

科学研究費助成事業 研究成果報告書

平成 29 年 6 月 13 日現在

機関番号：52101

研究種目：基盤研究(C) (一般)

研究期間：2014～2016

課題番号：26505013

研究課題名(和文) ESI-MS2スペクトルデータを用いた代謝物の化学構造推定

研究課題名(英文) Ring Structure Prediction of Chemical Structure of Metabolite Using ESI-MS2 Mass Spectral Data

研究代表者

蓬萊 尚幸 (HORAI, Hisayuki)

茨城工業高等専門学校・電子情報工学科・教授

研究者番号：80633346

交付決定額(研究期間全体)：(直接経費) 1,500,000円

研究成果の概要(和文)：既知物質の質量分析スペクトルデータを利用して、スペクトルデータと物質の化学構造の相関関係を算出し、その結果を用いて、未知物質の質量分析スペクトルデータからその部分的な化学構造を推定する方法を提案しその推定性能を検証した。とくに、この方法は、化合物の母核構造となることが多い環状部分構造の推定に有効であることがわかった。具体的には、スペクトル内のピークに対して、既知データからそのピークが現れたときに環構造を持つ確率(スコア)を算出し、未知物質のスペクトルに現れるすべてのピークについてスコアの平均を計算する。

研究成果の概要(英文)：Prediction method of chemical structure of an unknown chemical compound using mass spectral data of known chemical compounds is proposed and the predictability is validated. This method is effective to predict a ring substructure which is eventually a fundamental structure of a chemical compound. Each peak's score to appear in a spectrum of a known compounds which has a ring is calculated, and average of scores of peaks in a spectrum of an unknown compounds is calculated.

研究分野：Computer Science

キーワード：質量分析 部分構造予測 環構造 プロダクトイオン

1. 研究開始当初の背景

検出されても同定されずに放置されている代謝物：質量分析技術は、メタボロミクス分野での広く利用される標準的技術の一つになっている。最近の測定機器開発の進捗により微量の代謝物の検出も容易になり、多くの代謝物の存在が明らかになっている。しかしながら、その多くは同定されずに放置されている。

新規物質の構造推定技術が必要：従来、代謝物の同定は、既知代謝物のリファレンススペクトルをデータベース化し、測定結果を単純に検索（比較）することで行われてきた。多くの研究者の努力により、徐々にデータは拡充されてはいるが、検出代謝物の大量化への対応は難しい。このようなスペクトルの収集のみに基づく代謝物の同定には限界があり、スペクトルからデータベース化されていない新規物質を推定するための新たなインフォマティクス技術の開発が期待されている。

スペクトルと部分化学構造の相関：一方、ESI-MS2 技術で検出されるプロダクトイオンの生成過程（開裂過程）の研究など、ESI-MS2 スペクトルと化合物部分構造の関係を示唆する成果がもたらされている。さらに、本研究代表者も研究開発に参加した公開スペクトルデータベース MassBank の出現により ESI-MS2 スペクトルデータを共有できる環境が整い、広く利用されている。また、MassBank プロジェクトでは、蓄積されているスペクトルデータを用いて、プロダクトイオンの生成過程を人手で分析し開裂スキームにまとめ、化合物の部分構造に対応するプロダクトイオンのデータベースである Fragmentation Library を構築している。開裂スキームや Fragmentation Library はスペクトルの理解のために非常に重要な知識であるが、人手による分析には高度な専門知識が必要であり、専門家でもコスト（時間）がかかる作業である。本研究で目的とする部分構造推定技術は、このような専門家の分析技術の支援にもなると期待できる。

本研究は公開データベースの出現により可能となった：また、組合せ論の立場からみると、化合物の部分構造のように多様性が非常に大きなデータを用いて推定の信頼度を高めるためには大量のデータが必要である。MassBank の出現により、ESI-MS2 スペクトルと化合物の部分構造に関してインフォマティクス研究を行う状況が整ってきた。

2. 研究の目的

本研究では、化合物の ESI-MS2 リファレンススペクトルを用いて、未同定代謝物の部分化学構造を推定することを目的とする。特に、公開データベースに登録されている大量のリファレンススペクトルに対して、統計的手法やデータマイニング技術などを適用し、プロダクトイオンと部分化学構造の関連性

（共起関係）を明らかにし、この共起関係を利用して未同定物質の ESI-MS2 スペクトルからその部分構造を推定するインフォマティクス技術を開発する。天然物科学の研究の過程で未同定のまま放置されている代謝物に対して、本研究の成果を用いると、化学構造を推定できるようになる。本研究は、化学構造を通して代謝物研究の発展に寄与するものである。

ESI-MS2 スペクトルから中性脱離分子の推定を介して水酸基などの比較的小さな部分構造（官能基など）の推定は既に行われているが、本研究の成果を利用することで、それらの推定に加えて、化合物同定で重要な親構造の推定が可能となる。

本研究で開発する部分構造・プロダクトイオン関連度計算ツールや部分構造推定ツールにより、質量分析技術を用いて代謝物等の同定の作業を推進できる。

さらに、本研究の成果は、開裂スキームなどのプロダクトイオンの生成過程の研究を推進するための強力な支援手法とツールとなると予想される。

3. 研究の方法

本研究では、まず、環状部分構造と ESI-MS2 スペクトルの関係に注力する。MassBank に蓄積された化合物とその ESI-MS スペクトルに対して、複数の化合物に共通する環状部分構造を抽出する。次に、抽出した環状部分構造ごとに、それを含む化合物のスペクトルに現れるピークの環構造判定能力を計算する。具体的には、ピークを判定基準にした場合の情報検索における統計的評価値（精度や再現率など）を計算する。次に、判定能力が高いピークやピークの組合せを選び、それらの存在/非存在を判定基準とする環状部分構造推定法としてまとめる。

また、本手法を用いて実際に部分構造推定をするためのツールを開発する。その際、環状部分構造以外の部分構造への将来的な研究適用範囲の発展を見据えて、「部分構造・プロダクトイオン関連度」を知識として外在化し、それらを取捨選択して利用できるツールとする。

当初は環状部分構造に注力するが、環状部分構造以外の部分構造（鎖状部分構造、官能基、特徴基など）の推定に適用範囲を拡大してゆく。まず、環状部分構造の抽出、ピークの環構造判定能力の計算、判定能力の高いピークやピーク集合の選択を行う一方、部分構造推定ツールの設計を行う。次に、交差検定を用いた本手法の有効性の検証と部分構造推定ツールの開発を行う。その後、環状部分構造以外に適用範囲を拡大する。

4. 研究成果

マススペクトルに現れるピークと化合物の部分構造の間の相関関係を利用し、化合物同定問題の部分問題である部分構造同定問

題について、ESI マススペクトルデータを用いた統計的な化合物の部分構造推定法を提案する。

化合物の部分構造

一般に有機化合物は、中心的な骨格の構造として鎖式母核構造あるいは環式母核構造をもち、それに官能基が結合していると考えられる。有機化合物が骨格としてもつ鎖式母核構造や環式母核構造は、化合物の部分構造であり、それぞれを鎖状部分構造、環状部分構造という。また、官能基も部分構造である。代謝物は母核構造を用いて分類されており、未知代謝物に対してその母核構造を同定することは重要な意味をもつ。本研究では、マススペクトルデータを用いた環状部分構造の同定に関する情報学的手法を提案する。

Figure 1 に様々な環状部分構造を示す。単環以外は必ず複数の単環を部分構造として持ち、3 個以上の単環を持つ縮合環（例：アントラセン、テトラセンなど）は、複数のより小さな縮合環を部分構造として持つという意味で、環状部分構造は重疊的である。本手法では、ある化合物のマススペクトルは、それが重疊的に持つすべての環状構造と相関があると考えられる。

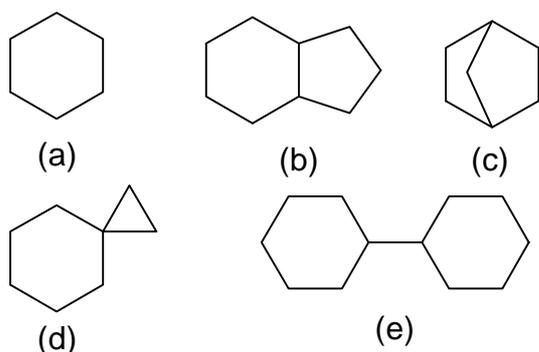


Figure 1: Types of Ring Structure: (a) single, (b) fused, (c) bridged, (d) Spiro, (e) assembly

ESI 質量分析法では、開裂時に、しばしば、環構造内の多重結合の位置が変化したり、結合の多重度が変化する。そこで、本手法で、マススペクトル内のピークと（部分構造としての）環構造の相関関係を考えるときには、環構造内の結合の多重度は無視して扱うこととする。

たとえば、本手法では、Figure 2 のように、化合物(a)は、左側の 4 個の単環を部分構造に持つとみなす。

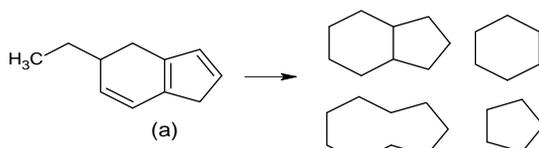


Figure 2: Extract Rings from a Compound

スペクトルデータとピーク

本手法では、スペクトル内のピークの m/z 値を以下のように 0.5 単位で量子化する。測定された m/z 値の小数部分を $frac$ とすると、

- ・0.0 $frac < 0.4$ のとき、小数点以下を切上げる。
- ・0.4 $frac < 0.6$ のとき、 $frac$ を 0.5 とする。
- ・0.6 $frac < 1.0$ のとき、小数点以下を切上げる。

本手法では、ピーク強度は利用しない。すなわち、個々のスペクトルデータは、0.5 単位で量子化した（1 個以上の） m/z の集合とみなす。

ピークと部分構造の相関

本手法では、様々な化合物について測定された ESI マススペクトルをもとに、ピークと部分構造が相関を数値化する。

部分構造とピークの関連付け

ある部分構造（環構造）を持つ化合物を測定して得られたマススペクトルにあるピーク（0.5 単位で量子化した m/z 値が m ）が現れたとき、その部分構造とピークは 1 回関連付けられたとする。言い換えると、そのピークがその部分構造に由来することを示唆する化合物が 1 個存在することが判明したとみなす。

部分構造とピークの相関

推定に用いる既知化合物のスペクトルデータすべてについて、前節で述べた関連付けを環構造ごとに m/z 値ごとに、それらを関連付けた化合物（スペクトルデータ）の個数を、その化合物がその環構造を持つかどうかの場合分けして、数え上げる。この結果を Table 1 のように環構造ごとに表にまとめたものを Voting 表と呼ぶ。たとえば、Table 1 において、 m/z 値が 10 であるピークが Imidazolidine 環を持つ化合物 a1 個のスペクトルに現れ、 m/z 値が 10 であるピークが Imidazolidine 環を持たない化合物 b1 個のスペクトルに現れたことを意味する。

Table 1: Voting Table

Structure	m/z	m/z 10	m/z 10.5	m/z 11	...	m/z 919
	Yes	a1	a2	a3	...	a _n
	No	b1	b2	b3	...	b _n
Voting Point		$\frac{a1}{a1 + b1}$	$\frac{a2}{a2 + b2}$	$\frac{a3}{a3 + b3}$...	$\frac{a_n}{a_n + b_n}$

この表は、たとえば、 $m/z=10$ のピークが現れたスペクトルは $a1+b1$ 個あり、そのうち $a1$ 個のスペクトルが Imidazolidine 環を持つ化合物のものであり、 $b1$ 個がそうでない化合物のものであることを表していると解釈できる。言い換えると、 $m/z=10$ のピークがスペクトルに現れれば $a1/(a1+b1)$ の確率で Imidazolidine 環を持っていたことになる。この確率を「 $m/z=10$ のピークの Imidazolidine に関する Voting 点」と呼ぶ。

未知化合物の部分構造の推定

本部分構造推定法では、ESI 質量分析法で測定して得られたマススペクトルに現れたピークが、自分自身のある部分構造に関する Voting 点をそのピークが「その未知化合物がその部分構造を部分構造としてもつこと」を肯定する得点であるとみなして肯定票を投じると考える。

この肯定票の平均が、その未知化合物がその部分構造をもつことを、測定で得られたすべてのピークによって肯定するスコアとする。

すなわち、ある未知化合物のマススペクトルがある部分構造を持つことのスコアは、そのスペクトルに現れるピークについてその部分構造に関する Voting 点の合計をピーク数で割った値である。

このスコアが高ければ高いほど、その化合物がその部分構造を持つ確率が高いと考えられる。

評価実験

MassBank に登録されている ESI マススペクトルを利用して、クロスバリデーションの一種である Leave-One-Out 法を用いて評価した。ここでは、Imidazolidine 環について評価結果を述べる。

実験に用いたデータ

本評価実験では、MassBank に登録されている ESI-MS 正モード慶応データ 438 個を用いた。Imidazolidine 環を持つ化合物は 63 個存在する。

Leave-One-Out 法を用いた評価

まず、Leave-One-Out 法を用いて、438 個すべてのスペクトルについて、そのスペクトルを除外して残りのスペクトルを用いて Voting 表を作成し、除外したスペクトルのスコアを計算する。

スコアが閾値 以上のとき Imidazolidine 環を部分構造を持つと判断するような部分構造推定を行う。閾値 を変えることで、Imidazolidine 環が存在すると推定されるスペクトルの集合(Positive 集合)も変化する。Positive 集合ごとに、True/False Positive/Negative の 4 個の値を決定し、精度 (precision) = $TP/(TP+FP)$ と再現率 (recall) = $TP/(TP+FN)$ を算出し、再現率精度グラフにプロットする。さらに、F 値 (精度と再現率の調和平均) を算出する。Figure 3 に得られた再現率精度グラフを示す。

一般的に精度と再現率は反比例の関係になることが多いが、本評価実験では、Figure 3 が示すように、再現度を上げてても、精度は下がらない。スコア上位 10 件についてはすべて Imidazolidine 環を持ち、精度 100% となった。すなわち、Imidazolidine 環を持つ化合物の 18% が精度 100% で推定できた。さらに、

精度 90% でイミダゾリジン部分構造を持つ化合物の約 48% で推定できた。よって、本評価実験を通して、本推定手法が有効な環状部分構造推定法であることを確認できた。

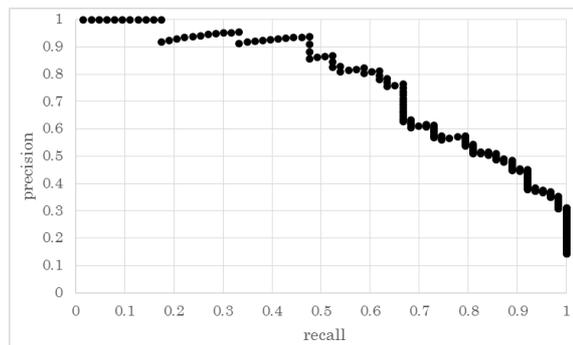


Figure 3: Recall-Precision Graph for Imidazolidine Prediction

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)

[学会発表](計2件)

川崎聡大, 蓬萊尚幸: ESI マススペクトルを用いた代謝物の環構造予測, 日本質量分析学会第 65 回質量分析総合討論会, 2017 年 5 月 19 日, つくば国際会議場エゴカルつくば (茨城県つくば市).

川崎聡大, 蓬萊尚幸: 質量分析スペクトルを用いた部分構造予測, 電子情報通信学会東京支部第 21 回学生会研究発表会, 2016 年 3 月 5 日, 東海大学高輪キャンパス(東京都港区).

6. 研究組織

(1) 研究代表者

蓬萊 尚幸 (HORAI, Hisayuki)

茨城工業高等専門学校・電子情報工学科・教授

研究者番号: 80633346

(4) 研究協力者

川崎 聡大 (KAWASAKI, Toshihiro)