を較正する新たな方法を開発することも目的とする。

３．研究の方法

まず、ペプチドごとの物理化学的特性値を利用した質量分析での検出されやすさの指標情報と、*de novo* シーケンス法による発現部分ペプチド鎖探索とを組み合わせることで、スプライス部位にまたがるペプチドを正確に検出する仕組みの理論構築を試みる。続いて、様々なデータで提案手法の有効性を検証する。最終的には、これまで明らかになって

いない選択的スプライシングの発見をねらう。全体の研究計画としては、(1) スプライス部位にまたがるペプチド同定の理論構築、(2) 人工的なテストデータでの有効性の検証、(3) 実験データでの有効性の検証、(4) 魚類の発生プロテオームデータを質量分析で取得、(5) (4) で起こっている選択的スプライシングの発見・検証、(6) 研究のとりまとめ、と段階を踏んで進める。

なお、質量分析によるタンパク質の同定では、質量分析機器自体の性質も研究に大きな影響を及ぼす。本研究では、連続してハイスループットな測定を行うことができるためプロテオームデータの測定に適しているエレクトロスプレーイオン化・飛行時間型タンデム質量分析（LC/ESI-TOF MS/MS）で取得したデータを解析対象とする。この装置で取得したデータの精度を高めるため、測定後のデータ較正法を新たに提案し、LC/ESI-TOF MS/MS で取得した実データを用いて検証する。

４．研究成果

まず、スプライス部位にまたがるペプチド同定の理論構築の研究を行った。発現ペプチドの検索に *de novo* シーケンス法を用いると、データベース中のデータに依存せずに発現タンパク質の断片的なアミノ酸の部分配列情報を直接明らかすることができる。そこで、(1) 発現ペプチドの事前探索としてゲノム DNA 配列に対して一般的なデータベース検索を行う、(2) 発現が認められたペプチドが複数個存在するゲノム上のエリアを抽出する、(3) そのエリアに対して、ペプチドとしての検出されやすさの情報と *de novo* で実際に検出されたアミノ酸の部分配列情報をもとにスプライス部位予測を行い、スプライス部位にまたがるペプチドを検出する、というフレームワークを構築した。なお、(3) の“ペプチドとしての検出されやすさの情報”は、次のように求めた。まず、ペプチドの C-N 結合エネルギーや、親水性など、ペプチドの物理化学的特性をなるべく数多く抽出しておき、それらを帰納学習で統合してペプチドのイオン強度を説明するモデルをつくっておく。そして、イオン強度予測を行った結果を“検出されやすさ”とした。

次に、実証のための実データを用意した。まず、線虫（*Caenorhabditis elegans*）のゲノム DNA 配列データを NCBI からダウンロードし、小規模な人工的な質量分析データを作成した。次に、選択的スプライシングがほとんど起こらないといわれている（つまり、必ず決まったスプライス部位でのみスプライシングが起きていることになり、テストデータとして評価しやすい利点がある）分裂酵母（*Schizosaccharomyces pombe*）の質量分析プロテオームデータを用意した。これら 2 種類のデータに対して、スプライス部位予測の提案手法の検証を行った。

分裂酵母 *S. pombe* でスプライス部位予測を行った事例を図 1 に示す。図 1 は、*S. pombe* の DNA（2 本鎖）の 2726000～2728400 付近を示している。1 つの鎖に付き 3 つの読み枠 (Reading Frame) があるので、合計 6 つの読み枠がある。それぞれの読み枠でストップコドンの箇所には短い黒線が引いてあるので、その線と線の間が ORF（Open Reading Frame）となる。図 1 は SPAC6F6.03 という ID がついている遺伝子のスプライス部位を予測した例であるが、赤い部分がプロテオームの実験データから同定されたペプチドを表し、空色の部分が予測された遺伝子を表す。遺伝子部分が途中で読み枠を飛び越え、あたかも分断されているように見えるのが、スプライス箇所である。

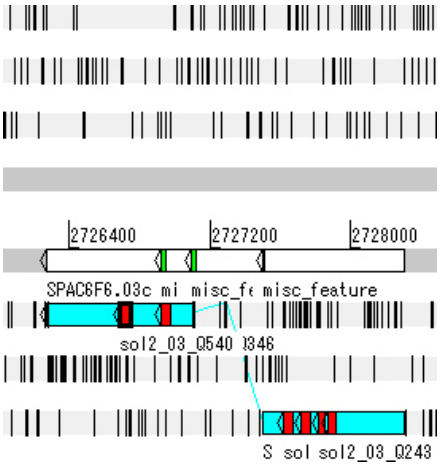

図１．*S.pombe* のスプライス部位予測例

これは、予測がうまくいった例であるが、全体からみるとスプライス部位の一部しか予測できず、良い結果とはいえなかった。

以上のことから、スプライス部位予測に必要な情報がまだ不足している可能性が示唆された。しかし、質量分析データの精度が十分でない場合もこのような結果を引き起こしうる。

そこで、次に、質量分析データの精度改善のために、測定前のキャリブレーションに加え、測定後のデータ較正方法を検討した。データ精度は、質量分析機器自体の性質に強く依存するので、本研究では、エレクトロスプレーイオン化・飛行時間型タンデム質量分析（LC/ESI-TOF MS/MS）に限定して検討した。LC/ESI-TOF MS/MS は、連続してハイスループットな測定を行うことができるためプロテオームデータの測定に適している。LC/ESI-TOF MS/MS では MS 測定モードと

MS/MS 測定モードが交互に切替わって実行されることに着目し、第一段階として MS/MS イオンを内部の較正物質として利用して飛行チューブ長の変動に起因する質量誤差を較正し、第二段階として、不感時間損失（Dead time loss）に起因する質量誤差を較正するという方法を新たに提案した。分裂酵母 *S. pombe* の質量分析プロテオームデータを用いて検証したところ、提案手法の有効性が実証された。

図２は、(A)データ較正を行う前、(B)第一段階の較正後、(C)第二段階の較正後、のそれぞれの段階での、同定されたペプチドの質量誤差の頻度分布を示している。第一段階の較正を行うことで、質量誤差のバラツキは明らかに改善されたが、誤差平均は大きくマイナス側に偏っている。そして、その偏りは、第二段階の較正により正された。

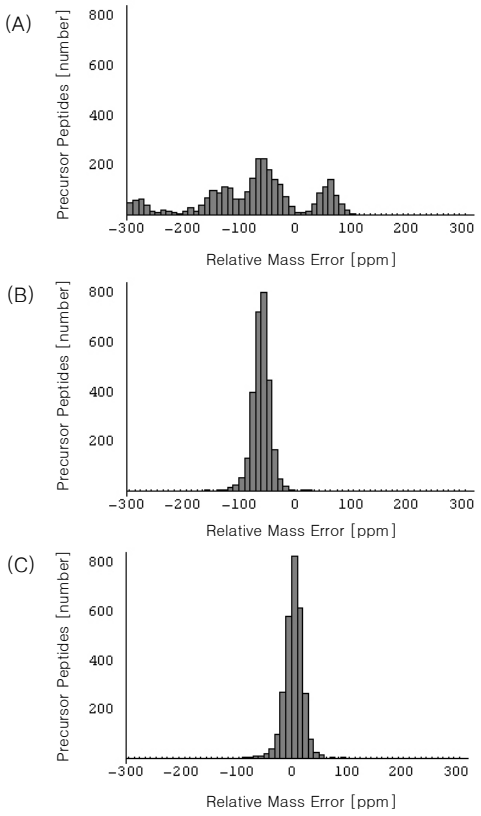

図２．同定されたペプチドの質量誤差の頻度分布

図３は、ペプチド同定の際の質量誤差範囲によって、真陽性（true positive）タンパク質の同定率（検出感度）がどう変化するかを示している。なお、真陽性タンパク質の検出感度は、次式で表す。

$$\text{検出感度} = \frac{(\text{全検出タンパク質数} - \text{偽陽性タンパク質の検出数})}{\text{全検出タンパク質数}}$$

図３の白四角はデータ較正をする前、黒三角は第一段階の較正後、黒丸は第二段階の較正後、それぞれのデータでタンパク質同定を行った結果を示している。100ppm の質量誤差を許したとき、データ較正を全くしていないと62%程度の真陽性タンパク質しか同定できない。一方、第二段階までの較正を行った場合は、45ppm で 100%の真陽性タンパク質を同定でき、20ppm でも 95%を同定できる。このことから、提案するデータ較正法は、真陽性タンパク質の検出感度を増大させることが示された。

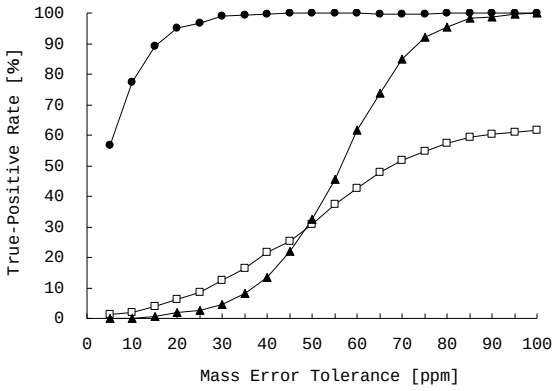

図３．タンパク質同定の検出感度

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計４件）

1. Yoko Ishino, Takanori Harada, In Silico 3D Structure Prediction of Activated GPCRs as a Drug Target, *Proceedings of the 2011 International Conference on Bioscience, Biochemistry and Bioinformatics*, 査読有, 巻無し, 2011, pp.261-265
2. Yoko Ishino, Hisaaki Taniguchi, Dead time loss correction of mass errors occurring in high-throughput proteomics based on electrospray ionization time-of-flight tandem mass spectrometry, *Rapid Communications in Mass Spectrometry*, 査読有, Vol.24, 2010, pp.1490-1495
3. 石野洋子，原田隆範，相田美砂子，実数値GAを用いた薬物標的GPCRの活性型立体構造の探索，人工知能学会論文誌，査読有, Vol.24, 2009, pp.386-396

4. 山内英美子，石野洋子，小西博昭，プロテオミクスによるタンパク質リン酸化の解析，生体の科学，査読有，Vol.60，2009，pp.151-157

〔学会発表〕（計６件）

1. Yoko Ishino, Takanori Harada, In Silico 3D Structure Prediction of Activated GPCRs as a Drug Target, the 2011 International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB2011), Feb. 26-28, 2011, Peninsula Excelsior Hotel, Singapore
2. Yoko Ishino, Hisaaki Taniguchi, An Accurate Mass Error Correction Method in LC-MS/MS Proteome Analysis, the 9th Asia Pacific Bioinformatics Conference (APBC2011), Jan. 11-14, 2011, Incheon, Korea
3. Yoko Ishino, Hisaaki Taniguchi, A Posteriori Calibration Method for Genome-Wide Proteomics Using LC/ESI-TOF MS/MS, the 2010 Annual Conference of the Japanese Society for Bioinformatics (JSBi2010), Dec. 13-15, 2010, Fukuoka, Japan
4. Yoko Ishino, Prediction of Eukaryotic Translation Initiation Sites Using Machine Learning, the 20th International Conference on Genome Informatics (GIW2009), Dec. 14-16, 2009, Yokohama, Japan
5. Yoko Ishino, Takanori Harada, Misako Aida, Search for 3D Structure of Drug Target GPCR Active Forms, BioInfo2009, CBI (the Chem-Bio Informatics Society) and KSBSB (the Korean Society for Bioinformatics) Joint Conference, Nov. 4-6, 2009, Busan, Korea
6. Yoko Ishino, Takanori Harada, Misako Aida, Conformation Search of Drug Target GPCR Using Real-Coded Genetic Algorithm, the 2008 Annual Conference of the Japanese Society for Bioinformatics, Dec. 15-19, 2008, Osaka, Japan

６．研究組織

(1)研究代表者

石野　洋子　（ISHINO　YOKO）

広島大学・産学・地域連携センター・研究員

研究者番号：９０３７３２６６

(2)研究分担者

なし

(3)連携研究者

なし