4N-08

質量分析データの機械学習による微生物同定

浅野　公平†　豊坂　祐樹‡　成　凱†

九州産業大学†　　九州工業大学‡

1. はじめに

食の安全を守るため、食中毒の原因になる微生物の同定を迅速にかつ正確に行うことは非常に重要である．微生物同定の手法には RNA 遺伝子解析により全ゲノムの配列を決定する方法も存在するが、時間とコストを要するだけではなく同属の別種間の配列の類似性が高いものは同定の精度が下がるという課題が挙げられる[1]．近年、質量分析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微生物を効率よく同定する手法が普及してきている．しかし、試料のイオン化に伴うノイズが存在し、高精度の微生物同定の実現が道半ばといえる[2]．我々は食品関連の微生物を中心として、機械学習等による高精度の微生物同定を目指して開発を行っている．本稿ではこれまで我々の取り組んだアプローチと途中結果をまとめて紹介する．

2. 質量分析とマススペクトル

2.1 微生物のマススペクトル

質量分析(Mass spectrometry)とは、原子、分子、クラスター等の粒子を気体状のイオンとし、真空中で運動させ電磁気力を用いて、あるいは飛行時間差によりそれらイオンを質量電荷比(m/z)に応じて分離・検出することである．質量電荷数比に応じて分離・検出されたイオンをもとに、横軸に m/z 、縦軸にイオンの相対強度をとった棒グラフをマススペクトル(Mass spectrum)と呼ぶ．マススペクトルは、どの分子がどのくらいの信号強度でどのタイミングで現れているのかを示されており、菌種によってパターンが異なる．そのため、この特徴を利用することで、微生物の同定が可能としている [1]．

2.2 同定に利用されるマススペクトルの特徴

微生物同定に利用されるマススペクトルの特徴は以下の通りである．(1)検出されたイオンの個数が多ければ多いほど、イオンの相対強度は高くなる．(2)イオンの質量電荷比が大きいものはイオン化しづらいため、マススペクトルの質量電荷比が小さい部分にピークが現れやすい．(3) 同位体を持つ原子が試料に含まれていた場合、マススペクトル上では隣り合ってピークが出現するため、同位体の数に応じてピークが固有な分布で現れる．

Identification of Bacteria by Machine Learning of Mass Spectrum, Kouhei Asano†　Yuki Toyosaka‡　Kai Cheng †
†　Kyushu Sangyo University
‡　Kyushu Institute of Technology

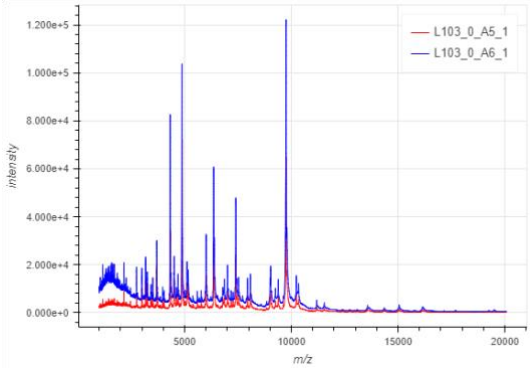

図 1 同一菌株の異なるマススペクトル

しかし、試料分子をイオン化する過程は要するに分子からの電子移動であるため、後続反応によって試料が分解するフラグメンテーションと知られる現象がしばしば起こる．また、近い m/z 値を持つ 2 つのピークを分離する分解能に差があり、空気、GC カラムの液相、マトリックスやフタル酸エステル類などのバックグラウンドピークも重なることがある．これらのことは未知試料のマススペクトルを読む際の障害となる場合が多く、精度の高い同定が困難な要因である．そのため、図 1 のように同一菌株に対して質量分析を二回行う場合にピークの出現パターンがかなり違うことがある．

3. ピークパターンによる菌種同定

3.1 ピーク・ピーク検出

データセット内の局所的最大値のことをピーク(Peak)といい<m/z, intensity>で表す．m/z は、質量電荷比（m/z 値）であり、intensity は信号強度である．データセットから局所的最大値とその位置を検出することは、ピーク検出(Peak Detection)といい、スムージングとベースライン補正を行なった後に行う．ピークと強度のパターンは菌種によって異なるため、ピークパターンが菌種同定のポイントとなっている．MALDI Biotyper[1]では、各菌株とその強度の平均値を算出する．まれにしか検出されないピークはあらかじめノイズとして除去後データベースに登録される．同定したい菌株がデータベースのどの菌株（菌種）と近縁であるかをスコア値で表現し、高い順に候補の菌種名が表示される．また VITEK MS[1]では、同一菌種の株を多重測定し、安定して検出されるピーク群のうち科、属、菌種に特

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

徴的なピークに重み付けを行い、データデータに登録される．同定したい菌株が科、属、菌種レベルの順に一致したピークのスコア値を加算し、合計スコア値で同定を行う．

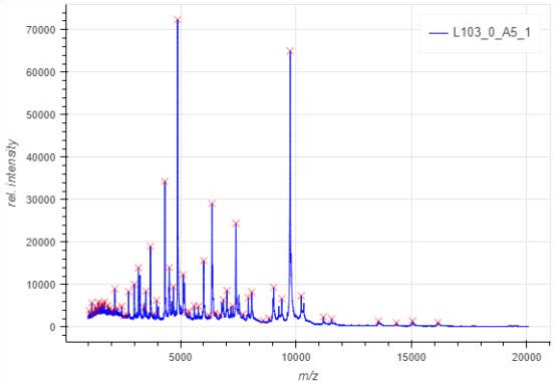

図 2 マススペクトルにおけるピーク検出

3.2 ピーク類似度に基づく微生物同定

本研究では、ピークパターンの類似度により菌種同定につながる試みとして、以下の類似度尺度を検討している．

Jaccard 類似度：データセット間のピークの共起度を Jaccard 類似度という．二つのデータセットにおいてピークの m/z 値の集合をそれぞれ A と B とするとき、Jaccard 類似度は次のように定義することができる．

$$\frac{|A \cap B|}{|A \cup B|}$$

Rank 類似度：データセット間の高さ（強度）を考慮して高さの順位の近いピークの共起度を Rank 類似度という．二つのデータセットにおいて、ピークの m/z 値の集合をそれぞれ A と B とし、各集合における要素の高さ順位を rank() で与えるとき、Rank 類似度は次のように定義する。

$$\frac{|\{m|rank(m,A) \approx rank(m,B), m \in A \cap B\}|}{|A \cup B|}$$

重み付け Rank 類似度：前述の Rank 類似度の各集合における要素の高さの順位を考慮し、重みものを重み w() を与えたもの Rank 類似度といい、次のように定義する。

$$\frac{\sum_{m\,\varepsilon A \cap B}\{w(m,A) + w(m,B)\}}{|A \cup B|},$$

ここで$w(m,A) = 1/rank(m,A)$

3.3 機械学習による微生物同定

ピークの有無や出現順序を特徴量として機械学習で菌株の分類を行うことが期待される．k-means 法による微生物同定を行うため、時系列モデルに適用する k-means 法を採用する．k-means 法は、対象のデータセットを k 個のグループに分割し、各クラスターの重心の算出と算出した重心による再クラスタリングを繰り返すことでデータのクラスターを構築する一般的なクラスタリングの一種である．時系列データの各時間項目の列を使用して、k-means のアルゴリズムを適用することで時系列データでのクラスタリングを実現している．

4. 評価実験

ピーク類似度に基づくピーク検出を用いた微生物同定とピークパターンの機械学習による微生物同定を行い、その結果を比較することでそれぞれの精度を検証する．

実験準備として、51 種の菌株を MADLI マススペクトルの実データを使い、2 回ずつ測定した計 102 個のマススペクトルデータを用いる．ピーク検出ではマススペクトル解析で最もよく使われている連続ウェブレット変換(CWT)[2]を採用した．

今回用いるモデルは Jaccard 類似度、Rank 類似度、k-means 法の 3 つとする．評価指標として、3 つのモデルによる各菌株のデータに対する同定を行い、予測した菌株の正解数によって評価を行う．今回使用したのは Python のライブラリである tslearn の timeseriesKmeans であり、ユークリッド距離を使用した k-means 法によってクラスタリングを行った．実験では、マススペクトルデータの系列データを時系列データと同様に扱い、k-means を適用して菌株の同定を行った．

ピークパターンの類似度を求めるとき、m/z 値に誤差があることに対処するため、あらかじめ m/z 値を丸めた．例えば、m/z 値が 1244.470，1237.432 を四捨五入して 1240 にすると誤差を吸収する．

表 1 各モデルの菌株同定の正解数（計 51）

	Jaccard	Rank	k-means
正解数	9/51	13/51	17/51

実験結果は表 1 に示す．単純な類似度による同定では、51 の菌株のうち正しく同定されたのはそれぞれ 9 と 13 しかない．重み付け Rank 類似度もほぼ同じ結果であった．これに比べて k-means 法による機械学習では正解数が 17 と高かった．

5. 終わりに

本研究では、質量分析データの解析による微生物同定を検討し、類似度評価と機械学習による同定を行い、精度向上のヒントが得られた．

謝辞

本研究の一部は，KSU 基盤研究費 K060069 の助成を受けたものである．

参考文献

[1] 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

[2] Chao Yang et al. 2009, Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis, BMC Bioinformatics. 2009; 10: 4

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.