メタボロミクスのための
高性能ピークピッキングと
キーワードレコメンド法の開発

提出先　　大阪大学大学院情報科学研究科

提出年月　2021 年 1 月

金 澤　慎 司

学位に関わる研究業績

1. 学術雑誌

[1-1] **Shinji Kanazawa**, Akira Noda, Arisa Ito, Kyoko Hashimoto, Akihiro Kunisawa, Tsuyoshi Nakanishi, Shigeki Kajihara, Norio Mukai, Junko Iida, Eiichiro Fukusaki, Fumio Matsuda, "Fake metabolomics chromatogram generation for facilitating deep learning of peak-picking neural networks", *Journal of Bioscience and Bioengineering*, Published online October 2020:S1389172320303625. (博士論文第2章)

2. 国際会議

[2-1] **Shinji Kanazawa**, Yohei Yamada, Hiroyuki Yasuda, Akihiro Kunisawa, Toru Shiohama, Shigeki Kajihara, Norio Mukai, Masaki Kakisako, Go Fujisawa, Yuzuru Yamakage, Junko Iida, Eiichiro Fukusaki, Fumio Matsuda "Deep learning methods applied to the analysis of metabolomics data" 67th ASMS Conference on Mass Spectrometry and Allied Topics, June 5th, 2019（博士論文第2章）

[2-2] **Shinji Kanazawa**, Yohei Yamada, Hiroyuki Yasuda, Fumio Matsuda, Samik Ghosh, Takeshi Hase, Nikolaos Tsorman, Yukiko Matsuoka, Shigeki Kajihara, Hiroaki Kitano, Eiichiro Fukusaki, Junko Iida, "Automated visualization of multiomics (metabolomics, proteomics, fluxomics and transcriptomics) data on Garuda, a connectivity platform for biological analytics." 66th ASMS Conference on Mass Spectrometry and Allied Topics, June 5, 2018（博士論文第3章）

[2-3] **Shinji Kanazawa**, Tsuji Shuma, Yohei Yamada, Hiroyuki Yasuda, Samik Ghosh, Takeshi Hase, Nikolaos Tsorman, Yukiko Matsuoka, Shigeki Kajihara, Hiroaki Kitano, Eiichiro Fukusaki, Junko Iida, Fumio Matsuda "Automated visualization of time series data on Garuda, a connectivity platform for biological analytics." 2nd International BMS Symposium 2018, October 27, 2018（博士論文第3章）

3. 出願特許

[3-1] **Shinji Kanazawa**, Yoshihiro Hayakawa, "識別器の生成方法", PCT/JP2018/036801 (博士論文第 2 章)

[3-2] **Shinji Kanazawa**, “波形解析方法及び波形解析装置”, PCT/JP2019/039018 (博士論文第 2 章)

[3-3] **Shinji Kanazawa**, Fumio Matsuda, “データ生成方法及び装置、並びに、識別器の生成方法及び装置”, 特願 2020-107606 (博士論文第 2 章)

[3-4] Yohei Yamada, **Shinji Kanazawa**, Hiroyuki Yasuda, Akihiro Kunisawa, “分析装置”, PCT/JP2019/018338 (博士論文第 2 章)

[3-5] **Shinji Kanazawa**, Fumio Matsuda, “解析装置および解析方法”, PCT/JP2020/018586 (博士論文第 3 章)

内容梗概

メタボロミクスは生物試料中に含まれる、多数の代謝中間体含量を測定する技術である。代謝は生命の基本システムであり、生物試料中の代謝中間体含量には、遺伝的な要因（体質など）に加え、外部環境（疾患や薬物投与、生活習慣など）が反映する。例えば、医学分野では、がんの疾患群と健常者群の血中代謝中間体含量を液体クロマトグラフィー質量分析装置（LC-MS）用いたメタボロミクスで測定し、疾患群で特異的に増減する成分の特定が行われている。さらに代謝中間体含量の増減と発病メカニズムとの関連を解釈することで、疾患の早期発見および病態の理解につなげる試みが行われている。食品分野でも同様に、高級緑茶と一般的な緑茶の代謝中間体含量の比較から、品質の客観的な評価が可能となるなど、メタボロミクスは様々な分野で重要な役割を担っている。

メタボロミクスは、主に 4 つの技術 [1. データ取得、2. 生データ処理、3. データ解析、4. データ解釈] から構成される。LC-MS で得られた信号データからピーク面積を決定する（ピークピッキング）作業では、自動ピークピッキング法の開発が進められているが、熟練作業者レベルの精度には達していないため、確認・手動修正に数時間から半日程度の時間が費やされているという課題がある。近年、深層学習を用いた自動ピークピッキングによる精度向上が試みられているが、未だ熟練作業者レベルには達していない。データ解釈では、まず、含量が増減した代謝物のリストから、関連するキーワードを連想する作業が行われる。しかし、「解釈担当者」の知識に大きく依存し、文献調査に日から週スケールの作業が必要であることも深刻な問題となっている。

そこで、本研究はメタボロミクスのためのピークピッキング作業の不要化とキーワード連想作業の省力化を目指すために、次の 2 点を実現する手法の開発を行った。

１．自動ピークピッキング作業の正確さを熟練作業者並みとし、修正作業を不要にすること。

２．キーワード連想作業を自動化するキーワードレコメンド手法を開発して、キーワード連想作業の属人性を解消し、作業時間を短縮すること。

本学位論文は、第 1 章から第 4 章より構成される。第 1 章では、本研究の背景と目的、本論文の構成について記述した。第 2 章では、ピークの正確な開始点と終了点でラベル付けされた疑似クロマトグラムを生成する新たな技術を開発し、生成した大量の疑似クロマトグラムを用いてピークピッキングニューラルネットワークの深層学習を行

った。その結果 U-Net がニューラルネットワークとして Single Shot MultiBox Detector より優れていることを示した。構築したピークピッキングニューラルネットワークと熟練者とで、LC-MS ワイドターゲットメタボロミクスのテストデータを用いて性能比較を行ったところ、両者は同等の性能を示した。こうした自動ピークピッキング法の開発により、実際に生データ処理にかかる時間は実質ゼロになった。

第 3 章では、第 2 章で得られた網羅的な測定データの統計解析で得られる「代謝物リスト」から、関連度の高いキーワードを提案する手法を開発した。アソシエーション分析の手法を活用することで、代謝物リストと「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」をレコメンドする新たな手法を開発した。さらに、レコメンドされたキーワードの妥当な閾値を設定するために、ランダムデータベースを用いた帰無分布を利用する手法を開発した。これにより、解釈担当者の知識に依存せず、より客観的な仮説生成を自動化することを可能とした。さらに、日から週スケールの調査時間を約 1.5 時間で必要なキーワードレコメンド法を構築できた。

第 4 章では、本研究の結果を総括し、開発した高性能ピークピッキング手法とキーワードレコメンド法の汎用性と今後の課題、メタボロミクスへの寄与について述べ、結言とした。

目次

第1章 緒言

1-1 代謝計測の役割

代謝は、ヒトをはじめとする全生物の基本システムである。外部から食事として取り込んだ炭素源、窒素源は各細胞内で代謝経路を通じて数百を超える代謝中間体に変換され、細胞の維持に必要な構成要素の合成や、エネルギーの供給に用いられている。代謝経路の各反応を担う酵素タンパク質は、ゲノム上の遺伝子の転写・翻訳機構を通じて生合成される (Crick, 1970)。このため、遺伝的な要因（体質など）に加え、外部環境（疾患や薬物投与など）の影響が遺伝子発現やタンパク質の活性化や抑制化に繋がり、その結果、各代謝中間体の細胞内存在量に反映されると考えられている。

このような代謝の特性を利用する試みは、医学、農学、工学、食品など多岐にわたる分野で行われている。農学分野では、代謝の特性を、品質の評価・予測に利用する試みが行われている (Figure 1-1)。例えば、高級緑茶と一般的な緑茶の 25 個の代謝中間体含量を比較することで、高級緑茶には、キナ酸やリン酸が多いことがわかり、緑茶の品質評価・予測に繋がることが期待される (Pongsuwan et al., 2006)。

工学分野では、生物を利用した有用物質生産の品質評価に利用する試みが行われている。例えば、培地成分を変えたチャイニーズハムスター卵巣由来細胞株を対象に 117 成分の培地由来の代謝中間体を比較することで、ウリジン二リン酸-N-アセチルグルコサミンが生産物質（ヒトモノクローナル抗体）の品質の影響に大きいことがわかり、効率の良い物質生産に繋がることが期待される (Zürcher et al., 2020)。その他の例として、異なる寿命をもつ出芽酵母を対象に、87 個の代謝中間体を比較することで、アスパラギン酸とグルタミン酸が長寿命な株と正の相関があることがわかり、効率の良い物質生産に繋がることが期待される (Yoshida et al., 2010)

その中でも、医学分野での研究が特に盛んである。PubMed で「metabolite（代謝物）& medicine（医学）」を検索すると 62,138 件ヒットし、medicine（医学）を agriculture（農学）や engineering（工学）、food（食品）に置き換えて検索すると、それぞれ 16,696 件、16,611 件、33,105 件であり、医学の検索件数が最も多かった（2020 年 12 月現在）。

医学分野では、代謝の特性を、診断や生体メカニズムの理解に利用することが試みられている。例えば、心筋梗塞や動脈硬化を発病した疾患群と、健常者群の血中代謝中間体含量の比較から、疾患群のコレステロール含量が高いことが見出された (Liu et al.,

1992)。この知見は、発病高リスク群を早期に診断する手法の開発につながった (Salonen et al., 1997)。また、コレステロールが心筋梗塞や動脈硬化を引き起こす生体メカニズムが解明され、コレステロール生合成阻害剤の開発にもつながっている (Edwards, 1998)。このような生体内状態を反映して増減する代謝中間体をバイオマーカーと呼ぶ。バイオマーカー量の測定から、病気の存在や進行度を判断できるため、治療の重要な指標の一つになっている。

現在、代謝メカニズムの理解が強く求められる分野の 1 つとしてがん研究があげられる (Hsu et al., 2008)。がんは早期発見が重要な疾患であり、患者の血液や唾液などの生体試料から、がんの存在を検知できるバイオマーカーの探索が盛んに行われている (Hou et al., 2014, Li et al., 2014., Wong, 2006)。例えば、がんのバイオマーカーの例として膵臓がん疾患群と健常者群の血中代謝中間体含量の比較から、グルタミン酸やヒスチジンなど 5 つの代謝中間体が疾患群と健常者群を識別できるバイオマーカーであることが示された (Itoi et al., 2017)。また、ステージが異なる乳がん疾患群と健常者群の唾液中代謝中間体含量を比較することで、ポリアミン類の含量ががん患者のステージと正に相関することが明らかとなった (Takayama et al., 2016, Tsutsui et al, 2013)。このように、がんを鑑別できるバイオマーカーの発見は、がんの早期発見につながると期待される。

がん以外の疾患として、糖尿病の研究でも代謝のもつ生体機能の解明が強く求められている (Guasch-Ferré et al., 2016)。糖尿病は早期診断や予後診断が重要な疾患であり、患者の尿や血液などの生体試料から糖尿病の存在を検知できるバイオマーカーの探索が行われている (Maclsaac et al., 2014)。例えば、糖尿病患者群と健常者群の血中代謝中間体含量の比較から、グルコースとフルクトースを含むいくつかの糖と糖新生の基質が疾患群で高いことが明らかになった (Würtz et al., 2012)。また、ステージの異なる糖尿病患者の尿と血液の代謝中間体含量を比較したところ、糖尿病の病期の進行と代謝中間体含量データに相関があることが示唆された (Hirayama et al., 2012)。このような知見を利用した糖尿病の検査方法が確立されることで、糖尿病の早期発見と予後診断可能になると期待される。

バイオマーカーの探索はがんや糖尿病の検査だけでなく、抗がん剤の作用機序や抵抗性獲得メカニズムの解明にも用いられている。抗がん剤はがん細胞の増殖を阻害することで、がんを死滅に導く。そのメカニズムが解明できれば、がん細胞を死に至らしめる新たなヒントが得られ、より高い抗がん作用を有する新薬の開発につながると期待され

る (Shajahan Haq et al., 2015)。既報では、乳がん細胞に対する抗がん剤 XD14 の投与有無で 3 日間の時系列実験を行った結果、アミノ酸代謝や脂質代謝などの 67 種類の代謝中間体含量が大きく変化しており、それががん細胞の増殖の低下と関係があると報告されている (Pan et al., 2016)。抗がん剤の投与はがんの代表的な治療法であるが、治療が長期間に及ぶと、がん細胞が薬剤に対して抵抗性を持ち、抗がん効果が減弱する場合がある。この抵抗性獲得の原因が分かれば、そのような抵抗性を獲得させないような多剤併用戦略の立案につながる。例えば、ドキソルビシンといった抗がん剤に対して、耐性を有する乳がん細胞株にドキソルビシンを投与有無で代謝中間体含量を比較する実験が行われた。その結果、グルタチオンとシステインの生合成経路の 24 種類の代謝中間体がドキソルビシンの耐性に重要な役割をもつバイオマーカーとして発見された (Zhang et al., 2016)。これらの代謝中間体の産生経路は、抗がん剤耐性化のためにがん細胞が代償的に活性化させた経路である可能性があることから、このような経路を阻害する薬剤を抗がん剤と併用する新たな治療戦略が有用であると示唆された。

このように、バイオマーカーの活用をさらに進めることで、疾患の早期発見・予後診断、新規治療薬の開発、薬効増進が期待でき、より健康な人間社会の実現に貢献できると考えられる。

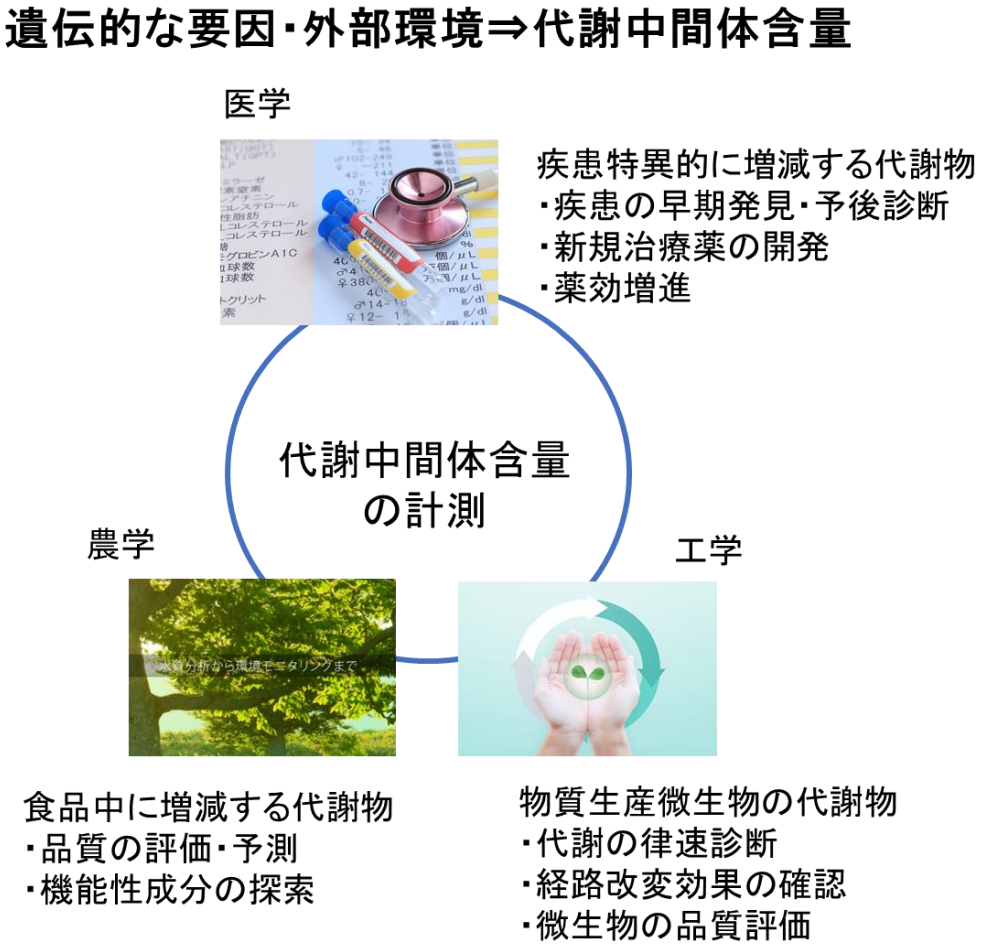


Figure 1-1　代謝中間体含量計測の役割

遺伝的な要因（肥満になりやすいなど）や外部環境（ストレスや食習慣など）が代謝中間体含量に影響していると考えられている。そこで、代謝中間体含量を計測することで、遺伝的な要因や外部環境がわかる。これら代謝の特性は、医学分野や農学分野、工学分野で利用されている。

1-2 メタボロミクスのバイオマーカー探索への利用

メタボロミクス法とは生体中の代謝中間体含量の網羅的な解析を目指す計測手法である (Fiehn, 2002, Fiehn et al., 2001)。メタボロミクスの用途はバイオマーカー探索を始め、システム生物学研究や、代謝の計算機シミュレーション法の開発などにも用いられている。PubMed で「metabolomics（メタボロミクス） & biomarker（バイオマーカー）」を検索すると 11,034 件ヒットし、「metabolomics（メタボロミクス） & simulation（シミ

ュレーション)」を検索すると、1,034 件であることから、メタボロミクスのバイオマーカー探索への利用が進んでいると考えられた（2020 年 12 月現在)。

生体から得たサンプルの代謝中間体含量を網羅的に測定した結果から、バイオマーカーを探索し、診断や生体メカニズムの理解に利用するメタボロミクスでは次のような 4 つの要素技術からなるワークフローで作業が行われている。

1．適切な実験計画をもとに作成したサンプル（例えば、投薬群／対照群の２群など）から計測機器をもちいて生データを取得する（データ取得)。
2．生データを処理して代謝中間体の含量を網羅的に決定する（生データ処理)。
3．代謝中間体含量を２群間比較し、有意に増加、減少した代謝中間体を同定する（データ解析)。
4．同定された代謝中間体のリストから、関連するキーワードを連想する（キーワード連想)。

以降では、ワークフローで用いられる要素技術について述べ、1．データ取得、3．データ解析の進展に比べ、2．生データ処理、4．キーワード連想に問題が残されていることを指摘する (Figure 1-2)。

本法のワークフローには汎用性がある。まず、食品の品質評価などのどの用途でも 1．データ取得、2．生データ処理、3．データ解析は必ず行う。さらに、種々のデータ解析、シミュレーションなどを実施した結果、得られた代謝中間体リストから、さまざまなキーワードを連想し、その解釈から生体メカニズムの理解、仮説生成をする作業は代謝工学分野などでも実施されている。つまり、本ワークフローの改善は多くの領域に波及する。

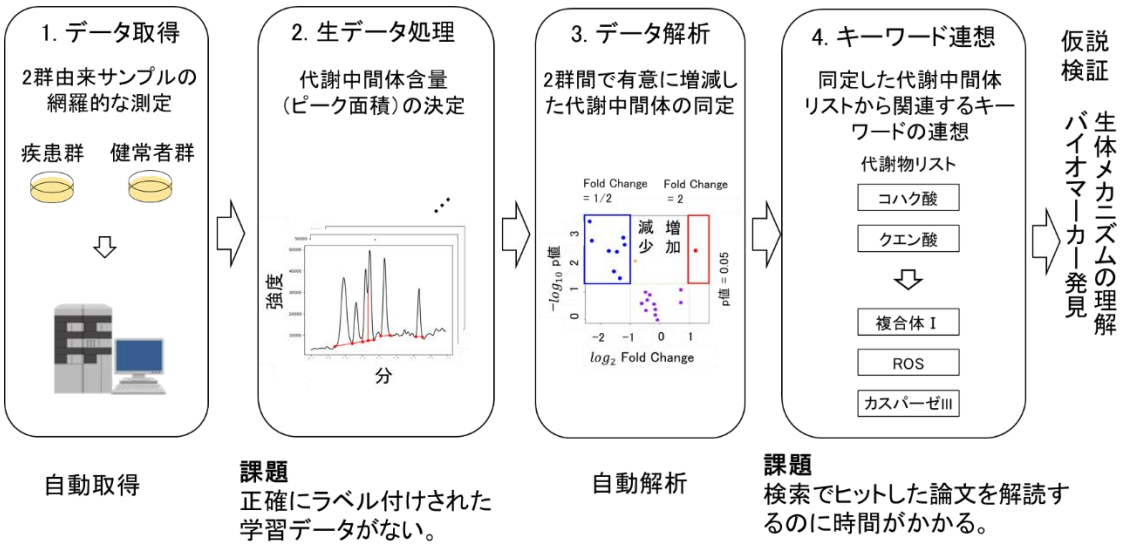


Figure 1-2 メタボロミクスのワークフロー

メタボロミクスを構成する４つの要素技術の説明と、それぞれの現状と課題を述べた。

1-2-1 メタボロミクスで用いるデータ取得法

メタボロミクスで用いるデータ取得とは、多数の代謝中間体含量を網羅的に測定することを目指した計測技術である。その際、クロマトグラフィーによる分離法と質量分析法 (Mass Spectrometry, MS) の選択的検出を組み合わせた手法を用いる。クロマトグラフィーとは、複雑な成分からなる生体試料を成分の物性に基づいて分離する手法である。生体試料は、カラムと呼ばれる筒状の流路に注入され、気体や液体などの流体にのってカラムの中を進む。カラムには、特定の物性を有する担体が充填されており、カラム担体との相互作用の強さに応じて、生体試料中の物質はカラムを通過する時間に差が生じる (Figure 1-3)。相互作用の小さいもの（代謝中間体A）はカラムから早く、相互作用の大きいもの（代謝中間体B）は遅くカラムから溶出し、質量分析計で検出される。得られるデータは、横軸がカラムから溶出した時間（保持時間とよぶ）、縦軸が各成分の量を表すイオン強度（イオンの個数）からなる (Figure 1-3) 。これをクロマトグラムと呼ぶ。また、クロマトグラム中の代謝中間体由来のシグナル（山）をピークと呼ぶ。ピークの面積（Figure 1-3 の斜線領域。黒のシグナル線の下側と赤い直線の上側の面積。単位は時間×イオンの個数。）は溶出した代謝中間体含量に比例して増減する。例えば、2群間で代謝中間体 A の含量を比較する場合、それらの面積値を除算し、比較する。その

他、濃度既知の代謝中間体 A のピーク面積と濃度未知の代謝中間体 A のピーク面積を比較することで、代謝中間体 A の濃度を決定する手法もある。

生体試料の分析では、補酵素、ヌクレオチド、アミノ酸、糖リン酸などの重要な生体分子を測定できる液体クロマトグラフィー (Liquid Chromatography, LC) 質量分析装置が最もよく用いられる (Banoub et al., 2005, Luo et al., 2007, Wellerdiek et al, 2009)。LC-MS を用いたデータ取得法は、近年の性能向上の結果、1 サンプル当たり、20 分で 200 個分の代謝中間体含量のデータを取得する能力がある (Peñaloza et al., 2020)。LC-MS を用いたバイオマーカーの発見には、80～400 のサンプル数を扱うため (Di Gangi et al., 2016, Rhee et al., 2011, Wang et al., 2011, Wang et al., 2013)、沢山のデータを取得する必要があるが、終夜自動でデータを取得できることと複数台での分析でカバーできる。このため、メタボロミクスの大きなボトルネックとはなっていない。

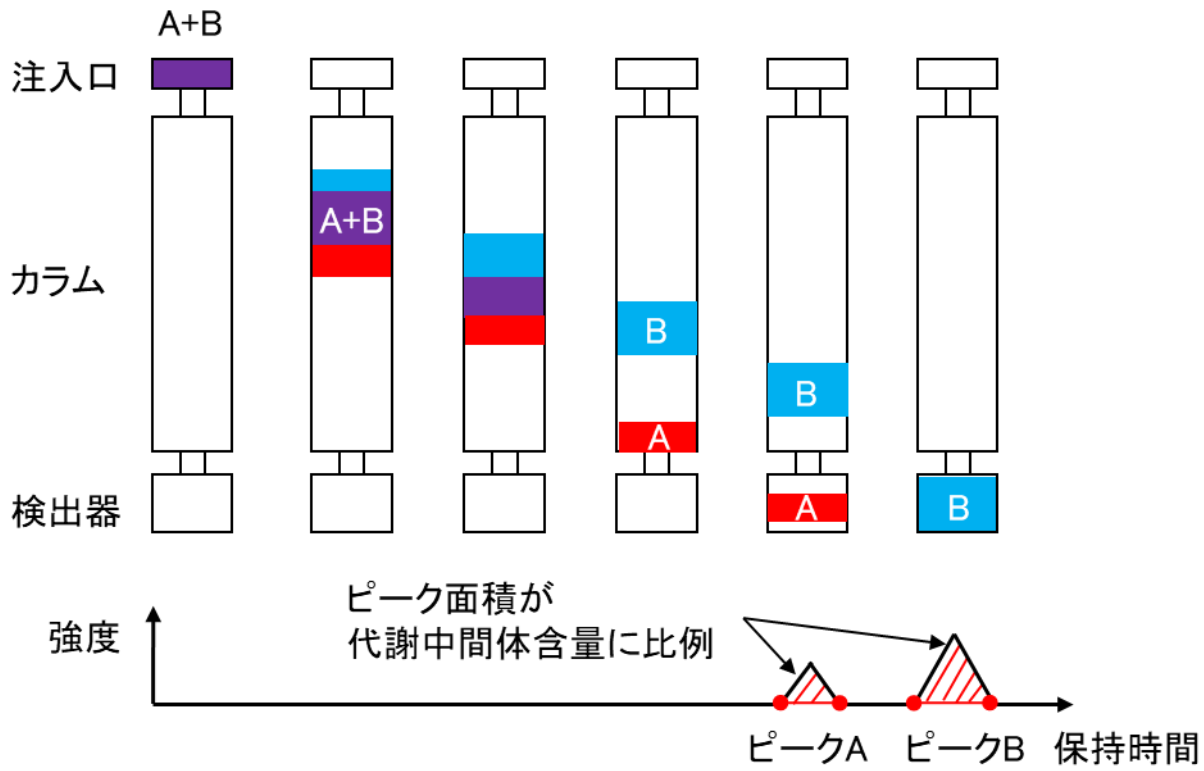


Figure 1-3 クロマトグラフィー

サンプルがカラムを通ることで成分 A と成分 B に分離される。ピークの面積は、黒のシグナル線の下側と赤い直線の上側の面積（斜線部分の領域）である。

1-2-2 メタボロミクスで用いる生データ処理（ピークピッキング）法

メタボロミクスの 2 番目の作業は生データ処理である。生データ処理とは得られたクロマトグラムのピーク領域を判定（ピークピッキング）し、面積値を求めることである (Figure 1-4)。この例では、ピーク面積値は、黒のシグナル線の下側と赤い直線の上側の面積を測定することで決まる。具体的には、黒のシグナル線から赤い線を引き算して得られたシグナルの各サンプリング点から矩形積分もしくは台形積分して求める。その際、赤い線は直線を仮定し、ピーク領域の重なりを許容しないことが多い。ピーク面積値はサンプル中の代謝中間体含量を反映していることから、ピークピッキングによって生体試料中の代謝中間体含量を知ることができる。ピーク面積以外にピーク高さを計算することもある。歪みのないガウス分布形状のピークの場合では、高さと面積値は代謝中間体含量に比例する。実際は、クロマトグラムの様々な要因により、歪んだ形状として溶出することが多い (Figure 1-4b)。この場合は、ピーク高さは代謝中間体含量に比例せず、ピーク面積値が代謝中間体含量に比例する。このため、本研究では面積値を扱うこととした。ピークピッキングは、非常に時間と労力を要する作業であり、高速化が求められている。1 人の作業者が扱える 20 サンプル（200 種類の代謝中間体含量）を測定することを考えると、4000 回のピークピッキングが必要である。

ピークピッキングにかかる時間を高速化するべく、これまでに様々な自動ピークピッキング手法が考案されてきた。ピーク形状がガウス分布であることを利用した方法では、ピークの勾配や変曲点が検出される。また、周波数解析技術も取り入れられている (Du et al., 2006, Lommen et al., 2009, Mo et al., 2010, Pluskal et al., 2010, Smith et al., 2006, Tautenhahn et al., 2008)。この方法は理想的なピークではうまくいく (Figure 1-4a)。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 (Figure 1-4b)。また、複数のピークが重なることもよくある (Figure 1-4c)。

これらのようなケースにおいては、従来法ではうまくピークピッキングできないことが多い。そこで、自動ピークピッキング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。経験的に約 2 割のピークを修正する必要がある。これらの手動によるピークピッキングには以下の 2 つの問題がある。

1 つ目は時間がかかる点である。4000 回のピークピッキングを行い、確認・手動修正するのに平均して 1 つ 6 秒かかるとすると約 6.5 時間が必要であると考えられる。これは、メタボロミクス研究を行う上でのボトルネックになる。

2 つ目は再現性の問題がある。作業者によって手動でピークピッキングする場合、明確な基準がないため、作業者によって結果が異なる。また、単純な見落としや操作ミスすることも避けられない。

そこで近年、ニューラルネットワークを利用し、深層学習技術を活用した、自動ピークピッキングの高性能化が試みられている (Melnikov et al., 2020, Woldegebriel et al., 2017)。その際の学習データ作成時に、再現性の問題が課題となっている。

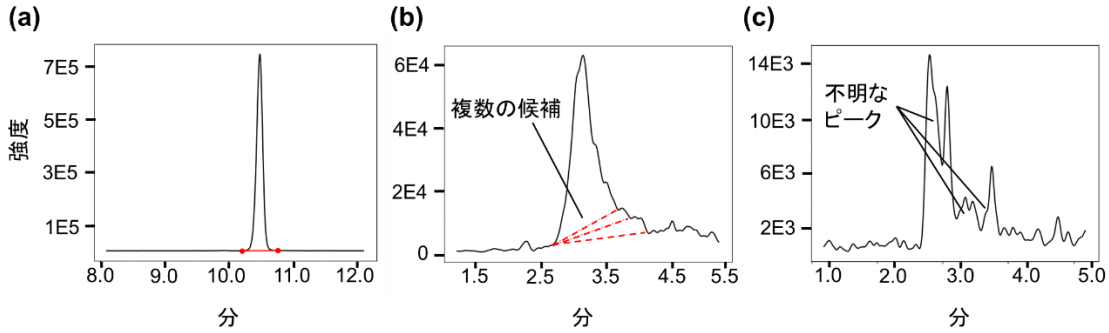


Figure 1-4 ピークピッキングの問題

(a) 理想的なピーク、(b) 歪んだピーク、(c) 複数のピークが重なった未分離ピーク（黒線がクロマトグラム、赤線がピーク領域）。(b) はテーリングにより、終了点の決め方が作業者により異なる。(c) は未分離ピークの微小な盛り上がりをピークとするかノイズとするかが作業者により異なる。

1-2-3 メタボロミクスで用いるデータ解析（統計処理）法

メタボロミクスの 3 番目の作業は、データ解析である。この作業は、比較したい群の間で特徴的に増減する成分を見つけ出すことが目的である。2 番目のピークピッキング作業が終わると、代謝中間体含量に対応したピーク面積値の表が得られる (Figure 1-5a)。例えば、20 サンプルから 200 代謝中間体含量を測定すると 20 行×200 列の表が得られる。データ解析では、この表を用いて、主成分分析、部分的最小二乗法、ボルケーノプロット法、階層的クラスター分析などの統計解析作業が行われる (Fan et al., 2020, Pan et

al., 2016, Zhang et al., 2016)。例えば、ボルケーノプロット法を行うことで、2 群間で有意に増減した代謝中間体を同定することができる (Figure 1-5b)。このように、ピークピッキングで得られた網羅的な測定データから、有意に増減した代謝物を探すデータマイニング法は数多くあり、オープンソースのコードやソフトウェアが普及していることから、このステップでプログラムの実行に要する時間はたった数秒である。したがって、データ解析段階は、大きなボトルネックではない。

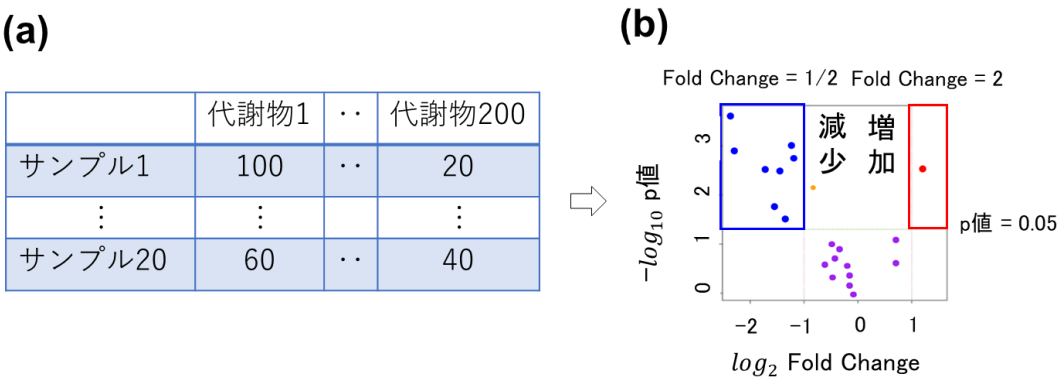


Figure 1-5　ボルケーノプロット法を用いた統計解析

(a) ピークピッキングで得られたピーク面積の表、(b) ボルケーノプロット法を適用した結果。

1-2-4 メタボロミクスにおけるキーワード連想

メタボロミクスの最後の作業は、網羅的な測定データから含量が変動した代謝中間体を同定し、生体メカニズムに関する仮説を生成することである。このために、次のようなキーワード連想作業が行われている。

例えば、抗がん剤の 4-ヒドロキシタモキシフェン (4-hydroxytamoxifen, 4-OHT) を処理したがん細胞、未処理のがん細胞を比較し、4-OHT 処理細胞で起きた生体内での出来事に関する仮説を得たいとする。そこで、これらの代謝中間体含量を測定、生データ処理し、統計解析を行い、コハク酸とクエン酸が 4-OHT 処理により有意に含量が変化していたとする。本研究では、このように見出した含量が変化した代謝中間体群を「代謝物リスト」と呼び、「代謝物リスト」中のそれぞれを「代謝物」と呼ぶ。また、この代謝物リストから、細胞内で起きた出来事のメカニズムを考察するキーワード連想作業に従

事する研究者を「解釈担当者」と呼ぶ。「解釈担当者」は抗がん剤 4-OHT を投与すると薬剤応答を示した後にアポトーシスに至るという先行研究 (Rafal et al., 2007) の知見があるとする。そこで、アポトーシスを「解釈担当者の既知のキーワード」と呼ぶ。「解釈担当者」は「解釈担当者の既知のキーワード」（アポトーシス）と「代謝物リスト」（コハク酸やクエン酸）との間をつなぐ生体メカニズムを調べることになる (Figure 1-6a)。

まず、解釈担当者は、「解釈担当者の既知のキーワード」の中から、アポトーシス、コハク酸やクエン酸と関連する生体分子やメカニズムの探索、つまり「キーワード連想」を行う。しかし、さまざまな代謝物とアポトーシスに関する知識を網羅することは困難である。このため、PubMed などの文献データベースを用いて「コハク酸 & アポトーシス」や「クエン酸 & アポトーシス」などをクエリとする検索をおこない、関連する論文を調査する。本研究ではこれを「論文検索」とよぶ。調査の結果、関連する論文によく出現する、複合体Ⅰ、活性酸素種 (Reactive Oxygen Species , ROS) 、カスパーゼ III などの生体分子やメカニズムが「キーワード」として見出されるだろう (Figure 1-6b)。このように文献上は既に知られた事実であるが、解釈担当者が知らないキーワードを「解釈担当者の知識外のキーワード」と本論文では呼ぶ。

この「キーワード連想」作業には次の問題がある。それは、時間と労力が必要な点である。例えば、「コハク酸 & アポトーシス」をクエリとして、生物医学系データベースの PubMed で論文検索を行うと、32 件（2020 年 6 月現在）の論文がヒットするが、要旨と本文を読んで概要をつかむのに、1 件あたり平均して 15 分かかるとすると、単純計算でも 8 時間と膨大な時間が必要とされることになる。この問題があるため、キーワード連想段階がメタボロミクスの大きなボトルネックになっている。そこでキーワード連想を自動化、支援する手法の開発が必要とされている。

キーワード連想の方法としては、代謝物エンリッチメント解析がある。代謝物エンリッチメント解析とは、注目する代謝物リストが、特定の代謝経路に有意に集中しているかどうかを統計的に解析する手法である。この手法は、マイクロアレイや RNAseq などの網羅的遺伝子発現データに対して、注目する遺伝子リストが特定の機能的分類に含まれるかどうかを統計検定する Gene enrichment 解析 (Subramanian et al., 2005) を代謝物へ応用したものである。遺伝子発現データの場合は、一回の解析ですべての mRNA の

データが得られるのに対し、代謝物の場合は計測対象が、原理上一部の限られた物性の代謝物のみとなるため、この解析はうまく機能しないという欠点がある。

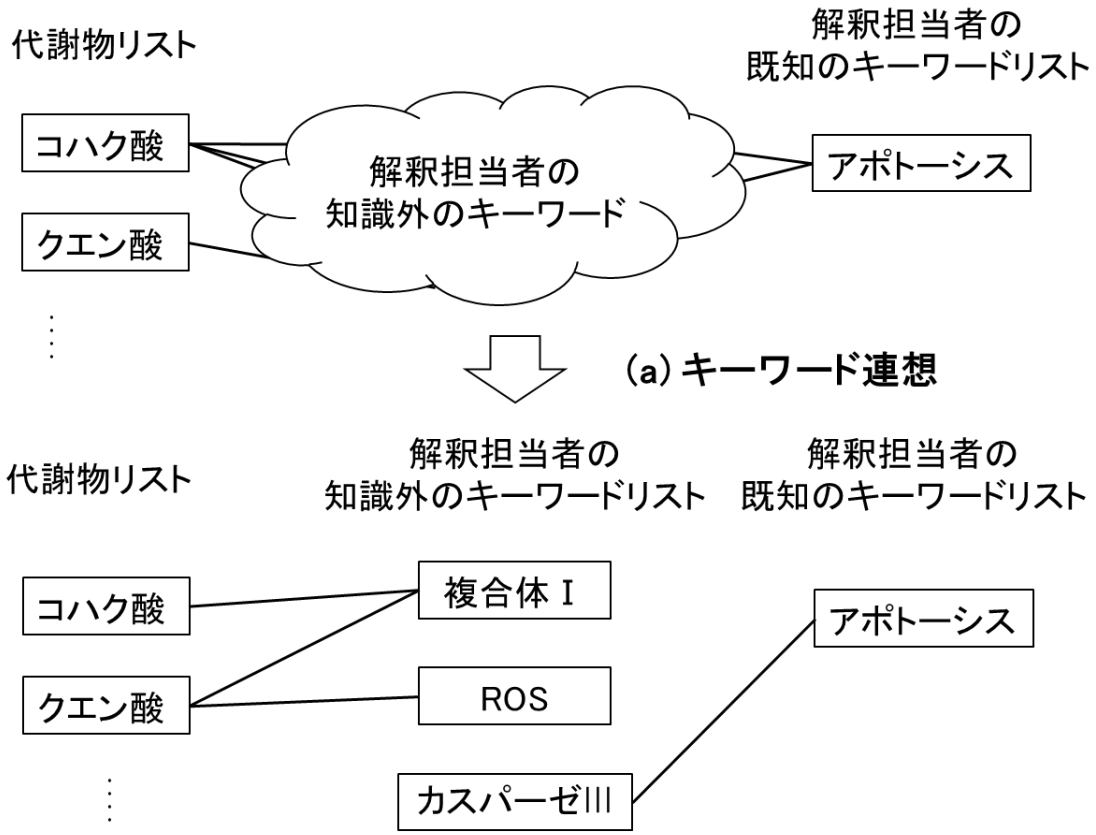


Figure 1-6 キーワード連想作業の例

(a) キーワード連想作業。測定されたデータから抽出した代謝物リスト（コハク酸とクエン酸）と解釈担当者の既知のキーワード（アポトーシス）から解釈担当者の知識外のキーワードリストを連想する。

1-2-5 現状のメタボロミクスにかかる時間と達成したいスループット

これまで 1-2-1 から 1-2-4 で見てきたように、現状のメタボロミクスでは、1 サンプルあたり実働時間が、１.データ取得には数秒（終夜運転・複数台測定可）、２.生データ処理には数時間オーダー、３.データ解析には数秒オーダー、４.キーワード連想には数時間オーダーの時間がかかることを述べた。LC-MS を用いたバイオマーカーの発見には、80～400 のサンプル数を扱う (Di Gangi et al., 2016, Rhee et al., 2011, Wang et al., 2011, Wang

et al., 2013)。1 人の作業者が扱える 20 サンプル（200 種類の代謝中間体含量）を測定することを考えると、下記の実働時間がかかる。

1. データ取得には、数秒で終わる。
2. 生データ処理には、20 サンプル、200 代謝物で 4,000 回のピークピッキングで、約 6.5 時間かかる。
3. データ解析は数秒で終わる。
4. 3 つの代謝物に注目すると、1 つの代謝物あたり 8 時間かかるので 24 時間かかる。

したがって、メタボロミクス作業の合計時間は、約 30.5 時間かかることになり、メタボロミクスのスループットは著しく低い。例えば、作業にかかる約 30.5 時間を 20 分の 1 にすることができれば、バイオマーカー研究が 20 倍に加速されることが期待される。そのためには、ボトルネックとなる生データ処理とデータ解析にかかる時間の約 30.5 時間を約 1.5 時間以内にする必要がある。

1-3 課題解決のアプローチ

上記のように、メタボロミクスの課題とは、

１．生データ処理（ピークピッキング）における、ピークピッキング法の自動化

２．「キーワード連想」の自動化

を通して、人がやったのと同じ信頼性を保ったままこれらのステップにかかる時間を約 30.5 時間から約 1.5 時間以内に短縮することである。そこで、本研究ではこの２つ課題を解決するため、「自動ピークピッキング法の性能向上」に向け、深層学習技術を活用したピークピッキングニューラルネットワークの構築を試みることとした。また、「キーワード連想」作業の自動化を目指し、共起情報に基づくアソシエーション分析を適用することとした。

1-3-1 深層学習技術を活用したピークピッキングニューラルネットワークの課題

ピークピッキングの時間がかかる問題を解決するためには、ピークピッキングの精度を向上させる必要がある。近年、ニューラルネットワークを利用した深層学習技術を活用した、自動ピークピッキングの高性能化が試みられている (Melnikov et al., 2020, Woldegebriel et al., 2017)。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ニューラルネットワークである。U-Net (Ronneberger et al., 2015) や

Single Shot MultiBox Detector (SSD) (Liu et al., 2016) と呼ばれるニューラルネットワークは、画像認識で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ピークピッキングニューラルネットワークの構築が試みられている。ピークピッキングに深層学習を活用する場合、人がピークピッキングした（ラベルをつけた）クロマトグラムを用いてニューラルネットワークを学習することで (Figure 1-7a)、学習済みモデルは、ピークピッキング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予測できるようになる (Figure 1-7b)。

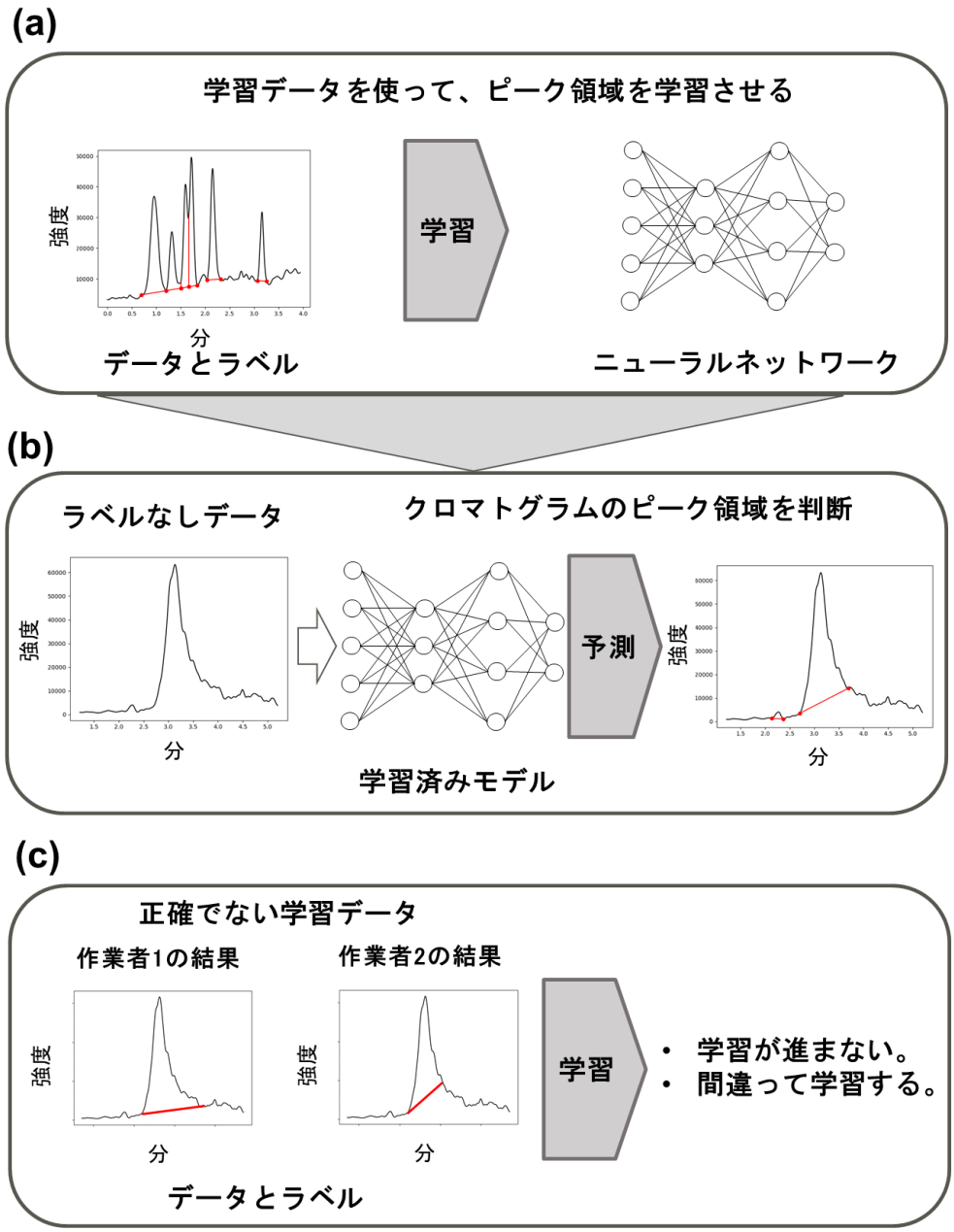


Figure 1-7 深層学習技術を活用したピークピッキングニューラルネットワーク

(a) 学習。ピークピッキング済みの学習データを用いて、ニューラルネットワークへピーク領域（ピークピッキングのやり方）を学習させる。(b) 予測。ラベルなし（ピークピッキングされていない）クロマトグラムを学習済みモデルへ入力することで、クロマトグラムをピークピッキングできるようになる。(c) 正確でないデータでの学習。学習が進まないもしくは間違って学習する。

一方、前述のように人がピークピッキングした学習データは正確性が不明なため、学習データとするのは不適切である。このため、「正確」にピーク領域がラベルされたクロマトグラムを学習データとして用意する必要がある (Figure 1-7c)。

近年、その解決策の一つとして、入力されたデータの特徴を含んだ疑似学習データを生成する敵対的生成ネットワーク (Generative Adversarial Network , GAN) と呼ばれる手法が検討されている (Goodfellow et al., 2014)。GAN は二つのニューラルネットワーク（疑似データを生成する生成器と疑似データを評価する識別器）を用いることで、実データに似た疑似データを生成することができる。具体的には、生成器は識別器に見破られないような疑似データを生成する。もし、識別器に見破られた場合には、より実データに近いものになるように生成器を学習させる。一方、識別器は疑似データもしくは実データのどちらかを分類する。もし、間違えた場合には正確に分類できるように学習する。このように、生成器と識別器は互いに学習しあうことで本物そっくりの疑似データを生成できる (Goodfellow et al., 2014, Yu et al., 2017)。クロマトグラムは成分由来のピークと溶媒由来のノイズ成分（ブランクサンプルクロマトグラム）から生成される。そこで疑似データ生成技術をピークに適用することで様々な形状のピークを生成することが可能である。さらに、生成したピークにブランクサンプルクロマトグラムを足し合わせることで、ピーク面積が既知なラベル付きクロマトグラムが生成できる。このアプローチによって、作成した疑似クロマトグラム生成器を用いてピークピッキングニューラルネットワークを学習させることで、手動での確認を必要としない正確さを担保しつつ、時間を数秒に短縮することを目指す。

1-3-2 共起情報に基づくアソシエーション分析の課題

キーワード連想作業の課題は、「解釈担当者の知識外のキーワード」を見つけるための論文検索に時間が非常にかかることである。本研究では、キーワード連想作業の自動化を目指した。

まず、キーワード連想作業を自動化するにあたって、PubMed の Medical Subject Headings (MeSH) を利用することにした。PubMed は世界で最もよく使用される生物医学系データベースであり、National Center for Biotechnology Information (NCBI) により開発および維持されている。2020 年 6 月現在では 5000 以上のジャーナルから、3000 万件

以上の文献が登録されている (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed)。PubMed の特徴として、文献の多さだけでなく、MeSH が挙げられる。

MeSH とは、生物医学系分野の様々なキーワードに対して一意な ID もしくはタームを割り当てたものである (http://www.nlm.nih.gov/mesh/meshhome.html)。例えば、アポトーシス、クエン酸というキーワードにはそれぞれ Apoptosis, Citric Acid というタームが割り当てられている。 PubMed には 2020 年 6 月現在約 3 万個の MeSH タームが登録されており、 PubMed に登録された各文献には、平均して 10 個程度の MeSH タームが紐づけられている。文献への紐づけ作業は、生物医学系の学士以上の学位を持つ専門メンバーらによって、文献の中身を読んだ上で行われている。また、MeSH タームはシソーラスの機能も兼ねている。例えば「プログラム細胞死」は「アポトーシス」の類義・同意語であるが、どちらも Apoptosis というタームが割り当てられている。このため、表記ゆれなどの問題を気にせずに、自分の知っているキーワードが紐づけられた文献を PubMed で検索することが可能となっている。例えば、クエン酸 (Citric Acid) とアポトーシス (Apoptosis) をキーワードとして論文検索（アンド検索）をすると、二つの MeSH タームの両方に紐づけられた論文として、Terasaki et al., 2018 などが得られる。

上記のように、PubMed は、キーワードと文献を紐づけるデータベースとして非常に良い特性を備えている。よって、PubMed の文献データベースに対してアソシエーション分析の手法を適用することで、MeSH ターム間の関連度を計算することができる。アソシエーション分析とは、売り上げ情報などの膨大なデータから、意味のある関連性を抽出し、関連性の強いものをレコメンドするために開発された手法である (Atluri et al., 2009, Chenel al., 2005)。例えば売り上げ情報の場合、ある２種の商品が、同一のレシート中に頻繁に出現（共起）する場合、関連性があると考えられる。本研究では、ある２種の MeSH タームが同一の論文に頻繁にキーワードとして付与（共起）されている場合、関連性があるとみなした。

これまでに、キーワード間の関連度を考える手法として、MeSH 間の意味的類似性をみる MeSH-Gram (Abdeddaïm et al., 2019) や MeSH 間の関連度を計算する MeSHSim (Zhou et al., 2015) が提案されている。しかし、メタボロミクスの「キーワード連想」では、「代謝物」と「解釈担当者の既知のキーワード」の間にあるキーワードを探索する手法は存在しない。そこで本研究では、「代謝物」と「解釈担当者の既知のキーワード」

の両者に関連度が高いキーワードを探索するキーワードレコメンド手法を開発することとした。

関連度などのスコアを用いた検索では、統計的に意味のある関連度スコアを判定する閾値を設定する必要がある。閾値を適切に設定できない場合、得られたキーワード連想に多くの誤り（偽陽性）を含むことになる。一般にデータベース検索では、キーワード間に全く関連がない場合に偶然起きてしまう関連度の集合（帰無分布）を設定し、偽陽性の出現率を推定することで、妥当な閾値の設定が行われている (Aggarwal et al.,2016, Wu et al.,2018)。つまり、キーワード連想手法を開発するには、妥当な閾値の設定法も必要となる。

1-4 本論文の目的

そこで、本研究はメタボロミクス作業の自動化を実現するために、次の２点を実現する手法の開発を行った。

１．自動ピークピッキング作業の正確さを熟練作業者並みとし、修正作業を不要にすること。

２．キーワード連想作業を自動化するキーワードレコメンド手法を開発して、キーワード連想作業の属人性を解消し、作業時間を短縮すること。

1-5 本論文の構成

本学位論文では、メタボロミクス自動化に向けた高性能生データ処理とキーワードレコメンド法を目的とし、4 章より構成される（Figure 1-8）。

第 1 章では、本研究の背景と目的、本論文の構成について記述した。

第 2 章では、ピークピッキングには時間と労力がかかるという課題と担当者により結果が異なるという再現性の課題を解決するため、自動ピークピッキング作業の正確さを熟練作業者並みとし、修正作業を不要にすることを目的とする。そこで、近年活用されている深層学習を用いたピークピッキングニューラルネットワークを利用する。しかし、深層学習では、正確にラベル付けされた（ピークピッキング済み）学習データを用意できないとう課題がある。そこで正確なラベル付けされた学習データを用いた高性能なピークピッキングニューラルネットワークを開発するため、下記 2 点を行う。

1. 疑似クロマトグラムの生成により正確にラベル付けされた学習データを準備する手法を開発すること。
2. 熟練した作業者と同等の性能を持つピークピッキングニューラルネットワークを構築すること（Figure 1-8、第 2 章)。

第 3 章では、第 2 章で得られた網羅的な測定データを利用する。網羅的な測定データを統計解析した結果からキーワードを連想する際、解析担当者の知識に依存することや時間と労力を要することが課題である。それを解決するため、下記 2 点を行う。

1. 「代謝物リスト」と「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」をレコメンドする手法を開発すること。
2. 「解釈担当者の知識外のキーワード」の妥当な閾値を設定する手法を確立すること (Figure 1-8、第 3 章)。

第 4 章では、本研究で開発した自動ピークピッキング手法とキーワードレコメンド法の汎用性と今後の課題、メタボロミクスへの寄与について述べ、結言とした。

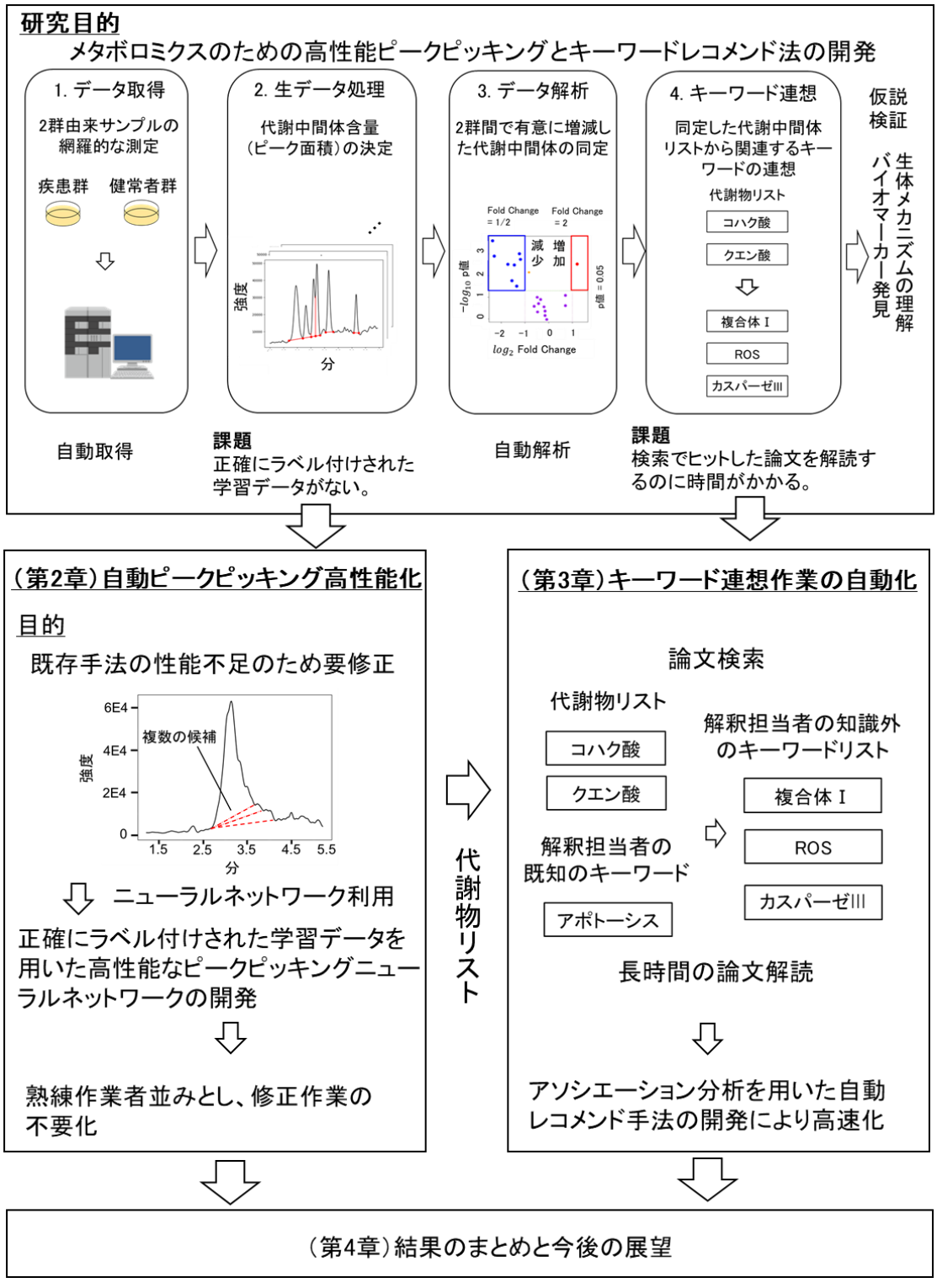


Figure 1-8 本論文の概要

第2章 熟練作業者に匹敵する自動ピークピッキング法開発

2-1 緒言

クロマトグラフィーを用いた化合物の定量分析は、バイオテクノロジーの基盤技術である (Matsuda et al., 2017)。クロマトグラムからピークを検出し、その開始点と終了点を決定すること（ピークピッキング）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 (Matsuda, 2016)。第 1 章で述べたように、ピークピッキングはメタボロミクスを効率的に実施するためのボトルネックとなっている。

これまでにもウェーブレット変換手法に基づく自動ピークピッキングツールが開発され、データ処理ソフトウェア（例:XCMS と i-PeakFinder）に実装されている (Du et al., 2006, Lommen et al., 2009, Mo et al., 2010, Pluskal et al., 2010, Smith et al., 2006, Tautenhahn et al., 2008,)。ウェーブレット変換に基づく自動ピークピッキング手法では、クロマトグラムを異なる平滑化具合の二次微分波形へ変換する (Figure 2-1)。それらの波形のうち曲率の高いところ（二次微分波形の最大値）をピーク位置として検出する。XCMS では、ピーク位置に近い変曲点（二次微分波形のゼロクロス位置）を開始時間と終了時間とし、ピーク開始強度とピーク終了強度を 0 とする。i-PeakFinder では、ノイズ値よりピークの立ち上がりと立下りの位置を推定し、ピークの開始点（時間と強度）と終了点（時間と強度）を求める。そして、クロマトグラムからピーク開始点と終了点の直線を引き算してピーク面積を測定する。

前述のツールでは、理想的なピークではうまくいくが、実際のクロマトグラムのピーク検出には失敗することが多い。実際のクロマトグラムでは、化合物ピーク間の分離が不十分なために、歪んだピークや重なり合ったピークが含まれているからである。そのため、熟練者による手動での修正が必要である。

第 1 章でも述べたように、手動によるピークピッキングには多量の労力が必要である。例えば、LC-MS を用いたバイオマーカーの発見には、80～400 のサンプル数を扱う (Di Gangi et al., 2016, Rhee et al., 2011, Wang et al., 2011, Wang et al., 2013)。1 人の作業者が扱える 20 サンプル（200 種類の代謝中間体含量）を測定することを考えると、4000 回のピークピッキングを行う必要があり、これらの確認・修正に平均して 1 つ 6 秒かかると約 6.5 時間の作業が必要である。

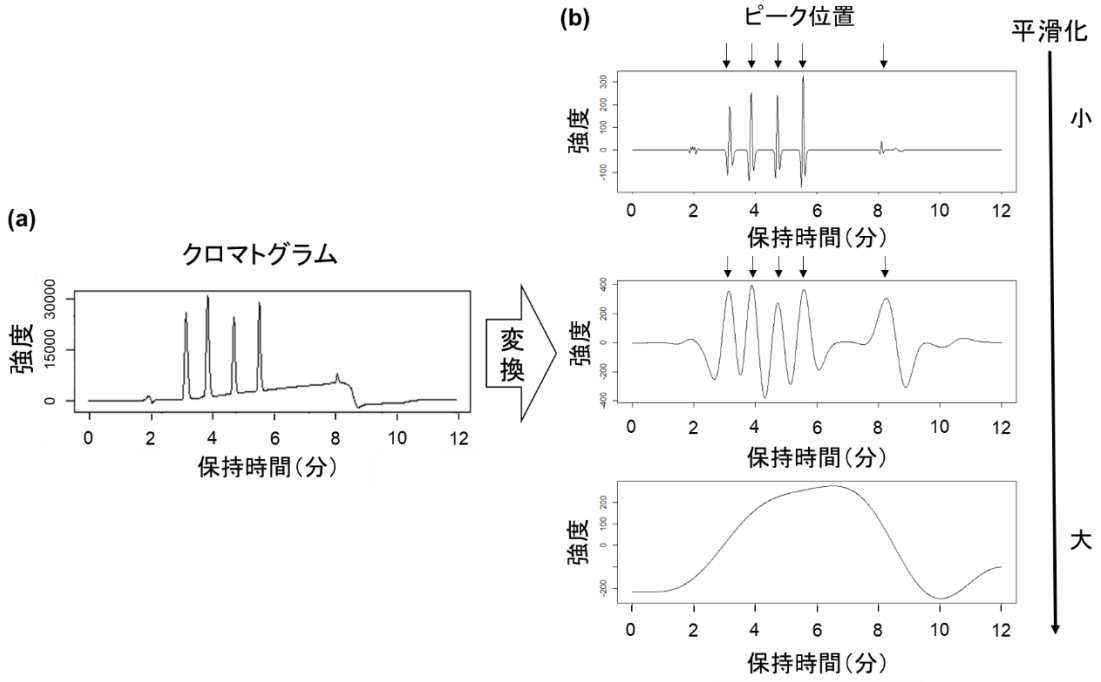


Figure 2-1　ウェーブレット変換手法に基づく自動ピークピッキング

(a) ウェーブレット変換前のクロマトグラム、(b) ウェーブレット変換後のクロマトグラム。ウェーブレット変換により、平滑化具合を変えた 2 次微分波形が得られる。平滑化具合が小さい波形から大きい波形の曲率が高いところ（2 次微分波形の最大値）を辿ることでピーク位置を検出する。

そこで、近年、ピークピッキングニューラルネットワークを用いた機械学習システムが報告されている (Melnikov et al., 2020, Woldegebriel et al., 2017)。Melnikov らの例では、ピーク開始強度とピーク終了強度を 0 として、ピーク開始時間とピーク終了時間をピークピッキングニューラルネットワークに学習させる。ピークピッキングニューラルネットワークの出力は、ピーク開始時間とピーク終了時間であり、それぞれの強度値を 0 として、ピーク面積値を計算する。これまでの報告例では、手動でピークピッキングされたクロマトグラムを学習データとして使用している。しかし、正確にピークピックされたクロマトグラム（ラベル付きクロマトグラム）の作成は、深層学習におけるボトルネックである。なぜなら、正確な開始点と終了点（真値）は、実クロマトグラムの複数のピークが重なった未分離ピークにおいてしばしば不明瞭だからである。第 1 章でも議論したように、例えば、Figure 1-4 のような場合、異なる作業者がピークピッキングを行

うと、異なる結果が得られ、そのどちらが正しいのかを判別することは難しい。クロマトグラムは成分由来のピークと溶媒由来のノイズ成分（ブランクサンプルクロマトグラム）から生成される。また、クロマトグラムでは成分由来のピークが歪むことがある。このため、Figure 1-4 (b) の場合、ピークの裾の部分がなだらかになり、そこにブランクサンプルクロマトグラム由来のノイズが重なることにより、ピーク終了点の判別が難しくなる。また、Figure 1-4 (c) の場合、複数のピークが重なることで、ピークによる山なのかブランクサンプルクロマトグラム中のノイズなのか判断がつかなくなる。その他、単純な見落としや操作ミスすることも避けられない。これは、深層学習を用いた自動ピークピッキングの学習データ作成時の課題となっている。

その解決策の一つとして、入力されたデータの特徴を含んだ疑似学習データを生成する GAN と呼ばれる手法が検討されている (Goodfellow et al., 2014)。GAN は二つのニューラルネットワーク（疑似データを生成する生成器と疑似データを評価する識別器）を用いることで、実データに似た疑似データを生成することができる。具体的には、生成器は識別器に見破られないような疑似データを生成する。もし、識別器に見破られた場合には、より実データに近いものになるように生成器を学習させる。一方、識別器は疑似データもしくは実データのどちらかを分類する。もし、間違えた場合には正確に分類できるように学習する。このように、生成器と識別器は互いに学習しあうことで本物そっくりの疑似データを生成できる (Goodfellow et al., 2014, Yu et al., 2017)。

また上述のように、クロマトグラムとは、さまざまな形状のピークが時間軸上に並んだものに、ノイズを足し合わせたものと考えることができる。つまり、本物そっくりの疑似ピークを生成し、配置した各ピークの開始点、終了点は正確に決定できる。これにノイズを足し合わせれば、正確な開始点、終了点でラベルされた疑似クロマトグラムが生成できる。そこで本章の目的を下記に２点に設定した。

1. 疑似クロマトグラムの生成により正確にラベル付けされた学習データを準備する手法を開発すること。
２. 熟練した作業者と同等の性能を持つピークピッキングニューラルネットワークを構築すること。

2-2 実験材料と実験方法

2-2-1 一次代謝産物のターゲットメタボロミクスによる実クロマトグラムデータ準備

市販の清酒、ビール、ワイン、および生物学的試料の分析試料は、文献に従って調製した (Kubo et al., 2011, Ogura et al., 2013)。試料は、LC/MS/MS メソッドパッケージ一次代謝物 ver. 2 (Shimadzu) のイオン対試薬と PFPP カラムと LC-MS (Nexera X2 システムとトリプル四重極型質量分析計 LCMS-8050, Shimadzu) を用いてメタボロミクスデータ（99 種類）を取得した。LabSolution (ver 5.97, Shimadzu) を用いて、アミノ酸、有機酸、ヌクレオチドを含む代謝物の 12,411 の実クロマトグラムを収集した。得られた実クロマトグラムピークを学習用 (90%) とテスト用 (10%) に分割した。同じ方法でブランクサンプル分析を繰り返し行い、全 7,615 ブランクサンプルクロマトグラムを得た。作成したブランクサンプルのクロマトグラムを学習用 (80%) とテスト用 (20%) に分割した。

2-2-2 ピーク形状生成器

GAN (Goodfellow et al., 2014) を用いてピーク形状生成器を構築した。ピークの高さおよび幅に関する情報は、疑似クロマトグラム生成器に加えられた。GAN は二つのニューラルネットワーク（疑似データを生成する生成器と疑似データを評価する識別器）を用いて実データに似た新しい疑似データを生成する。本研究では、歪んだピーク形状を生成するために GAN を用いた。時間の歪み関数 D(t)を生成するため、生成器を構築した。D(t)を用いて、ピークを $f(D(t)):= \exp(-D(t)^2)$ として表す (Figure 2-2a)。例えば、D(t) = t の場合、正規分布の曲線になる。D(t) = t ＋ Δt（Δt = 0, t < 0, Δt = -1/4t, t >= 0）のとき、テーリングで歪んだピーク形状が表現できる (Figure 2-2b)。例えば、t=4 を 2 つの場合で考えると D(t) = t のとき D(t) = 4 であるのに対し、D(t) = t ＋ Δt（Δt = 0, t < 0, Δt = -1/4t, t >= 0）のとき、D(t) = 2 であり、時間 2 だけ遅れている。この時間の遅れがテーリングを表現し、歪んだ形状のピークになる。

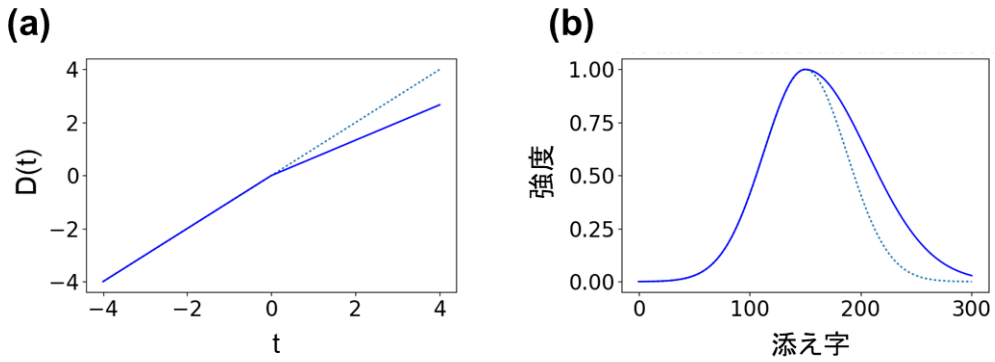


Figure 2-2 時間の歪み関数

(a) 時間の歪み関数 D(t)、 (b) 疑似ピーク f(D(t)):= exp(-D(t)^2)。D(t) = t の場合が点線であり、D(t) = t ＋ Δt（Δt = 0, t < 0, Δt = -1/4t, t >= 0）が実線である。

GAN では、生成器は画像または時系列データを生成するが、提案した方法は時間歪み関数 D(t)を生成する。ここでは、実際の単一ピークは 0 から 1 の範囲の強度値に正規化された。識別器は、正規化された実単一ピークと生成器から出力されたピーク関数 f(D(t))を分類する。生成器および識別器は、実際のメタボロミクスデータから取得した 665 個の実単一ピークデータにより学習された (Figure 2-3)。665 個の単一ピーク（他のピークと重複しないピーク）の抽出を熟練者が行った。学習方法は次のステップで行われる。まず、生成器は識別器に見破られないような疑似ピークを生成する。もし、識別器に見破られた場合には、より実ピークになるように生成器を学習させる。一方、識別器は疑似ピークもしくは実ピークのどちらかを分類する。もし、間違えた場合には正確に分類できるように学習する。このように、生成器と識別器は互いに学習しあう。

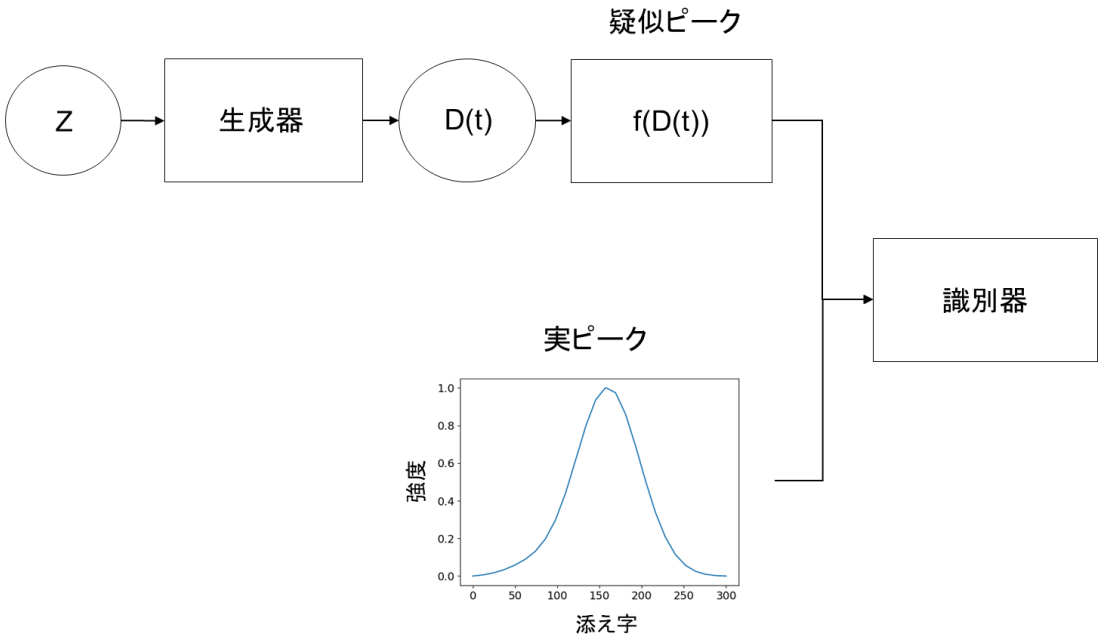


Figure 2-3　疑似ピークの学習

潜在変数 z を用いて時間の歪み関数 D(t)を生成する、時間の歪み関数から生成した疑似ピーク $f(D(t)) := \exp(-D(t)^2)$と実ピークを用いて、疑似ピーク生成器と疑似ピーク識別器の学習を行う。

生成器は、Figure 2-4 に示すように、三次元標準正規分布 z からピークを発生させる。入力層において、潜在変数 z を 1×301×3 次元の潜在変数 zz まで拡張し、1×301×1 次元の時間 t:(-4, ..., 4)とコピーと結合を行った。1×301×4 次元の初期情報 h は、潜在変数 zz と時間 t を結合したものである。隠れ層では、 初期情報 h は 1×301×30 次元の出力を持つ全結合層に入力される。全結合層とはニューラルネットワークのノード同士を全て結合したものである。ピーク形状は正規分布であることが多いので、全結合層の出力と初期情報 h をコピーおよび結合した。4 つの同じ隠れ層を用いた。出力層では、時間 t と隠れ層の出力との時間差 Δt（1×301×1 次元）を出力する。活性化関数として Exponential Linear Unit (ELU)を用いた。

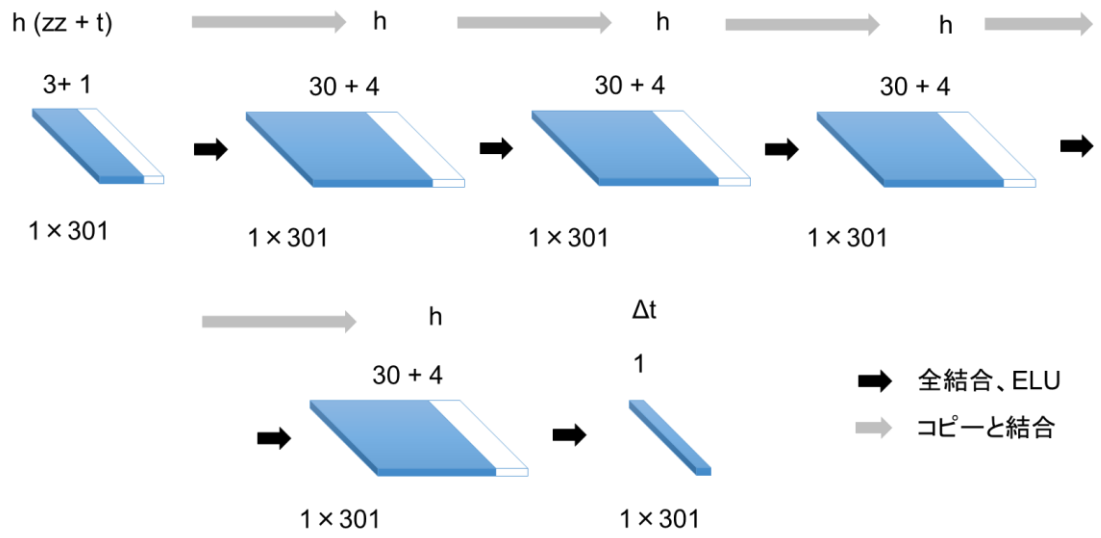


Figure 2-4　ピーク形状生成器のネットワーク構成

拡張された潜在変数 zz が潜在変数 z を 1×301×3 次元まで拡張したもの。時間 t が 1×301×1 次元の時間 t:(-4, ..., 4)。初期情報 h が拡張された潜在変数 zz と時間 t を結合したもの。 Δt が時間 t との差分である。4 つの隠れ層をもち、活性化関数として ELU を用いた。

ニューラルネットワークは複数のノードから構成される (Figure 2-5a)。1 つのノードは、それぞれの入力値に重みを掛けたものとバイアス項との線形和に対して、活性化関数を適用したもので構成される (Figure 2-5b)。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、1 つ 1 つのノードがニューロンと呼ばれる神経細胞に対応する。活性化関数や複数のノードがあることで、複雑な関数を表現することができる。活性化関数 ELU は、入力値をxとすると、

$$\mathrm{f}(x) = \begin{cases} e^x - 1, & x < 0 \\ x, & x \geq 0 \end{cases}$$

の式で表される。

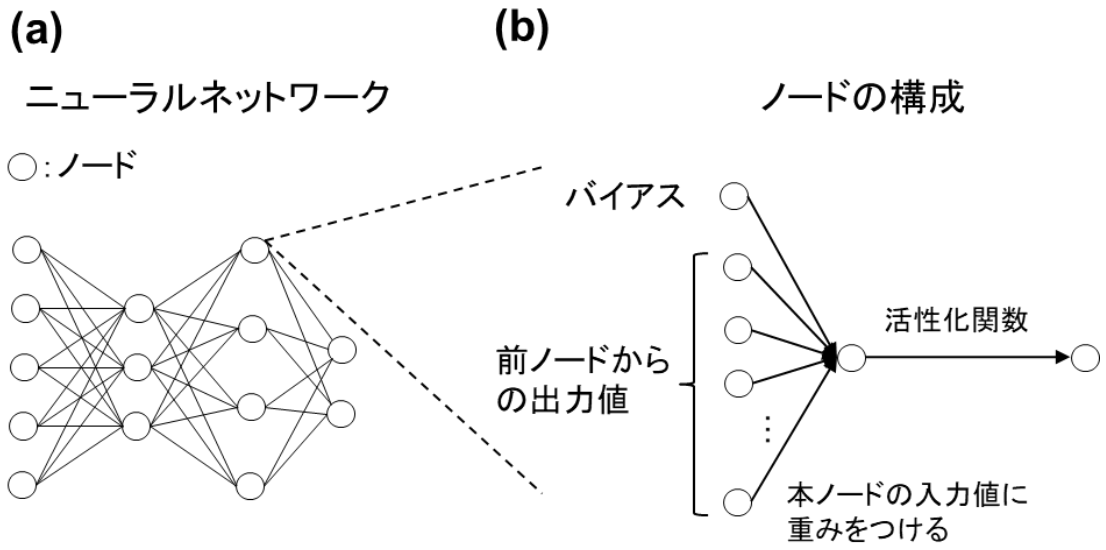


Figure 2-5　ニューラルネットワークの構成

(a) ニューラルネットワーク。複数のノードから構成される。(b) ノードの構成。1 つのノードは、全ノードからの出力値にそれぞれ重みを掛けたものとバイアス項との線形和に対して、活性化関数を適用したもので構成される。

潜在変数と生成ピークを 2 段階で評価した。1. 潜在変数 z がテストデータのピークに近いかどうかを調べた (Figure 2-6)。距離の指標として最小二乗誤差を用い、テストデータに近い潜在変数 z を見つける方法として勾配法を用いた。損失関数は、最小二乗誤差と正規化項 λ* $\sum_i z_i^2$ との和である。実験結果に基づき、λ として 1.0E-4 を用いた。さらに、潜在変数 z が標準正規分布に従うかにおいても評価した。

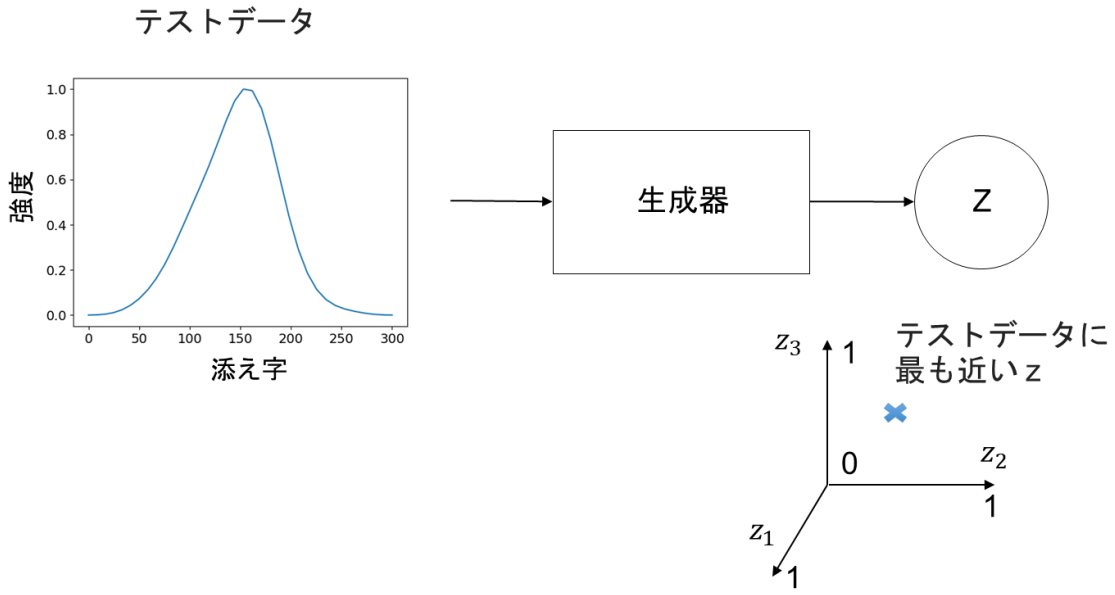


Figure 2-6　潜在変数 z の評価方法

潜在変数 z から疑似ピーク生成器から生成したデータとテストデータの最小二乗誤差を計算し、勾配法を用いて最適な潜在変数 z を探索した。

831 個の単一ピーク（他のピークと重複しないピーク）の抽出を熟練者が行った。得られた実際の単一ピークを学習 (80%) とテスト (20%) 目的のデータに分割した。生成器および識別器は、実メタボロミクスデータから取得した 665 の実単一ピークデータにより学習した。勾配法を用いて、損失関数を 1.0E-4 * $\sum_i z_i^2$の和として表し、テストデータのピークに近い潜在変数 z を求める。学習モデルから生成したピークとテストピークの間の平均二乗誤差は 99.4%が 1.87 E-3 以下であった (Figure 2-7a, b)。潜在変数 z が標準正規分布に従わない場合、最適化手法または損失関数が不適切である。テストデータに対して最適化した標準正規分布に従う潜在変数 z をプロットした。3 次元の標準正規分布に従う潜在変数 z は、中心の点が多く、中心から離れると点が少なくなる標準正規分布からサンプリングされていることがわかった (Figure 2-7c)。

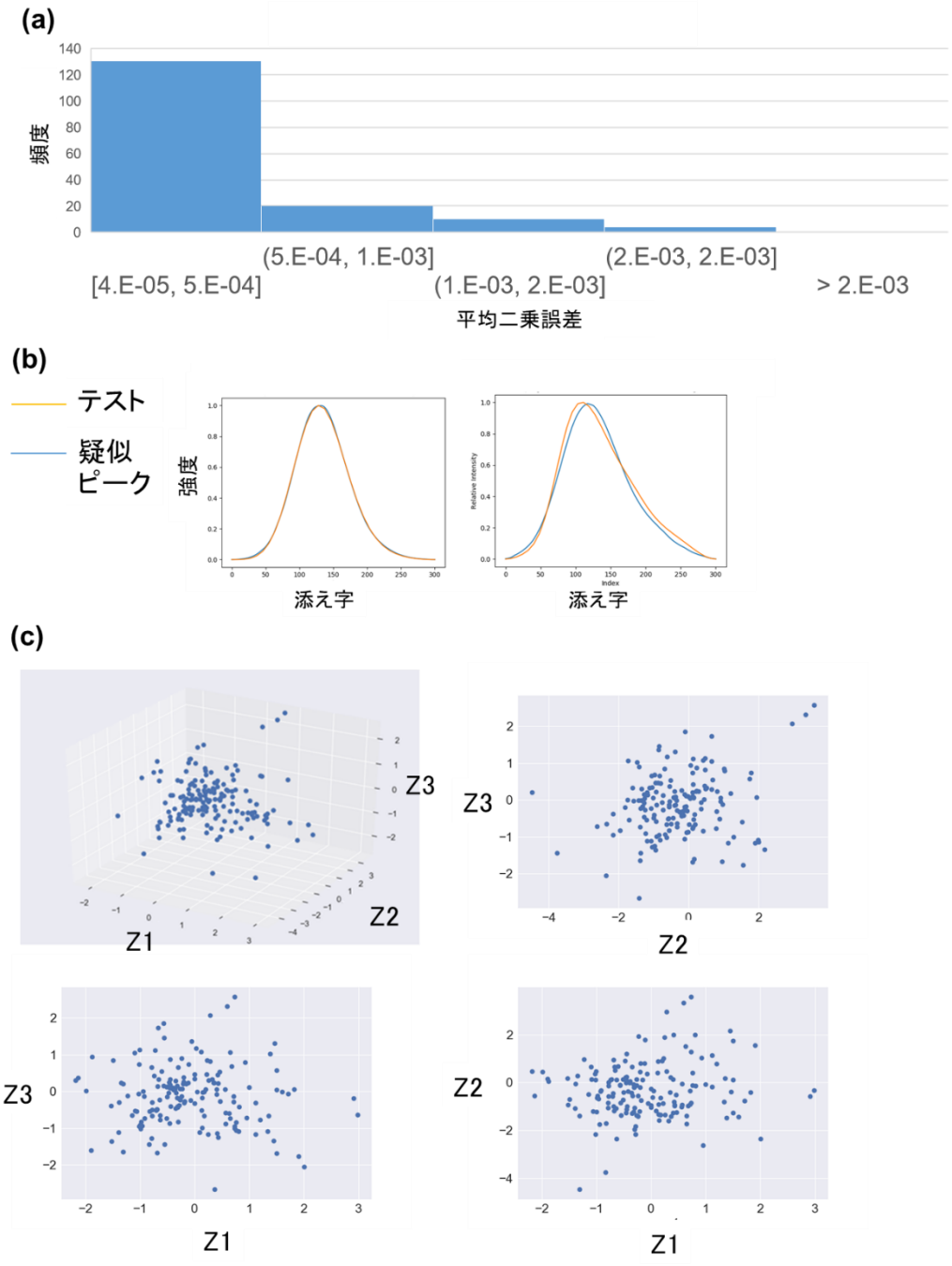


Figure 2-7　潜在変数 z の評価例

(a) 学習済みモデルから生成された疑似ピークとテストピークの平均二乗誤差、(b) 疑似ピークとテストピークの例（平均二乗誤差が小さい 99.4%の中で、左：最も小さい平均二乗誤差 1.20E-4、 右：最も大きい平均二乗誤差 1.84E-3）、(c) テストデータに対応する潜在変数 z の 3 次元散布図と各軸に射影した図。

2. 潜在変数 z から発生するピークの形状を調べた。テストデータから形状の異なる 2 つのピークを選択し、中間ピークが生成できることを確認した。具体的には、データ A のピーク近傍にピークを発生する潜在変数z_Aと、データ B のピーク近傍にピークを発生する潜在変数z_Bとを用いて、潜在変数空間上をz_Aからz_Bに移動する際に発生するピークが中間ピークであるか否かを確認した (Figure 2-8)。

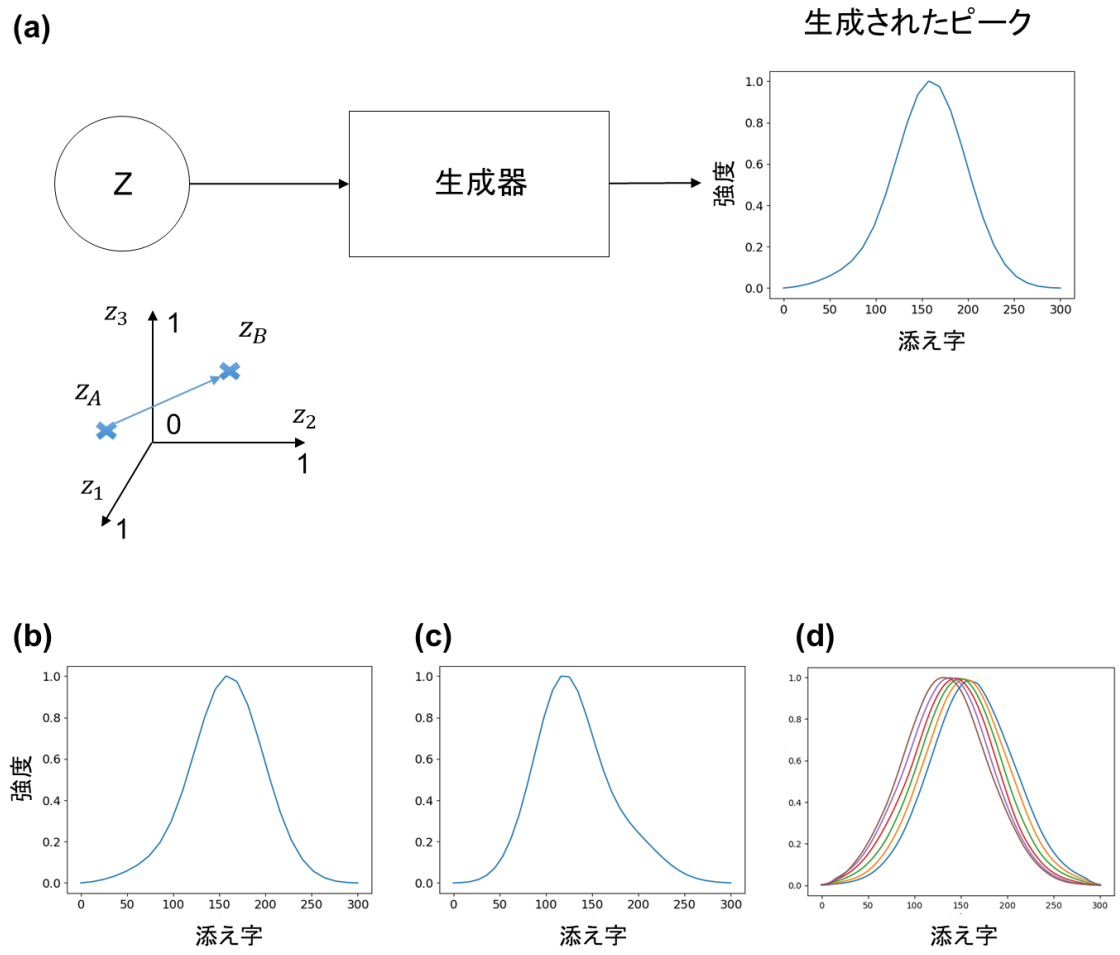


Figure 2-8　生成されたピークの評価方法と評価例

(a) 評価方法、(b) 理想的なピーク、(c) テーリングにより歪んだピーク、(d) (b)理想的なピークを生成する潜在変数z_Aから(c)歪んだピークを生成する潜在変数z_Bを移動させて生成された中間ピークの例

2-2-3 疑似クロマトグラム生成器

疑似クロマトグラム生成器は、ピーク形状生成器と特徴量分布、ブランクサンプルクロマトグラムから生成する。そこで、自作で作成した Python スクリプトを用いて、11,011 個の実クロマトグラムデータから、5 つのクロマトグラムに関連するパラメータ（クロマトグラムのピーク数、中心からのピーク距離、ピーク間距離、信号対ノイズ比 (S/N) 、半値全幅 (full width at half maximum, FWHM)）の特徴量の頻度分布「特徴量分布」を得た (Figure 2-9)。

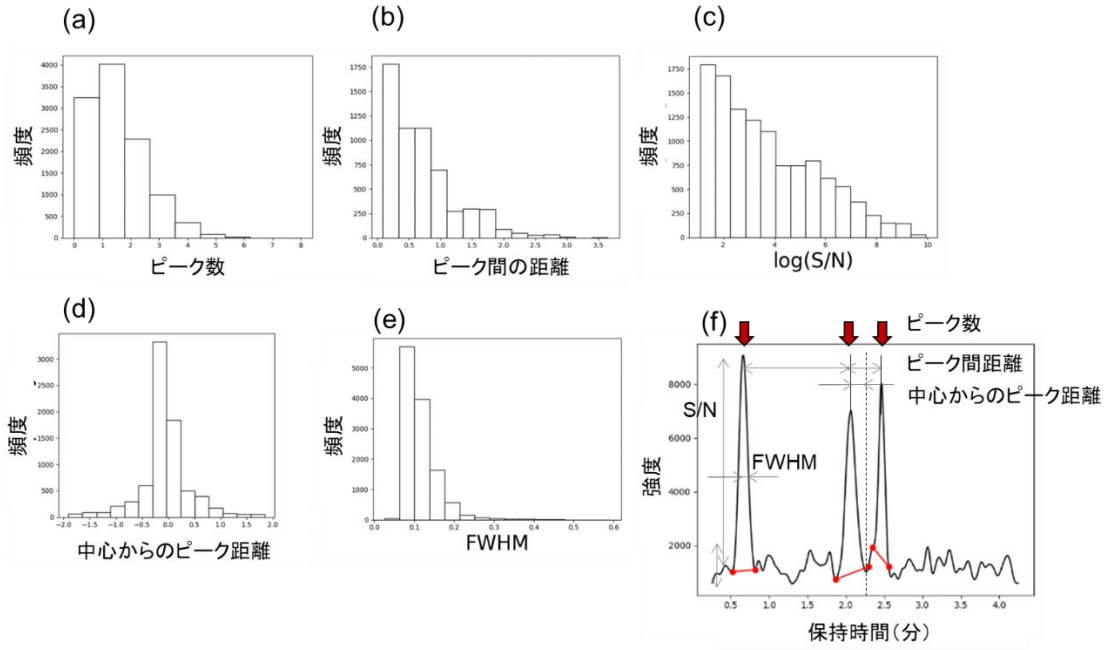


Figure 2-9　5 つのクロマトグラム特徴量分布

(a) クロマトグラムのピーク数、(b) ピーク間距離、(c) 信号対ノイズ比 (S/N)、(d) 中心からのピーク距離、(e) 半値全幅 (FWHM)、(f) 5 つのクロマトグラム特徴量の説明

ノイズの定義は、ASTM International (https://www.astm.org/) の規格を用いた。具体的には、0.5 分ごとに最小二乗法により直線を求める。その直線を上下に平行移動させ、その直線間に各区間のクロマトグラムが収まるように平行線を決定する。そして、求まった平行線の強度方向の距離を全区間で平均したものをノイズとする (Figure 2-10)。

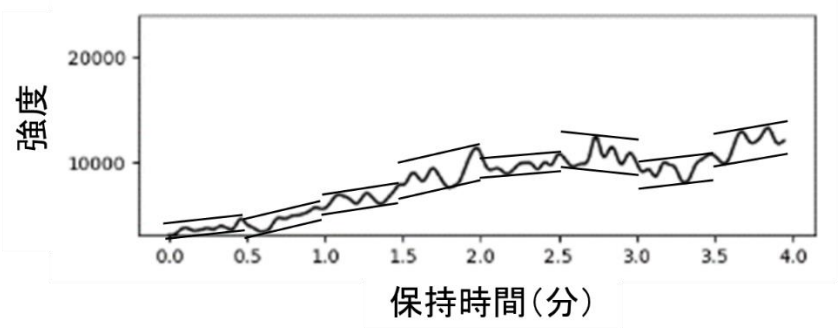


Figure 2-10　ASTM 法でノイズを計算するイメージ図

0.0 分から 4.0 分までを 0.5 分刻みでノイズを計算した場合の図である。0.5 分間隔の区分に分割して、それぞれの領域で平行線を求める。

ピーク形状生成器、特徴量分布、及びブランクサンプルクロマトグラムを用いて、以下のステップにより疑似クロマトグラムを生成した (Figure 2-11)。(1) ピーク位置は、3 つの特徴量分布から、パラメータを無作為抽出することで決定した。(2) 二つの特徴量分布からパラメータを無作為抽出することによりピーク高さと幅を決め、ピーク形状生成器から出力されたピークを拡大もしくは縮小し高さと幅を調整し、疑似ピークを生成した。(3) 疑似ピークを整列させてノイズレスピークを生成した。(4) 各ピークについて正確な開始点及び終了点を特定した。ピーク面積の誤差を最小化することにより、未分離ピークを分割した。(5) ノイズレスピークとブランクサンプルクロマトグラムを統合して疑似クロマトグラムを生成した。ノイズレベルは一定と仮定し、正確な開始時間と終了時間は変えず、正確な開始強度と終了強度はブランクサンプルクロマトグラム（真値）上の対応する時間の強度とした。

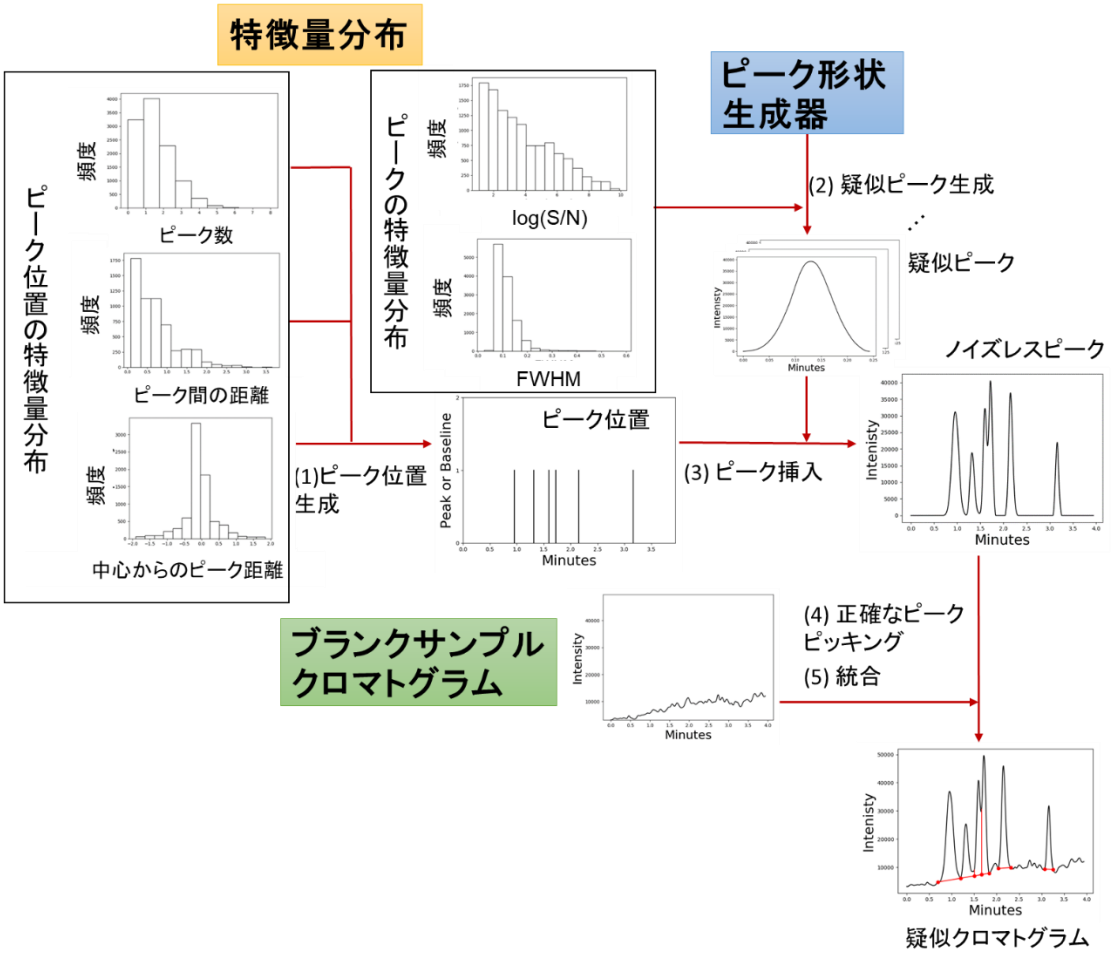


Figure 2-11　疑似クロマトグラム生成器の構築

(1) ピーク位置生成。3 つの特徴量分布からパラメータを無作為抽出してピーク位置を決める。(2) 疑似ピーク生成。ピーク形状生成器からピーク形状を決定し、ピークの特徴量分布からパラメータを無作為抽出してピーク高さとピーク幅を決定し、疑似ピークを生成する。(3) ピーク挿入。生成したピーク位置に生成した疑似ピークを挿入し、ノイズレスピークを生成する。(4) ピーク面積が既知なので、誤差が小さくなるように、未分離ピークを分割する。(5) 統合。ノイズレスピークとブランクサンプルクロマトグラムを統合する。

疑似クロマト生成器を用いて、6,088 個のノイズレスピークを生成し、6,088 個のブランクサンプルクロマトグラムと統合して擬似クロマトグラムを生成した。これを 10 回の手順を繰り返すことにより 60,880 個の疑似クロマトグラムを生成した。生成した疑

似クロマトグラムの特徴量分布は実クロマトグラムの特徴量分布と似ていることを確認した (Figure 2-12)。

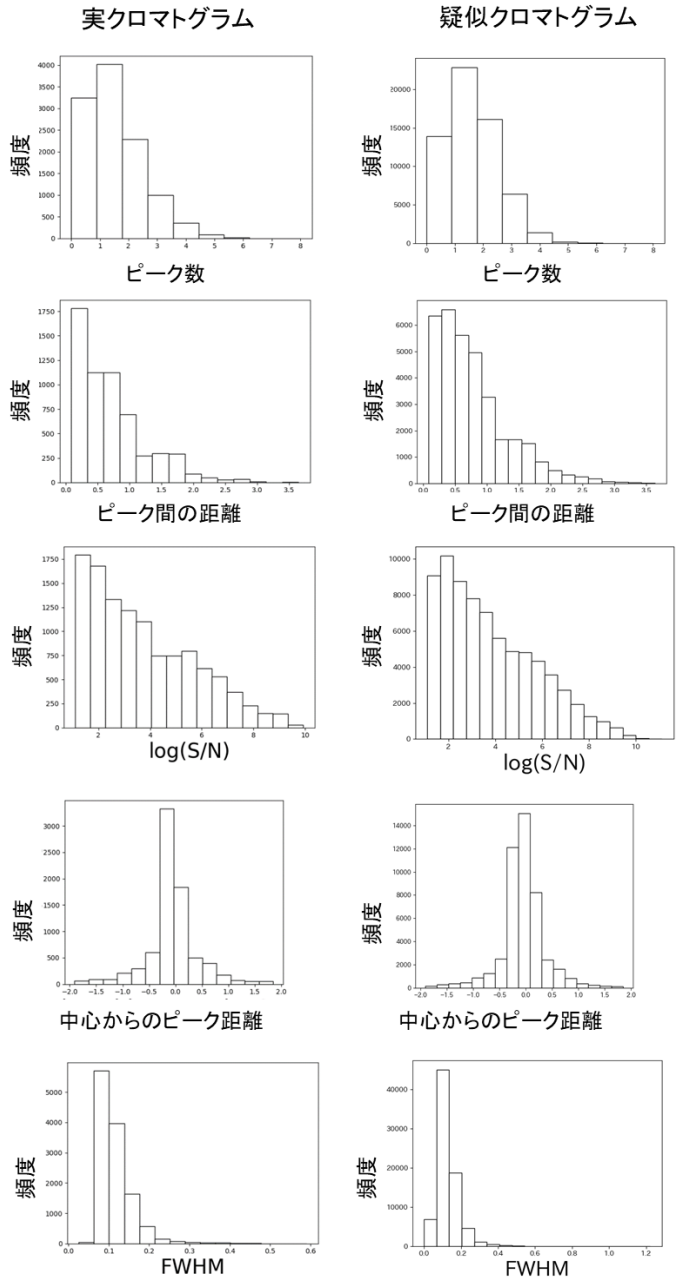


Figure 2-12　実クロマトグラムと疑似クロマトグラムの特徴量分布

2-2-4 U-Net を用いたピークピッキングニューラルネットワークの構築

PeakOnly (Melnikov et al., 2020) は、セマンティックセグメンテーションの問題として、ピークピッキングの問題を定式化した。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。PeakOnly ではピークのみを扱うが、本研究ではクロマトグラム全体を入力として扱い、ベースラインのトレンドを学習できるようにした。その他異なる点は、クロマトグラムとクロマトグラムの対数値に加えて、ピークの開始点と終了点の位置を容易に決定するためにクロマトグラムの一次微分を加え、ピークトップの位置を容易に決定するために二次微分を加えたことである。具体的には、1,024 点になるようにアップサンプリングしたクロマトグラム$\mathbf{x}(\mathrm{n})\ (\mathrm{n} = 1, 2, \ldots, 1024)$を用いた。4 つの 1 次元時系列データとして、元波形$\mathbf{x}(\mathrm{n})$と一次微分波形$x^{(1)}(\mathrm{n}) = \mathrm{x}(\mathrm{n}) - \mathrm{x}(\mathrm{n}-1)$、二次微分波形$x^{(2)}(\mathrm{n}) = x^{(1)}(\mathrm{n}) - x^{(1)}(\mathrm{n}-1)$、対数波形$\log(\mathrm{x}(\mathrm{n}) + 1)$を、-1 から 1 の強度に正規化 (元波形の場合、$\frac{\mathbf{x(n)}}{\mathbf{max(x(n))}} \times \mathbf{2} - \mathbf{1}$) したもの（4 チャンネルの 1 次元時系列データ）を入力値とした。出力値は、クロマトグラムを 5 つのクラスに分類した（0:ピークなし、1:単一ピーク、2:未分離ピーク、3:始点、4:分離点または終点）。具体的には、出力値$\mathbf{y}(\mathrm{n})\ (\mathrm{n} = 1, 2, \ldots, 1024)$は、5 つのラベルの確率のマトリックス (1024×5) である（0:ピークなし$y_0(n)$、1:単一ピーク$y_1(n)$、2:未分離ピーク$y_2(n)$、3:始点$y_3(n)$、4:分離点または終点$y_4(n)$）のうち最大の確率であるラベル$\mathbf{y}(\mathrm{n}) = \max(y_i(n)), (\mathrm{i} = 0,1,..,4)$にまとめて、配列 (1024×1) にした。ここでは、i チャンネルの波形を$y_i(n)$として表記した。$y_i(n)$は横軸が時間で縦軸が確率である。出力値は、ピーク開始時間とピーク終了時間であるため、それぞれのピークの強度値は、クロマトグラム上の点とした。但し、未分離ピークの分割点の強度値は、未分離ピークの開始点と終了点を直線で結んだ強度値とした (Figure 2-13)。このようにすることで、ベースラインがドリフトするときに正しく動作するようにした。各領域はクロマトグラムにおいて出現頻度が異なるので、ピークのない領域と単一ピーク、垂直分割ピーク、開始点、終了点の損失をそれぞれ 1 と $4, 8, 30, 30$ と重み付けした (Figure 2-14)。開始点と終了点の間のピーク領域（単一ピーク領域および未分離ピーク領域）のうち、一部分がピークなし領域になることがあるので、開始点から終了点までの半分以下がピークなし領域の場合には、単一ピーク領域もしく未分離ピーク領域の多い方に開始点と終了点の間の領域を置き換えた。

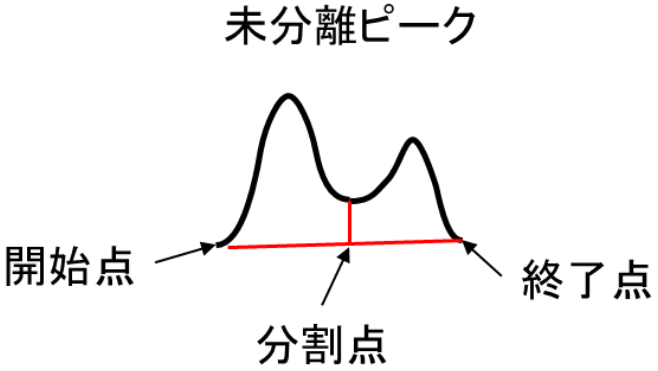


Figure 2-13　未分離ピークのピーク開始点と終了点、分割点の決め方

未分離ピークの開始時間、分割時間、終了時間がわかっているとき、ピーク開始点とピーク終了点の強度値は、クロマトグラム上の点とした。未分離ピークの分割点の強度値は、未分離ピークの開始点と終了点を直線で結んだ強度値とした。分割点はクロマトグラム上にないので、分割時間を変えずクロマトグラム上の点まで直線で結ぶ。

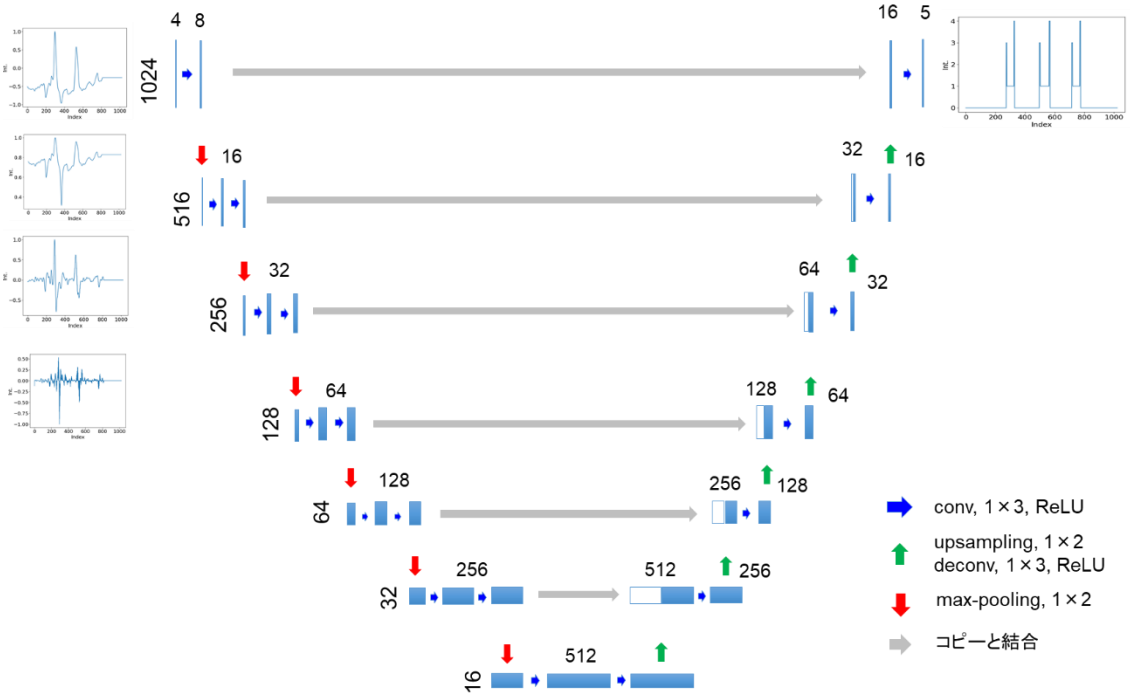


Figure 2-14　本研究で用いた U-Net の構成

ニューラルネットワークの入力は、(1) クロマトグラム、(2) その一次微分、(3) 二次微分、(4) 対数値から成る時系列データのマトリックス (1024 データ点数)のクロマトグラム (1024×4)である。フィルタサイズ 1×3 の畳み込み層 (conv) 、1×2 でアップサンプリング (upsampling) 、フィルタサイズ 1×3 デコンボリューション (deconv) 、1×2 でマックスプーリングを用いた。活性化関数として ReLU を用いた。U-Net の出力は、5 つのラベルの確率のマトリックス (1024×5) である: 0:ピークなし、1:単一ピーク、2:未分離ピーク、3:始点、4:分離点または終点のうち最大の確率であるラベルにまとめて、配列 (1024×1) にした。

2-2-5 SSD を用いたピークピッキングニューラルネットワークの構築

ピークピッキングの問題を物体検出として定式化した。物体検出は画像と時系列データを扱い、特定の物体位置を検出する。物体検出の典型的な方法である SSD を参考にした (Liu et al., 2016)。SSD の入力は U-Net と同じである。SSD の出力はピークの開始時間とピーク終了時間とした。具体的には、1,024 点のクロマトグラムに対して、出力された m 個のピーク範囲 [s(i), e(i)] (i=1,2,…,m) から、ピーク開始時間とピーク終了時間を求めた。出力値は、ピーク開始時間とピーク終了時間であるため、それぞれのピー

クの強度値は、クロマトグラム上の点とした。但し、未分離ピークの分割点の強度値は、未分離ピークの開始点と終了点を直線で結んだ強度値とした (Figure 2-13)。SSD は、物体を囲むデフォルトバウンディングボックスと呼ばれる矩形領域を使用してオブジェクトを検出する。デフォルトバウンディングボックスは、最小ピークおよび最大ピーク範囲内に収まるように調整した (Figure 2-15)。

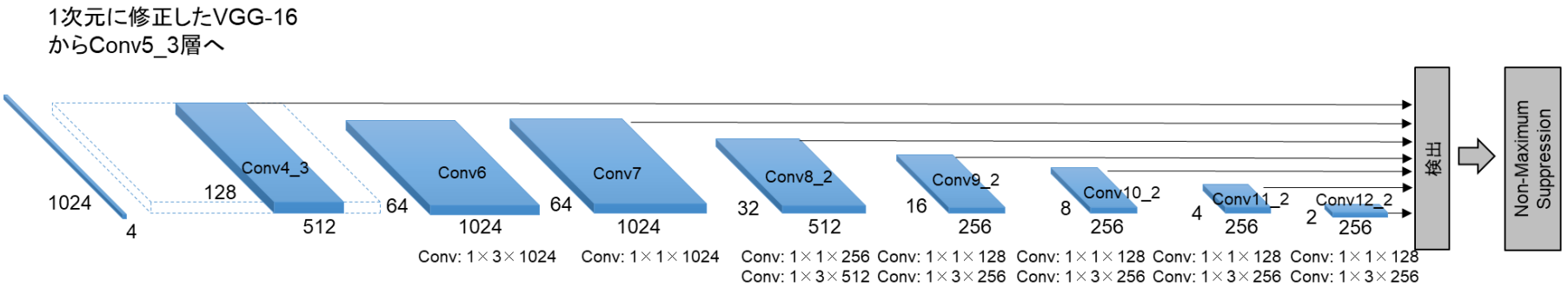


Figure 2-15　本研究で用いた SSD の構成

ニューラルネットワークの入力は、(1) クロマトグラム、(2) その一次微分、(3) 二次微分、(4) 対数値から成る時系列データのマトリックス (1,024 データ点数) のクロマトグラム (1024×4) である。畳み込み層の 2 次元配列を 1 次元配列に変えた以外 SSD と同じ畳み込みニューラルネットワークである。出力値は開始時間と終了時間である。

2-2-6 評価

ピークピッキングニューラルネットワークの性能を、適合率、再現率、 F 値（適合率と再現率の調和平均）、およびピーク面積に関して評価した。セマンティックセグメンテーションの基準と同様に、「真陽性」（正しく検出できた）は、予測したピーク範囲と正確なピーク範囲の間で Intersection over Union (IoU) が 50%以上として定義した (Figure 2-16)。

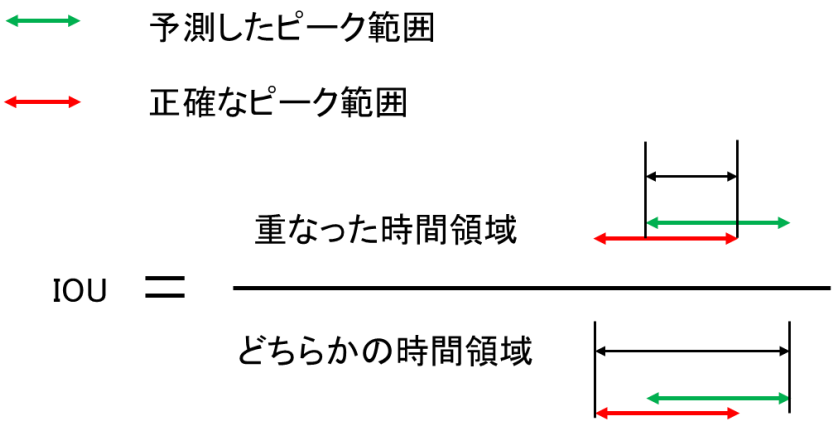


Figure 2-16 IoU

予測したピーク範囲と正確なピーク範囲が重なった時間領域をどちらかの時間領域で割った値。

適合率は、正確な検出数（真陽性）を予測ピーク数（真陽性+偽陽性）で割って計算した。再現率は、正解検出数（真陽性）を正解ピーク数（真陽性+偽陰性）で割って計算した (Figure 2-17)。

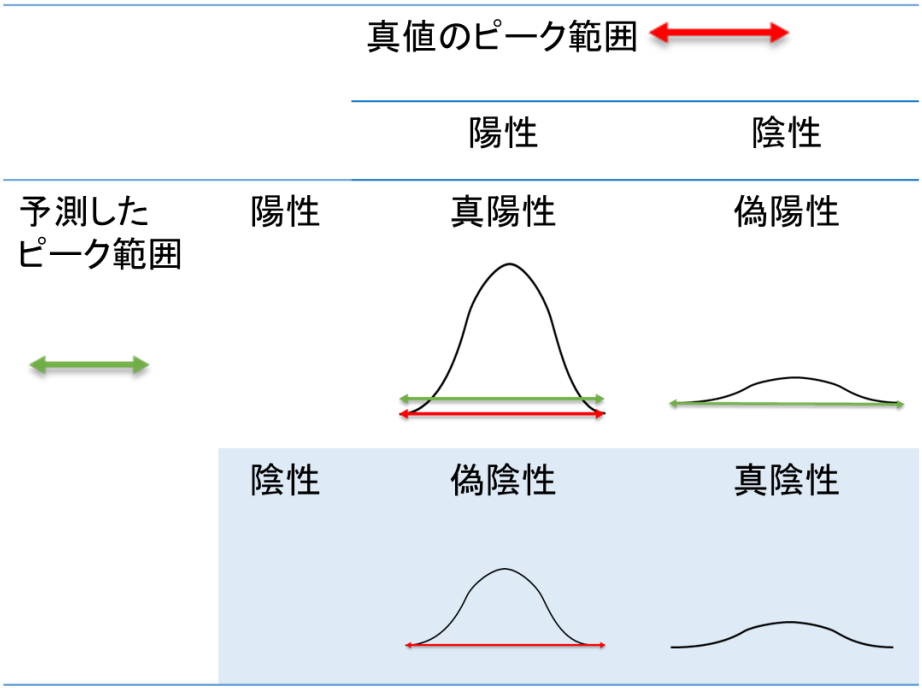


Figure 2-17　真陽性、偽陽性、偽陰性、真陰性

予測したピーク範囲と真値のピーク範囲の重なっている（IOU が 50%以上の）ときに真陽性とした。予測したピーク範囲が真値のピーク範囲に対して重なっていない（IOU が 50%未満の）ときに偽陽性、真値のピーク範囲が予測したピーク範囲に対して重なっていないときに偽陰性とした。

ピーク面積の誤差は、正確なピーク範囲から計算されたピーク面積（正確なピーク面積）と予測したピーク開始点と終了点から計算された面積の絶対値の差を正確なピーク面積で割った値で定義し、10%以下である場合、ピーク面積が正しく予測できたとした(Figure 2-18)。

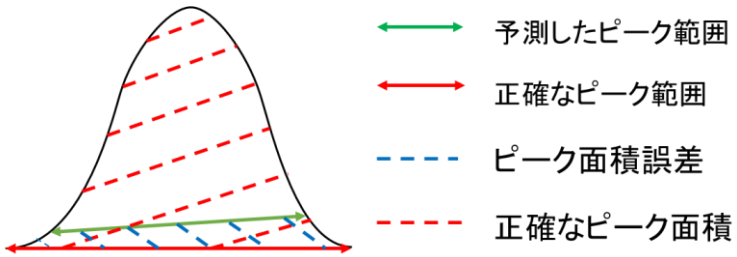


Figure 2-18　ピーク面積の評価

ピーク面積誤差（正確なピーク面積と予測したピーク面積の差の絶対値）を正確なピーク面積で割った値が 10%以下である場合にピーク面積が正しいとした。

2-2-7 コードとデータ利用

Python のクラスと、疑似クロマトグラム生成器と U-Net ベースのピークピッキングニューラルネットワークの例示的なスクリプト、ならびに本研究で使用される疑似クロマトグラムと実クロマトグラムの一部は、http://www-symbio.ist.osaka-u.ac.jp/software.html から入手可能である (データ S 1~S 3)。

2-3 結果と考察

2-3-1 実行環境

全てのニューラルネットワークと疑似クロマトグラム生成器は、Windows PC (NVIDIA Quadro K1200 GPU, Intel (R) Xeon (R) CPU 3.50 GHz, 16.0 GB RAM, 64-bit Windows 10 Professional Operating System)　で実行される Python 3 の Chainer v.6.2 (Tokui et al., 2019) 上で構築された。

2-3-2 疑似クロマトグラム生成器の開発

Figure 2-19 は本研究の作業フローである。Figure 2-19a は、目的 1 である正確にラベル付けされた疑似クロマトグラムの生成手法を示している。Figure 2-19b は、疑似クロマトグラムを用いた深層学習によるピークピッキングニューラルネットワーク手法を示す。以降は、その詳細を述べる。

本章で作成するピークピッキングニューラルネットワークは、次のような作業を行うこととした。

１．クロマトグラム中に観察されるすべてのピークのピークピッキングを行う。

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。

これらは現在使用されているピークピッキングソフトウェアの基本動作となっている。さらに、技術開発をおこなうターゲットとして、LC-MS の多重反応モニタリングモードで一次代謝物を測定する、ワイドターゲットメタボロミクス法を選択した。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。ワイドターゲットメタボロミクス法は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が、大量のクロマトグラムのピークピッキングが作業のボトルネックとなっている。

まず、血液、日本酒、ワインなどの生体試料から一次代謝物ワイドターゲットメタボロミクス法で 99 回の分析を行い、12,411 個の実クロマトグラムを含むデータセットを収集した (Figure 2-19a と Figure 2-19b、データ S1)。合計 7,615 個のブランクサンプルクロマトグラムも同じ方法で取得した (Figure 2-19b)。

学習データとして実クロマトグラムを用い、疑似クロマトグラムを生成する GAN の構築を試みた (Goodfellow et al., 2014) 。しかし、学習がうまく進まないため断念した

（資料未記載）。それは、勾配法で学習する際、目的関数が滑らかに変化する必要があるからである。今回、学習がうまく進まなかった原因は、ピーク位置や高さ等の複数のパラメータを学習させる際、それらの目的関数が離散的であり、学習がうまくできなかったためだと考えられた。

そこで、クロマトグラムの特徴量分布からランダムサンプルして、ピーク形状を GAN で生成するためルールベースの方法を開発した (Figure 2-19a)。クロマトグラムにおけるピーク数と中心からのピーク距離、隣接ピーク間の距離、信号対ノイズ比 (S/N)、ピークの半値全幅 (FWHM) など、クロマトグラムの性質を表す 5 つの特徴量を用いた。11,011 個の実クロマトグラムから特徴量分布を得た (Figure 2-9)。12,411 個の実クロマトグラムを学習データ (11,011、90%) およびテストデータ (1,400、10%) に分割した。

さらに、クロマトグラム中のピークで観測される歪みは、パラメータを用いてモデル化できないので、実際のピークに類似した疑似ピークの形状を生成する GAN を構築した。学習データとして用いるため、実クロマトグラムから、隣接ピークと重複していない 665 個の単一ピークを得た (Figure 2-19a、データ S2)。GAN は二つのニューラルネットワーク（疑似データを生成する生成器と疑似データを評価する識別器）を用いて実ピークに似た新しい偽ピークを生成する。時間の歪み関数 D(t)を生成するため生成器を構築した (Figure 2-3)。学習データとして実ピークを用いて、ピーク形状生成器を開発した（実験材料と実験方法を参照）。

以下の手順で疑似クロマトグラムを生成した (Figure 2-19a)。クロマトグラム中のピーク位置とピークの高さと幅を特徴量分布からの無作為抽出により決定した。次に、ピーク形状生成器によって生成された疑似ピークをクロマトグラムの対応する位置に置き、それらの開始点および終了点によってラベル付けした。最後に、この方法により 6,088 個のノイズレスピークを生成し、6,088 個のブランクサンプルクロマトグラムと統合して擬似クロマトグラムを生成した (Figure 2-19b)。目的達成のため、7,615 個のブランクサンプルクロマトグラムを学習データ (6,088、80%) とテストデータ (1,527、20%) に分割した。本研究では、10 回の手順を繰り返すことにより 60,880 個の疑似クロマトグラムを生成した。加えて、テスト目的のために用意した 1,527 個のブランクサンプルクロマトグラムを用いて 1,527 個の疑似クロマトグラムを、同じ手順により生成した (Figure 2-19b, データ S3)。

開発した手法で生成した代表的な疑似クロマトグラムを Figure 2-20 に示す。その結果、もっともらしい疑似クロマトグラムの生成に成功した。疑似クロマトグラムのピークは正確な開始点と終了点でラベル付けされていることに注意されたい。ここでは、これらのラベル付けを真値と考えた。

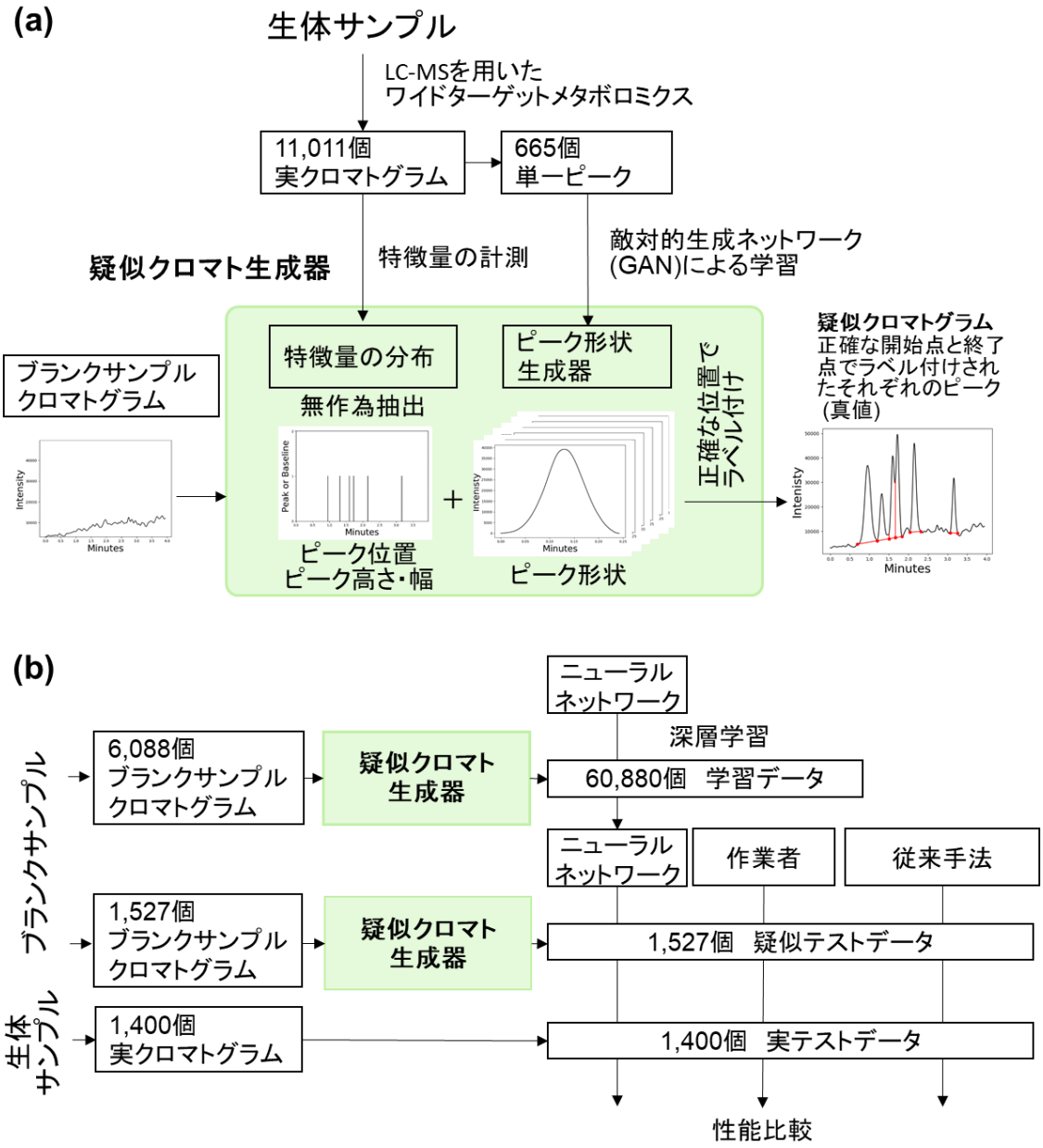


Figure 2-19　本章での作業フロー

(a) 疑似クロマトグラム生成器。クロマトグラム特徴量分布と GAN によるピーク形状生成器を構築するために生物サンプルを測定した。特徴量分布からの無作為抽出により、疑似クロマトグラムのピーク位置と高さ、幅を作成した。ピーク形状生成器によって生成された疑似ピークをブランクサンプルクロマトグラムと組み合わせ、それらの正確な開始点および終了点（真値：赤い線）を決定した。(b) ピークピッキングニューラルネ

ットワークの構築。7,615 個のブランクサンプルクロマトグラムを学習データ（6,088 個）とテストデータ（1,527 個）に分けた。6,088 個のブランクサンプルクロマトグラムと疑似クロマトグラム生成器を用いて、60,880 個の学習データを作成した。60,880 個の学習データを用いて、ピークピッキングニューラルネットワークを学習した。1,527 個の疑似テストデータを 1,527 個のブランクサンプルクロマトグラムを用いて同じ手順で作成した。ニューラルネットワーク、作業者および従来法の性能を、実テストデータおよび疑似テストデータの 1,400 個の実クロマトグラムを用いて評価した。

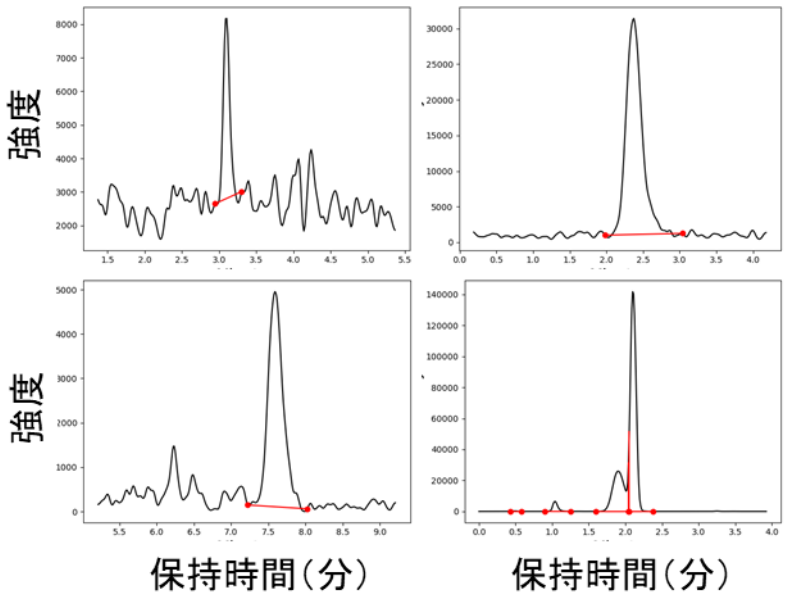


Figure 2-20　疑似クロマトグラムの例

ピークの正確な開始点と終了点がラベル付けされた疑似クロマトグラムが生成された。

2-3-3 作業者に匹敵するピークピッキングニューラルネットワークの構築

SSD と U-Net を用いてピークピッキングニューラルネットワークの構築を試みた (Figure 2-21)。SSD は、単一の畳み込みニューラルネットワークを使用して、画像内の物体を検出する (Liu et al., 2016)。U-Net は画像内の物体を、高速で正確にセグメンテーションするための畳み込みニューラルネットワークである (Ronneberger et al., 2015)。ニューラルネットワークの入力は、(1) クロマトグラムと (2) その一次微分、(3) 二次微分、(4) 対数値から成る時系列データのマトリックス (1,024 個のデータ点数) のクロマトグラム (1024×4) で、SSD の出力はピーク開始時間と終了時間とし、U-Net の出力は、5 つのラベルの確率のマトリックス (1024×5) である: 0:ピークなし、1:単一ピーク、2:未分離ピーク、3:始点、4:分離点または終点のうち最大の確率であるラベルにまとめて、配列 (1024×1) にした (Figure 2-21)。

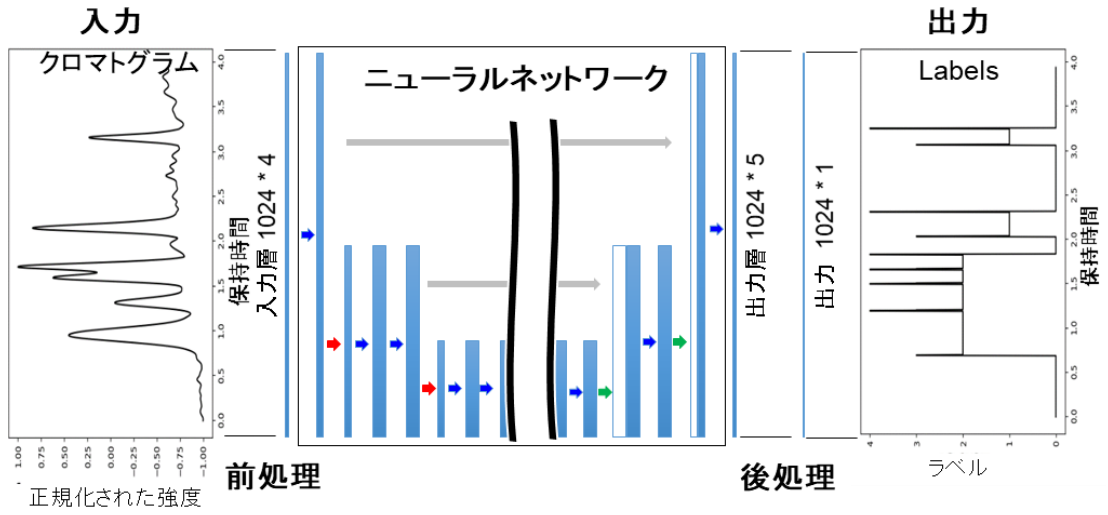


Figure 2-21　疑似クロマトグラムデータ中の正確なラベルを用いたニューラルネットワークの構築。1 次元時系列であるクロマトグラム、その 1 次微分波形、2 次微分波形、対数値を-1 から 1 へと正規化し、ニューラルネットワークへ入力した。出力値として得られる、5 つのラベル（0:ピークなし、1:単一ピーク、2:未分離ピーク、3:始点、4:分離点または終点）の値が最大であるラベルを出力した。

前項で生成した擬似クロマトグラムを、学習データ (45,540 個)、検証データ (15,340 個) 、テストデータ (1,527 個) に分けた。ピークピッキングニューラルネットワークは、深層学習による疑似クロマトグラム上の正確な開始点と終了点のラベルを用いて学習された。SSD では、疑似クロマトグラム上の正確な開始点（時間と強度）と終了点（時間と強度）をピーク開始時間とピーク終了時間に加工して学習させ、U-Net では、疑似クロマトグラム上の正確な開始点と終了点を 0:ピークなし、1:単一ピーク、2:未分離ピーク、3:始点、4:分離点または終点に加工して学習させた。

市販のピークピッキングソフトウェアの i-PeakFinder とクロマトパックアルゴリズムを用いて比較した。i-PeakFinder は、ウェーブレット変換を用いて、クロマトグラムを異なる平滑化具合の二次微分波形へ変換してピーク位置を検出する。その後、ノイズ値よりピークの立ち上がりと立下りの位置を推定し、ピークの開始点と終了点を求める。一方クロマトパックアルゴリズムは、ピークの立ち上がりと立ち下がりの傾きを閾値として設定することで、閾値以上の傾きになった箇所をピーク開始点とピーク終了点にするアルゴリズムである。

テストデータのピークピッキング性能は、セマンティックセグメンテーションの基準（50%以上の重複を真陽性とする）に従って評価した (Figure 2-22a)。同じテストデータを用いたピークピッキング性能を、三人の作業者（作業者 1 と 2、3）と市販のピークピッキングソフトウェア（最適なパフォーマンスにパラメータを調整した i-PeakFinder およびクロマトパック）について評価した。結果は、U-Net と SSD ベースのピークピッキングニューラルネットワーク（適合率と再現率の調和平均である F 値が 0.959 と 0.930）が、作業者（0.914 と 0.427、0.904）と二つの従来のソフトウェア（0.853 と 0.749）よりも優れていることを示した (Figure 2-22a)。i-PeakFinder では、ベースラインのトレンドを誤検出する傾向と、ノイズ情報の少ないピークを検出できない傾向があるが、U-Net ではそれが改善された (Figure 2-23)。この原因として、ウェーブレット変換（周波数解析）を用いた i-PeakFinder ではベースラインのトレンドの周波数とピークの周波数を区別することが難しいこと、またノイズの少ないクロマトグラムからピークを検出することが難しいことが考えられる。

U-Net ベースのニューラルネットワークの F 値が最も良く、ピークピッキングの有望なニューラルネットワーク構成であることを示唆した。SSD ベースのニューラルネットワークは適合率が低いことから、この方法では多くの偽陽性が生じることが示唆された (Figure 2-22a)。SSD は物体検知という物体の位置を大まかに予測するために考えられた手法である。一方、U-Net はセマンティックセグメンテーションという領域を予測するために考えられた手法である。後者の手法の方がピーク領域を求めるというピークピッキングの問題設定に適していたと考えられる。より低い適合率は、作業者のピークピッキング結果でも観察された。これは、作業者が偽陽性を生成する傾向があることを示し、作業者によって手動でラベル付けされた実クロマトグラムは、深層学習のための学習データセットとして不適切であることを示唆する。

U-Net より測定されたピーク面積を比較すると、ピーク面積が真値に非常に類似していることを示した (Figure 2-20, Figure 2-22b)。U-Net より測定されたピーク面積の誤差は作業者 1 と作業者 3 のそれと同等またはそれ以下であった。詳細に分析すると、真値よりピーク面積の誤差が 20%以上であるピーク数の割合は、U-Net と作業者 1、作業者 3 が約 6%であることを確認した (Figure 2-22 c)。U-Net では閾値が 5%の場合ピーク数の割合が 17.7%であるのに対し、作業者 1 と作業者 3 はそれぞれ、22.3%と 22.4%であった (Figure 2-22c)。

疑似クロマトグラムで学習することが実クロマトグラムと比べて意義があるかを調べるため、11,011 個の実クロマトグラムを用いて U-Net を学習させた。それを 1,527 個の疑似テストデータで比較したところ、疑似クロマトグラムで学習した U-Net（0.959 の F 値）は実クロマトグラムで学習した U-Net（0.817 の F 値）より優れていることを示した (Figure 2-22d)。

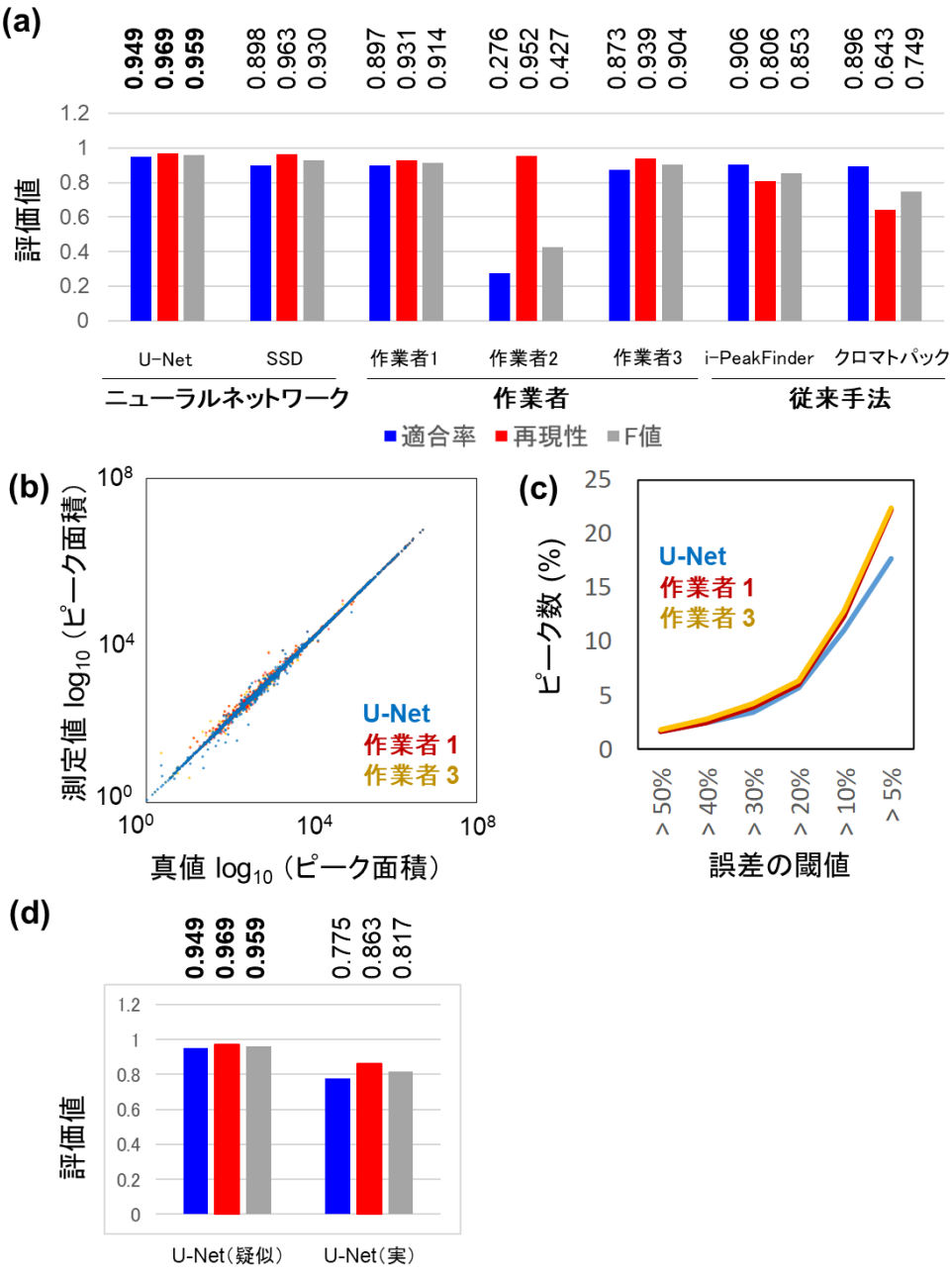


Figure 2-22　ピークピッキングニューラルネットワークの深層学習

(a) 疑似クロマトグラムを用いた学習済みモデルと熟練者、従来ピークピッキングソフトウェアの評価。(b) 真値のピーク面積と測定値（U-Net と作業者 1、作業者 3）のピーク面積の散布図。(c) 閾値以上のピーク面積誤差をもつピークピッキングの数。(d) 疑似クロマトグラムを用いた学習済みモデル（疑似）と実クロマトグラムを用いた学習済みモデル（実）。

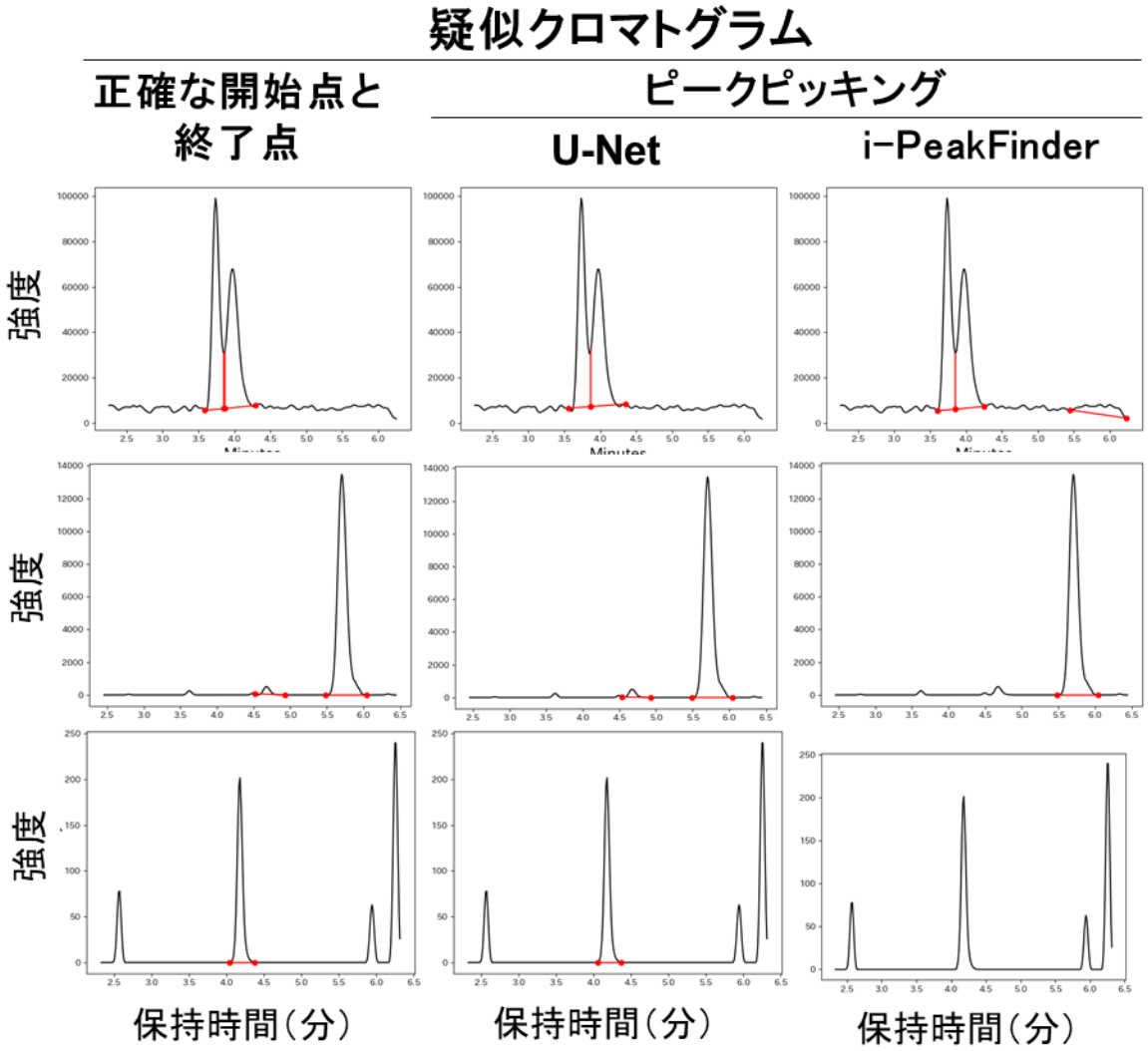


Figure 2-23　i-PeakFinder の課題が U-Net で課題された例

ベン図を用いて、正確な開始点・終了点（真値）がわかるようになった疑似クロマトグラムと真値がわからない実クロマトグラムに対して、ピークピッキング結果を比較した (Figure 2-24)。まず、作業者 1 と作業者 3 が行った疑似クロマトグラムのピークピッキング結果をみると、作業者 1 の偽陽性率は 9.1%、作業者 3 の偽陽性率は 11.8%だった。また、作業者 1 と作業者 3 で共通する偽陽性の割合は 5.8%だった (Figure 2-24a)。つまり、熟練作業者のピークピッキング結果のうち、約 1 割は偽陽性であり、さらに、そのうち約半分については、熟練作業者間でも結果が異なっていた。さらに

、疑似クロマトグラムのピークピッキング結果を U-Net と作業者 3 で比較した場合、作業者 3 の偽陽性率は U-Net に比べ高くなっていた (Figure 2-24b)。これは、熟練作業者のピークピッキング結果には、U-Net のものに比べ、偽陽性が多く含まれること示唆する。

また、疑似クロマトグラムをピークピッキングした結果の作業者 1 と作業者 3 の間の一致率は約 83%だった (Figure 2-24a)。同等の一致率（約 84%）は、実クロマトグラムとピークピッキングした結果でもが観測された。この結果は、疑似クロマトグラムが実クロマトグラムと類似した特性を持つことを示唆している (Figure 2-24c と d)。

さらに、もし、作業者 3 が作業者 1 のピークピッキング結果の修正作業を行ったとすると、約 16%の結果を受け入れず手動で修正することになる (Figure 2-24c)。しかし、修正結果全体の偽陽性、偽陰性数は改善しないことになる。同様に、作業者 3 が U-Net のピークピッキング結果を確認すると約 23%の結果を受け付けず手動で修正することを示唆している (Figure 2-24d)。しかし、その修正は明らかな間違いを除くと、全体としてピークピッキングの品質が改善されることはないと推測された。

作業者 1 と作業者 3 の差異が最も小さく、SSD と作業者 3 の差異、U-Net と作業者 3 の差異、i-PeakFinder は作業者 3 の差異の順に大きかった (Figure 2-24cdef)。Figure 2-22a の疑似クロマトグラムの結果において、作業者 3 を基準とすると、適合率より再現性が高いことが、作業者 1 と SSD、U-Net でいえるが、i-PeakFinder は逆であることから、i-PeakFinder と作業者 3 の差異が最も大きいと考えられる。適合率と再現率の傾向が同じである作業者 1 と SSD、U-Net の差異は、適合率と再現率の調和平均である F 値と作業者 3 の F 値の差と比例している。このことからも、疑似クロマトグラムと実クロマトグラムのピークピッキング結果が同様の傾向を示していることがわかる。

正確な開始点と終了点と U-Net、作業者 3 間での疑似クロマトグラムのピークピッキング結果を比較した (Figure 2-25)。上 4 つの例で U-Net が正確な開始点と終了点と一致しており、上から 3 番目の例では、作業者 3 が誤検出していることがわかる。一番下の例では、U-Net が検出できていないピークがあり、端のピークを検出できない傾向にある。これは、学習データに端のピークが少ない傾向があるためだと考えられる。また、U-Net と作業者 1、作業者 3 間での実クロマトグラムのピークピッキング結果を比較した (Figure 2-26)。1 番上は全てノイズの例であり、U-Net と作業者 1 は正しく処理しているが、作業者 3 はピークを誤検出している。一番下は、U-Net は複数の重な

ったピークとしたが、作業者 2 と作業者 3 は単一のピークとした。しかし、実クロマトグラムだけではどちらの結果が正しいかはわからない。

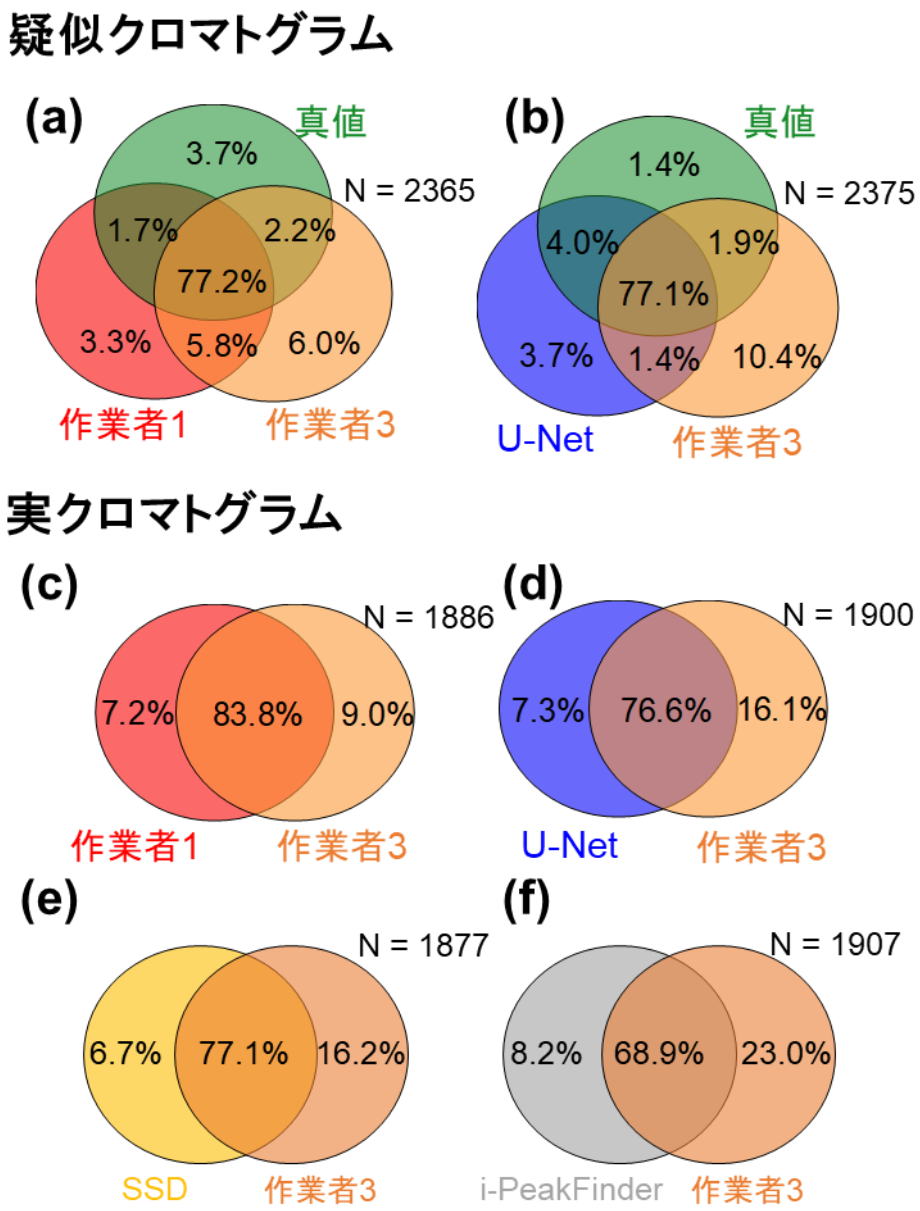


Figure 2-24　真値と U-Net、作業者 1 と作業者 3 のピークピッキング結果の比較
(a) 真値と作業者 1、作業者 3、(b) 真値と U-Net、作業者 3 の疑似クロマトグラムでの比較。(c) 作業者 1 と作業者 3、(d) U-Net と作業者 3、(e) SSD と作業者 3、(f) i-PeakFinder と作業者 3 の実クロマトグラムでの比較。

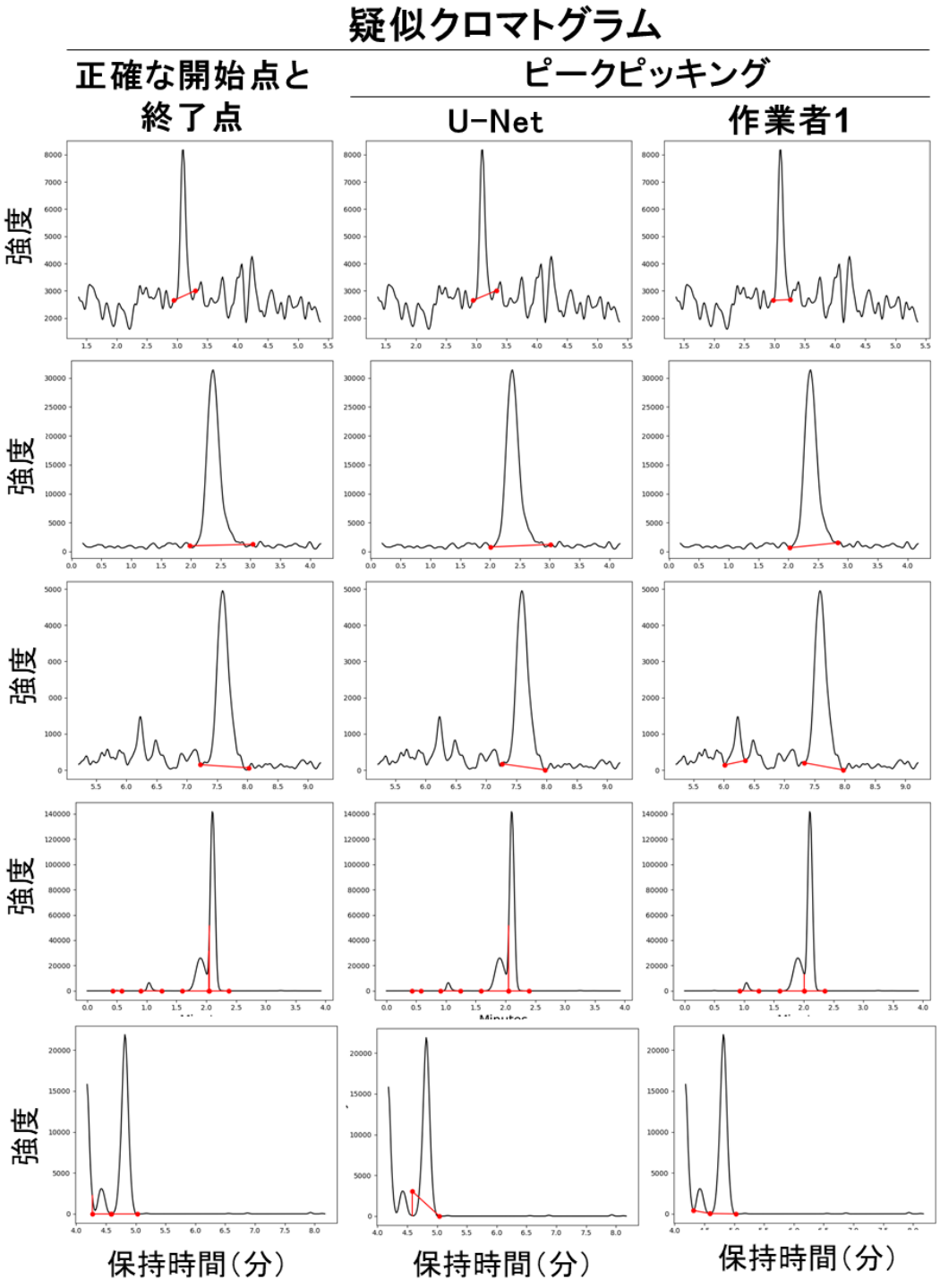


Figure 2-25　正確な開始点と終了点と U-Net、作業者 3 間での疑似クロマトグラムのピークピッキング結果の比較例

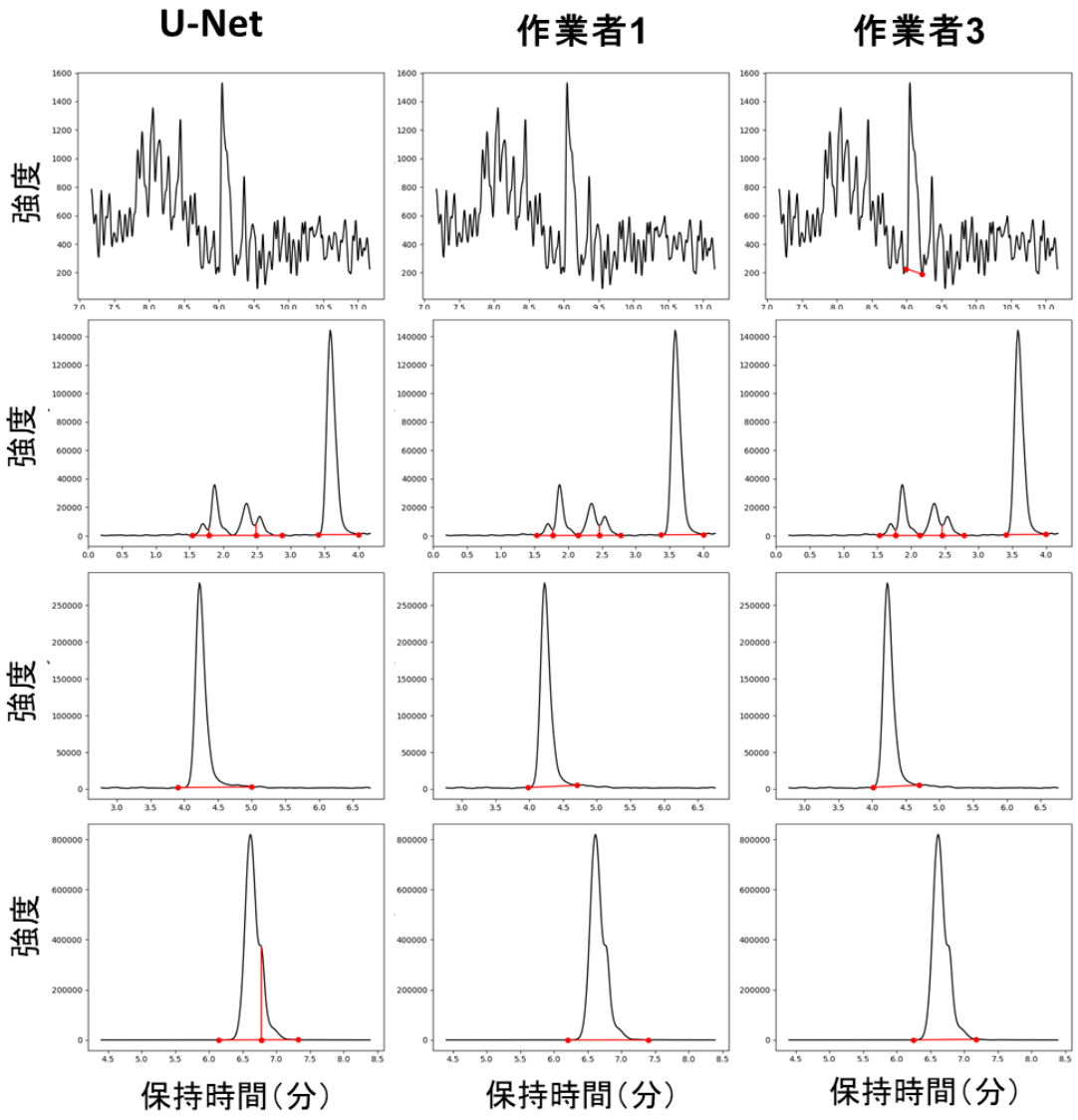


Figure 2-26　U-Net と作業者 1、作業者 3 間での実クロマトグラムのピークピッキング結果の比較例。一番上は全てがノイズの例であり、U-Net と作業者 1 は正しく処理しているが、作業者 3 はピークを誤検出している。一番下は、U-Net は複数の重なったピークとしたが、作業者 2 と作業者 3 は単一のピークとしているが、実クロマトグラムだけではどちらが正しいかはわからない。

2-4 結言

本章では、ピークの正確な開始点と終了点でラベル付けされた疑似クロマトグラムが、ピークピッキングニューラルネットワークの学習と評価に有効であることを実証した (Figure 2-20 と Figure 2-22)。これにより緒言で述べた２つの目的を達成した。

本章では、5 つのクロマトグラムの特徴量分布に沿って GAN で生成した疑似ピーク形状を配置し、ブランクサンプルクロマトと組み合わせることで、疑似クロマトグラムを生成する手法を構築した。本章では LC-MS を用いたワイドターゲットメタボロミクスを題材としたが、特徴量分布、疑似ピークを生成する GAN、ブランクサンプルクロマトグラムを用意することで、他の分析手法にも適用できる汎用性がある。例えば、ガスクロマトグラフィー (Gas Chromatography, GC) 質量分析法ワイドターゲットメタボロミクス法の疑似クロマトグラムの生成器を構築することも可能である (Figure 2-19a)。これにより、目標１の疑似クロマトグラムの生成により正確にラベル付けされた学習データを準備する手法を開発することができた。

本章では、SSD と U-Net を用いたピークピッキングニューラルネットワークを構築した。多数の疑似クロマトグラムを用いることで、ピークピッキングニューラルネットワークの偏りのない学習を可能にした。また、ピークの正確な開始点と終了点の情報を用いた異なるピークピッキング法の性能評価が可能となった (Figure 2-22) 。構築したピークピッキングニューラルネットワークは、従来のピークピッキングソフトウェアよりも優れており、熟練者と同等の性能を示した (Figure 2-22、Figure 2-24)。これにより、目的２の熟練した作業者と同等の性能を持つピークピッキングニューラルネットワークを構築することを達成した。

また、SSD と U-Net の結果を比較したところ、U-Net のほうが高い性能を示した。SSD は物体検知という物体の位置を大まかに予測するために考えられた手法である。一方、U-Net はセマンティックセグメンテーションという領域を予測するために考えられた手法である。ピークピッキングはピーク領域を求めるという問題なので、セマンティックセグメンテーションの領域を予測する問題設定の方が適していたと考えられる。さらに、二人の熟練者によるピークピッキングの結果は約 16%が一致せず、作業者により手動でラベル付けされた実クロマトグラムは深層学習の学習データセットとして不適切であることを示した (Figure 2-24c)。

本章の結果は、疑似クロマトグラムの生成がピークピッキングニューラルネットワークの開発とピークピッキングニューラルネットワークの更なる改善のための重要な技術であることを示す。

今回は、クロマトグラムの特徴量分布からランダムサンプルして、ピーク形状を GAN で生成するためルールベースの方法を開発したが、実クロマトグラム全体からラベル付き疑似クロマトグラムを生成することで、より簡便にラベル付き疑似クロマトグラムが生成できると考えられる。そのためには、学習がうまく進まなかった原因を解決する必要がある。原因としてピーク位置や高さ等の複数のパラメータを学習させる際、目的関数が離散的で局所解に陥っていたと考えられる。平滑化などを用いた局所最適解に関する研究が行われており (Kleinberg et al., 2018)、それらの研究を組み合わせることで、うまく学習する可能性がある。

セマンティックセグメンテーションのニューラルネットワークには新しい構造として PSPNet (Zhao et al., 2017) や DeepLab (Chen et al., 2018) などが開発されている。これらは日々進歩している技術なので、それを適用することが性能向上につながると考えられる。それらの技術の進歩により、端のピークが検出される可能性はあるが、実際にうまくいくかは今後の課題である。また、ブランクサンプルクロマトグラムを自動生成してバリエーションを増やすことで、学習データの網羅性が上がり、ピークピッキングの性能向上が見込める。その他、代謝中間体が溶出される保持時間の情報や繰り返し分析した複数のクロマトグラム、フラグメントイオンなどの情報を組み合わせることでより性能が上がり、熟練作業者を超えることが期待される。

また、LC-MS を用いたバイオマーカーの発見には、80～400 のサンプル数を扱う (Di Gangi et al., 2016, Rhee et al., 2011, Wang et al., 2011, Wang et al., 2013)。1 人の作業者が扱える 20 サンプル（200 種類の代謝中間体含量）を測定することを考えると、4000 回のピークピッキングが行われるため、確認・修正するのに平均して 6 秒かかるので約 6.5 時間が必要である。第 2 章で、人と同等のピークピッキングニューラルネットワークの構築に成功したため、「6 秒×代謝物数×サンプル数」の修正作業を実質ゼロにすることができた。

正確にピークピッキングされたデータが得られると、第 1 章で述べたように次の作業として、増減した代謝物のリストから、関連するキーワードを連想する作業を行う必要がある。

第3章 キーワードレコメンド法の開発

3-1 緒言

クロマトグラフィーを用いた代謝中間体の定量分析は、バイオテクノロジーの基盤技術である (Matsuda et al., 2017)。メタボロミクス分野では、クロマトグラフィー質量分析法で数百の代謝物を一斉定量する事が一般的になってきている。網羅的な測定データの統計解析から、薬剤投与の結果含量が変化した代謝物や、バイオマーカー候補となる代謝物の「代謝物リスト」が得られる。同様の作業は農学、工学分野でも行う。高級緑茶と一般的な緑茶の比較からは、品質の差に寄与する代謝物候補のリストが得られる。代謝工学分野では、異なる寿命をもつ出芽酵母の比較から、長寿命な株と正に相関する代謝物のリストが得られる。「代謝物リスト」と、薬剤名や細胞名などの「解釈担当者の既知のキーワード（知識）」だけから、薬剤作用機構の推定や、疾患とバイオマーカー候補となる代謝物の関連を推定することは困難である。そこで、「代謝物リスト」と「解釈担当者の既知のキーワード（知識）」から「解釈担当者の知識外のキーワード」を連想する作業が、論文検索などを用いて行われている。

第 1 章で述べたように、論文検索でのキーワード連想作業には次の問題がある。時間と労力が必要な点である。例えば、「コハク酸 & アポトーシス」をクエリとして、PubMed 検索を行うと、32 件（2020 年 6 月現在）の論文がヒットするが、要旨と本文を読んで概要をつかむにも膨大な時間（1 件あたり平均して 15 分かかるとすると 8 時間）が必要とされることになる。そこで、本研究では「代謝物リスト」と「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」をレコメンドする技術の開発を試みることにした。

近年、レコメンドアルゴリズムは盛んに研究されており、様々な手法が提案されている。特にオンラインショッピングなどの分野では、過去の購買履歴などから、ユーザーに新しい商品（アイテム）をレコメンドする手法として協調フィルタリングが用いられている。協調フィルタリングでは、他のユーザーが過去にアイテムを評価した結果を基にレコメンドモデルを構築し、新たにユーザーが興味を持つアイテムを提示する (Resnic et al,. 1994)。協調フィルタリング手法としては、アイテムベース手法、行列分解手法などが提案されている。

アイテムベース手法は、過去のデータからアイテム同士の関連度を計算し、その値が大きいものをレコメンドする方法である (Sarwar et al,. 2001)。関連度の計算には Cosine や Lift などの値が用いられている。例えば、過去の売り上げデータにおいて、同時に購入されたアイテムがいくつかあったとすると、同じ傾向で購入したユーザーのアイテムをレコメンドする。

次に、行列分解手法では、レコメンドを「ユーザー×アイテムの評価行列の欠損部分を予測する」という問題として表現している (Koren et al,. 2009) 。評価行列とは、ユーザーが過去に購入したアイテムに対して付けた評価からなる行列である。ユーザーがまだ買ったことのないアイテムに対する評価値は欠損値となるため、欠損値を予測すれば、予測された値が高いアイテムをレコメンドできるという手法である。行列分解手法では、行列のスパース性をうまく利用し、元の行列を近似することで欠損値の予測を行う。

協調フィルタリング手法として近年主流となっているのは、行列分解手法や、それをさらに発展させた Factorization Machines (Rendle et al,. 2010)、ニューラルネットワークを用いた手法 (He et al,. 2017) である。しかし、本研究の目的は分析データから得られたキーワードに対して別のキーワードをレコメンドすることであるため、これらの手法を適用することはできない。すなわち、本研究の問題設定に行列分解手法を適用しようとすると、共起行列の欠損を埋める（既存論文に対して新たな MeSH をレコメンドする）という問題を解くことになり、目的に合わない。

最後に、共起情報に基づくアソシエーション分析を用いたレコメンド法がある。アソシエーション分析とは、売り上げ情報などの膨大なデータから、意味のある関連性を抽出し、関連性の強いものをレコメンドするために開発された手法である (Atluri et al.,2009, Chen el al., 2005)。例えば売り上げ情報の場合、ある２種のアイテムが、同一のレシート中に頻繁に出現（共起）する場合、関連性があると考えられる。協調フィルタリングのアイテムベースでは、複数のアイテムからアイテムの関連度を計算して、レコメンドするが、共起情報に基づくアソシエーション分析では、2 つのアイテム間の関連度を基にレコメンドするので、複数のアイテムの関連度を計算するか２つのアイテムの関連度を計算するのかの違いである。アソシエーション分析は因果関係の推定にも使われることとキーワード連想でも将来的には因果関係の推定が必要なことから、本研究ではアソシエーション分析を適用する。

そこで、キーワードをレコメンドする作業を、1) 「代謝物リスト」および「解釈担当者の既知のキーワード」と関連度の高い「解釈担当者の知識外のキーワード」を探索し、2) 「解釈担当者の知識外のキーワード」の候補について、関連度合いから重要度（スコア）を決定し、3) ある閾値をもとに要否を判断するという作業で実現できると考えた（Figure 3-1a）。そのための、キーワードおよび関連度を計算するための共起情報データとして、PubMed のデータベースと MeSH ID に注目した。PubMed は世界で最もよく使用される生物医学系データベースであり、NCBI により開発および維持されている。2020 年 6 月現在では 5,000 以上のジャーナルから、3,000 万件以上の文献が登録されている (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed)。

PubMed の特徴として、文献の多さだけでなく、MeSH が挙げられる。MeSH とは、生物医学系分野の様々なキーワードに対して一意な ID もしくはタームを割り当てたものである (http://www.nlm.nih.gov/mesh/meshhome.html)。例えば、アポトーシス、クエン酸というキーワードにはそれぞれ Apoptosis, Citric Acid というタームが割り当てられている。Pubmed には現在約 30,000 の MeSH タームが登録されており、PubMed に登録された各文献には、平均して 10 個程度の MeSH タームが紐づけられている。文献への紐づけ作業は、生物医学系の学士以上の学位を持つ専門メンバーらによって、文献の中身を読んだ上で行われている。また、MeSH タームはシソーラスの機能も兼ねている。例えば「プログラム細胞死」は「アポトーシス」の類義・同意語であるが、どちらも Apoptosis という MeSH タームが割り当てられている。このため、表記ゆれなどの問題を気にせずに、自分の知っているキーワードが紐づけられた文献を PubMed で検索することが可能となっている。例えば、クエン酸 (Citric Acid) とアポトーシス (Apoptosis) をキーワードとして論文検索（アンド検索）をすると、二つの MeSH タームの両方に紐づけられた論文として Terasaki et al., 2018 などが得られる。

上記のように、PubMed は、キーワードと文献を紐づけるデータベースとして非常に良い特性を備えている。また、２つの MeSH タームが同一論文にキーワード付与（共起）されている数から、アソシエーション解析の手法を用いて、２つの MeSH ターム間の関連度合いを定量化することが可能となる。MeSH を用いた手法としては、MeSH 間の意味的類似性をみる MeSH-Gram (Abdeddaïm et al., 2019) や MeSH の階層構造を利用して関連度を計算する MeSH Sim (Zhou et al., 2015)、アソシエーション分析 (Agrawal et al., 1993, Agrawal et al., 1994) で用いられる類似尺度を用いた手法 (Ono T, 2014) が提案さ

れている。しかし、代謝物リストと解釈担当者の既知のキーワードから解釈担当者の知識外のキーワードをレコメンドする研究はない。さらに、レコメンドされたキーワードが統計的に信頼できるものかがわからない。

そこで本章の目的を下記の２点とした。

1. 「代謝物リスト」と「解釈担当者の既知のキーワード」から、PubMed の MeSH ID（MeSH タームと 1 対 1 対応）の共起情報を用いて、「解釈担当者の知識外のキーワード」をレコメンドする手法を開発すること
2. 「解釈担当者の知識外のキーワード」の重要度（スコア）の閾値を決定する手法を確立すること

本章の目的は「代謝物」の MeSH ID と「解釈担当者の既知のキーワード」の MeSH ID の両方と、PubMed 文献中で高頻度に共起する MeSH ID を抽出する手法を開発することである。したがって、文献上で関連することが知られている「解釈担当者の知識外のキーワード」をレコメンドする。それは、解釈担当者の知識不足を補足することを意味している。したがって、PubMed の文献中でまだ共起していない新たな関連を見出すことはできない。また、レコメンドしたキーワードの意外性や、生物学的意義を保証するものでもない。

本章では、まず、PubMed と MeSH ID の共起情報を用いて、アソシエーション分析の手法を適用して、２つの MeSH ID 間の関連度合いを算出することを可能とした。次に「代謝物の MeSH ID」と「解釈担当者の既知のキーワードの MeSH ID」の両方と関連度合いの高い MeSH ID の重要度（スコア）を計算する手法を開発した。最後に、「解釈担当者の知識外のキーワード」としてレコメンドするスコアの閾値を合理的に決定する手法を開発した。これは、ランダムデータベース (Data Base , DB) を作成し、閾値を決定することで、統計的に信頼できるキーワードをレコメンドする手法を開発した。また、これらの作業は、抗がん剤 4-OHT を処理した乳がん細胞株 MCF-7 のメタボロミクスデータをテストデータとして用いた (Figure 3-1b)。

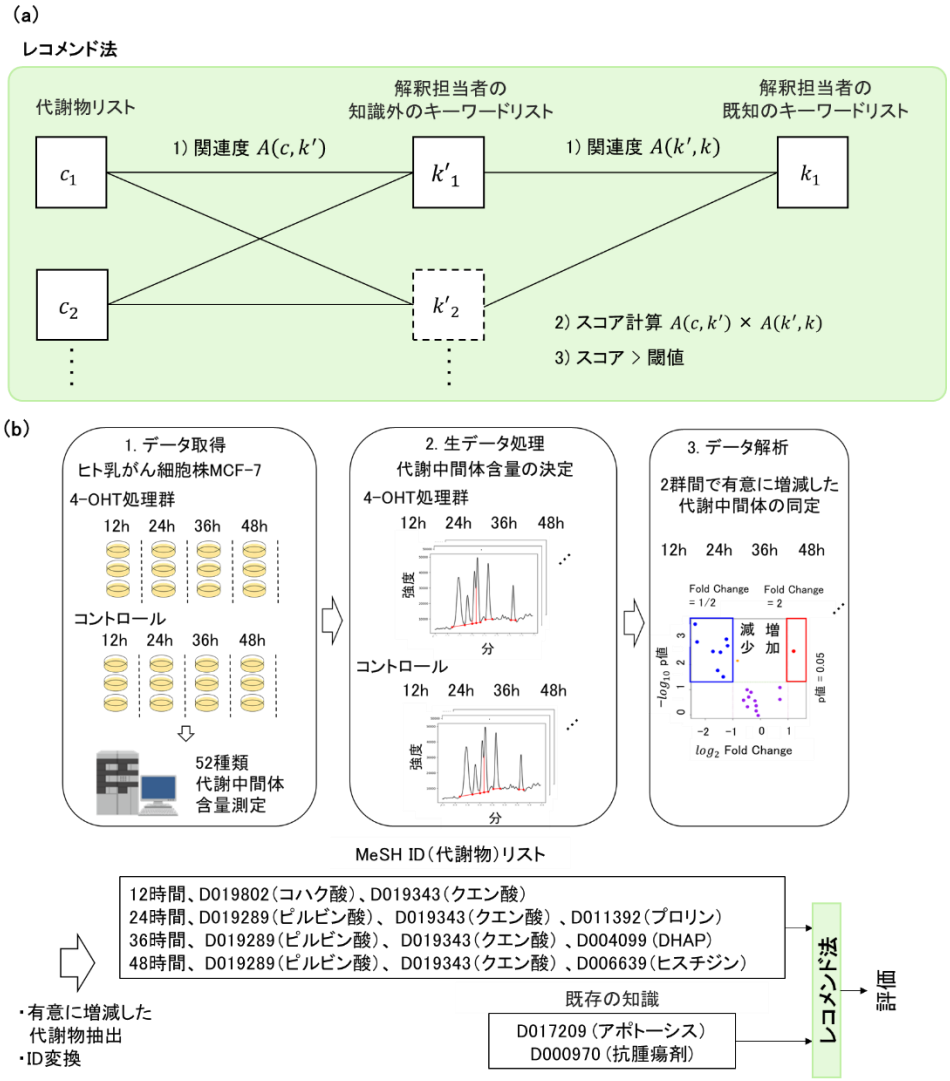


Figure 3-1 研究目的。(a) 統計的に有意なキーワードをレコメンドする手法の開発。1) 代謝物リストと解釈担当者の知識外のキーワードの関連度合いを決定することと解釈担当者の知識外のキーワードと解釈担当者の既知のキーワードの関連度合いを決定すること、2) 関連度合いから解釈担当者の知識外のキーワードの重要度（スコア）を決定すること、3) 解釈担当者の知識外のキーワードの要否を判断するためのスコアの閾値を決定する手法を開発した。(b) レコメンド機能の評価。テストデータとして、ヒト乳がん細胞 MCF-7 に対して抗がん剤 4-OHT の処理有無で測定した 52 種類の代謝中間体含量の 12 時間、 24 時間、36 時間、48 時間の時系列データを用いた。各時刻において、有意に増減した代謝物リストを作成し、解釈担当者の既知のキーワードとしてアポトーシスもしくは抗腫瘍剤を入力し、レコメンドした。

3-2 実験材料と実験方法

3-2-1 PubMed、MeSH ID データの準備

PubMed ファイルサーバー（ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline）から 2020 年までの全論文データとそれに紐づく MeSH ID のデータをダウンロードした。MeSH ID はカテゴリ分けされているため、一次代謝物のターゲットメタボロミクスに関連するカテゴリである、Cells と Neoplasms, Nutritional and Metabolic Diseases, Chemically-Induced Disorders, Chemicals and Drugs, Phenomena and Processes を選び、13,985 個の MeSH ID を本章では用いた。

3-2-2 テスト用メタボロミクスデータ

テスト用メタボロミクスデータとして、4-OHT を処理した MCF-7 ヒト乳がん由来細胞株（TCC No.HTB-22）の時系列メタボロミクスデータを用いた（西本ら、2020）。本データセットは、抗がん剤 4-OHT を処理した 12 時間、24 時間、36 時間、48 時間後の MCF-7 中の 52 種の代謝中間体含量のデータが含まれている。

3-2-3 アソシエーション分析を用いたスコアの決定

アソシエーション分析は、膨大なデータから意味のある関連性を抽出し、関連性の強いものをレコメンドするために用いられる (Agrawal R et al., 1993, Agrawal R et al., 1994)。ここでは、PubMed の論文と MeSH ID から、代謝物リストと解釈担当者の既知のキーワードリストと関連性の強いキーワードをレコメンドした。キーワード同士の関連度の指標として、Cosine, Simpson, Confidence, Lift を用いた (Azevedo et al., 2007, Murguía et al., 2003)。

全部で M 個（2020 年 6 月現在、29,054 個中メタボロミクスに関係する 13,985 個使用）の MeSH ID の集合$\mathrm{W} = \{w_1, \ldots, w_M\}$があるとき，代謝物の MeSH ID リスト$\mathrm{C} = \{c_1, \ldots, c_n;\ c_i \in \mathrm{W}\}$，解釈担当者の知識外のキーワードの MeSH ID リスト$\mathrm{K}' = \{k'_1, \ldots, k'_m;\ k'_j \in \mathrm{W}\}$，解釈担当者の既知のキーワードの MeSH ID リスト$\mathrm{K} = \{k_1, \ldots, k_m;\ k_l \in \mathrm{W}\}$で$\mathrm{C}, \mathrm{K}', \mathrm{K} \subset \mathrm{W}$であり、$\mathrm{C} \cap \mathrm{K}' = \phi$，$\mathrm{K} \cap \mathrm{K}' = \phi$，$\mathrm{C} \cap \mathrm{K} = \phi$である（$\phi$は空集合，$\subset$は真部分集合）。代謝物リストの候補は、10,141 個である。代謝物リストの候補 10,141 個から分析で得られた代謝物リストを決め、13,985 個の MeSH ID から、解釈担当者の既知のキーワード選び、それ以外が解釈担当者の知識外のキーワードリスト

になる。また、全部でU個（2020年6月現在、31,840,483個中メタボロミクスに関連するMeSHが付与されている20,159,576個を使用）の論文の集合D = $\{d_1, \ldots, d_U\}$があるとき, あるMeSH IDaが存在する論文の集合を$X_a \subset$ Dとする($X_a = \{d_1, \ldots, d_n;\ d_m \in$ D かつ d_mにはaが存在する})。ある代謝物の MeSH ID c_iの含まれる論文の集合，解釈担当者の知識外のキーワードMeSH ID k'_jを含む論文の集合，解釈担当者の既知のキーワードの MeSH ID k_lを含む論文の集合は，それぞれX_{c_i}, $X_{k'_j}$, $X_{k_l} \subset$ Dである。ある代謝物のMeSH ID c_iと解釈担当者の知識外のキーワード MeSH ID k'_jの関連度はA(X_{c_i}, $X_{k'_j}$)とし、

Cosineは

$$\text{Cosine}\ (X_{c_i}, X_{k'_j}) := \left|X_{c_i} \cap X_{k'_j}\right| / \sqrt{\left|X_{c_i}\right| \times \left|X_{k'_j}\right|},$$

Simpsonは

$$\text{Simpson}\ (X_{c_i}, X_{k'_j}) := \left|X_{c_i} \cap X_{k'_j}\right| / \min\left(\left|X_{c_i}\right|, \left|X_{k'_j}\right|\right),$$

集合 X に対する Confidence ($X_{c_i} \rightarrow X_{k'_j}$)は

$$\text{Confidence}\ (X_{c_i} \rightarrow X_{k'_j}) := \left|X_{c_i} \cap X_{k'_j}\right| / \left|X_{c_i}\right|,$$

Liftは

$$\text{Lift}\ (X_{c_i}, X_{k'_j}) := |\text{D}| \times \left|X_{c_i} \cap X_{k'_j}\right| / \left|X_{c_i}\right| \times \left|X_{k'_j}\right|$$

で計算された。具体的に、Table 3-1 に対して、c_1がD019802（コハク酸）、k_1がD042967（複合体Ⅰ）を例として考える。論文集合X_{c_1}はコハク酸を含む論文1が該当し、論文集合X_{k_1}は複合体Ⅰを含む論文1と2が該当し、論文集合$X_{c_1} \cap X_{k_1}$ はコハク酸と複合体Ⅰを含む論文1が該当する。したがって、Cosine (X_{c_1}, X_{k_1})= $\left|X_{c_1} \cap X_{k_1}\right| / \sqrt{\left|X_{c_1}\right| \times \left|X_{k_1}\right|}$は、それぞれの論文数から$1/\sqrt{1 \times 2}$である。ただし以降はA($X_{c_i}$, $X_{k'_j}$)を簡単のため$A(c, k')$と表記する（Cosine, Simpson等も同様）。また、c_iの添え字を省略し、cと表記する。

Table 3-1 論文と MeSH ID のマトリックス

行が論文であり、列が MeSH ID である。論文に MeSH ID が含まれる場合には 1 で、含まれない場合には 0 を表す。

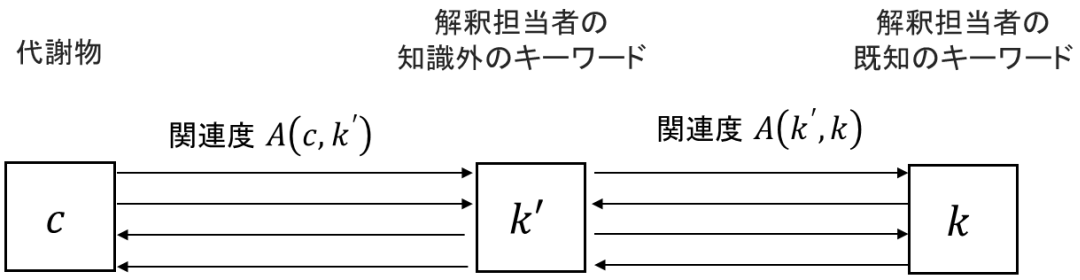
	D019802（コハク酸）	D042967（複合体Ⅰ）
論文 1: Chouchaniet et al, 2014	1	1
論文 2: Li et al., 2019	0	1
論文 3: Pistritto et al., 2016	0	0

これらの関連度の指標を用いて、代謝物リストの MeSH ID cと解釈担当者の知識外のキーワードk′の関連度$A(c,k')$と解釈担当者の知識外のキーワードk′と解釈担当者の既知のキーワードkの関連度$A(k',k)$が計算された (Figure 3-1a)。関連度$A(c,k')$の Confidence に関しては、向きがあるため、代謝物リストのキーワード cが存在する論文集合X_cに対する Confidence (c → k')と解釈担当者の知識外のキーワードk'が存在する論文集合$X_{k'}$に対する Confidence (k' → c)の 2 通りが計算された。同様に、関連度 A(k', k)に対しても Confidence (k' → k)と Confidence (k → k')の 2 通りが計算された。

次に、関連度 A(c, k')と関連度 A(k', k)のスコアを

スコア(c, k′, k):= 関連度$A(c,k')$×関連度$A(k',k)$

として計算した。Confidence には向きがあり、どの向きが良いかを調べるために、関連度$A(c,k')$では、Confidence (c → k')と Confidence (k' → c)の 2 通りが計算され、関連度$A(k',k)$に対しても Confidence (k' → k)と Confidence (k → k')の 2 通りが計算された。スコアに関しては、関連度$A(c,k')$×関連度$A(k',k)$の 4 通りが計算された。Confidence のスコアに関しては、Confidence (c → k')と Confidence (k' → k)の積を、Confidence (c → k' → k)と書くことにし、さらにキーワードを省略し、Confidence (→,→)と書くことにする。その他の confidence に関してもキーワードを省略し、Confidence (→,←), Confidence (←,→), Confidence (←, ←) と書くことにする (Figure 3-2)。



$$\text{Confidence}(\rightarrow,\rightarrow) := \text{Confidence}\ (\text{c} \rightarrow k') \times \text{Confidence}\ (k' \rightarrow \text{k})$$

$$\text{Confidence}(\rightarrow,\leftarrow) := \text{Confidence}\ (\text{c} \rightarrow k') \times \text{Confidence}\ (k' \leftarrow \text{k})$$

$$\text{Confidence}(\leftarrow,\rightarrow) := \text{Confidence}\ (\text{c} \leftarrow k') \times \text{Confidence}\ (k' \rightarrow \text{k})$$

$$\text{Confidence}(\leftarrow,\leftarrow) := \text{Confidence}\ (\text{c} \leftarrow k') \times \text{Confidence}\ (k' \leftarrow \text{k})$$

Figure 3-2　Confidence のスコア値

3-2-4 ランダム DB の作成方法

上記の方法で推定したスコアの、統計的有意性の検定を行う。その際生成する帰無分布に用いるランダム DB は Figure 3-3 に示す手法で作成した。本章で取り扱う文献情報の実データとは、各文献に付与されたキーワードのリストである (Figure 3-3a)。例えば文献 1 にはキーワード a, b, …, c が付与されている。

次に、ランダム DB とは、論文・キーワードの数と、各キーワードの登場回数、各論文に割り当てられたキーワードの数が、実 DB と同じになるように、キーワードをランダムシャッフルした DB である。まず、

1. ランダム DB の論文リストを初期化する。論文リストの長さは実 DB の論文数と同じで、ランダム DB の各論文は実 DB の論文と対応している。この時点ではランダム DB のいずれの論文に割り当てられたキーワードの数は 0 である (Figure 3-3a)。
2. 実 DB の全キーワードからなるキーワードリストを作る。キーワードリストは実 DB における各キーワードの登場回数で降順ソートしておく (Figure 3-3b)。
3. キーワードリストの先頭のキーワードを X、その（実 DB における）登場回数を A とする。Figure 3-3b の例では、X がキーワード e にあたり、e の個数を A とする。

4. 論文リストから一様ランダムに A 個の論文を選ぶ。ただし、論文リストの長さが A 以下の場合はリスト中の全ての論文を選ぶ。もし、割当先がなかったらキーワードは破棄する。選ばれた各論文にキーワード X を追加する。論文に割り当てられたキーワードが実 DB と同じ値に達した（割り当て先がなくなった）論文は、論文リストの割り当て対象から削除する (Figure 3-3c)。
5. キーワードリストの先頭から X を削除する。
6. キーワードリストが空であれば終了、空でなければ 3. に戻る。

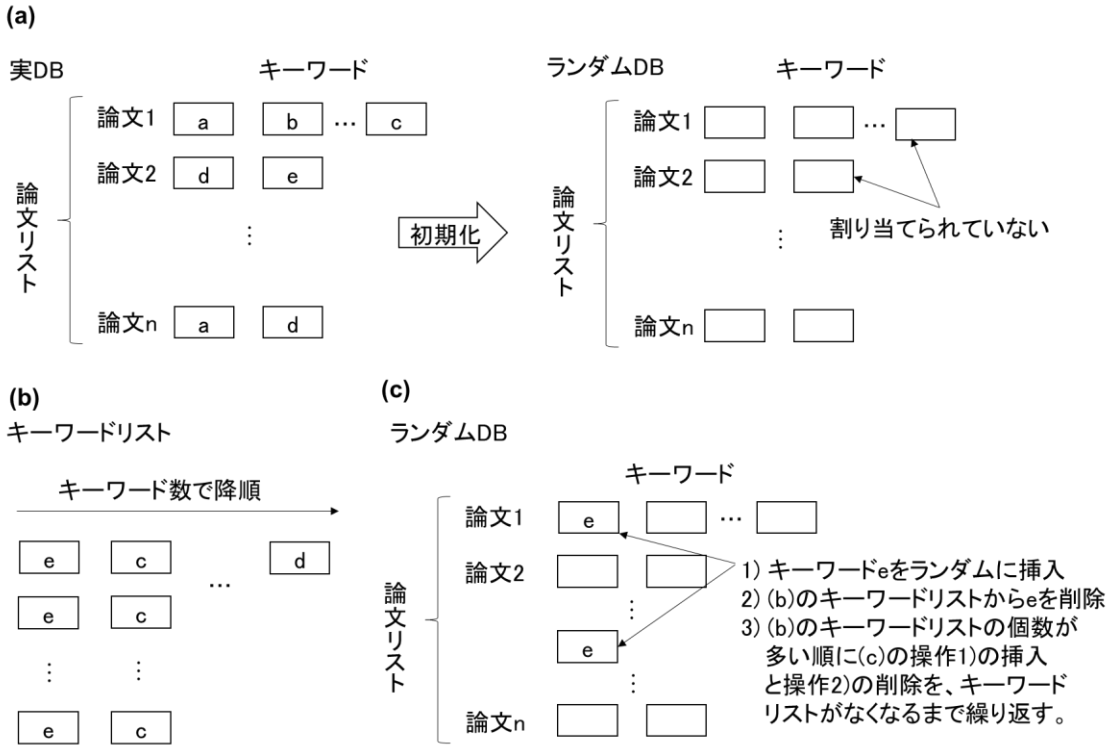


Figure 3-3 ランダム DB 作成方法

(a) 初期化手順。実 DB では、各論文に対応したキーワードが割り当てられている。論文の数と割り当て可能なキーワード数は変えず、キーワードを抽出する。(b) キーワードリスト。(a) で抽出したキーワードをキーワード数の降順に並べる。(c) (b) のキーワードリストからランダム DB へキーワードをランダムに挿入する。

3-2-5 統計的に有意なキーワードの閾値を求める手法の開発

作成したランダム DB を用いて下記のように帰無分布を生成した。3-2-4 で説明したあるスコアリング法の帰無分布は、一様ランダムに 3 つのキーワードを選び（代謝物、解釈担当者の既知のキーワード、解釈担当者の知識外のキーワードに相当する）、その 3 つのキーワードの関連性のスコアを計算するというサンプリング操作を 1e+8 回行った。このスコアの分布を帰無分布とした。（ランダム DB に起因する偏りを防ぐため、実際は５個のランダム DB それぞれで 2e+7 回のサンプリングを行い、合計した）。３つのキーワードから実 DB を用いて計算したスコアを S とする。S に対し

p 値 ＝ （帰無分布中の、S 以上のスコアの個数）/（帰無分布の全点数）

で p 値を求めた。多重検定における p 値の補正方法として Benjamini & Hochberg 法 (BH 法) (Benjamini et al., 1995) を用いた。BH 法は、False Discovery Rate (FDR) を調整する方法である。FDR とは、真の帰無仮説を誤って棄却（α エラー）した数に対して、帰無仮説を棄却した総数で割った値である。BH 法は次の手順で計算する。まず、全ての p 値が一様分布に従うと仮定して、q 値は p 値を p 値以下である確率（p 値以下の一様分布の面積）で割った値として計算する (Figure 3-4ab)。

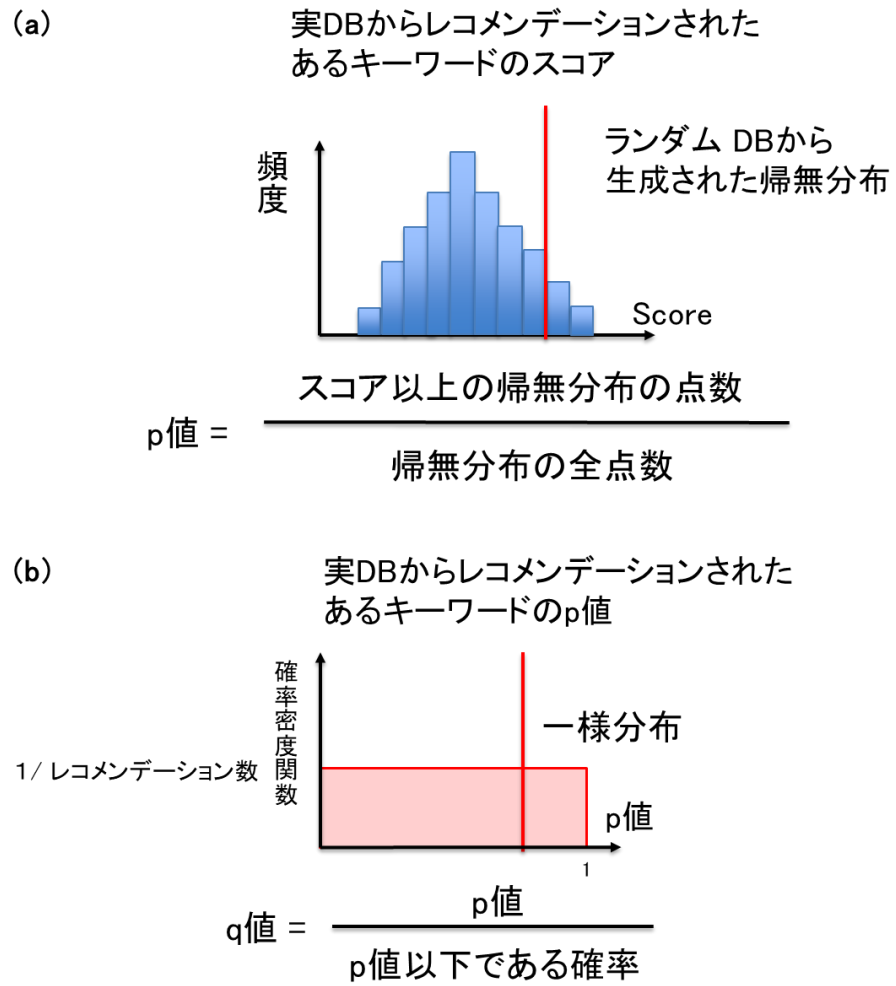


Figure 3-4　p 値と q 値の計算方法

(a) p 値の計算方法。ランダム DB から帰無分布を生成した（青の頻度分布）。実 DB (PubMed) とキーワードを用いてレコメンドすると、1 つのキーワードに対して、スコアが計算される。そのスコアと帰無分布を用いて、p 値を計算する。(b) 全ての p 値が一様分布に従うと仮定して、q 値は p 値を p 値以下である確率（p 値以下の一様分布の面積）で割った値として計算される。

次に、求まった q 値に対して FDR の閾値を定め、BH 法で統計的に有意な閾値を決定した。具体的には、

1. p 値を昇順ソートする。
2. p 値を q 値へ変換する。

3. p 値の大きいものから FDR の閾値を順番に検討し、q 値が閾値以下になる最初のキーワードを閾値として設定する。

というステップで閾値を決める。例えば、Figure 3-5 では、p 値 0.19, 0.09, 0.13, 0.01, 0.02 に対して FDR0.01 として適用すると、番号 4 と 5 が該当する。BH 法で用いられる q 値と閾値の判定はライブラリ statsmodels.stats.multitest (Seabold et al., 2010) を用いて計算した。

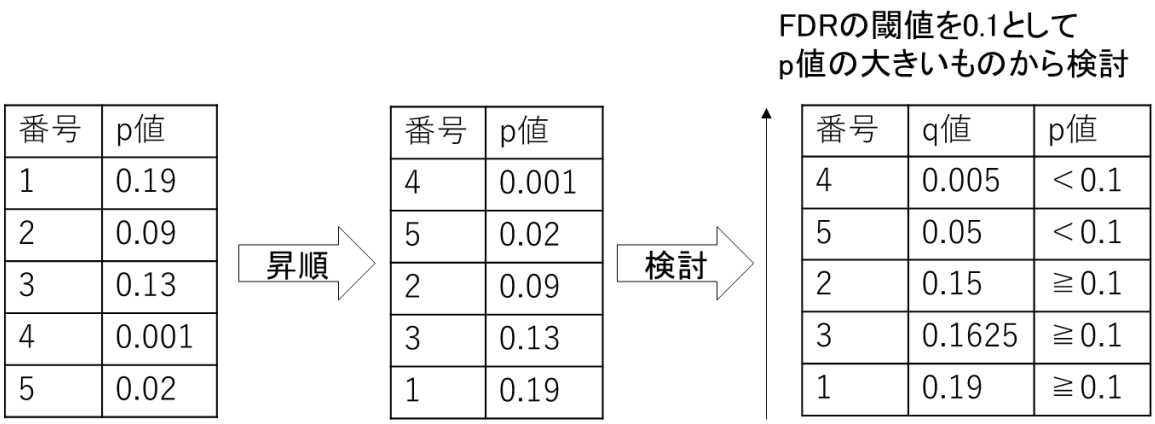


Figure 3-5　統計的に有意な閾値の決め方

p 値を昇順になるように並び替えて、q 値へ変換する。FDR の閾値を設定し、p 値の大きいものから検討する。

本研究では、5 個のランダム DB から 2e+7 点ずつサンプリングして（合計 1e+8 点）帰無分布を作成したが、ランダム DB の個数が 5 個（サンプリング点数が 1e+8）で十分かを検証した。まず、あるランダム DB から 2e+7 点のスコアをサンプリングした帰無分布 D1 と、別のランダム DB から 2e+7 点のスコアをサンプリングした帰無分布 D2 を用意した。そして、入力キーワードとして 3 つのキーワード（代謝物と解釈担当者の既知のキーワード、解釈担当者の知識外のキーワード）の組み合わせを、1e+7 通りランダムに生成した。1e+7 通りのキーワードを実 DB に入力し、スコアを計算した。得られたスコアの集合を S とする。

1. S の各要素に対し、帰無分布 D1 を用いて p 値を計算する。得られた p 値の集合を P1 とする。

2. S の各要素に対し、帰無分布 D2 を用いて p 値を計算する。得られた p 値の集合を P2 とする。
3. P1 と P2 の二乗平均平行根誤差が 1e-5 以上 1e-6 未満であった。

同じ手順を 10 回行ったが、二乗平均平行根誤差は下記の通り、1e-5 以上 1e-6 未満であり、1 つの DB でも p 値のばらつきが少ないことがわかる。

Table 3-2 1 個ランダム DB から生成した帰無分布の誤差（二乗平均平方根誤差）

回数	二乗平均平行根誤差
1	2.46e-06
2	2.30e-06
3	1.88e-06
4	2.94e-06
5	3.34e-06
6	2.75e-06
7	4.43e-06
8	1.63e-06
9	2.41e-06
10	3.15e-06

また、1 個のランダム DB から 2 つの帰無分布を作成していたが、2 個のランダム DB から 2 つの帰無分布を生成して同様の比較をしたところ、1e-5 以上 1e-6 未満であることからランダム DB は 1 個以上あれば十分であることがわかる (Table 3-3)。

Table 3-3 2 個のランダム DB から生成した帰無分布の誤差(二乗平均平方根誤差)

回数	二乗平均平行根誤差
1	1.33e-05
2	2.47e-06

3-3 結果と考察

3-3-1 実行環境

全てのレコメンド機能は、DGX-Station (CPU Intel Xeon (R) CPU E5 2698 v4 @ 2.20GHz × 40, 20 physical Cores (40 with Hyper Threading), RAM 256GB, OS Ubuntu 18.04) で実行される Python 3 上で実装および実行された。それ以外のソフトウェアは Windows10（CPU：Intel Core i5 (2.40GHz), メモリ：8.00GB, OS：Windows 10 Pro 64 ビット）上で使用した。

3-3-2 実験ワークフロー

本研究は 4 段階からなる。その実験デザインを Figure 3-6 に示した。

1　抗がん剤 4-OHT を処理したヒト乳がん細胞 MCF-7 の時系列メタボロミクスデータをテストデータとして収集した。

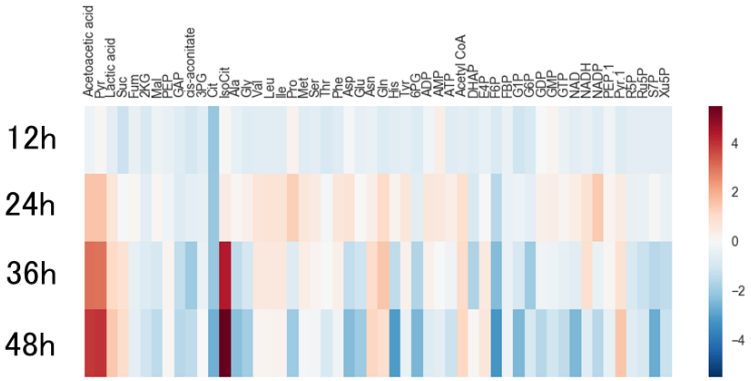
2　テストデータの各時刻で 4-OHT 処理群とコントロール群をボルケーノプロット法で比較し、有意に含量が変動した代謝物について、MeSH ID へ変換することで「代謝物リスト」を作成した。

3　各時刻で「代謝物リスト」と「解釈担当者の既知のキーワード（MeSH ID: D017209 (アポトーシス)と MeSH ID: D000970（抗腫瘍剤））」をテストデータとすることで、有用なレコメンド手法を開発した。

4　開発した手法の有用性を評価するために、１．模擬的状況での評価。２．模擬的状況で、PubMed 検索結果を絞り込めるかの評価。３．テストデータを用いた「解釈担当者の知識外のキーワード」のレコメンド結果の評価。を実施した。

(1)テストデータ収集

乳がん細胞株MCF-7
4-OHT処理群とコントロール
52 種類の代謝物含量



ボルケーノプロット法適用

(2)代謝物リスト作成

MeSH ID(代謝物)リスト

経過時間	有意に増減した代謝物名	MeSH ID（代謝物名）
12時間	コハク酸、クエン酸、G1P	D019802（コハク酸）、D019343（クエン酸）
24時間	ピルビン酸、クエン酸、プロリン、アセト酢酸、F6P	D019289（ピルビン酸）、D019343（クエン酸）、D011392（プロリン）
36時間	ピルビン酸、クエン酸、DHAP, 6PG, G6P、アセト酢酸、F6P	D019289（ピルビン酸）、D019343（クエン酸）、D004099(DHAP)
48時間	ピルビン酸、クエン酸、ヒスチジン、イソクエン酸、アセト酢酸、F6P, S7P	D019289（ピルビン酸）、D019343（クエン酸）、D006639（ヒスチジン）

(3) 統計的に有意なキーワードをレコメンドする手法の開発・評価

(4) 1. 模擬的状況での評価
2. 模擬的状況でPubMed検索結果を絞り込めるかの評価
3. テストデータを用いた「解釈担当者の知識外のキーワード」のレコメンド結果の評価

Figure 3-6 実験のワークフロー

3-3-3 テストデータ

テストデータとして、抗がん剤 4-OHT を処理したヒト乳がん細胞 MCF-7 のワイドターゲットメタボロミクスデータを用いた（西本ら、2020）。抗がん剤 4-OHT を投与すると薬剤応答を示した後にアポトーシスに至ることが知られている (Rafal et al., 2007)。MCF-7 に 10 μ M の 4-OHT を処理し、処理群と非処理群の 12h, 24h, 36h, 48h 後の 52 種類の代謝中間体含量をワイドターゲットメタボロミクス法で測定したものである。実験は３連で行われたため、2 群×4 時点×3 連×52 代謝物＝1,248 データポイントからなる (Figure 3-7)。各経過時間での非処理区に対する処理区の代謝中間体含量の比 (Fold Change) を見ると、多くの代謝物で含量が増減していることがわかる (Figure 3-7)。

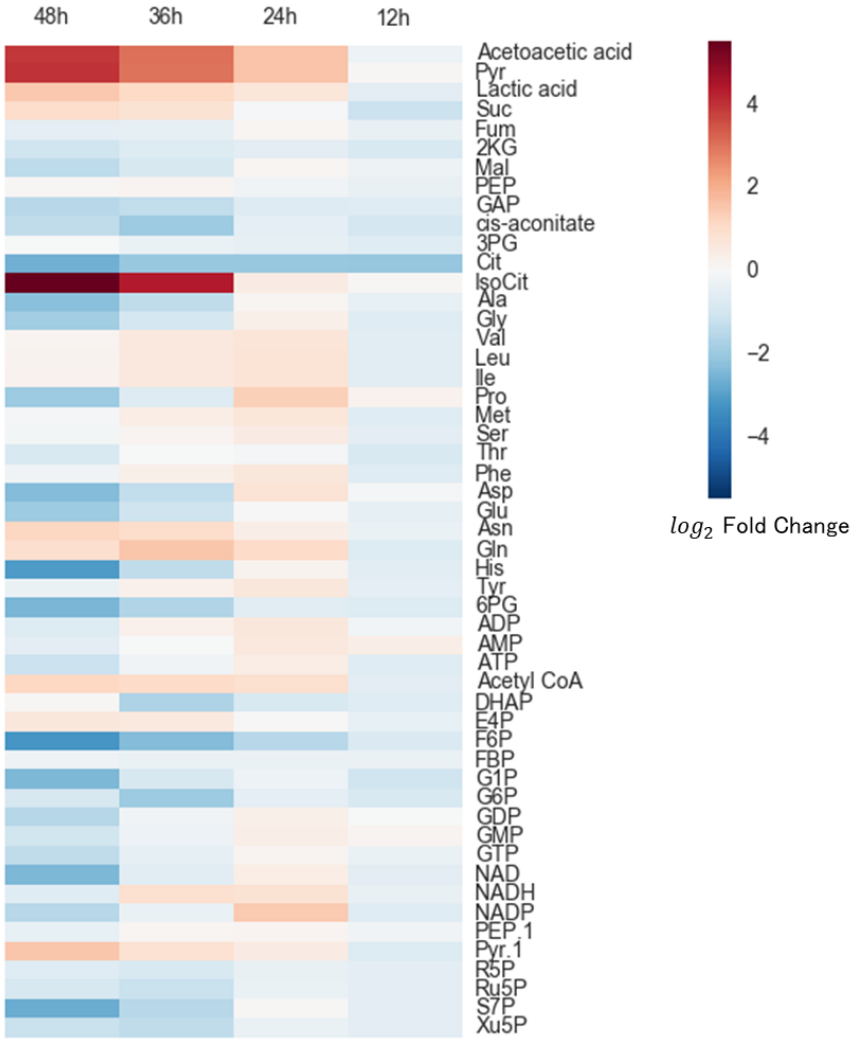


Figure 3-7 4-OHT の投与有無で測定した時系列データ

24 サンプル（時間数:4、2 群、n=3）×52 種類の代謝中間体含量における、4-OHT を処理した群に対して、4-OHT を処理していない群の Fold Change の対数値

3-3-4 代謝物リストの作成

まず、4-OHT 処理で増加、減少した代謝物の MeSH ID リストを得るため、ボルケーノプロット法を用いることとした。ボルケーノプロット法とは、2 群間で有意に含量が増減した代謝物をリスト化できる方法である。前項のテストデータを用いて、12 時間後の 4-OHT 処理群と非処理群の代謝中間体含量をボルケーノプロット法（閾値 p 値 0.05,

Fold Change 2.0）で比較したところ、コハク酸 (Suc) とクエン酸 (Cit)、グルコース-1-リン酸 (Glucose 1-phosphate, G1P) という代謝物が有意に増減した。但し、ここでの Fold Change 2.0 は、非処理区に対する処理区の代謝中間体含量が増加した場合の 2 倍または減少した場合の 2 分の 1 が閾値となる。MeSH ID の変換機能のみを使用し、D019802（コハク酸）、D019343（クエン酸）へ変換した。ただし、G1P に該当する MeSH ID はないので今回は用いなかった。ここで、代謝物と MeSH ID の対応関係を調べるため、解糖系、ペントースリン酸回路などの主要な代謝物 133 個を MeSH ID へ変換したところ、85 個（約 64.0%）が変換された。

次に 24 時間後のデータで同様の作業を行ったところ、多数の代謝物（ボルケーノプロット法で 7 個、該当する MeSH ID が 6 個）が有意に増減した。代謝物数が多いと、レコメンドの探索幅が広くなりすぎるため、MeSH ID の個数が 3 個になるように、Fold Change の閾値を 0.5 刻みで調整した（12 時間、24 時間、36 時間、48 時間の Fold Chagne の閾値はそれぞれ 2.0, 2.5, 3.0, 6.0）。Table 3-4 に、この方法で抽出した 12 時間、24 時間、36 時間、48 時間後に増減した代謝物のリストと MeSH ID を示す。この MeSH ID（代謝物名）は、Figure 3-1b の代謝物リストに相当する。

Table 3-4 有意に増減した代謝物と MeSH ID

経過時間	Fold Chagne の閾値	有意に増減した代謝物名	MeSH ID（代謝物名）
12 時間	2.0	コハク酸、クエン酸、G1P	D019802（コハク酸）、D019343（クエン酸）
24 時間	2.5	ピルビン酸、クエン酸、プロリン、アセト酢酸、フルクトース-6-リン酸	D019289（ピルビン酸）、D019343（クエン酸）、D011392（プロリン）
36 時間	3.0	ピルビン酸、クエン酸、ジヒドロキシアセトンリン酸 (Dihydroxyacetone Phosphate，DHAP), 6-ホスホグルコン酸、グルコース-6-リン酸、アセト酢酸、フルクトース-6-リン酸	D019289（ピルビン酸）、D019343（クエン酸）、D004099 (DHAP)
48 時間	6.0	ピルビン酸、クエン酸、ヒスチジン、イソクエン酸、アセト酢酸、フルクトース-6-リン酸、セドヘプツロース-7-リン酸	D019289（ピルビン酸）、D019343（クエン酸）、D006639（ヒスチジン）

3-3-5 統計的に有意なキーワードをレコメンドする手法の開発

ここまでに用意した「代謝物リスト」と「解釈担当者の既知のキーワードリスト」の MeSH ID から「解釈担当者の知識外のキーワードリスト」の MeSH ID をレコメンドすることが本章の目的である。そこで、文献データベース PubMed の全論文に付与されたキーワードである MeSH ID の共起情報を用いた手法を開発した。

まず、PubMed に掲載されている 31,840,483 報の論文から、題材とする一次代謝物のターゲットメタボロミクスに関連する MeSH が付与されている 20,159,576 報の論文を収集した（2020 年 6 月現在）。同様に 29,054 個の MeSH ID のうち、代謝に関係する化合物の 13,985 個の MeSH ID を用いた。そのうち、代謝物リストの候補となる総数は

8,889 個である。つぎに、MeSH ID 間の関連度の計算方法として、アソシエーション分析を適用した。キーワード同士の関連度の指標として、Cosine, Simpson, Confidence, Lift を用いた (Azevedo et al., 2007, Murguía et al., 2003)。これらの関連度の指標を用いて、代謝物リストの MeSH ID cと解釈担当者の知識外のキーワードk′の関連度$A(c, k')$と解釈担当者の知識外のキーワードk′と解釈担当者の既知のキーワードkの関連度$A(k', k)$が計算された (Figure 3-1a)。関連度$A(c, k')$の Confidence に関しては、向きがあるため、代謝物リストのキーワード cが存在する論文集合X_cに対する Confidence (c → k')と解釈担当者の知識外のキーワードk'が存在する論文集合$X_{k'}$に対する Confidence (k' → c)の 2 通りが計算された。次に、関連度 A(c, k')と関連度 A(k', k)の積をスコアとして計算した。Confidence のスコアに関しては、Confidence (c → k')と Confidence (k' → k)の積を、Confidence（c → k' → k) と書くことにし、さらにキーワードを省略し、Confidence (→,→)と書くことにする。その他の Confidence に関してもキーワードを省略し、Confidence (→,←), Confidence (←,→), Confidence (←, ←)と書くことにする。

次に、レコメンドされたキーワードのスコアのうち、信頼できる値を決定するため統計的に有意な閾値を計算した。そこで、あるキーワードを入力したときに、あるレコメンドされたキーワードのスコアが偶然起きるかを検定した。まず、論文中に現れる MeSH 数を変えずにランダムに MeSH ID を入れ替えたランダム DB を作成した（詳細は方法を参照)。ランダム DB 用い、一様ランダムに 3 つのキーワードを選び (代謝物、解釈担当者の既知のキーワード、解釈担当者の知識外のキーワード)、その 3 つのキーワードの関連性のスコアを計算するというサンプリング操作を 2e+7 回行った。1e+8 個のスコアを計算して、帰無分布を生成した。帰無分布を利用し p 値を計算し、BH 法 (Benjamini et al., 1995) を用いて統計的に有意な閾値を決定する手法を開発した。

まず、7 種の異なるスコア計算法を用いて、12 時間後のデータで有意に増減した代謝物リスト (MeSH ID) と、解釈担当者の既知のキーワード（MeSH ID: D017209 (アポトーシス)）からレコメンドされた、「解釈担当者の知識外のキーワード」の数を調べた (Figure 3-8a-1)。統計的に有意なキーワード数の割合（FDR が 0.01 以下の数を全レコメンド数で割った割合（%））を調べたところ、Confidence(←, →) で 1.59%（有意なキーワード数 445 個）が最も高く、ついで, Confidence(→, ←) 0.16%（有意なキーワ

ード数 45 個)、Cosine 0.14%（有意なキーワード数 38 個）の順で減少した（Figure 3-8a-1 は 12 時間、24 時間、36 時間、48 時間の平均値が高い順に左から並んでいる)。Confidence (→, →), Simpson, Lift, Confidence (←, ←) の割合は 0.00%であった。また、同様の作業を 24 時間,36 時間, 48 時間で増減した代謝物についても行ったところ、統計的に有意なキーワード数の割合は、Confidence (→, ←)と Cosine に変化はあったが、同様の傾向を示した。最も高かった Confidence (←,→) は、1.6%～10.3%（有意なキーワード数 445 個～4,334 個）となり、統計的に有意なキーワードがレコメンドされるスコアの計算方法が考えられた。

次に、12 時間後のデータで有意に増減した代謝物リスト (MeSH ID) と、解釈担当者の既知のキーワードとして MeSH ID: D000970（抗腫瘍剤）を用いて、同様の評価を行った (Figure 3-8a-2)。統計的に有意なキーワード数の割合（FDR が 0.01 以下の数を全レコメンド数で割った割合（%））を調べたところ、Confidence (←, →) で 0.48%（有意なキーワード数 135 個）が最も高く、ついで, Confidence (→, ←) 0.09%（有意なキーワード数 24 個）の順で減少した。Cosine , Confidence (→, →), Simpson, Lift, Confidence (←, ←) の割合は 0.00%であった。また、同様の作業を 24 時間、36 時間、48 時間で増減した代謝物についても行ったところ、統計的に有意なキーワード数の割合は、同様の傾向を示した。最も高かった Confidence (←,→) は、0.25%～5.45%（有意なキーワード数 135 個～2,287 個）となり統計的に有意なキーワードがレコメンドされるスコアの計算方法が考えられた。

さらに、「解釈担当者の既知のキーワードリスト」の有効性を検討した。２つの MeSH ID の関連度を定義するために Cosine や Simpson を用いる手法 (Ono, 2014) は既に存在する。Ono らの手法は、PubMed の論文と MeSH ID を用いて、疾患名から関連する別の疾患名を探すためにそれらの関連度を計算している。そこで、「解釈担当者の既知のキーワードリスト」を用いずに、「代謝物リスト」のみから統計的に有意なキーワード数の割合を計算した（Figure 3-9a)。その結果、全ての指標で 0.00%（有意なキーワードは 0 個）と、解釈担当者の既知のキーワードを用いた場合よりも非常に低い結果となった（Figure 3-9b)。

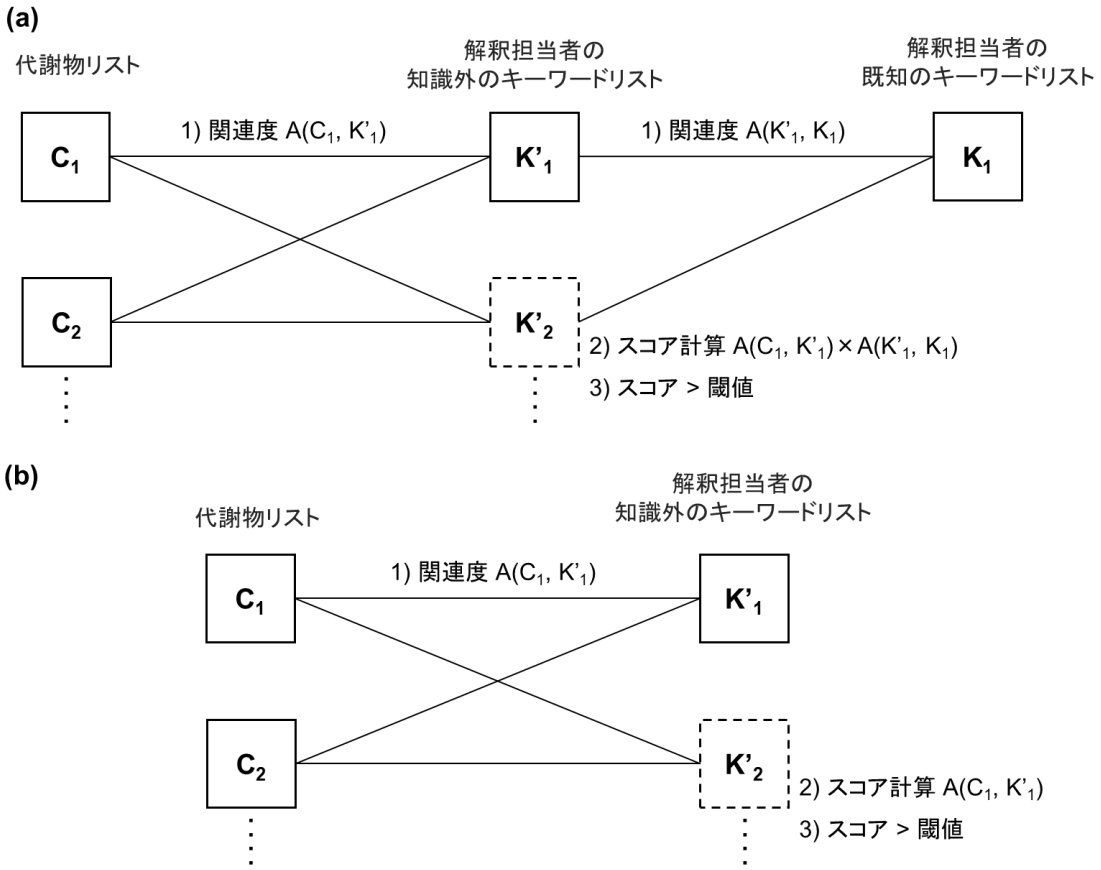


Figure 3-8　(a) 解釈担当者の既知のキーワードリストありの場合のレコメンド手法と (b) 解釈担当者の既知のキーワードリストなしの場合。

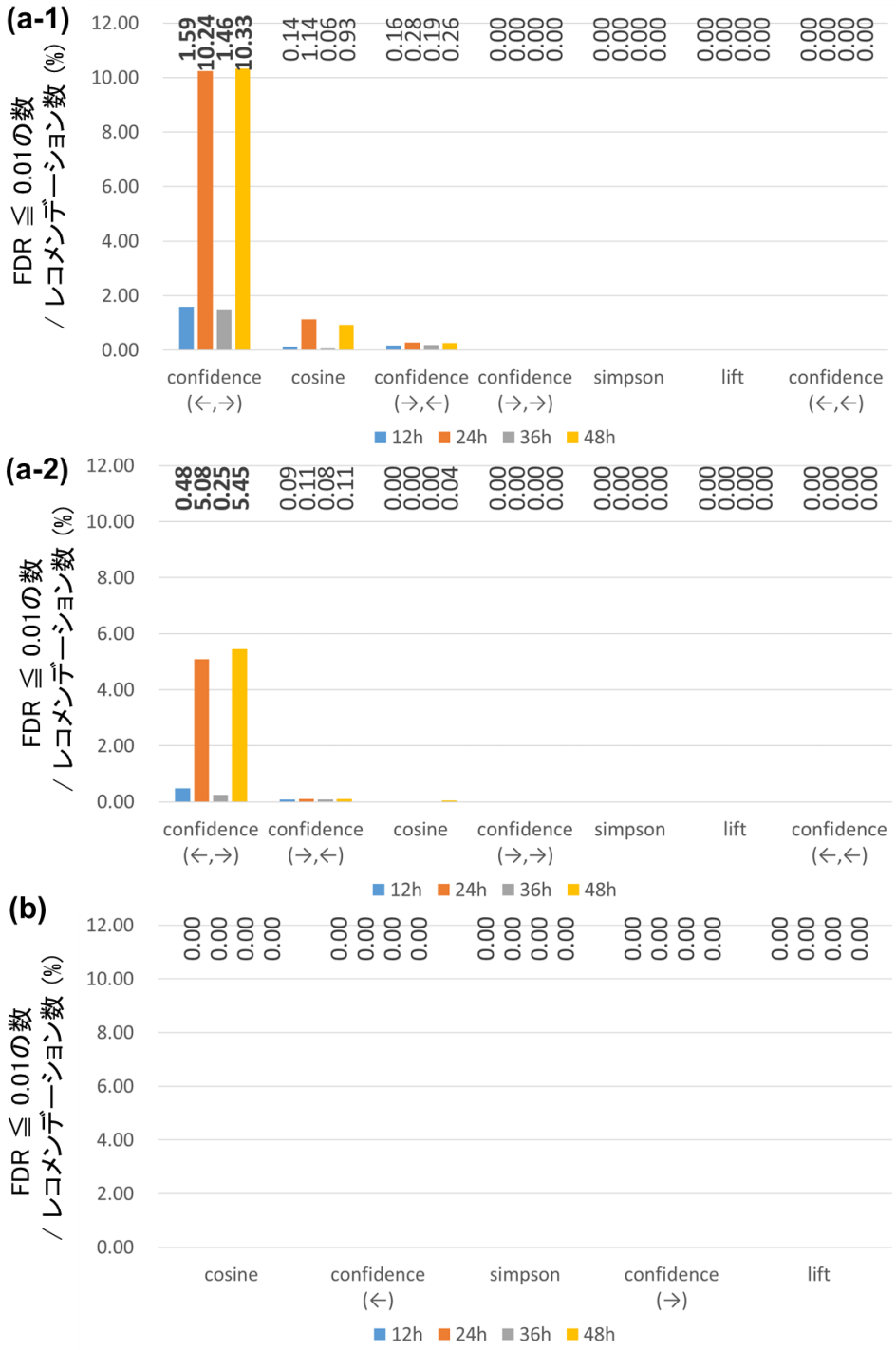


Figure 3-9.各関連度で計算した統計的に有意な個数の割合

(a) 解釈担当者の既知のキーワードあり（(a-1)：アポトーシス、(a-2)：抗腫瘍剤）で各関連度を比較した結果。(b) 解釈担当者の既知のキーワードなしで各関連度を比較した結果。

3-4 評価

本研究で開発した手法が、実際のデータ解釈に有効かどうかを評価するには、正解のがわかっているテストデータを用いた模擬評価が有用である。具体的には、知見を得る前の状況を想定して、当時の知識外のキーワードがレコメンドされるかを検証した。そこで、テストデータを用いた３項目の評価を行った（Figure 3-10）。

1. テストデータに関する「解釈担当者の既知のキーワード」を、模擬的に「解釈担当者の既知のキーワード」と「解釈担当者の知識外のキーワード」に分け、3-3-2 項で作成した「代謝物リスト」と「解釈担当者の既知のキーワード」から、「解釈担当者の知識外のキーワード」をレコメンドできるかを評価した。
2. 「代謝物」＆「解釈担当者の既知のキーワード」を用いてヒットする PubMed 文献検索結果と、「代謝物」＆「解釈担当者の既知のキーワード」＆レコメンドされた「解釈担当者の知識外のキーワード」を含む論文を比較し、参考とする文献の絞り込みが可能かを評価した。
3. 「代謝物リスト」と「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」のレコメンドを行い、どのようなキーワードがランキング上位にレコメンドされるかを評価した。

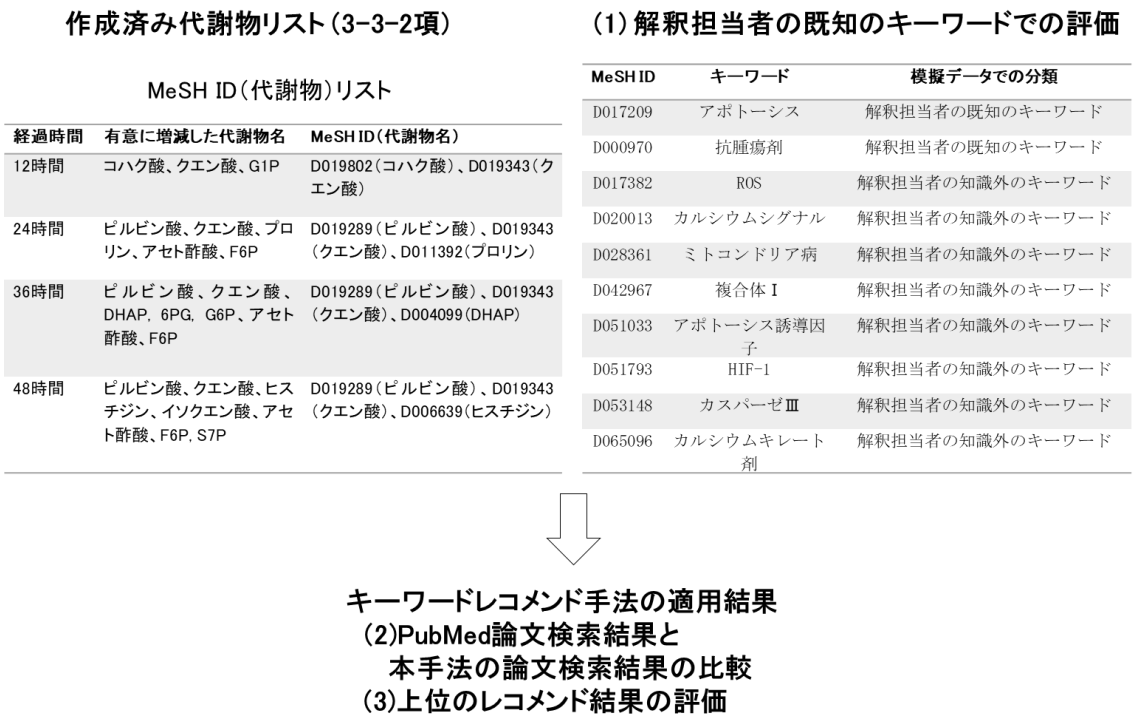


Figure 3-10　実施例のワークフロー

3-4-1 模擬データによる評価

上述のように、抗がん剤 4-OHT を投与すると薬剤応答を示した後にアポトーシスに至ることが知られている (Rafal et al., 2007)。模擬的なテストを行うため、代謝物リストと既知のキーワードであるアポトーシスの間を埋めるキーワード（正しい例）をマニュアルで探した。2 週間の文献調査の結果、下記の論文と MeSH ID を取得した (Table 3-5)。

1. 抗がん剤 4-OHT を投与すると複合体Ⅰが不活性化される (Paula et al., 2005)。
2. 抗がん剤 4-OHT を投与すると通常低い値であるカルシウムシグナルが増加し、その後、ROS が発生し、アポトーシスの重要なメディエータであるカスパーゼⅢが活性化する (Lee et al., 2000, Paul et al., 1994, Rafal et al., 2007, Zhang et al., 2000)。
3. ROS の発生により HIF-1 が活性化することも知られている (Yu et al., 2015)。

Table 3-5 解釈担当者の既知のキーワードと模擬データでの分類

MeSH ID	キーワード	模擬データでの分類
D017209	アポトーシス	解釈担当者の既知のキーワード
D000970	抗腫瘍剤	解釈担当者の既知のキーワード
D017382	ROS	解釈担当者の知識外のキーワード
D020013	カルシウムシグナル	解釈担当者の知識外のキーワード
D028361	ミトコンドリア病	解釈担当者の知識外のキーワード
D042967	複合体Ⅰ	解釈担当者の知識外のキーワード
D051033	アポトーシス誘導因子	解釈担当者の知識外のキーワード
D051793	HIF-1	解釈担当者の知識外のキーワード
D053148	カスパーゼⅢ	解釈担当者の知識外のキーワード
D065096	カルシウムキレート剤	解釈担当者の知識外のキーワード

次に、知見を得る前の状況を想定して、当時の知識外のキーワードがレコメンドされるかを検証した。そこで、Table 3-5 の「解釈担当者の既知のキーワード」を模擬的に「解釈担当者の既知のキーワード（アポトーシス、D017209）」と「解釈担当者の知識外のキーワード（８個）」に分けた。そして、3-3-2 項で作成した「代謝物リスト」と「解釈担当者の既知のキーワード」を用いてレコメンドを行うと、「解釈担当者の知識外のキーワード」をレコメンドできるかを検証した。統計的に有意なキーワード数の割合が多かった Confidence (←,→) を用いて、アポトーシスが「解釈担当者の既知のキーワード」である。その結果、FDR が 0.01 以下で連想されたキーワードには「解釈担当者の知識外のキーワード」８個のすべて（複合体Ⅰ、カルシウムシグナル、カルシウムキレート剤、ROS, HIF-1, カスパーゼⅢ、ミトコンドリア病、アポトーシス誘導因子）が含まれていた (Table 3-6)。

次に、「解釈担当者の既知のキーワード」を（抗腫瘍剤、D000970）と「解釈担当者の知識外のキーワード（８個）」に分けたときの結果を Table 3-7 に示した。レコメンドされたキーワードには「解釈担当者の模擬知識外のキーワード」８個のうち、ミトコンドリア病とカスパーゼⅢを除く６個が含まれていた。これらの例は、本研究で開発した手法を用いることで、人手を用いて探索したキーワードをレコメンドしうることを示唆している。

また、「代謝物リスト」のみを用いてレコメンドを行ったところレコメンドされたキーワードはなかった。この結果から、解釈担当者の既知のキーワードを加味して関連度を計算することが重要であると示唆された。

Table 3-6 解釈担当者の模擬既知のキーワードあり（アポトーシス）

MeSH ID	キーワード	q 値
D017382	ROS	24h：0.0018
D020013	カルシウムシグナル	24h：0.0067
D028361	ミトコンドリア病	12h：0.0097，24h：0.0020
D042967	複合体 I	12h：0.0041
D051033	アポトーシス誘導因子	12h：0.0047，24h：0.0016
D051793	HIF-1	24h：0.0019
D053148	カスパーゼⅢ	36h：0.0045
D065096	カルシウムキレート剤	12h：0.0029，24h：0.0010

Table 3-7 解釈担当者の模擬既知のキーワードあり（抗腫瘍剤）

MeSH ID	キーワード	q 値
D017382	ROS	24h：0.0052
D020013	カルシウムシグナル	24h：0.0
D042967	複合体 I	12h：0.0089
D051033	アポトーシス誘導因子	24h：0.0054
D051793	HIF-1	24h：0.0042
D065096	カルシウムキレート剤	12h：0.0063，24h：0.0027

3-4-2 PubMed 文献検索結果の絞り込みによる評価

PubMed と MeSH ID を用いる利点の一つに、レコメンドの結果から読むべき論文をリストアップできる点にある。例えば前節の模擬的な状況では、「代謝物」のコハク酸と「解釈担当者の既知のキーワード」であるアポトーシスから、「解釈担当者の知識外のキーワード」として複合体 I をレコメンドできた。そこで、コハク酸とアポトーシスの MeSH ID を用いて PubMed で論文検索（アンド検索）すると、32 報の論文が検索された (Figure 3-11a)。つぎに複合体 I の MeSH ID も加味すると、3 つの MeSH ID が同時に付与された 2 報の論文が見つかった。この 2 論文は解釈に関連のある論文だった。1 つ目の論文は、グルコースを用いて細胞内の ROS を抑制していることや、ROS を減少させるために NADPH を使うことでアポトーシスを阻害することが述べられている

(Martens et al., 2005)。2 つ目の論文はα-ケトグルタル酸とコハク酸が HIF-1 の活性に関係のあることが述べられていた (Gasparre et al, 2011)。

同様に、ピルビン酸とアポトーシスの MeSH ID を用いて PubMed で論文検索（アンド検索）すると、86 報の論文が検索された (Figure 3-11b)。さらに複合体 I の MeSH ID も加味すると、3 つの MeSH ID が同時に付与された論文が見つからなかった。このように論文を絞り込めない例も見られた。同様の実施例を「クエン酸」「抗腫瘍剤」「ROS」で実施したところ、35 報の論文を 1 報に絞り込めた。

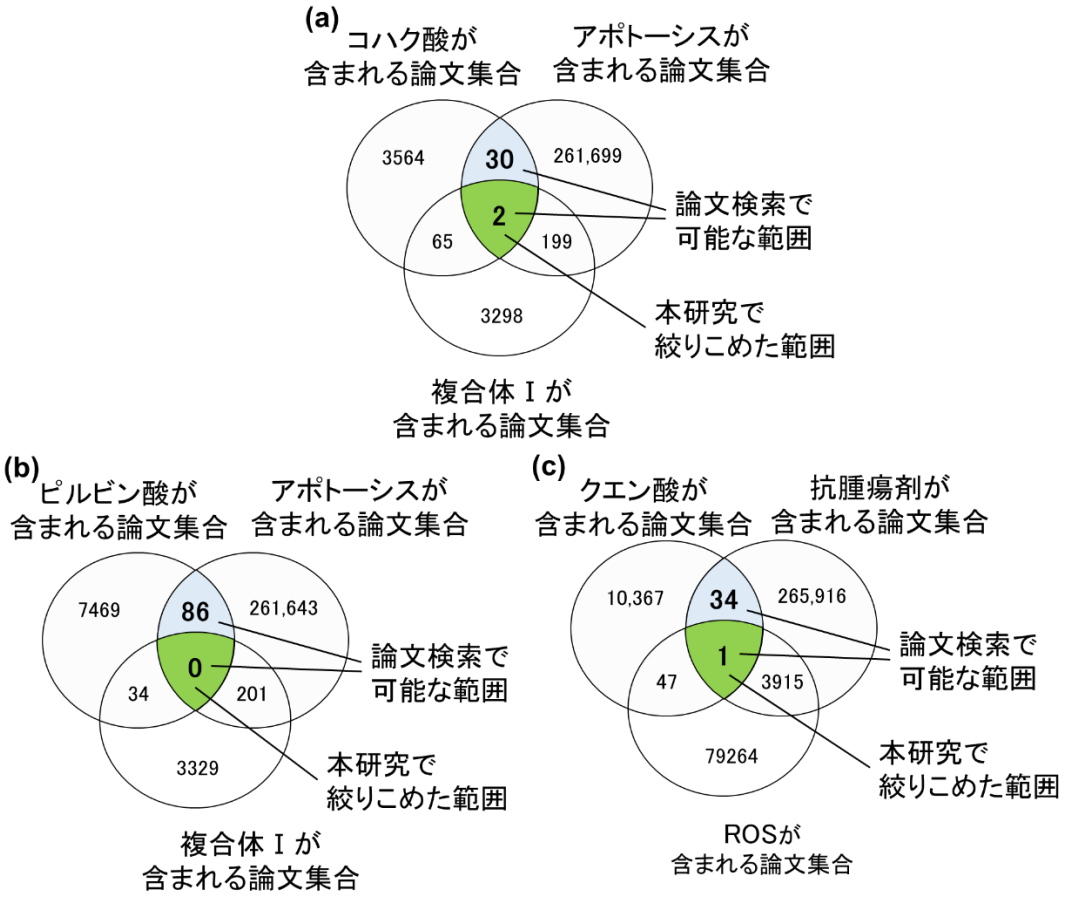


Figure 3-11 代謝物、解釈担当者の既知のキーワード、解釈担当者の知識外のキーワードを含む論文数

水色が論文検索で可能な論文集合であり、緑色が本研究で検索可能になった論文集合である。

3-4-3 「解釈担当者の知識外のキーワード」のレコメンド結果の評価

最後に、本研究で用いたテストデータを用いたレコメンドを実際に行い、Table 3-6 に示した「解釈担当者の既知のキーワード（アポトーシス）」以外のどのようなキーワードが上位にレコメンドするかを検証した。時系列データの 12 時間目で有意に増減した代謝物リスト（D019802（コハク酸）、D019343（クエン酸））と（Table 3-4）、解釈担当者の既知のキーワード（アポトーシス、D017209）を用いてレコメンドを行った（Table 3-8）。その結果、トップ５は１位 D065096（カルシウムキレート剤）、２位 D013804（テノイルトリフルオロアセトン）、３位 D053478（アポトソーム）、４位 D000154（アコニット酸ヒドラターゼ）、５位 D012402（ロテノン）であり、１位 D065096（カルシウムキレート剤）の以外が「解釈担当者の知識外のキーワード」と異なっていた。同様の解析を時系列データの 24、36、48 時間目を用いて行った（Table 3-4）。上位のキーワード（重複なし 13 個）のうち 12 個が「解釈担当者の知識外のキーワード（Table 3-5）」と異なっていた。

次に、各時刻のトップ２の結果を精査した。D013804（テノイルトリフルオロアセトン）は既知キーワードである複合体Ⅱの阻害剤である。24 時間の 1 位プロリンオキシダーゼと 2 位 1-ピロリン-5-カルボン酸デヒドロゲナーゼはプロリンの代謝とアポトーシスに関連する分子だった。36 時間の 1 位アデニンヌクレオチドトランスロケーター3 と 2 位アデニンヌクレオチドトランスロケーター2 はミトコンドリアのエネルギー生産やアポトーシスに関わっていた。一方、48 時間の 1 位 B30.2-SPRY ドメインとがん細胞や代謝との関連はわからなかった。全体として新たにレコメンドされた重複無しの 7 個のキーワードのうち、5 個は研究に役立つ可能性があるキーワードだった。

Table 3-8「解釈担当者の知識外のキーワード」のレコメンド結果の評価

経過時間	12 時間	24 時間	36 時間	48 時間
代謝物リスト	D019802（コハク酸)、D019343（クエン酸）	D019289（ピルビン酸)、D019343（クエン酸)、D011392（プロリン）	D019289（ピルビン酸)、D019343（クエン酸)、D004099 (DHAP)	D019289（ピルビン酸)、D019343（クエン酸)、D006639（ヒスチジン）
解釈担当者の既知のキーワード	D017209 (アポトーシス)	D017209 (アポトーシス)	D017209 (アポトーシス)	D017209 (アポトーシス)
1 位	D065096（カルシウムキレート剤）	D011394（プロリンオキシダーゼ）	D033781（アデニンヌクレオチドトランスロケーター3）	D000071197（B30.2/SPRY ドメイン）
2 位	D013804（テノイルトリフルオロアセトン）	D050842（1-ピロリン-5-カルボン酸デヒドロゲナーゼ）	D033742（アデニンヌクレオチドトランスロケーター2）	D033781（アデニンヌクレオチドトランスロケーター3）
3 位	D053478（アポトソーム）	D033781（アデニンヌクレオチドトランスロケーター3）	D065096（カルシウムキレート剤）	D000071476（カスパーゼ活性化誘引ドメイン）
4 位	D000154（アコニット酸ヒドラターゼ）	D050698（グルタミン酸セミアルデヒドデヒドロゲナーゼ）	D000081382（ピルビン酸デヒドロゲナーゼ キナーゼ）	D033742（アデニンヌクレオチドトランスロケーター2）
5 位	D012402（ロテノン）	D033781（アデニンヌクレオチドトランスロケーター3）	D053478（アポトソーム）	D065096 (カルシウムキレート剤)

3-5 結言

本章では、メタボロミクスデータを解釈する作業を支援するための技術を開発した。本章の作業を実施するには、ピークピッキング済のメタボロームデータを統計解析することで得た「代謝物リスト」が必要である。したがって、高精度なピークピッキングによる、正確な代謝物定量値が妥当なキーワード連想を行うための必要条件の一つである。

まず、アソシエーション分析の手法を活用し、PubMed の MeSH ID の共起情報を用いることで、「代謝物リスト」と「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」をレコメンドする手法を開発し、1 つ目の目標を達成した。また、ランダム DB を用いて関連度の帰無分布を作成して関連度スコアの p 値を得ることを可能とし、そこから妥当な閾値を設定する手法を開発した。これにより 2 つめの目標も達成した。

つぎに、本手法の有用性を評価するため、知見を得る前の状況を想定して、当時の知識外のキーワードがレコメンドされるかを検証した。テストデータに関する「解釈担当者の既知のキーワード」を、「解釈担当者の既知のキーワード」と「解釈担当者の知識外のキーワード」に分け、模擬的なレコメンドを行った。その結果「知識外のキーワード」をレコメンドできていた (Table 3-6)。本研究で構築した手法がレコメンドするキーワードとは、「代謝物」の MeSH ID と「解釈担当者の既知のキーワード」の MeSH ID の両方と、PubMed で高頻度に共起する MeSH ID である。このため、まだ PubMed の文献中で共起が起きていない新たな関連を見出すことはできない。また、レコメンドするキーワードが、解釈担当者にとって新しいか、すなわち未知かどうかは、解釈担当者の知識に依存する。つまり、解釈担当者にとって新しいキーワードを必ずレコメンドすることを保証するものではない。一方、文献上で関連する「解釈担当者の知識外のキーワード」のリストを得ることで、解釈担当者の知識レベルに依存せず関連するキーワードを網羅的に列挙することが可能となった。

さらに、コハク酸とアポトーシスに関する論文検索（アンド検索）では 32 報の論文がヒットしたのに対し、模擬的に「複合体 I 」をさらに付け加えることで、該当論文を 2 報にまで絞り込むことができた（Figure 3-11 (a)）。これも必ず絞り込みできることは保証されないが、「代謝物」と「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」を得て、これを用いて文献検索することで、優先的に読むべき論文を絞り込める、すなわち大幅なデータ解釈時間短縮につながる可能性が示唆された。

最後に、テストデータを用いて、「解釈担当者の知識外のキーワード」をレコメンドした (Table 3-8)。その結果、上位にランキングされたキーワード計 13 個のうち、12 個が「解釈担当者の知識外のキーワード」と異なっていた。次に、各時刻のトップ２の結果を精査した。全体として新たにレコメンドされた重複無しの 7 個のキーワードのうち、5 個は研究に役立つ可能性があるキーワードだった (Table 3-8)。これは研究者が当初思いつかなかったキーワードを 5 個提示できたことを示している。これらのキーワードを、本手法を用いずに従来通り、論文検索を通じて見出すには、数日レベルでの調査時間が必要であると推測された。

キーワード連想作業は従来では、研究者の知識レベルと、長時間の文献検索作業に依存してきた。本研究の成果を定量的に示すことは困難だが、例えば、キーワード連想部分の自動化を実現することで、高度な知識を持たない研究者でも文献上で関連する「解釈担当者の知識外のキーワード」のリストを得ることができるようになる。また、解釈担当者の知識外のキーワードを活用した文献検索を行うことで、読むべき優先度の高い文献を絞り込みうることも示した。例えば、要旨と本文を読んで概要をつかむのに 1 報あたり平均して 15 分かかるとすると、Figure 3-12(a)の例では、日から週スケールの調査時間（約 24 時間）が必要なキーワード連想を、数時間（約 1.5 時間）減少しうることになる。

本研究では、1 つのテストデータでしか評価を行うことができなかったが、今後は評価例を増やすことで、改善を進める必要がある。本研究で開発したレコメンド法の課題は、キーワードの組み合わせによっては、統計的に有意にレコメンドされるキーワード数が多い場合で数千個になる点である（Figure 3-13.）。そのような場合に Confidence と Cosine の両方で有意とされたキーワードをレコメンドすることや解釈担当者の知識外のキーワードの MeSH ID をグループ分けして、グループとの関連性が強いキーワードをレコメンドするなどの改善方法の検討が考えられる。

本手法を適用できる範囲は、MeSH ID に登録されているものである。代謝中間体 133 個のうち、85 個（約 64.0%）の MeSH ID が対応していることを述べた。代謝物の ID 不足は、ケミカルアブストラクトサービスのデータベース (https://www.cas.org/covid-19-dataset-properties) を活用することで補足できると考えられる。

第4章 結言

4-1 結果のまとめ

本研究の目的は、メタボロミクス自動化に向けた高性能ピークピッキングとキーワードレコメンド法を実現し、時間短縮を図ることである。

第 1 章では、まずメタボロミクス法が生命科学分野におけるバイオマーカー探索や生体メカニズム解明などで果たす役割を述べた。次いで、メタボロミクス法の課題（ボトルネック）を整理した。メタボロミクス法を構成する 4 つの作業のうち、大量のクロマトグラム中から代謝中間体に由来するピークを同定し、開始点と終了点を決定するピークピッキング作業（生データ処理）では、自動ピークピッキング結果の、確認・修正作業が多くの時間を要することを述べた。例えば、20 サンプル中の 200 種類の代謝中間体を測定した場合には、約 6.5 時間を要する。さらに、含量が増減した代謝物のリストから、関連するキーワードを連想する作業は、作業の成果が解釈担当者の知識に依存し、さらに文献調査に日から週スケールの時間を要することを指摘した。

そこで、本研究の次の２点を実現することを目標にした。

1．自動ピークピッキング作業の正確さを熟練作業者並みとし、修正作業を不要にすること。

2．キーワード連想作業のキーワード連想作業を自動化するキーワードレコメンド手法を開発して、キーワード連想作業の属人性を解消し、作業時間を短縮すること。

第２章では、まず、ピークの正確な開始点と終了点でラベル付けされた疑似クロマトグラムを生成する新たな技術を開発した (Figure 2-11)。5 つのクロマトグラムの特徴量分布に沿って GAN で生成した疑似ピーク形状を配置し、ブランクサンプルクロマトグラムと組み合わせることで、疑似クロマトグラム生成する手法を構築した。生成した大量の疑似クロマトグラムを用いてピークピッキングニューラルネットワークの深層学習を行った。その結果、U-Net がニューラルネットワークとして SSD より優れていることを示した。構築したピークピッキングニューラルネットワークと熟練者で、LCMS ワイドターゲットメタボロミクスのテストデータを用いて性能比較を行ったところ両者は、同等の性能を示した (Figure 2-22、Figure 2-24)。これにより目標 1 を達成し、修正作業が実質ゼロになった。

第３章では、統計解析で得られる「代謝物リスト」から、関連度の高いキーワードを提案する手法を開発した。アソシエーション分析の手法を活用することで、代謝物リストと「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」をレコメンドする新たな手法を開発した。さらに妥当な閾値を設定するために、ランダムＤＢを用いた帰無分布を利用する手法を開発した。その結果、日から週スケールの調査時間（24 時間）が必要なキーワード連想を、数時間（1.5 時間）に支援する手法を構築できた。これにより２つめの目標を達成した (Figure 4-1)。

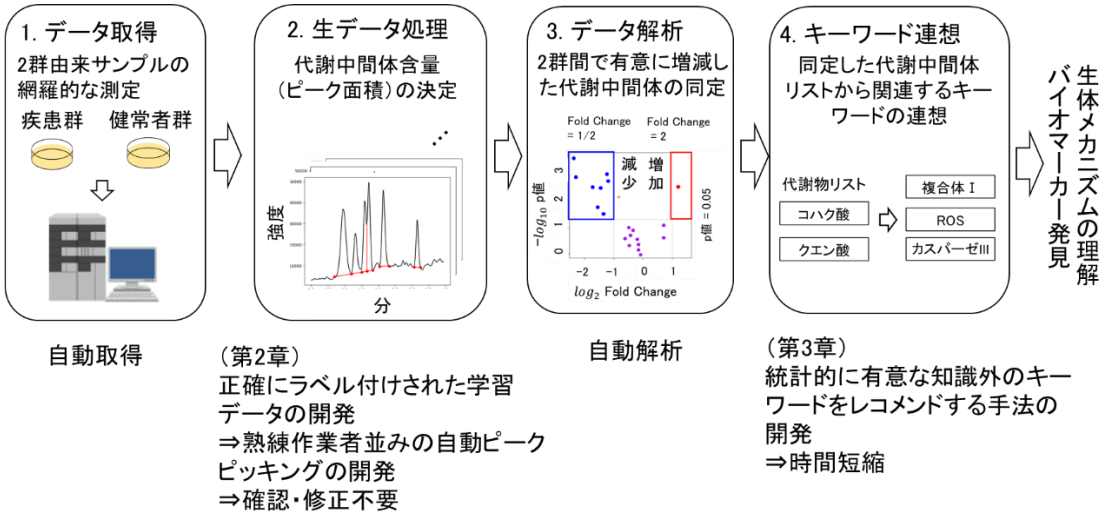


Figure 4-1　結果のまとめ

本研究の成果により、メタボロミクス法のボトルネックのうち２つを解消する方法論を構築できた。構築した方法論の、メタボロミクス自動化への寄与、さらなる高度化、汎用化に向けた課題と、今後の展望について述べる。

4-2 メタボロミクス法自動化への寄与

本研究の成果はメタボロミクス法の２つのボトルネックを解消するものである。LC-MS を用いたバイオマーカーの発見には、80～400 のサンプル数を扱う (Di Gangi et al., 2016, Rhee et al., 2011, Wang et al., 2011, Wang et al., 2013)。1 人の作業者が扱える 20 サンプル（200 種類の代謝中間体含量）を測定することを考えると、4000 回のピークピッキングが行われ、確認・修正するのに平均して 6 秒かかるので約 6.5 時間が必要である。

第 2 章で、人と同等のピークピッキングニューラルネットワークの構築に成功したため、「6 秒×代謝中間体数×サンプル数」が必要な修正作業を実質ゼロにすることができた (Figure 4-2a)。その結果、サンプルに何の代謝中間体がどれだけ入っているかがすぐわかるようになった。また、熟練作業者並みの生データ処理を可能にしたことで、作業者の人件費の削減につながり、人に依存しない再現性のあるデータを確保できるようになった。

第 3 章で、PubMed 全体の文献調査を行い、関連するキーワードを見出す作業を自動化することに初めて成功した。4-OHT 処理したヒト乳がん細胞 MCF-7 から取得したテストデータからレコメンドされたキーワード計 7 個を調べたところ、研究者が当初思いつかなかったキーワードを 5 個提示できた。これにより、解釈担当者の知識に依存せず、より客観的なキーワード連想作業自動化することを可能とした。また、新たなキーワードを用いて論文検索を行うことで、32 個の文献から、重要性が高い 2 個（約 6.3%）の文献を絞りこむことができた。論文検索では、要旨と本文を読んで概要をつかむのに 1 件あたり平均して 15 分かかる。正確な定量化は難しいものの、約 6.3%に文献を絞り込むことができたことから、日から週スケールの調査時間（24 時間）が必要なキーワード連想を、数時間（1.5 時間）に支援する手法を構築できた。その結果、解釈担当者は遺伝的要因や外部環境情報からバイオマーカー発見・生体メカニズムの解明につながる仮説検証に集中できるようになった。

さらに、今までは、初学者は仕事の質が低く、週単位の時間がかかっていた (Figure 4-2b)。既存自動化手法では仕事の質が低く、数時間～実質ゼロの作業時間であった。しかし、仕事の質が低いため、その結果を基に熟練者が週単位の時間をかけて仕事の質を高めていた。2 章と 3 章の本手法を適用することで、仕事の質が高く、数時間～実質ゼロで作業ができることが可能になった。

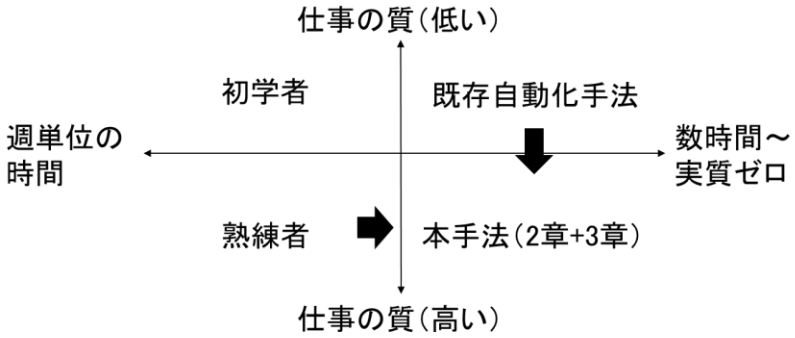
その結果、20 サンプル、約 200 種類の実働時間の約 30.5 時間のうち 20 分の 1 の約 1.5 時間に短縮された (Figure 4-2c) 。これは、医療バイオマーカーではより多くの外部環境を考慮した分析が可能になると考えられる。甲状腺がんのコホート研究では、アメリカや中国、韓国など最大約 200 検体のコホート研究が行われている (Khatami et al, 2019)。実働時間が 20 分の 1 になったことから、20 倍の 4000 検体のコホート研究が可能になると考えられる。これは、地域の違いだけでなく、食習慣、ストレスなどを反映した診断に繋がると期待される。さらに、医学分野のみならず、農学分野や工学分野で

も、共通の代謝中間体を計測していることと、ピークピッキング処理とキーワード連想を行うことから、農学分野や工学分野へも水平展開が可能な技術である。農学分野では、グレードと地域が異なる日本酒 40 検体の香味研究が行われている (Mimura et al., 2014)。これが 20 倍の 800 検体の香味研究を行うことが可能になるので、グレードや地域だけでなく、米の種類や生育環境、熱燗や冷酒などの飲み方を反映した品質チェックが可能になると期待される。工学分野では、寿命の異なる出芽酵母 6 検体の微生物の品質に関する研究が行われている (Yoshida et al., 2010)。これが 20 倍の 120 検体の研究を行うことが可能になるので、寿命以外に、培地の違いや複数の遺伝子改変などを反映した微生物の品質をチェックすることが可能になると期待される。

(a)

	時間短縮	仕事の質
2章	サンプルに何の代謝中間体がどれだけ入っているかがすぐわかる。	熟練作業者並みの生データ処理 • 人件費（生データ処理作業者）の削減 • 人に依存しない再現性のあるデータ
3章	遺伝的要因や外部環境情報からバイオマーカー発見・生体メカニズムの解明につながる仮説検証に集中できる。	知識外のキーワードのレコメンド 物知りでなくても可能

(b)



(c)

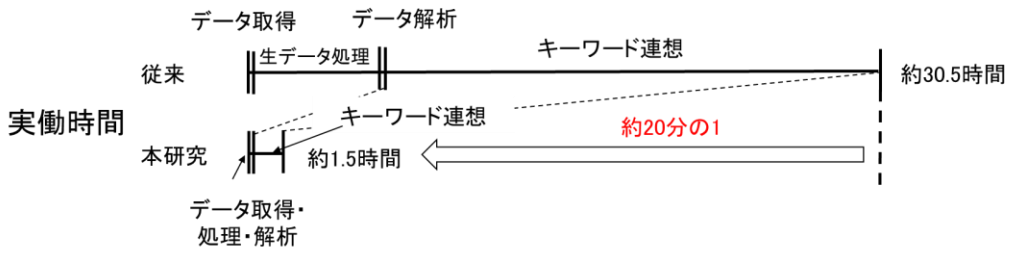


Figure 4-2 メタボロミクスに対する本研究の寄与

(a) 2 章と 3 章のそれぞれの寄与。(b) 本手法の適用前後の状況。(c) 従来手法と本研究の実働時間の違い

4-3 ピークピッキングニューラルネットワークの汎用化、性能向上に向けた課題

ピークピッキングニューラルネットワークを LC-MS クロマトグラム以外へ水平展開する方法と熟練作業者並みのピッキングニューラルネットワークが熟練作業者を超える方法について述べる。

本章では、5 つのクロマトグラムの特徴量分布に沿って GAN で生成した疑似ピーク形状を配置し、ブランクサンプルクロマトグラムと組み合わせることで、疑似クロマトグラム生成する手法を構築した。LC-MS を用いたワイドターゲットメタボロミクスを題材としたが、特徴量分布、疑似ピークを生成する GAN, ブランクサンプルクロマトグラムを用意することで、他の分析手法にも適用できる汎用性がある。例えば、GC-MS ワイドターゲットメタボロミクス法の疑似クロマトグラムの生成器を構築することも可能である (Figure 2 17a)。その他、同様に、スペクトル（横軸が *m/z*, 縦軸が強度の波形）を生成することも可能である。しかし、仮に LC-MS クロマトグラムと GC-MS クロマトグラム、LC-MS スペクトル、GC-MS スペクトルの計 4 通りに対応する場合、4 通りのピークを準備しなければならない。今回は自動生成のために GAN を用いたが、GAN の応用例として馬の画像をシマウマの画像へ変換する CycleGAN (Zhu et al., 2017) という技術がある。今回の GAN を CycleGAN に置き換えると、LC-MS クロマトグラムピークと GC-MS クロマトグラムピークから LC-MS から GC-MS へスタイル変換器を構築し、LC-MS クロマトピークから LC-MS スペクトルピークへスタイル変換器を構築しておくことで、それぞれの変換器を組み合わせると LC-MS クロマトグラムピークから GC-MS スペクトルピークへ変換することができる。CycleGAN を用いた手法は、本技術を水平展開する際、装置や波形の種類など組み合わせが多い場合に、それぞれのピークを用意することが不要になるため、有用だと考えられる。

今回は、クロマトグラムの特徴量分布からランダムサンプルして、ピーク形状を GAN で生成するためルールベースの方法を開発したが、実クロマトグラム全体からラベル付き疑似クロマトグラムを生成することで、より簡便にラベル付き疑似クロマトグラムが生成できると考えられる。そのためには、学習がうまく進まなかった原因を解決する必要がある。原因としてピーク位置や高さ等の複数のパラメータを学習させる際、目的関数が離散的で局所解に陥っていたと考えられる。平滑化などを用いた局所最適解に関する研究が行われており (Kleinberg et al., 2018)、それらの研究発展を組み合わせることで、うまく学習する可能性がある。

セマンティックセグメンテーションのニューラルネットワークには新しいアルゴリズムとして PSPNet (Zhao et al., 2017) や DeepLab (Chen et al., 2018) などが開発されている。これらは日々進歩している技術なので、それを適用することが性能向上につながると考えられる。また、ブランクサンプルクロマトグラムを自動生成してバリエーションを増やすことで、学習データの網羅性が上がり、ピークピッキングの性能向上が見込める。その他、繰り返し分析した複数のクロマトグラムやフラグメントイオンの情報を組み合わせることでより性能が上がり、熟練作業者を超えることが期待される。そのためには、複数の正確な学習データが必要であることから、正確な学習データを生成する疑似クロマトグラム生成器が必要になる。また、先程述べた CycleGAN も有用だと考えられる。

4-4 キーワード連想支援の汎用化、性能向上に向けた課題

データ解釈支援を簡便に使える方法と統計的に有意なキーワード数を絞り込む方法について述べる。

本章では、メタボロミクスデータを解読する作業を支援するための技術を開発した。統計解析で得られる「代謝物リスト」と「解釈担当者の既知のキーワード」から「解釈担当者の知識外のキーワード」をレコメンドする手法を開発した。今後は、誰でも簡便にレコメンド機能を利用できるようにする必要がある。そこで、特定非営利活動法人システム・バイオロジー研究機構が構築を進めている、統合データ解析環境の Garuda Platform (http://www.garuda-alliance.org/) が有用だと考えられる。Garuda Platform は、世界の研究機関コミュニティ (Garuda Alliance) により開発が進められている共通プラットフォームであり、バイオメディカル、バイオテクノロジー研究用の様々な解析アプリケーションやデータベース を接続するフレームワークを提供している。ここでは、MS データの読み込みから統計解析ソフトが用意されており、それぞれのソフトウェアをシームレスにつなぐことができる。Garuda Platform 上に本機能を実装することで、誰でも使えるソフトウェアになる。

レコメンド機能を性能向上させるには更なる絞り込みを行う必要がある。今回提案したレコメンド手法では、統計的に有意とされるキーワードが、多い場合で数千個レコメンドされることがわかっているが、目視で確認することを考えると数十個に絞る必要がある。そのような場合に解釈担当者の知識外のキーワードの MeSH ID をグループ分け

して、グループとの関連性が強いキーワードをレコメンドするという改善方法の検討が考えられる。その際には、MeSH ID の階層構造の情報が有用である。例えば、D042967（複合体Ⅰ）の兄弟には、D014450（複合体Ⅲ）や D003576（複合体Ⅳ）があり、親には、D017494（プロトンポンプ）がある。これより、複合体Ⅰと複合体Ⅲ、複合体Ⅳは、プロトンポンプのグループであることがわかる。また、協調フィルタリングのアイテムベース手法という、過去のデータからアイテム同士の関連度を計算し、その値が大きいものをレコメンドする方法がある (Sarwar et al,. 2001)。協調フィルタリングのアイテムベースでは、複数のアイテム「代謝物リスト」から 1 つのアイテム「解釈担当者の知識外のキーワード」の関連度を計算して、レコメンドするが、共起情報に基づくアソシエーション分析では、1 つのアイテム「代謝物」と 1 つのアイテム「解釈担当者の知識外のキーワード」間の関連度を基にレコメンドするという違いがある。そのまま、協調フィルタリングのアイテムベースを適用すると代謝物リストと解釈担当者の知識外のキーワードの関連度を計算することになるが、今回は化合物リストを 3 個以下に絞っていることから、解釈担当者の知識外のキーワードでグループ分けして、分けられたグループと代謝物との関連度を計算し、統計的に有意でないグループを除外することが有用だと考えられる。例えば、先程の例のようにプロトンポンプのグループと代謝物との関連度を計算し、統計的に有意な閾値以下なら全キーワードを除外し、閾値以上なら除外しないことが考えられる。その他「Confidence と Cosine の両方で有意とされたキーワードをレコメンドする」や、「解釈担当者に関係のある論文もしくはインパクトファクターの高い論文、引用数が多い論文で絞り込む」などを組み合わせることで改善が期待される。

4-5 メタボロミクスの今後の展望

「代謝物リスト」と「解釈担当者の既知のキーワード」、「解釈担当者の知識外のキーワード」から自動で因果関係を推定する課題が残っている。これは因果探索と呼ばれる分野で研究が進んでおり、そのうちの 1 つであるベイジアンネットワークでは、複数の変数と変数間が矢印の依存関係で示されている (Pearl, 2009)。今回の例だと、キーワード間の依存関係が示されることになる。これらの技術を用いて、キーワード間の因果の向きがわかるようになれば、因果関係の推定（仮説生成）が可能になり、メタボロミク

ス作業が自動化される。したがって、測定が終われば、仮説の生成までを短時間で作業できるようになる。

本技術は、食品の品質管理、工学分野などにもそのまま適用することが可能な汎用的な技術である。様々な分野で自動化が進むことで、分析そのものへのニーズが高まり、汎用的な測定機器が求められる。それらの測定機器が開発されると、家電や体温計のような身近なものになり、食品安全と、身近な医学、低炭素社会の実現が期待される (Figure 4-3)。

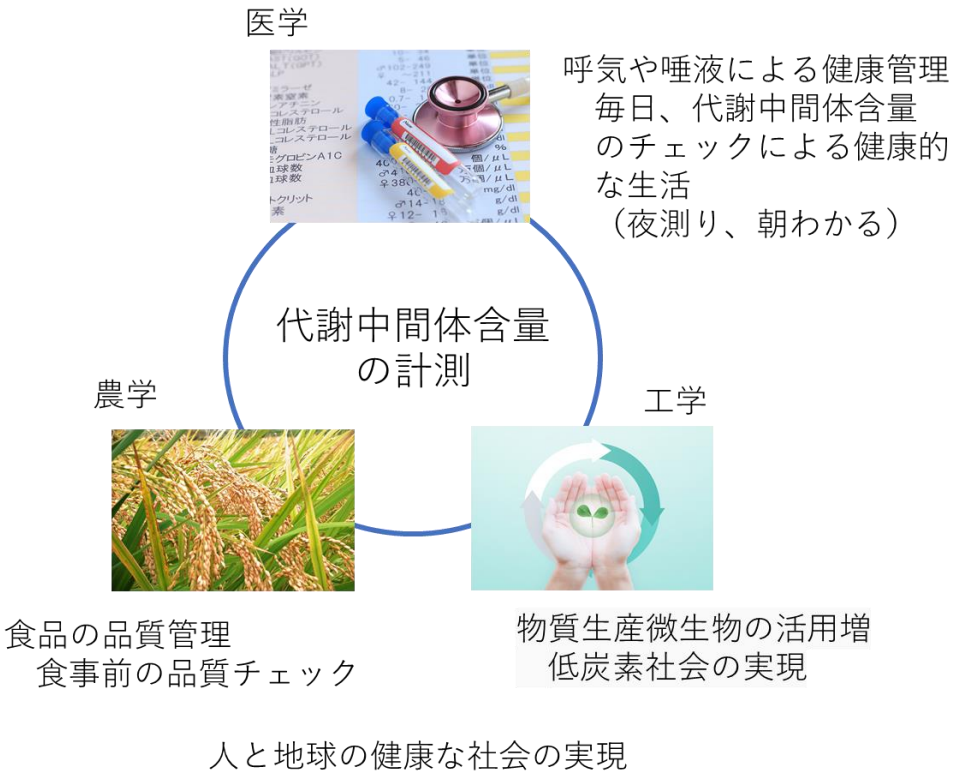


Figure 4-3　今後の展望

例えば、食品安全に関する問題としては、保存条件の異なるだだちゃ豆のメタボロミクスデータを比較することで、保存条件と成分変動を明らかにした例がある (Sugimoto et al., 2010)。身近な測定機器が開発されると、食品の安全が確保された品種が確立されることや、保存状況が悪い食品を食事前にチェックされると考えられる。

医学分野においては、患者の体質や病気の特徴に合わせて治療する個別化医療につながる。現在、個別化医療は、遺伝子検査を行うことでがんなどの病気の治療に用いら

れており (Hamburg et al., 2010)、将来的には病気の予防や早期発見が期待されている。各代謝中間体の細胞内存在量は、遺伝的な要因（体質）に加え、外部環境（疾患や薬物投与など）の影響が反映されていると考えられている。そこで、個別化医療の遺伝子検査に加えて、代謝中間体含量の情報を加えることで、外部要因の影響を加味した診断や予後に活かす研究がされている (Jacob et al., 2019)。身近な測定機器を用いて、夜に呼気や唾液などを測定して朝に診断結果がわかることで、健康に気をつけた生活を送ることができ、健康寿命が延びて QOL の向上が期待される。本研究のメタボロミクス自動化に関する研究は、夜測定した結果、朝健康チェックができるまでの測定後の処理の高速化に寄与したと考えられる。

本研究で開発した高性能ピークピッキング及びキーワードレコメンド法は、これらのメタボロミクス応用の技術的基盤となる汎用性がある。さらなるメタボロミクス自動化を進め、汎用的な装置の開発と連携することで、環境問題の改善、精度の高い医療が可能となり、人と地球の健康な社会の実現につながることが期待される。

参考文献

Abdeddaïm, S., Vimard, S., Soualmia, F., L., 2019. The MeSH-Gram Neural Network Model: Extending Word Embedding Vectors with MeSH Concepts for Semantic Similarity, Studies in Health Technology and Informatics 264, 5-9.

Aggarwal S., Yadav A.K. 2016. False Discovery Rate Estimation in Proteomics, Jung K. (eds) Statistical Analysis in Proteomics. Methods in Molecular Biology, vol 1362.

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD ’93. ACM Press, 207-216.

Agrawal, R., and Srikant, R., 1994. Fast algorithms for mining association rules. In VLDB '94, 487-499.

Arase, S., Horie, K., Kato, T., Noda, A., Mito, Y., Takahashi, M., and Yanagisawa, T. 2016. Intelligent peak deconvolution through in-depth study of the data matrix from liquid chromatography coupled with a photo-diode array detector applied to pharmaceutical analysis. Journal of Chromatography A, 1469.

ASTM International: https://www.astm.org/ (Last accessed on November 2020)

Atluri, G., Gupta, R., Fang, G., Pandey, G., Steinbach, M., Kumar, V. 2009. Association analysis techniques for bioinformatics problems. In International Conference on Bioinformatics and Computational Biology. pp. 1-13.

Azevedo, P. J., Jorge, A., M., 2007. Comparing rule measures for predictive association rules. J. N. Kok, J. Koronacki, R. L. de Mantaras, S. Matwin, D. Mladenič, & A. Skowron (eds), Machine Learning: ECML 2007 (Vol. 4701, pp. 510‑517). Springer Berlin Heidelberg.

Banoub, J. H., Newton, R. P., Esmans, E., Ewing, D. F., Mackenzie, G., 2005. Recent developments in mass spectrometry for the characterization of nucleosides, nucleotides, oligonucleotides, and nucleic acids. Chemical Reviews, 105(5), 1869–1916.

Benjamini, Y., and Hochberg Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 57: 289-300.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Proceedings of the European Conference on Computer Vision (ECCV), pp. 801-818.

Chen, Y. L., Tang, K., Shen, R. J., Hu, Y. H. 2005. Market basket analysis in a multiple store environment. Decision support systems, 40(2), 339-354.

Chouchani, T., E., Pell, R., V., Gaude, E., Aksentijevic, D., Sundier, Y., S., Robb, L., E., Logan, A., Nadtochiy, M., S., Ord, J., N., E., Smith, C., A., Eyassu, F., Shirley, R., Hu, C., Dare, ., J., A., James, M., A., Rogatti, S., Hartley, C., R., Eaton, S., Costa, H., S., A., Brookers, S., P., Davidson, M., S., Duchen, R., M., Saeb-parsy, K., Shattock, J., M., Rovinson, J., A., Work, M., L., Frezza, C., Krieg, T., Murphy, P., M., 2014. Ischaemic accumulation of succinate controls reperfusion injury through mitochondrial ROS. Nature 515(7527):431-435.

Crick, F., 1970. Central dogma of molecular biology. Nature, 227(5258), 561–563.

Di Gangi, I. M., Mazza, T., Fontana, A., Copetti, M., Fusilli, C., Ippolito, A., Mattivi, F., Latiano, A., Andriulli, A., Vrhovsek, U., Pazienza, V. (2016). Metabolomic profile in pancreatic cancer patients: A consensus-based approach to identify highly discriminating metabolites. Oncotarget, 7(5), 5815–5829.

Du, P., Kibbe, W. A., and Lin, S. M., 2006. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics 22, 2059-2065.

Edwards, P., A., Ericsson, J., 1998. Signaling molecules derived from the cholesterol biosynthetic pathway: Mechanisms of action and possible roles in human disease: Current Opinion in Lipidology, 9(5), 433–440.

Fan S, Shahid M, Jin P, Asher A, Kim J., 2020. Identification of metabolic alterations in breast cancer using mass spectrometry-based metabolomic analysis. Metabolites. 10(4):170.

Ferrannini, E., Natali, A., Camastra, S., Nannipieri, M., Mari, A., Adam, K.-P., Milburn, M. V., Kastenmüller, G., Adamski, J., Tuomi, T., Lyssenko, V., Groop, L., Gall, W. E., 2013. Early metabolic markers of the development of dysglycemia and type 2 diabetes and their physiological significance. Diabetes, 62(5), 1730–1737.

Fiehn, O., 2002, Metabolomics--the link between genotypes and phenotypes. Plant Mol. Biol. 48: 155–171.

Fiehn O, Kloska S, Altmann T., 2001, Integrated studies on plant biology using multiparallel techniques. Current Opinion in Biotechnology 12(1):82-86.

GARUDA: http://www.garuda-alliance.org/ (Last accessed on November 2020)

Gasparre, G., Kurelac, I., Capristo, M., Iommarini, L., Ghelli, A., Ceccarelli, C., Nicoletti, G., Nanni, P., De Giovanni, C., Scotlandi, K., Betts, C. M., Carelli, V., Lollini, P. L., Romeo, G., Rugolo, M., and Porcelli, A. M. 2011. A mutation threshold distinguishes the antitumorigenic effects of the mitochondrial gene mtnd1, an oncojanus function. Cancer Research, 71(19), 6220–6229.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., 2014. Generative Adversarial Nets, Adv Neur In, 27, 2672-2680.

Gordienko, Y., Gang, P., Hui, J., Zeng, W., Kochura, Y., Alienin, O., Rokovyi, O., and Stirenko, S. 2019. Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. Advances in Intelligent Systems and Computing, 754.

Guasch-Ferré, M., Hruby, A., Toledo, E., Clish, C. B., Martínez-González, M. A., Salas-Salvadó, J., Hu, F. B. 2016. Metabolomics in prediabetes and diabetes: A systematic review and meta-analysis. Diabetes Care, 39(5), 833–846.

Hamburg, M. A., Collins, F. S., 2010. The path to personalized medicine. New England Journal of Medicine, 363(4), 301–304.

Hirayama, A., Nakashima, E., Sugimoto, M., Akiyama, S., Sato, W., Maruyama, S., Matsuo, S., Tomita, M., Yuzawa, Y., Soga, T. 2012. Metabolic profiling reveals new serum biomarkers for differentiating diabetic nephropathy. Analytical and Bioanalytical Chemistry, 404(10), 3101–3109.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T. S. 2017. Neural collaborative filtering. In Proceedings of the 26th international conference on world wide web pp. 173-182.

Hou, Y., Yin, M., Sun, F., Zhang, T., Zhou, X., Li, H., Zheng, J., Chen, X., Li, C., Ning, X., Lou, G., Li, K., 2014. A metabolomics approach for predicting the response to neoadjuvant chemotherapy in cervical cancer patients. Mol. BioSyst., 10(8), 2126‑2133.

Hsu, P.P., Sabatini, D.M., 2008. Cancer cell metabolism: Warburg and beyond. Cell 134, 703-7.

Itoi, T., Sugimoto, M., Umeda, J., Sofuni, A., Tsuchiya, T., Tsuji, S., Tanaka, R., Tonozuka, R., Honjo, M., Moriyasu, F., Kasuya, K., Nagakawa, Y., Abe, Y., Takano, K., Kawachi, S., Shimazu,

M., Soga, T., Tomita, M., Sunamura, M., 2017. Serum metabolomic profiles for human pancreatic cancer discrimination. International Journal of Molecular Sciences, 18(4), 767.

Jacob, M., Lopata, A. L., Dasouki, M., Abdel Rahman, A. M., 2019. Metabolomics toward personalized medicine. Mass Spectrometry Reviews, 38(3), 221–238.

Khatami, F., Payab, M., Sarvari, M., Gilany, K., Larijani, B., Arjmand, B., and Tavangar, S. M. 2019. oncometabolites as biomarkers in thyroid cancer: A systematic review</p>. Cancer Management and Research, 11: 1829–1841

Kleinberg, R., Li, Y., Yuan, Y., 2018. An Alternative View: When Does SGD Escape Local Minima?, In International Conference on Machine Learning, pages 2703–2712

Koren, Y., Bell, R., Volinsky, C. 2009. Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

Kubo, A., Ohmura, M., Wakui, M., Harada, T., Kajihara, S., Ogawa, K., Suemizu, H., Nakamura, M., Setou, M., and Suematsu, M., 2011. Semi-quantitative analyses of metabolic systems of human colon cancer metastatic xenografts in livers of superimmunodeficient NOG mice, Anal. Bioanal. Chem. 400, 1895-1904.

Lee, Y., S., Kang, Y., S., Lee, S., H., Kim, J., A., 2000. Role of NAD(P)H oxidase in the tamoxifen-induced generation of reactive oxygen species and apoptosis in HepG2 human hepatoblastoma cells. Cell Death Differ. 7(10):925-32.

Li, X., Mao, M., Zhang, Y., Yu, K., Zhu, W., 2019. Succinate modulates intestinal barrier function and inflammation response in pigs. Biomolecules. 9(9):486.

Li, Y., Song, X., Zhao, X., Zou, L., Xu, G., 2014. Serum metabolic profiling study of lung cancer using ultra high performance liquid chromatography/quadrupole time-of-flight mass spectrometry. Journal of Chromatography B, 966, 147–153.

Liu, K., Cuddy, T., E., Pierce, G., N., 1992. Oxidative status of lipoproteins in coronary disease patients. American Heart Journal, 123(2), 285–290.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C., 2016. SSD: Single Shot MultiBox Detector, pp. 21e37, in: Leibe, B., Matas, J., Sebe, N. and Welling, M. (Eds.), Computer vision e ECCV 2016. Lecture notes in computer science vol. 9905.

Lommen, A., 2009. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing, Anal. Chem. 81, 3079-3086.

Luo, B., Groenke, K., Takors, R., Wandrey, C., Oldiges, M., 2007. Simultaneous determination of multiple intracellular metabolites in glycolysis, pentose phosphate pathway and tricarboxylic acid cycle by liquid chromatography–mass spectrometry. Journal of Chromatography A, 1147(2), 153–164.

MacIsaac, R. J., Ekinci, E. I., Jerums, G., 2014. Markers of and risk factors for the development and progression of diabetic kidney disease. American Journal of Kidney Diseases, 63(2), S39–S62.

Martens, G. A., Cai, Y., Hinke, S., Stangé, G., Van de Casteele, M., and Pipeleers, D. 2005. Glucose suppresses superoxide generation in metabolically responsive pancreatic β cells. Journal of Biological Chemistry, 280(21), 20389–20396.

Matsuda, F., Toya, Y., and Shimizu, H., 2017. Learning from quantitative data to understand central carbon metabolism, Biotechnol Adv, 35, 971-980.

Matsuda, F., 2016. Technical challenges in mass spectrometry-based metabolomics, Mass spectrometry 5, S0052.

Melnikov, A. D., Tsentalovich, Y. P., and Yanshole, V. V., 2020. Deep Learning for the Precise Peak Detection in High-Resolution LC-MS Data, Anal. Chem. 92, 588-592.

MeSH: Medical Subject Headings http://www.nlm.nih.gov/mesh/meshhome.html (Last accessed on November 2020)

Mimura, N., Isogai, A., Iwashita, K., Bamba, T., and Fukusaki, E. 2014. Gas chromatography/mass spectrometry based component profiling and quality prediction for Japanese sake. Journal of Bioscience and Bioengineering, 118(4), 406–414.

Mo, F., Mo, Q., Chen, Y., Goodlett, D. R., Hood, L., Omenn, G. S., Li, S., and Lin, B., 2010. WaveletQuant, an improved quantification software based on wavelet signal threshold de-noising for labeled quantitative proteomic analysis, BMC Bioinformatics 11, 219.

Morley, P., Whitfield, J., F., 1994 Jan 1. Effect of tamoxifen on carbachol-triggered intracellular calcium responses in chicken granulosa cells. Cancer Res. 54(1):69-74.

Murguía, M., Villaseñor, L., J., 2003. Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classifications, Annales Botanici Fennici Vol. 40, No. 6, pp. 415-421

Nazarewicz RR, Zenebe WJ, Parihar A, Larson SK, Alidema E, Choi J, Ghafourifar P., 2007. Tamoxifen induces oxidative stress and mitochondrial apoptosis via stimulating mitochondrial nitric oxide synthase. Cancer Res. 1;67(3):1282-90.

Ogura, T., Bamba, T., and Fukusaki, E., 2013. Development of a practical metabolite identification technique for non-targeted metabolomics. J. Chromatogr. A 1301, 73-79.

Ono T, Kuhara S., 2014. A novel method for gathering and prioritizing disease candidate genes based on construction of a set of disease-related MeSH® terms. BMC Bioinformatics. 15(1):179.

Pan, D., Kather, M., Willmann, L., Schlimpert, M., Bauer, C., Lagies, S., Schmidtkunz, K., Eisenhardt, U., S., Jung. M., Günther, S., Kammerer, B., 2016. Metabolic response to xd14 treatment in human breast cancer cell line mcf-7. IJMS. 17(10):1772.

Pearl, J., 2009. Causal inference in statistics: An overview. Statistics Surveys, 3(0), 96–146.

Peñaloza, E., Holandino, C., Scherr, C., Araujo, P. I. P. de, Borges, R. M., Urech, K., Baumgartner, S., Garrett, R., 2020. Comprehensive metabolome analysis of fermented aqueous extracts of viscum album l. By liquid chromatography−high resolution tandem mass spectrometry. Molecules, 25(17), 4006.

Pongsuwan, W., Fukusaki, E., Bamba, T., Yonetani, T., Yamahara, T., and Kobayashi, A. 2007. Prediction of japanese green tea ranking by gas chromatography/mass spectrometry-based hydrophilic metabolite fingerprinting. Journal of Agricultural and Food Chemistry, 55(2), 231–236.

Rendle, S, 2010. Factorization Machines, 2010 IEEE International Conference on Data Mining, pp. 995-1000.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work pp. 175-186.

Rhee, E. P., Cheng, S., Larson, M. G., Walford, G. A., Lewis, G. D., McCabe, E., Yang, E., Farrell, L., Fox, C. S., O'Donnell, C. J., Carr, S. A., Vasan, R. S., Florez, J. C., Clish, C. B., Wang, T. J., Gerszten, R. E., 2011. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. Journal of Clinical Investigation, 121(4), 1402–1411.

Pistritto, G., Trisciuoglio, D., Ceci, C., Garufi, A., D'Orazi, G., 2016. Apoptosis as anticancer mechanism: function and dysfunction of its modulators and targeted therapeutic strategies. Aging. 8(4):603-619.

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M., 2010 MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics 11, 395.

PubMed search engine: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed (Last accessed on November 2020)

PubMed file server: ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline (Last accessed on November 2020)

Ronneberger, O., Fischer, P., and Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation, pp. 234e241, in: Navab, N., Hornegger, J., Wells, W. M. and Frangi, A. F. (Eds.), Medical Image Computing and Computer- Assisted Intervention e MICCAI 2015. Proceedings of the 18th International Conference, Munich, Germany, October 5-9, 2015, Part III. Lecture notes in computer science, vol. 9351. Springer International Publishing.

Salonen, J., T., Nyysso¨nen, K., Salonen, R., Porkkala-Sarataho, E., Tuomainen, T., P., Diczfalusy, U., Bjo¨rkhem, I. 1997. Lipoprotein oxidation and progression of carotid atherosclerosis. Circulation, 95(4), 840–845

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web pp. 285-295.

Seabold, S., and Josef, P., 2010 statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference, 92-96.

Shajahan-Haq, A., Cheema, M., Clarke, R., 2015. Application of metabolomics in drug resistant breast cancer research. *Metabolites*, *5*(1), 100–118.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G., 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78, 779-787.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43).

Sugimoto, M., Goto, H., Otomo, K., Ito, M., Onuma, H., Suzuki, A., Sugawara, M., Abe, S., Tomita, M., Soga, T., 2010. Metabolomic profiles and sensory attributes of edamame under various storage duration and temperature conditions. Journal of Agricultural and Food Chemistry, 58(14), 8418–8425.

Takayama, T., Tsutsui, H., Shimizu, I., Toyama, T., Yoshimoto, N., Endo, Y., Inoue, K., Todoroki, K., Min, J. Z., Mizuno, H., Toyo'oka, T. 2016. Diagnostic approach to breast cancer patients based on target metabolomics in saliva by liquid chromatography with tandem mass spectrometry. Clinica Chimica Acta, 452, 18–26.

Tautenhahn, R., Bottcher, C., and Neumann, S., 2008. Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics 9, 504.

Terasaki, M., Mima, M., Kudoh, S., Endo, T., Maeda, H., Hamada, J., Osada, K., Miyashita, K., Mutoh, M. 2018. Glycine and succinic acid are effective indicators of the suppression of epithelial-mesenchymal transition by fucoxanthinol in colorectal cancer stem-like cells. Oncology reports, 40(1), 414–424.

Tokui, S., Okuta, R., Akiba, T., Niitani, Y., Ogawa, T., Saito, S., Suzuki, S., Uenishi, K., Vogel, B., and Vincent, H. Y., 2019. Chainer: A Deep Learning Framework for Accelerating the Research Cycle, Kdd'19: Proceedings of the 25th Acm Sigkdd International Conferencce on Knowledge Discovery and Data Mining 2002-2011.

Tsutsui, H., Mochizuki, T., Inoue, K., Toyama, T., Yoshimoto, N., Endo, Y., Todoroki, K., Min, J. Z., Toyo'oka, T., 2013. High-throughput lc–ms/ms based simultaneous determination of polyamines including n-acetylated forms in human saliva and the diagnostic approach to breast cancer patients. Analytical Chemistry, 85(24), 11835–11842

Yoshida, R., Tamura, T., Takaoka, C., Harada, K., Kobayashi, A., Mukai, Y., and Fukusaki, E. 2010. Metabolomics-based systematic prediction of yeast lifespan and its application for semi-rational screening of ageing-related mutants: Semi-rational screening of ageing-related mutants. Aging Cell, 9(4), 616–625.

Yu, Y., Gong, Z., Zhong, P., Shan, J. 2017. Unsupervised representation learning with deep convolutional neural network for remote sensing images. Y. Zhao, X. Kong, D. Taubman (eds), Image and Graphics (Vol. 10667, pp. 97‑108). Springer International Publishing.

Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., Lewis, G. D., Fox, C. S., Jacques, P. F., Fernandez, C., O'Donnell, C. J., Carr, S. A., Mootha, V. K., Florez, J. C.,

Souza, A., Melander, O., Clish, C. B., Gerszten, R. E. 2011. Metabolite profiles and the risk of developing diabetes. Nature Medicine, 17(4), 448–453.

Wang, T. J., Ngo, D., Psychogios, N., Dejam, A., Larson, M. G., Vasan, R. S., Ghorbani, A., O'Sullivan, J., Cheng, S., Rhee, E. P., Sinha, S., McCabe, E., Fox, C. S., O'Donnell, C. J., Ho, J. E., Florez, J. C., Magnusson, M., Pierce, K. A., Souza, A. L., Yu, Y., Carter, C., Light, E., P., Melander, O., Clish, B., C., Gerszten, R. E., 2013. 2-Aminoadipic acid is a biomarker for diabetes risk. The Journal of Clinical Investigation, 123(10), 4309–4317.

Wellerdiek, M., Winterhoff, D., Reule, W., Brandner, J., Oldiges, M. 2009. Metabolic quenching of Corynebacterium glutamicum: Efficiency of methods and impact of cold shock. Bioprocess and Biosystems Engineering, 32(5), 581–592.

Woldegebriel, M. and Derks, E., 2017. Artificial Neural Network for Probabilistic Feature Recognition in Liquid Chromatography Coupled to High-Resolution Mass Spectrometry, Anal. Chem. 89, 1212-1221.

Wong, D., T., 2006. Towards a simple, saliva-based test for the detection of oral cancer. Expert Review of Molecular Diagnostics, 6(3), 267–272.

Woo, Y., M., Shin, Y., Lee, E., J., Lee, S., Jeong, S., H., Kong, H., K., Park, E., Y., Kim, H., K., Han, J., Chang, M., Park, J., H., 2015. Inhibition of Aerobic Glycolysis Represses Akt/mTOR/HIF-1α Axis and Restores Tamoxifen Sensitivity in Antiestrogen-Resistant Breast Cancer Cells. PLoS One. e0132285.

Wu, G., Wan, X., Xu, B. 2018. A new estimation of protein-level false discovery rate. BMC genomics, 19(Suppl 6), 567.

Würtz, P., Tiainen, M., Mäkinen, V.-P., Kangas, A. J., Soininen, P., Saltevo, J., Keinänen-Kiukaanniemi, S., Mäntyselkä, P., Lehtimäki, T., Laakso, M., Jula, A., Kähönen, M., Vanhala, M.,

Ala-Korpela, M., 2012. Circulating metabolite predictors of glycemia in middle-aged men and women. Diabetes Care, 35(8), 1749–1756.

Zhang, R., Zhuang, X., Zong, L., Liu, S., Liu, Z., Song, F., 2016. Investigations on the cell metabolomics basis of multidrug resistance from tumor cells by ultra-performance liquid chromatography–mass spectrometry. Anal Bioanal Chem. 408(21):5843-5854.

Zhang, W., Couldwell, W., T., Song, H., Takano, T., Lin, J., H., Nedergaard, M., 2000. Tamoxifen-induced enhancement of calcium signaling in glioma and MCF-7 breast cancer cells. Cancer Res. 60(19):5395-400.

Zhao, H., Shi, J., Qi, X., Wang., X., Jia, J., 2017. Pyramid Scene Parsing Network Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881-2890.

Zhou J, Shui Y, Peng S, Li X, Mamitsuka H, Zhu S., 2015. MeSHSim: An R/Bioconductor package for measuring semantic similarity over MeSH headings and MEDLINE documents. J Bioinform Comput Biol. 13(06):1542002.

Zhu, JY., Park, T., Isola, P., Efros, A., A., 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223-2232

西本和生、岡橋伸幸、丸山正晴、松田史生、2020、時系列メタボローム解析によるがん細胞薬剤代謝応答の解明、大阪大学工学部応用自然科学科　応用生物工学コース

略称一覧

装置名

GC: Gas Chromatography

LC: Liquid Chromatography

MS: Mass Spectrometry

代謝物名

4-OHT: 4-hydroxytamoxifen

ROS: Reactive Oxygen Species

G1P: Glucose 1-Phosphate

DHAP: Dihydroxyacetone Phosphate

ATP: Adenosine Triphosphate

深層学習関連

SSD: Single Shot MultiBox Detector

GAN: Generative Adversarial Network

IoU: Intersection over Union

データベース関連

NCBI: National Center for Biotechnology Information

MeSH: Medical Subject Headings

FDR: False Discovery Rate

BH 法: Benjamini & Hochberg 法

DB: Data Base

謝辞

本研究の遂行にあたり、魅力的なテーマと研究の場を与えて下さり、終始暖かくご指導、ご助言を頂きました大阪大学大学院情報科学研究科バイオ情報工学専攻の松田史生教授に心より感謝申し上げます。

博士論文審査委員として、有益なご指導、ご助言を頂きました大阪大学大学院情報科学研究科バイオ情報工学専攻の清水浩教授、戸谷吉博准教授、瀬尾茂人准教授に心より感謝申し上げます。

研究に関するニーズ・問題点・解決策の提供をしていただきました大阪大学大学院工学研究科の福崎英一郎教授に心より感謝申し上げます。

進捗報告時に貴重なご意見をくださいました大阪大学大学院工学研究科の新間秀一准教授に心より感謝申し上げます。

研究に関する考え方や論文執筆など、本研究を遂行する上で有益なご助言、ご指導頂きました大阪大学大学院情報科学研究科バイオ情報工学専攻の岡橋伸幸准教授に心より感謝申し上げます。

論文執筆にあたり、有益なご助言を頂きました大阪大学大学院情報科学研究科バイオ情報工学専攻の清家泰介助教、丸山正晴氏に心より感謝申し上げます。

ウェットの面で情報提供をしていただいた大阪大学大学院情報科学研究科バイオ情報工学専攻の西本和生氏に心より感謝申し上げます。

研究の幅を広げる機会を与えてくださいました株式会社島津製作所（大阪大学・島津分析イノベーション協働研究所）飯田順子特任教授、河野真一招聘准教授に心より感謝申し上げます。

ピークピッキングをして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の伊東有紗特任研究員、大阪大学工学研究科の橋本恭子特任研究員に心より感謝申し上げます。

アルゴリズムの相談に乗っていただいた梶原茂樹主幹研究員と野田陽主任、清水悟司主任、ウェット関係で相談に乗っていただいた内方崇人主任、國澤研大副主任をはじめ、株式会社島津製作所のメンバーには、心より感謝申し上げます。

最後に、研究生活において様々な面でお世話になりました大阪大学大学院情報科学研究科バイオ情報工学専攻バイオ情報計測学講座の諸先輩方、後輩のみなさん、応援して

くれた友人、精神的に支えて下さった両親、兄妹、楽しい時間を提供し、明るく励ましてくれた妻に心より感謝いたします。