質量分析データの解析による微生物同定の実験評価

浅野　公平[1]　豊坂 祐樹[1,a)]　成　凱[1,b)]

概要：食品の安全を守るために，腐敗や食中毒を起こす微生物の種類を特定すること，いわゆる「微生物同定」が必要不可欠である．微生物同定のために様々な技術が開発されてきたが，時間やコストを要することや，菌種によっては同定できないか同定精度が低いという問題が存在する．本研究では，情報科学のアプローチとして，マススペクトルの適切な前処理手法，ピークパターンの類似度計算及びそれに基づく微生物同定の手法を提案し，実験データを用いた実験により提案手法の評価を行った．実験では，類似度計算の方式，正解と認める類似度の許容範囲 N，ピークアライメントの閾値 δ，重み付け関数の種類，それぞれによる，正解率への影響を評価した．

Experimental Evaluation of Microbial Identification by Quantitative Analysis of Mass Spectrometry Data

1. はじめに

食品の腐敗や食中毒を起こす微生物から食品の安全性を守るために，原因となる微生物の種類を特定する微生物同定法が必要不可欠である．微生物同定のタスクとして，菌種同定と菌株識別が存在する．**菌種同定**とは，未知の分類菌株がすでに記載されたどの菌種に最も近いかを決定する作業である．一方，**菌株識別**とは，登録された複数の菌株に最も近い菌株を識別する作業である．

現在は，遺伝子解析に基づく微生物同定手法が良く知られているが，時間とコストを要する等の問題点が存在する．近年，質量分析法（Mass Spectrometry: MS）で得られるマススペクトル (Mass Spectrum: MS) と呼ばれるデータを用いて効率的な微生物同定が期待されている [4]．しかし，マススペクトルを取得する際に試料のイオン化過程で生じるノイズや分子構造の変化が未知試料のマススペクトルを読む際の障害になることや，メーカー提供の同定を行うシステムは同定を行う仕組みがブラックボックスになっていることや同定を行う際に利用されるデータベースに依存していること，菌種よっては同定できないか同定精度が低いという課題が挙げられる．

[1] 九州産業大学
Kyushu Sangyo University,
2-3-1, Higashi-ku, Fukuoka, 813-8503, Japan
[a)] toyosaka@ip.kyusan-u.ac.jp
[b)] chengk@is.kyusan-u.ac.jp

本研究では，我々は情報科学的手法として，マススペクトルに対するピーク検出とピークアライメント，ピークの類似度計算を用いた微生物同定法を提案し，実験による評価を行った．微生物同定を行う上で適切なピーク検出とピークアライメントを行い高精度の微生物同定方法を開発することで微生物同定の精度向上のヒントを得ることが期待される．また，本研究では細菌やウイルスがもつ表面抗原の型に基づき，菌種からさらに細かい部類に分類したものを示す**血清型**に基づき，菌種同定を実施する．

2. 質量分析法と微生物同定

2.1 質量分析法

質量分析による微生物同定は微生物に含まれるタンパク質が各々固有の重さを持っているという性質を利用している．本研究ではマススペクトルを取得する際に利用されているマトリックス支援レーザー脱離イオン化飛行時間型質量分析計（MALDI-TOF-MS）を例に質量分析を用いてマススペクトルを取得するまでの流れは以下の通りである．

(1) 調査対象の微生物（試料）内の分子等に対してレーザー光で照射してイオン化させる．
(2) イオンを電磁気力を用いて，装置内のイオンの検出部分に向けて飛ばす．
(3) 飛ばされたイオンの飛行時間差によって分離し，検出部分で検出する．

2.2 マススペクトル

(1)〜(3) の手順後，検出部ではイオンが検出された際には信号が発生した際の結果を横軸を質量電荷比（m/z），縦軸を信号強度（intensity）として記録されたものがマススペクトルになる．マススペクトルの形状は異なる菌種または菌株によって異なる性質を利用して微生物同定が行われる．また，マススペクトル内に現れている信号をピークといい信号強度はイオン化した試料のイオン量に対応しているため，試料の同定や構造解析等に活用される．

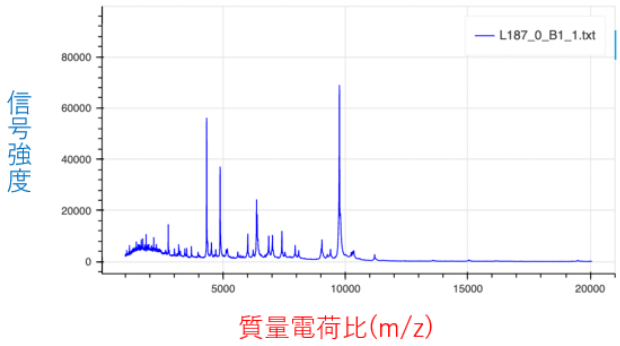


図 1: マススペクトルの例

同定に利用されるマススペクトルの特徴を以下で紹介する．

(1) イオンの質量電荷比が小さいものほどイオン化しやすいため，m/z が小さい範囲にピークが現れやすい．
(2) 同位体の存在によりマススペクトル上には隣合う固有のピークが現れる．

試料内の分子をイオンする前に装置内の気相にて発生する分解反応であるフラグメンテーションが発生する場合がある [6]. 通常，MALDI を用いたイオン化は発生しづらく，プロトン付加分子等のみが出現する場合が多いが [5]，マススペクトルを取得する実験環境の違いによっては予想外のフラグメンテーションの発生やマススペクトル内にノイズが発生する場合がある．また，出現位置が近いピークを分離する分解能に差があることから同じ菌種，菌株のマススペクトルを複数回取得した場合でもマススペクトルの形状に大きな違いが出る恐れがある．

3. ピークパターンの類似度に基づく微生物同定

本研究では微生物同定の精度向上のヒントを得るための試みとして各マススペクトルのピークパターンに着目し，それらの類似度を Jaccard 類似度等の類似度計算を用いて微生物同定を行う．まず，前処理としてピーク検出方法の一種である CWT を用いてマススペクトル内から微生物同定に有効なピークの検出を行った．加えて，マススペクトル内のピークの m/z のズレの修正を行うピークアライメントをペアワイズとグローバルの 2 つの手法を用いてを行った．ピークアライメント後，Jaccard 類似度と Rank 類似度，2 つの重み付き Rank 類似度を用いてピークの類似度計算を行った．計算は本研究で扱う 102 個の菌株別のマススペクトルのピークパターンの類似度を総当たりで求めた．

3.1 ピーク検出

ピーク検出（Peak Detection）とはマススペクトルからピークの信号強度と m/z を検出することである．ピーク検出は MALDI によって得られた質量分析データの解析で利用されているが [2]，他分野でも利用される技術である．例えば，医療分野では心電図（ECG）や脳波（EEG）などの生体信号解析において利用される．財務分析では，株価や市場データのピークを特定して，トレンド分析や予測を行う際に活用される．地震学分野では地震波形の解析において，重要な地震イベントの検出に使用される．

このようにピーク検出は様々な研究分野で利用されており，ピークは時系列データのようなデータの分析に欠かせない存在だといえる．一般的なピーク検出までの過程は以下の通りである [2].

(1) **平滑化（Smoothing）**：隣接値に比べて異常な値を除去する．
(2) **ベースライン補正**：イオン化に伴うノイズを除去する．
(3) **ピーク検出**：生物情報学分野では連続ウェーブレット変換（CWT）が基本とされている．

一般的にピークの検出は平滑化とベースライン補正後のデータに対して行う処理であるが，平滑化とベースライン補正を行う手法の組み合わせによってはピークの検出結果に違いがでてしまう．しかし，CWT によるピーク検出は平滑化とベースライン補正を連続ウェーブレット変換にて行うことが可能であるため，ピーク検出結果に一貫性を持たせることが可能である [2]．そのため，本研究では CWT を用いてピーク検出を行う．

CWT について説明していく．連続ウェーブレット変換（Continuous Wavelet Transform: CWT）とは対象のデータや信号から特徴的な信号を捉えるために使用されるウェーブレットと呼ばれる信号内の小さな波を用いた解析方法である．連続ウェーブレット変換は信号から切り取ったウェーブレットの伸縮（スケーリング）を行っていく．圧縮した場合は信号の位置と時間の分解能を高くすることができ，逆に伸縮した場合は信号の位置と時間の分解能が低くなるが周波数の分解能が向上する [8]．スケーリング後は元データに対してシフトを行い，異なる時点で現れている特徴的な信号を検出を行っていく．

CWT を用いたピーク検出はウェーブレット空間上でのパターンマッチングに基づいて行う．このパターンマッチングはウェーブレットが元データ内の信号とどの程度一致しているかを示す数値であるウェーブレット係数を用いて，信号の特定のパターンや特徴とどの程度一致する

かを分析を行う．ウェーブレット係数が高い場合，信号とウェーブレットの間の良好な一致を示し，異なるスケールでのウェーブレット変換を通じて，信号の異なる特性を明らかにすることができる．そのため複雑なパターンやノイズのあるピークの識別に有効な手段だとされている [1][2].

3.2 ピークアライメント

ピークアライメントとは，実験対象の複数の MS データセットにおいて，ピークの位置である m/z の真の値を求め，各データセットのピーク位置 m/z のずれを訂正する処理である．ピーク位置の誤差は本研究対象である微生物同定精度に悪影響を及ぼす可能性がある．そこで，対象の MS データセットに対してピークアライメントを行いピーク位置のずれを訂正することで，微生物同定の精度向上を図る．本研究におけるピークアライメントは，MS データに対してピーク抽出を行い，各 MS データのピークリスト（m/z 値）を作成した後に，ピークリストに対してピークアライメントを行う方法をとる．以下に本研究で実施したピークアライメントを紹介する．

3.2.1 ペアワイズピークアライメント

ペアワイズピークアライメント（Pairwise Peak Alignment）とは二つのピークリスト内のそれぞれの m/z を比較し，条件に基づき同じピークかどうかを判定後，同じ m/z であるピークで構成されるピークリストを作成する手法のことである．この手法はアライメントを行うピークリストの組み合わせに応じて，条件を変えることで柔軟にアライメントを行うことができる．しかし，異なるピークリスト間でのアライメントの一貫性を保つことが困難である．

本研究ではピークアライメント時に各ピークリスト内のピークの m/z の差がある閾値 δ の絶対値以内ならば同じピークと判定している．

3.2.2 グローバルピークアライメント

グローバルピークアライメント（Global Peak Alignment）はペアワイズピークアライメントと異なり，全てのピークリストを同時にアライメントを行っていく手法である．全てのピークリスト内のそれぞれの m/z を比較し，共通の条件に基づき同じピークかどうかを判定後，同じ m/z であるピークで構成されるピークリストを作成する．この手法は複数のピークリストのピークを共通の条件でアライメントを行うため異なるピークリスト間でのアライメントの一貫性を保つことができる．

しかし，大規模なデータセットが対象の場合，計算コストや処理時間がかかることや全てのピークリストのピークを同時に処理するため，ペアワイズアライメントと比較して複雑な処理になってしまう．本研究ではピークアライメント前に全てのピークリストのピークの m/z を一次元化後，カーネル密度推定（Kernel Density Estimation: KDE）を用いてピークアライメントを行う．

カーネル密度推定とは，各データから対応する確率密度関数（カーネル関数）を求め，それらを足し合わせることでデータの分布の推定を行っていくものである．図 2 はデータセットのヒストグラムに対し，カーネル密度推定を行なった結果を描画したものである．カーネル密度推定量を求めることでカーネル密度推定を行うことができるが，その際には各データに対応するカーネル関数とカーネル関数の幅を定義するバンド幅が必要である．カーネル関数は正規分布（ガウス分布）が最も使用されており，本研究で行うカーネル密度推定を用いたピークアライメントでも採用している．

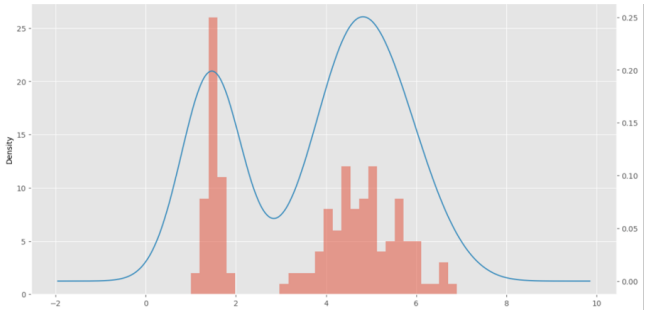


図 2: カーネル密度推定の例

カーネル密度推定を行うことでデータの分布を滑らかに近似し，それぞれのデータ分布を区別する「谷」を特定することができる．その谷は、密度の極小値となる点となり，異なるデータグループを最適に分割することが可能となるため，密度推定の結果を用いてアライメントが可能である．

3.3 ピークパターンの類似度計算手法

2 つのピークリストをそれぞれ A,B とした場合のピークリストの類似度 s を以下のように定義する [7].

$$s(A,B) = \frac{\Sigma_{x \in A \cap B} \sigma(x)}{|A \cup B|} \tag{1}$$

式（1）の評価関数 σ の定義によって各類似度計算を行うことができる．以下で本研究で扱うマススペクトルデータ内のピークデータ集合の類似度計算を行う際に利用する類似度計算手法を紹介する．

1．Jaccard 類似度

Jaccard 類似度（Jaccard Similarity）とは二つのピークリストの類似度を測定するために使用される手法のことで，ピーク間の出現位置のみを考慮して類似度計算を行う．2 つのピークリスト A と B があり，共通ピークを区別しない場合，評価関数は $\sigma(x) = 1$ となり，以下の式で Jaccard 類似度を求めることができる．

$$s_{jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

2．Rank 類似度

Rank 類似度（Rank Similarity）とは二つのピークリストの類似度を測定するために使用される手法のことで，ピークの高さ順位を与えるランク関数 $r()$ を用いて，各ピークの高さ順位のみを考慮して類似度計算を行う．2 つのピークリスト A と B があり，各データセットのピークのランク順位の差の絶対値が Δ 以下であれば 1，条件を満たさない場合は 0 とするステップ関数 $\sigma()$ で定義する．2 つのピークリスト間のピークで同じとものと判定された場合，評価関数は $\sigma(x) = 1$ となり，以下の式で Rank 類似度を求めることができる．

$$s_{rank}(A,B) = \frac{|\{x : x \in A \cap B \wedge |r_A(x) - r_B(x)| \leq \Delta\}|}{|A \cup B|} \tag{3}$$

3．重みつき Rank 類似度

重みつき Rank 類似度（Weighted Rank Similarity）とは前述の Rank 類似度によって各ピークの高さ順位のみを考慮して同じピークの組を求めた後に，それぞれのピークに対して重み w() を付与することでピーク順位の重要度を高める手法である．2 つのピークリスト A と B があり，それぞれのデータセットのピークのランク順位が近ければそれぞれのピークに対してピークの高さ順位を考慮した異なる重み w_A，w_B を与えることで重みつき Rank 類似度を求めることができる．

$$s_{weighted}(A,B) = \frac{\Sigma_{x \in A \cap B}(w_A(x) + w_B(x))}{|A \cup B|} \tag{4}$$

本研究で使用する重みは以下の 2 つの重みを用いて重みつけを行った．一つ目は，**逆数関数重み**である．ピークリスト A と B の共通ピークを求めた後，それぞれピークの高さ順位の逆数 $w_A(x) = 1/r_A(x)$ と $w_B(x) = 1/r_B(x)$ を重みとする．二つ目は，**シグモイド関数重み**である．ピークリスト A と B の共通ピークを求めた後，それぞれの高さ順位を入力値としたシグモイド関数を重みとする．シグモイド関数（sigmoid function）とはネイピア数 e を用いた式で入力値の大きさによらず出力値は 0 または 1 のどちらかに収束していく関数である．シグモイド関数の式は以下のように定義される．

$$w(r) = \frac{1}{1 + e^{ar}} \tag{5}$$

ピークのランク順位 r を引数とし，高いランク順位のピークの重みをより大きくしたいため，e のべき乗を正の値にし，シグモイド関数の滑らかさに関わるパラメータ $a > 0$ にしている．

4. 評価実験

まず前処理で CWT を用いてマススペクトル内からピークの検出を行い，ノイズのピークの除去を行う．次に¡m/z, intensity¿で構成されるピークリストを作成後，m/z のズレの修正をペアワイズピークアライメントまたはグローバルピークアライメントで行う．前処理後，全 102 個の各菌株のマススペクトルの類似度を Jaccard 類似度と Rank 類似度，重みつき Rnk 類似度をを用いて類似度計算を総当たりで行う. その後，後述する実験にて，各類似度計算手法によるピークパターンの類似度計算の評価を行う．

4.1 実験データ

本研究では 102 のリステリア菌 (Listeria) のマススペクトル実データを使用する．リステリア菌は河川水や動物の腸管内などの環境に多く存在しており，食品では生ハムなどの食肉加工品やナチュラルチーズなどの乳製品，魚介類加工食品など多くの食品で増殖する細菌である．リステリアによる食中毒に罹患した場合，妊婦や高齢者が特に重症化しやすく致死率が高いことから注意が必要な細菌の一種である．

下記の図 3 は本研究で使用するリステリア菌のメタデータの一部である．主に Listeria Monocyotogenes（LM）の各菌株から抽出した情報が xlsx ファイルでまとめられている．以下に本研究で用いる項目の説明を行う．

MALDITOFMS	No.	Srotype	MLST	hly	inlA	clpC	plcA	菌株名
L001	1	1/2a	33	6	3	9	13	LM1
L002	2	1/2a	45	5	6	2	16	LM3
L003	3	4b	21	3	14	10	8	LM4

図 3: リステリアのメタデータ

MALDITOFMS Listeria serial No. は L1〜L184 まである一意の ID であり，菌株 ID として使用する．菌株の総数は 184 個であり，No. 40. 41. 43. 116 は LM ではなく，No. 72 は血清型が間違っている可能性がある．Srotype は血清型のことである．リステリア属菌は O 抗原（菌体）と H 抗原（鞭毛）により 17 の血清型に分類されていて，LM では 13 の血清型が知られている．本研究では血清型を用いて菌種同定を行う．

実データはリステリア菌からマトリックス支援レーザー脱離イオン化法（MALDI-TOF）によって抽出された m/z と Intensity の 2 列の系列データ (MS データ) で，テキストファイルでまとめられており，102 個存在する．1 つのファイルには約 7 万行のデータで 1 列目は m/z，2 列目は intensity で構成されており，m/z の範囲は 996.800〜20070.096，intensity の範囲は 0〜172302 である．

また，各菌株にはいずれかの血清型（**1/2a, 1/2b, 1/2c, 3a, 3b, 3c, 4b, UT, 4e, 4c**）の情報が付与されているもので 94 個存在し，不明なものは 8 つである．毒性が強い血清型は 1/2a，4b，1/2b であり，リステリア症を引き起こす種類となっている．

4.2 前処理

前処理の一環として，ピーク検出を行う．CWT を用いてピーク検出を行った．使用ライブラリは scipy ライブラリで提供されている関数の 1 つである find_peaks_cwt で定義される．

```
find_peaks_cwt(z, [a])
```

- z：各菌株の intensity のデータ．
- $[a]$：ウェーブレットの幅．

今回は例として”L001_0_F3_1”のマススペクトルを用いて，CWT によるピーク検出を行なった．図 4 はピーク検出後のマススペクトルであり，ピークの検出結果を赤×印で描画した．L001_0_F3_1 のマススペクトルは 73,216 個のピークが存在するが，CWT によって 204 個のピークを検出した．

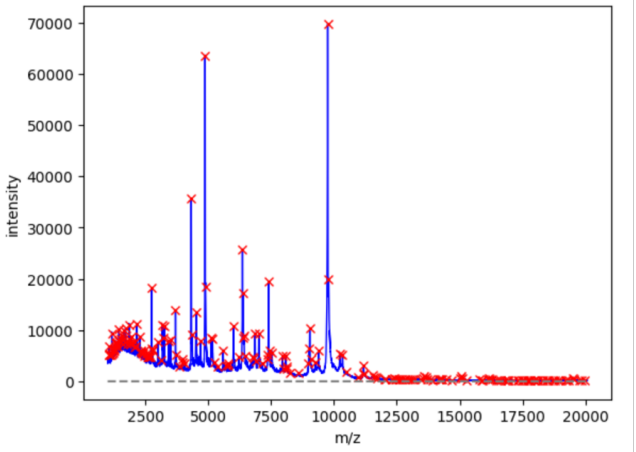


図 4: マススペクトル内からのピーク検出例

次にアライメントを行う．これまでマススペクトル内のノイズを除去するために，CWT を用いたピーク検出と同時にマススペクトルの縦軸である intensity の修正を行った．次の過程では横軸である m/z の修正を行なう．本研究では 2 つの手法によるピークアライメントを実施した．

（1）ペアワイズアライメント

ピーク検出後の各菌株の m/z のデータを用いて各ペアごとの m/z のデータをペアワイズピークアライメントの閾値である δ を基準に align 関数を用いて行う．

```
align(pk1, pk2, δ)
```

- pk1：ピークアライメント対象の菌株の m/z データ
- pk2：pk2 以外の菌株の m/z データ
- δ: ペアワイズピークアライメントの閾値

今回は L185_0_A9_1 と L101_0_A1_1 のマススペクトルを例に各 m/z データと $\delta = 3$ に設定し，ピークアライメントを行う．図 5 はペアワイズピークアライメント後の結果である．左の列から順にピーク ID，信号強度，m/z のデータ列が並んでいる．ピーク ID の付与条件は各マススペクトルの m/z を比較して，値の差が絶対値 δ 以内ならば同じピーク ID を付与している．そのため，図 5 の赤丸で囲まれた m/z の差は絶対値 3.0 以内のため，同じピークと判定し，今回はピーク ID の 4 を付与している．逆に条件を満たさないピークには一意のピーク ID を付与している．

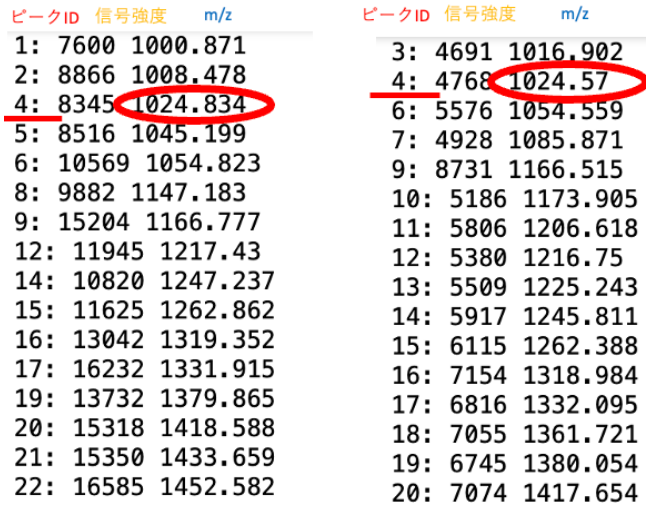


図 5: ペアワイズピークアライメントの実施例

（2）グローバルピークアライメント

ペアワイズピークアライメントでは各菌株の組み合わせで m/z のアライメントを行なったが，グローバルピークアライメントではすべての菌株の m/z を一次元化した後にカーネル密度推定を用いて同じ基準で m/z のアライメントを行う．各菌株の m/z と intensity が格納されているピークをすべての菌株の m/z のみで構成された一次元データに加工する．その後，このデータを用いてピークアライメントを行う．

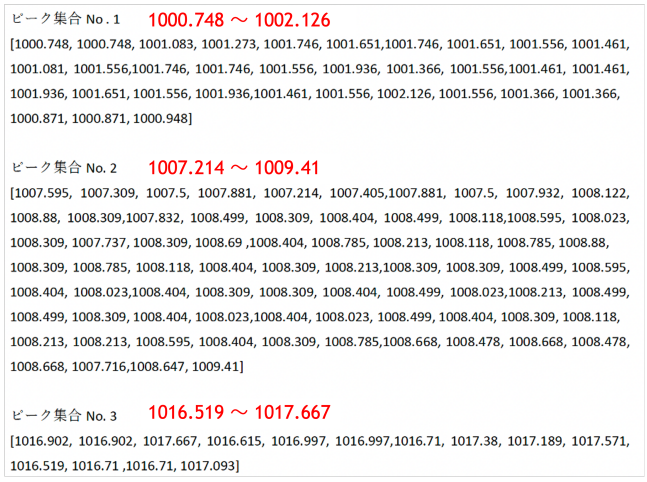


図 6: グローバルピークアライメントの結果の一部

図 6 はカーネル密度推定を用いて，m/z の極大値と極小値を基準に同一の m/z であるピーク集合を作成したものである．ピーク集合の作成基準は各マススペクトルの m/z がピーク集合の極大値～極小値の範囲内ならば同じピークと判定している．そのため，1 つ目のピーク集合の極大値は

1002.126，極小値は 1000.748 であるため m/z の 1001.083 はこのピーク集合に含まれ，同じピークとして判定された．

後述する評価指標による類似度計算による微生物同定の評価を行うために，ペアワイズもしくはグローバルピークアライメント後のピークリストを用いて Jaccard 類似度と Rank 類似度，２つの重みつき Rank 類似度による各類似度計算を式（1）を用いて求めた．

4.3 微生物同定の評価実験

類似度計算による微生物同定の評価は菌種別のマススペクトル 94 個または菌株別では 102 個のデータに占める同じ菌種または菌株が N 以内となった割合（正解率）を用いて評価を行う．数式の定義は以下のように示す．

$$\text{正解率} = \frac{\text{類似度が N 位以内の同一菌種・菌株数}}{\text{菌種・菌株総数}} \quad (6)$$

N は同じ菌種または菌株と判断する際の条件として，(1) 同じ菌種または菌株だと判断する際の候補となる菌種・菌株数，(2) 正解の候補となる菌種または菌株データの数 N 以内に正解のデータが含まれるならば正解とする．N が小さいほど正解の候補数が少なくなるため，より厳密な正解となり，正解率の価値は高くなる．

本研究では以下の２つの評価実験を行い，各類似度計算を用いて各菌株または菌種同士のピーク類似度の精度の評価を行なった．

TOPN：N 変化に伴う正答率変化を確かめる実験

$\delta = 3$ に固定するペアワイズアライメントとグローバルアライメント後，Jaccard 類似度と Rank 類似度，２つの重みつき Rank 類似度のピーク類似度の菌株識別と菌種同定の正答率を式（6）で評価する際に N を変化させた場合の影響を検証する．そのため本実験では $1 \leq N \leq 5$ に設定する．

DELTA：適切な閾値 δ を調べる実験

ペアワイズピークアライメントで使用する閾値である δ を変化させた場合の正解率への影響を検証する．また，N は 2 に固定する．

4.4 実験結果

4.4.1 TOPN の実験結果

TOPN の結果を図 7 と 8 に示す．ペアワイズピークアライメントを行ったマススペクトルのピーク類似度による菌株識別では N が 1〜3 の時，重みをシグモイド関数にした重みつき Rank 類似度の正解率が 0.35〜0.5 と最も高い結果となり，Jacard 類似度によるピーク類似度計算が最も低い結果となった. この結果からピーク類似度による菌株識別においては各ピークの m/z のみならず intensity を考慮することが正確な菌株識別を実現する上で重要な要素だと分かった．

同様にピーク類似度による菌種同定では $N = 1$ の時，Rank 類似度計算によるピーク類似度計算の正解率が 0.69 と最も高く，続いて２つの重みつき Rank 類似度の正答率がそれぞれ 0.68 となった．しかし，N が 2 以上の時の正答率は Jaccard 類似度の正解率が最も高い結果となった．この結果からピーク類似度による菌種同定においては各ピークの intensity を考慮することも不可欠な要素であるが，菌株識別に比べ各ピークの m/z の位置が正答率に関わってくることが分かった．

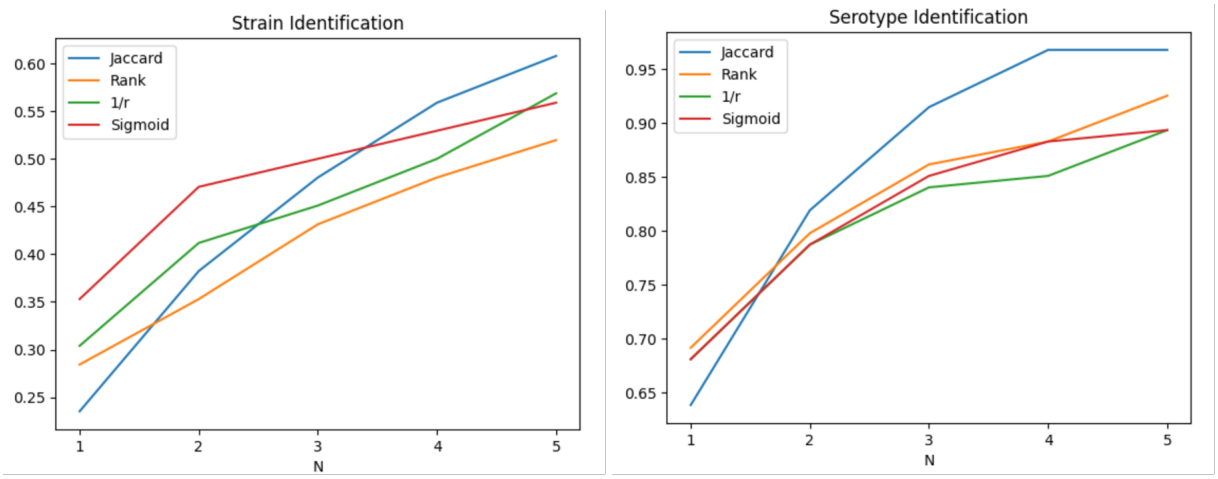
グローバルピークアライメントを行ったマススペクトルのピーク類似度による菌株識別では N が 1〜4 の時，重みをシグモイド関数にした重みつき Rank 類似度の正解率は 0.33〜0.54 となり，$N = 4$ の時まで 2 つの重みつき Rank 類似度が他手法の正答率と比較して高い結果となった．そのため，同様にグローバルピークアライメントの場合でもピーク類似度による菌株識別においては各ピークの m/z のみならず intensity を考慮することが正確な菌株識別を実現する上で重要な要素だと分かった．

同様にピーク類似度による菌種同定では N が 1 の時から，Jaccard 類似度の正解率が 0.67 と最も高い結果となった．また，ペアワイズピークアライメント後の $N = 2$ の時の各類似度計算の正解率は Rank 類似度は 0.79 で，1/ピーク順位を重みとした重みつき Rank 類似度は 0.78 であったが，グローバルピークアライメント後の各正答率は 0.76 と 0.77 になり低下する結果になってしまった．また，シグモイド関数を重みとした重みつき Rank 類似度の正解率は 0.78 から 0.81 と向上した．しかし，N が 3 以上でもシグモイド関数を重みとした重みつき Rank 類似度以外は正答率が 0.1〜0.4 ほど低下する結果となった．

4.4.2 DELTA の実験結果

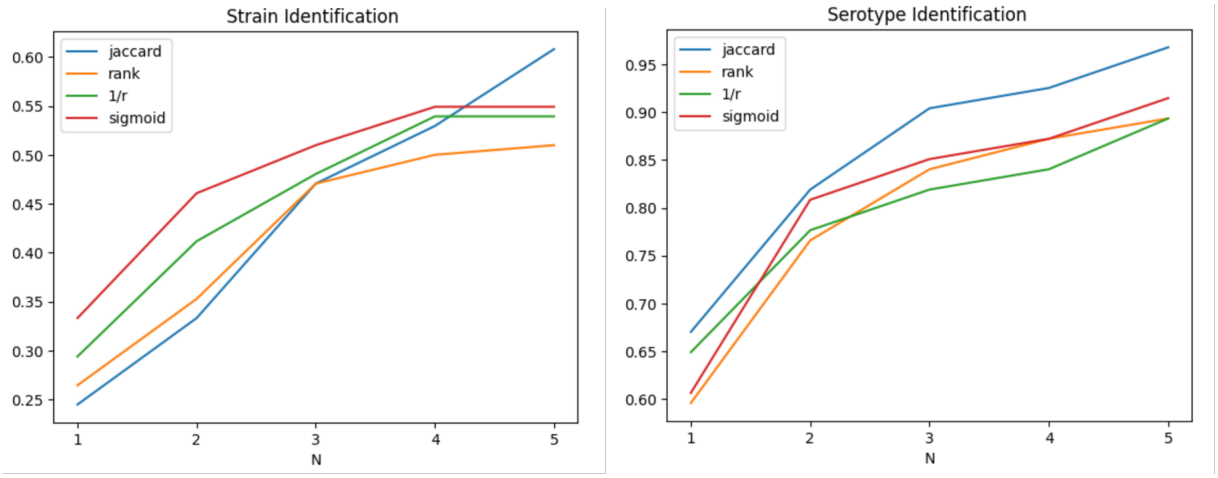
DELTA の結果を図 9 に示す．$1 \leq \delta \leq 5$ を 0.5 ずつ変化させた場合の菌株識別の正答率の推移は，シグモイド関数を重みつきした重みつき Rank 類似度は $\delta = 3.0, 3.5$ の時の正答率が 0.47 と最も高く，ピーク順位の逆数を重みつきした重みつき Rank 類似度は $\delta = 2.0$ の時の正解率が 0.43 であった．Jaccard 類似度と Rank 類似度の正解率は $\delta = 1$ の時，0.42 と 0.44 であった．そのため，ペアワイズピークアライメントでは各類似度手法に応じた δ の設定が必要である．

同様に正答率の推移は Jaccard 類似度計算は $\delta = 2.0$ の時に正答率が 0.86 と最も高く，シグモイド関数とピーク順位の逆数を重みつけした重みつき Rank 類似度はそれぞれ 0.81 と 0.8 であり，Rank 類似度の正解率は $\delta = 2.5$ の時，0.83 であった．しかし，$\delta > 2.0$ の Jaccard 類似度の正答率は $\delta = 4.5$ の時を除いて低下する結果となった．この結果から Jaccard 類似度はピークの m/z のみを考慮するため，適切な δ を設定しなければ，本当は m/z が異なる菌種のピークにも関わらず同じピークだと判定されていることが考えられる．また，他手法の正解率は多少の変動がある



(a) TOPN における菌株識別の正答率の推移　(b) TOPN における菌種同定の正答率の推移

図 7: ペアワイズピークアライメント後



(a) TOPN における菌株識別の正答率の推移　(b) TOPN における菌種同定の正答率の推移

図 8: グローバルピークアライメント後

ものの菌株識別時の Rank 類似度と Jaccard 類似度のような大きな減少傾向はなかったため，ピークの高さ順位が近いもののピーク位置が異なるピークが周辺に存在していると考えられる．

この結果からペアワイズピークアライメントの閾値 δ は菌株識別では 1〜3 の間で，菌種同定では 2〜2.5 の間で設定することが最適だと結論付けた．

5. 終わりに

本研究では，微生物同定の精度向上に向けて，マススペクトルの適切な前処理手法から，ピークパターンの類似度による微生物法を考案し，実データによる評価を行った．マススペクトルに潜む特徴的なピークの検出を CWT を用いて行い，m/z のズレの修正を行うピークアライメントをペアワイズとグローバルの 2 通りによる前処理の仕組みを開発した．各ピークアライメントと類似度計算の評価を 2 つの実験を通して行なった．

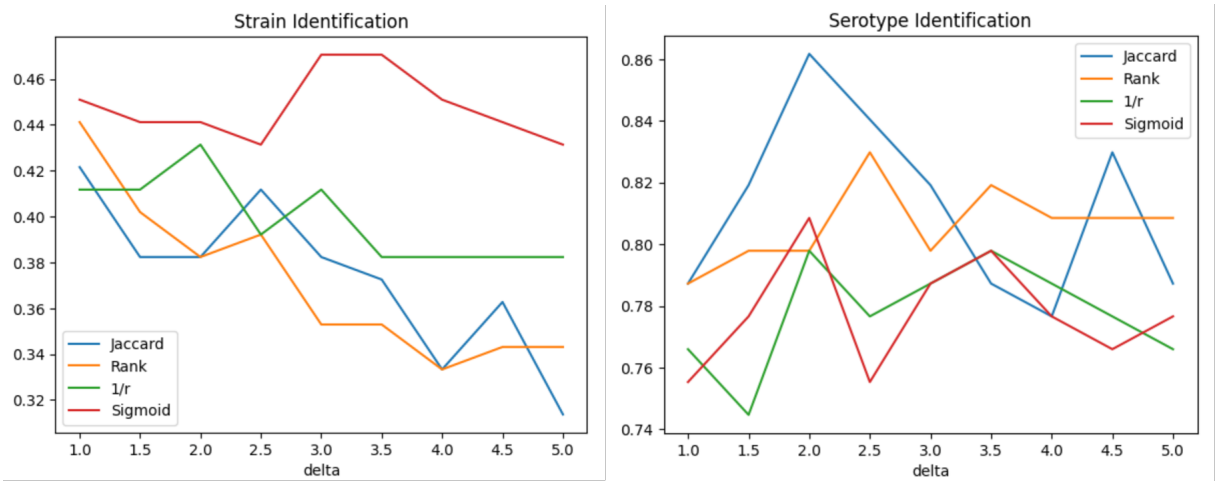
結果として，適切なパラメータを調べたうえ，Rank 類似度，Jaccard 類似度がより良い結果が得られることがわかった．また，重み付け関数としてシグモイド関数が適切で，前処理の一環となるピークアライメントとして，ペアワイズアライメントがグローバルアライメントよりもシンプルで結果もよいことがわかった．

今後の課題として，密度推定以外のピークアライメント法の調査と評価や機械学習や深層学習を用いた微生物同定と比較することがあげられる．

謝辞　本研究の遂行にあたり，実験データを提供してくださった本学総合機器センターの方々に感謝する．

参考文献

[1] Pan Du et al, (2006), Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics, Volume 22, Issue 17, September 2006, Pages 2059 – 2065, https://doi.org/10.1093/bioinformatics/btl355



(a) DELTA における菌株識別の正答率の推移　　(b) DELTA における菌種同定の正答率の推移

図 9: 実験 DELTA の結果

[2] Chao Yang et al, (2009), Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics 10, 4 (2009). https://doi.org/10.1186/1471-2105-10-4
[3] Roland Bamou et al, (2021), Using MALDI-TOF mass spectrometry to identify ticks collected on domestic and wild animals from the Democratic Republic of the Congo. Exp Appl Acarol. 2021 Jul;84(3):637-657. doi: 10.1007/s10493-021-00629-z. Epub 2021 Jun 19. PMID: 34146230; PMCID: PMC8257524.
[4] 大楠清文，質量分析技術を利用した細菌の新しい同定法，モダンメディア 58 巻 4 号 2012, pp.113-122
[5] 一般社団法人　日本医用マススペクトル学会，医療学生のための医用質量分析学テキスト，株式会社　診断と治療社，2019.
[6] 豊田岐聡，質量分析学ー基礎編ー，株式会社国際文献社，2020.
[7] 浅野公平，豊坂祐樹，成凱，質量法データの類似度評価による菌株識別，第 76 回電気・情報関係学会連合九州支部大会，2023 年 9 月
[8] 田原鉄也，連続ウェーブレット変換の基礎，IEEJ Journal vol.129 No.10 pp.660-663 2009 特集「ウェーブレット解析とその産業応用　1」
[9] 植木優夫, 笛田薫, カーネル密度推定におけるカーネル関数の比較，日本計算機統計学会大会論文集 17 巻 pp.147-150 2003