

深層学習を用いたGC/MS波形処理ソフトウェアの開発

金澤 慎司

1 はじめに

クロマトグラフィーを用いた化合物の定量分析は、オミックス解析などのライフサイエンス分野の基盤技術として活用されている¹⁾。クロマトグラムからピーク領域を判定（波形処理）し、面積値を求めることは、対象となる化合物の量を測定するための主要な作業の一つである²⁾。しかし、メタボロミクスデータ取得法として最も広く用いられているワイドターゲットメタボロミクス法において、波形処理は、非常に時間と労力を要する作業であり、効率化が求められている。ワイドターゲットメタボロミクス法では、1回の分析で、測定対象代謝物数分のクロマトグラムを同時並行で取得するため、大量のクロマトグラムの波形処理が必要である。具体的には、1人の作業者が10サンプル（500種類の代謝中間体含量）を測定することを考えると、5000回の波形処理が必要である。

波形処理にかかる作業時間を短縮するべく、これまでに様々な自動波形処理手法が考案されてきた。考案された波形処理手法は、次のことが基本動作となっている。

1. クロマトグラム中に観察されるすべてのピークの波形処理を行う。
2. その際に、保持時間や追加のクロマトグラム（例えばフラグメントイオン）のような追加情報を用いない。

代表的な波形処理アルゴリズムとしては、ピークの立ち上がり^{しきい}と立ち下りの傾きの閾値（Slope）とピーク

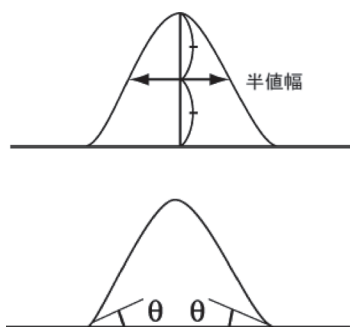


図1 従来アルゴリズムのパラメータ

幅の閾値（Width）を基にピークが検出されるクロマトパックアルゴリズムがある（以降、従来アルゴリズムと呼ぶこととする）。具体的には、Width以上の半値幅のピークを検出し、設定した幅よりも狭いピークをノイズとする。SlopeのパラメータではSlope以上の傾きをピークの開始点と終了点として検出し、Slopeより小さな変化をベースラインのドリフトとする（図1）。

この方法は理想的なピークではうまくいく（図2a）。しかし、実際のクロマトグラムでは、テーリング（ピーク後部が裾を引いている場合）、リーディング（ピーク前部が裾を引いている場合）などの理由でピーク形状がゆがむことがある（図2b）。

これらのようなケースにおいては、従来アルゴリズムではうまく波形処理できないことが多い。そこで、自動波形処理結果を目視で確認し、問題がある場合には、手動で修正する作業が行われている。メタボロミクスでは、経験的に約2割のピークを修正する必要がある。そのため、手動で修正するのに平均して一つ18秒かかるとすると、5000回の自動波形処理のうち、2割の1000回分を修正する必要がある、約5時間かかる。こ

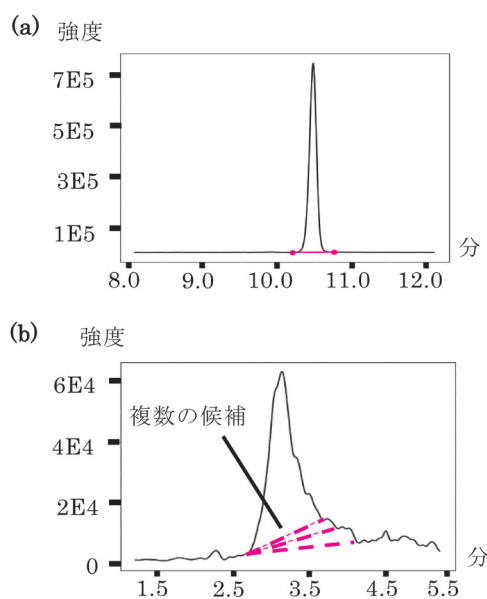


図2 波形処理の問題

(a) 理想的なピーク、(b) テーリングピーク。

れは、定量分析を行う上でのボトルネックになる。

そこで近年、ディープニューラルネットワークを利用し、深層学習を活用した、自動波形処理の高性能化が試みられている³⁾⁴⁾。深層学習は機械学習手法の一種であり、深層学習で扱われる計算アルゴリズムがディープニューラルネットワークである。ディープニューラルネットワークは脳をモデル化したものであり、一つ一つのノードがニューロンと呼ばれる神経細胞に対応し、複数のノードを用いることで、複雑な関数を表現することができる。深層学習は、画像認識分野において、従来手法と比べて高い精度を出している。これらの深層学習手法は、データとラベルを使って機械に知識を学習させる手法である。例えば、顔のラベルがついた画像と顔でないラベルがついた画像を、機械に学習させることで、学習済みの機械（学習済みモデル）は、顔のラベルがついていない画像を入力すると顔と判断できるようになる。同様の技術を用いることで、ピーク領域を同定する波形処理ディープニューラルネットワークの構築が可能である。波形処理に深層学習を活用する場合、人が波形処理したクロマトグラムを用いてディープニューラルネットワークを学習することで（図 3a）、学習済みモデルは、波形処理していないクロマトグラムを入力するとピークの領域を推論できるようになる（図 3b）。

このような深層学習を用いた波形処理ソフトウェアは LC/MS 用として開発されている（PeakintelligenceTM）⁵⁾。GC-MS から取得されるクロマトグラムは、LC-MS と比べピーク幅が狭いなどクロマトグラムの特性が違うため、LC/MS 用に開発されたものを GC/MS 用として使

用することができない。そのため GC/MS では、深層学習を用いた波形処理ソフトウェアがないため、波形処理作業に時間がかかっている。そこで、本稿では、GC/MS 用に深層学習を用いた波形処理ソフトウェア Peakintelligence for GCMSTM を開発することで波形処理作業を短縮することを目的とした。

2 実験材料と方法

2.1 実データの準備

茶葉やほうれん草などの食品サンプル、および生物学的試料の分析試料は、Bligh & Dyer の方法をベースとした前処理プロトコルにて抽出し、メトキシム-TMS 誘導体化した。試料は、GCMS-TQ8050 NX (NexisTM GC-2030 システムとトリプル四重極型質量分析計 GCMS-8050, 島津) を用いて、Smart Metabolites DatabaseTM Ver. 2 (島津) のメソッドで測定した。LabSolutionsTM (島津) を用いて、一次代謝物および二次代謝物の数万のクロマトグラムを収集した。そして、得られたクロマトグラムを、熟練者によって手動で波形処理（ラベル付け）し、学習用（90%）とテスト用（10%）に分割した。

2.2 ディープニューラルネットワークの学習方法

複雑な形状のクロマトグラムに対して、堅牢な手法を開発するため、波形処理の問題をセマンティックセグメンテーションの問題として定式化した。セマンティックセグメンテーションは、画像または時系列データのすべての要素でクラス分類する手法である。学習データは、各点の保持時間と強度値のクロマトグラムデータと、ピークの開始時間と終了時間のラベルデータから成る。

2.3 評価方法

熟練作業者が波形処理した結果に対して、未検出の少なさを評価するため、適合率を用いて評価した。適合率は、正確な検出数（真陽性）を推論ピーク数（真陽性 + 偽陽性）で割って計算した。未検出ピークの割合は、適合率から 1 を引いた値になる。

3 結果

3.1 Peakintelligence for GCMS のアルゴリズム評価

食品サンプルおよび生物学的試料からワイドターゲットメタボロミクス法で分析を行い、収集した学習データを用いて、Peakintelligence for GCMS の波形処理ディープニューラルネットワーク（アルゴリズム）を構築した。テストデータを用いて、Peakintelligence for GCMS のアルゴリズムと市販のクロマトグラフィーデータシステムの（最適なパフォーマンスにパラメータを調整した）従来アルゴリズムを比較した。

結果は、Peakintelligence for GCMS のアルゴリズム

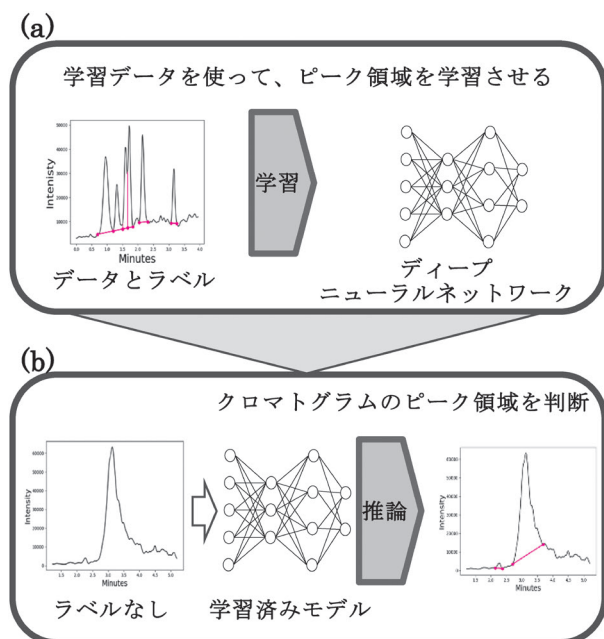


図 3 深層学習を活用した波形処理ディープニューラルネットワーク

(a) 学習, (b) 推論.

表 1 ピーク検出結果

アルゴリズム名	適合率	未検出率
Peakintelligence for GCMS	0.922	0.078
従来アルゴリズム	0.767	0.233

(0.078) が、従来アルゴリズム (0.233) よりも未検出の割合が少ないことを示した (表 1)。

Peakintelligence for GCMS は、従来アルゴリズムに比べて、未検出率が約 3 分の 1 に減少していることから、未検出ピークを手動で修正する回数が少なくなり、波形処理にかかる時間が短縮されることが示唆された。また、熟練者が作成したテストデータとの差異が少ないことから、熟練者を模倣した波形処理が可能なが示唆された。

3.2 パラメータフリーな波形処理ソフトウェア

Peakintelligence for GCMS の評価

従来アルゴリズムより高性能な波形処理アルゴリズムが構築できたので、構築したアルゴリズムを用いてパラメータフリーな波形処理ソフトウェア Peakintelligence for GCMS を開発した (図 4)。

本ソフトウェアは、LabSolutions Insight™ DB GCMS 上でオプションとして使用することができ、パラメータを設定することなく波形処理をすることを可能である。但し、学習機能は未搭載である。前節と同じテストデータを用いて波形処理結果を比較した。

従来アルゴリズムでは、ベースラインのドリフトとノイズによりピークの開始点と終了点が広がっている。一



図 4 パラメータフリーな波形処理ソフトウェア Peakintelligence for GCMS

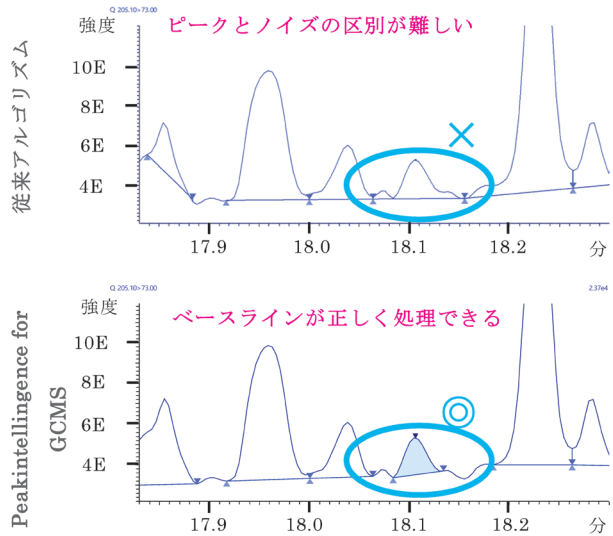


図 5 従来アルゴリズムとの比較 1

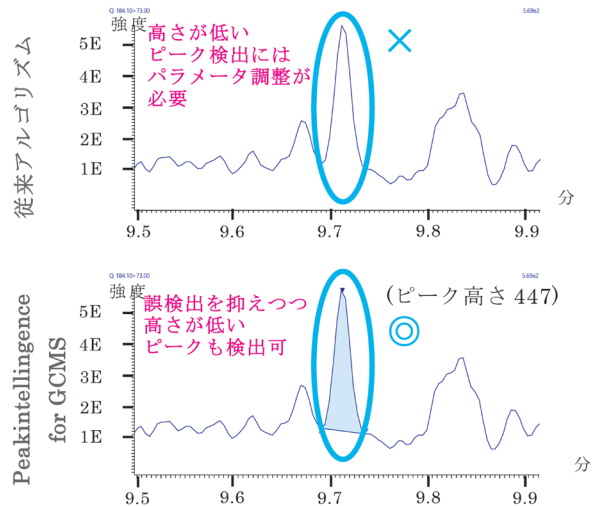


図 6 従来アルゴリズムとの比較 2

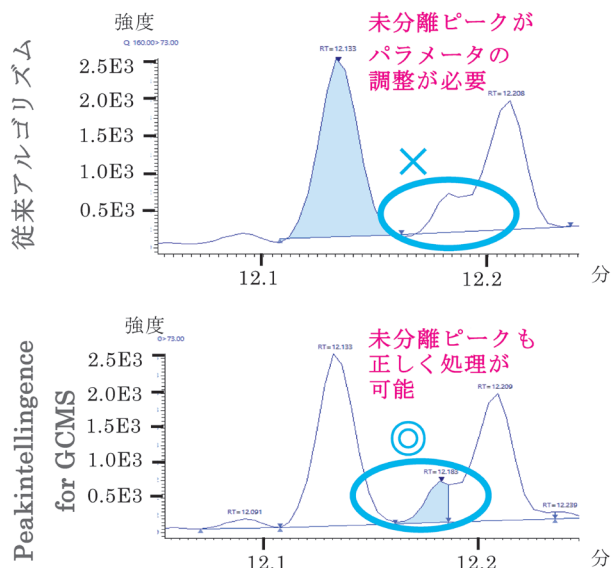


図 7 従来アルゴリズムとの比較 3

方, Peakintelligence for GCMS では, ベースラインを正しく処理できている (図 5).

また, 高さが低いピークに関しては, 従来アルゴリズムでは, パラメータの調整が必要だったが, Peakintelligence for GCMS では, ピークの誤検出を抑えつつ, 高さが低いピークを検出できた (図 6).

さらに, 未分離ピークに関しても従来アルゴリズムではパラメータの調整が必要だったが, Peakintelligence for GCMS では, 未分離ピークも正しくピークを分割できることを確認した (図 7).

従来アルゴリズムでは, ピークの高さと幅が変化することにより, ピークとノイズの区別が難しく, 未分離のピークに対応することも難しかったが, Peakintelligence for GCMS ではそれらが, 解決された.

これらの結果から, Peakintelligence for GCMS は従来アルゴリズムと比べて, パラメータレスで正しく波形処理できることが確認できた. パラメータの設定時間がなくなることで波形処理の修正時間が減ることから, 作業時間が短縮されることが示唆された.

4 む す び

深層学習を用いた波形処理ソフトウェア Peakintelligence for GCMS が既存のソフトウェアより有効であることを実証した (表 1, 図 5, 図 6, 図 7). 本稿の結果は, Peakintelligence for GCMS が波形処理にかかる時間を短縮し, クロマトグラフィーを用いた化合物の定量分析のボトルネックの解消につながるソフトウェアであることが示唆された. また, 熟練者の波形処理を再現できていることから, 担当者に依らず, 質の高い波形処理が

実現でき, 属人性が解消できることが示唆された. さらに, パラメータ不要の波形処理により, 難しい設定や, 波形処理の修正をすることなく, 意図する波形処理が実現でき, 恣意的なデータの捏造防止につながる.

最後に GC/MS における深層学習を用いた波形処理ソフトウェア Peakintelligence for GCMS の開発に際して, 学習データ作成や試作版評価にご協力いただいた大阪大学・島津分析イノベーション協働研究所に感謝の意を表す.

文 献

- 1) F. Matsuda, Y. Toya, H. Shimizu : *Learning : Biotechnol Adv*, **35**, 971 (2017).
- 2) F. Matsuda : *Mass spectrometry*, **5**, S0052 (2016).
- 3) A. D. Melnikov, Y. P. Tsentelovich, V. V. Yanshole : *Anal. Chem.*, **92**, 588 (2020).
- 4) S. Kanazawa, A. Noda, A. Ito, K. Hashimoto, A. Kunisawa, T. Nakanishi, S. Kajihara, N. Mukai, J. Iida, E. Fukusaki, F. Matsuda : *J. Biosci. Bioeng.*, **131**, 207 (2020).
- 5) 金澤慎司, 山田洋平, 安田弘之, 國澤研大, 塩浜 徹, 梶原茂樹, 向 紀雄, 垣迫正輝, 藤澤 剛, 山影 譲, 飯田順子, 松田史生 : 島津評論, **78**, 275 (2021).



金澤 慎司 (Shinji KANAZAWA)

株式会社島津製作所 IT ソリューションビジネスユニット兼 SolutionsCOE ヘルスケアソリューションユニット (〒210-0821 神奈川県川崎市川崎区殿町 3-25-40). 大阪大学大学院情報科学研究科博士後期課程修了. 博士 (情報科学). 《現在の研究テーマ》分析装置から取得されたデータ解析の自動化. 《趣味》テニス, 散歩, お酒.

E-mail : s-k@shimadzu.co.jp

会社ホームページ URL :

<https://www.shimadzu.co.jp/>

関連製品ページ URL :

<https://www.an.shimadzu.co.jp/products/gas-chromatograph-mass-spectrometry/gc-ms-software/peakintelligence-for-gcms/index.html>