ラマンスペクトルデータの AI 利用による微生物細胞分析

重　藤　　真　介

1　はじめに

化学の諸分野の中でも分析化学はとくに，統計学的なデータ解析との関係性が深い分野である．濃度，温度，pH，スペクトル強度など様々な変量に依存するデータ（多変量データ）を解析し，そこから最大限の情報を抽出するために，ケモメトリックスと呼ばれる新しい融合分野が 1960～1970 年代に生まれた[1]．以来，分析装置とコンピューターの急速な進歩に支えられ，ケモメトリックスはクロマトグラフィー，近赤外分光，振動分光などにおけるデータ解析に広く応用されてきた．

赤外吸収およびラマン散乱により得られる振動スペクトルは，分子の基準振動に由来する多くのピークからなるパターンを示し，分子構造や分子間相互作用などに関する豊富な情報を含む．多種多様な分子から構成される細胞や組織などの生体試料は，各成分の振動スペクトルが異なる割合で重なり合った複雑なスペクトルを示すことがほとんどである．細胞を構成する生体物質の種類と相対濃度は細胞の個性を鋭敏に反映するため，ラマンスペクトルパターンに基づいて，がん細胞と正常細胞，病原菌と非病原菌といった細胞の識別が原理的には可能である．しかし，たとえ細胞の種類が違っていても，細胞の構成成分そのものは似たり寄ったりであり，また極めて多数の細胞を取り扱う必要性があることから，細胞の識別は純物質の識別と比べると遥かに難しいことが容易に想像される．

近年，人間の目では違いを判別し難い大量のラマンスペクトルデータを人工知能（artificial intelligence, AI）技術を用いて解析し，病原菌検出や病理診断などに応用する研究が活発化している．筆者らは，環境中の未知微生物の探索と解明に資する新たな微生物解析技術の一つとして，微生物 1 細胞のラマンスペクトルデータの機械学習による微生物の高精度識別法の開発に取り組んできた．本稿では，深層学習（ディープラーニング）を用いた研究など当該分野の最近の動向について簡単に述べた後，微生物（細菌とアーキア）の種の識別を行った筆者らの研究成果[2)3)]について紹介する．

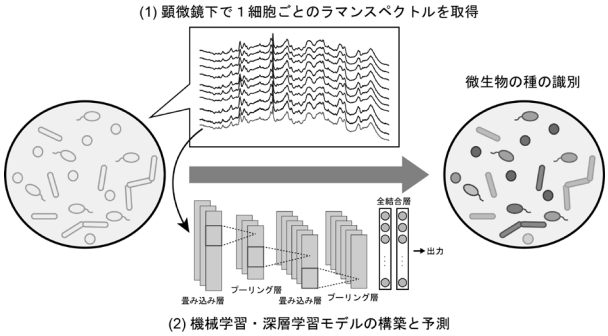


図 1　AI を利用した 1 細胞ラマンスペクトルデータに基づく微生物種識別の概念図
深層学習アルゴリズムの一つである CNN を例として示す．

2　ラマン分光と AI 技術の融合の現状

機械学習は AI による学習方法に応じて，教師なし学習，教師あり学習，強化学習の 3 種類に大別される．そのうち強化学習は，ラマンイメージング計測の迅速化に応用した最近の研究[4)]を除いて，まだあまり例がない．正解となるラベル付けされたデータを与えずに学習させる教師なし学習には，ケモメトリックスでもお馴染みの主成分分析や線形判別分析，クラスタリングなどがある．これらの手法は未知のデータに対する学習が可能であるという大きな特長を持つ．しかし，学習精度・速度の点では教師あり学習の方が優位であるため，ラベル付きデータを学習できる既知微生物種の識別においては教師あり学習，とくに深層学習がよく用いられている．図 1 に AI を利用した細胞のラマンスペクトルデータに基づく微生物種識別の概略を示す．自ら測定したデータ，もしくはデータベースから入手したデータを訓練データとして機械学習・深層学習モデルを構築し，テストデータでの予測を行ってその性能を評価するのが一連の流れである．

Ho らは ResNet アーキテクチャーを採用した畳み込みニューラルネットワーク（convolutional neural network, CNN）を用いて，メチシリン耐性黄色ブドウ球菌を含む感染症の原因となる主要な細菌および酵母の分離株 30 種から取得したラマンスペクトルを学習させ，82 % を超える平均正解率での識別に成功したことを報告した[5)]．同様のアプローチは，ESKAPE 病原菌と呼ばれる高度多剤耐性菌[6)]や食中毒の原因となる *Arcobacter* 属細菌[7)]など，多様な微生物群に適用されている．これらの応用例で使用されたラマンスペクトルデータには自発ラ

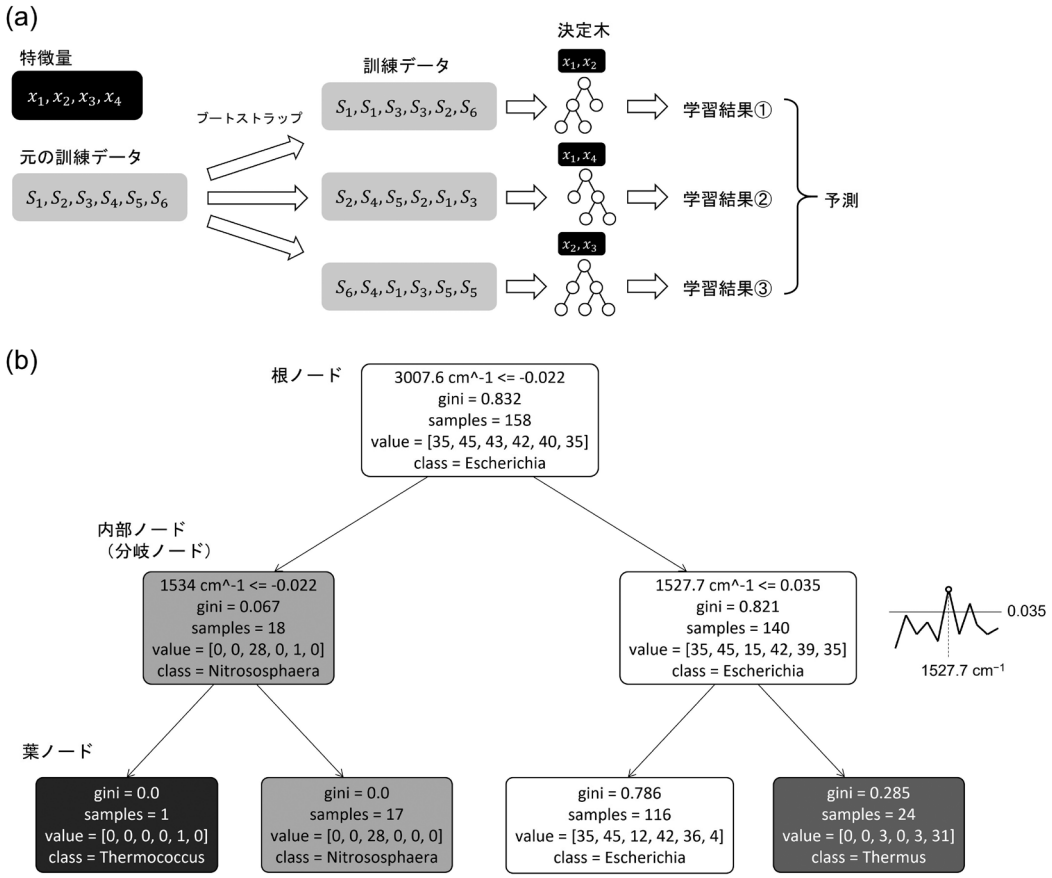


図 2　ランダムフォレストの原理

（a）アンサンブル学習の構造．（b）決定木の分岐（実際の Python の出力を示す）．

マンスペクトルだけでなく，表面増強ラマンスペクトルも含まれる[8]．また，限られたラマンスペクトルデータからより大きなデータセットを人為的に生成するためのデータ拡張（data augmentation）法の開発も並行して行われている[9]．

CNN は畳み込み層とプーリング層を複数重ねることによってデータの複雑な特徴を効率よく抽出し，高度な識別・予測を可能とするが，その過程や結果の根拠が不明であるという深層学習の「ブラックボックス問題」が存在する．もちろんブラックボックスであることが致命的な欠点とならない応用分野も多くあるだろう．しかし，AI の思考過程を理解することは，ラマンスペクトルデータへの応用に最適化した機械学習手法の開発に加えて，細胞の個性と生体物質の関係を明らかにするうえでも重要であると考えられる．そこで，今後さらに盛んになることが期待されるラマンスペクトルデータの AI 分析の基礎となる知見を得ることを目的として，筆者らは AI が「ラマンスペクトルをどのように捉えて細胞の識別を行っているか」を容易に可視化できる，よりシンプルな仕組みを持つ機械学習アルゴリズム，ランダムフォレストを用いた微生物種識別を行った．

3　1 細胞ラマンデータのランダムフォレストを用いた微生物種の識別

3･1　ランダムフォレスト

ランダムフォレストは 2001 年に提案された[10]，古典的と言ってもよい機械学習アルゴリズムである．ランダムフォレストは弱識別器である決定木を並列に作成し，それらの出力結果の平均または多数決により予測を行うアンサンブル学習モデルの一種である．各決定木での学習の際に，訓練データのサンプリング（バギング）を行う，一部の特徴量のみで決定木の分割を行うなどの工夫を施して決定木間の相関を小さくすることで，過学習を防ぎ高い汎化性能を得ることができる（図 2（a））．決定木は Gini 不純度が最小となるような特徴量で分割を行う（図 2（b））．今回の場合，選択されたラマンシフトにおけるスペクトル強度の値でデータを分割している．この Gini 不純度の値から，各特徴量の重要度を算出することができ，どのラマンバンドが識別に寄与したかを定量的に評価できるのである．

3･2　データセット

この研究では，系統分類学的に多様な 6 種の微生物（細菌 3 種とアーキア 3 種）を用いた．細菌 3 種のうち，

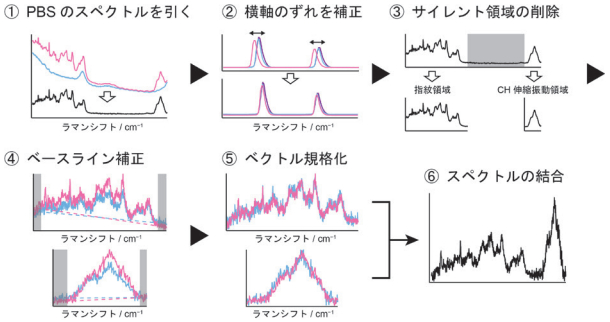


図 3　機械学習に用いた 1 細胞ラマンスペクトルデータの処理方法

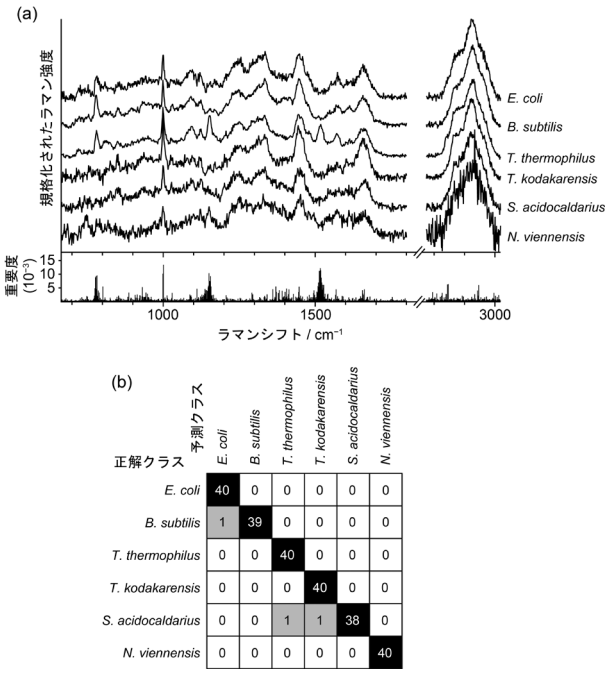


図 4　ランダムフォレストによる 6 種の微生物の識別
(a) 40 細胞の平均ラマンスペクトル（前処理済み）と各特徴量の重要度．(b) 識別結果を表す混同行列．（文献 2 より許可を得て転載）

Escherichia coli と *Thermus thermophilus* はグラム陰性細菌であるのに対して，*Bacillus subtilis* はグラム陽性細菌である．各微生物について，He−Ne レーザーの 632.8 nm の発振線を励起光とする研究室自作の共焦点ラマン顕微鏡[3)11)]を用いて，リン酸緩衝液（PBS）に懸濁させた細胞一つ一つのラマンスペクトルを測定した．その際，細胞は高い開口数（本研究では 1.3）の 100 倍対物レンズ（ニコン CFI Plan Fluor DLL）でタイトに集束したラマン励起光による光ピンセット効果で捕捉した．浮遊性の動物細胞も含め，数 μm 程度の細胞はレーザー光による光ピンセットで簡便に集光位置にとどめておくことができる．1 種につき 40 細胞からラマンスペクトルを得た．

得られた 1 細胞のラマンスペクトルは PBS のラマン散乱信号，細胞内物質由来の自家蛍光，ノイズなどを含む．上述のように，学習過程において各ラマンシフトにおける強度値が決定木の構築に用いられるため，実測のラマンスペクトルをどのような形で機械学習に供するのかは結果に大きな影響を及ぼす．言い換えると，スペクトルの前処理が極めて重要となる．図 3 に筆者らが行ったスペクトル処理の手順を示す．まず細胞のみのスペクトルを得るために PBS のスペクトルを減算した後，装置の設定によって実験ごとにわずかに異なるスペクトルの横軸（波数範囲）を，ネオンランプの輝線を利用して補正した．続いて，通常ラマンバンドが観測されないサイレント領域（1800〜2700 cm^{-1}）を除去し，スペクトルを指紋領域と CH 伸縮振動領域に分割した．分割されたそれぞれのスペクトルのベースラインを，直線を用いて補正した．ベースライン補正には多項式フィットやペナルティ最小二乗スムージングなどが用いられるが，筆者らは分光学を専門としない研究者でも処理が簡単かつ任意性の少ない，直線による大まかなベースラインの除去を採用した．最後に，ベクトル規格化を行い，指紋領域と CH 伸縮振動領域を結合した．

3･3　識別精度と特徴量の重要度の評価

3･2 で述べた方法で処理した各微生物種の平均ラマンスペクトル（n=40）を図 4（a）に示す．人間の目でも種間の違いが明らかなところもあれば，区別がつけ難いところもある．また，細胞サイズの違いを反映して，アーキア（とくに *Nitrososphaera viennensis*）のスペクトルは細菌のそれと比べてノイズが相対的に大きいことがわかる．平均ではない個々の細胞のスペクトルではそれがより顕著となるが，スペクトル処理をできる限りシンプルなものに留めるという方針のもと，特異値分解を利用したノイズ除去などは行わなかった．

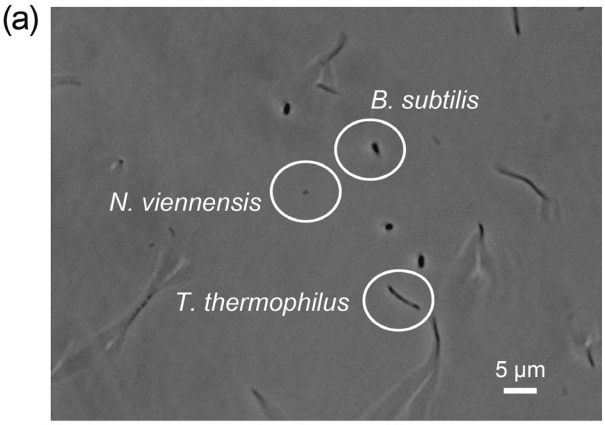
前処理済みのラマンスペクトルデータを用いてランダムフォレストモデルの構築を行った．ランダムフォレストは Python の scikit-learn で実装した．1 種につき 40 細胞分，計 240 スペクトルを使用した．データを 10 分割し，9 割のデータを用いて識別モデルの訓練を行い，残りの 1 割のデータをテストデータとして用いて識別精度を評価した（10 分割交差検証）．決定木の数や特徴量の個数などの重要なハイパーパラメーターのチューニングはグリッドサーチを用いて行った．モデル構築の詳細については文献 3 を参照していただきたい．その結果，図 4（b）に混同行列の形で示すように，微生物 6 種を平均正解率 98.8±1.9% で識別可能なモデルを構築することができた．

では，ランダムフォレスト機械学習モデルは微生物細胞のラマンスペクトルのどのような特徴を捉えてそれらを識別しているのであろうか？　本当に意味のある特徴量（ラマンシフト）で識別しているのか？　これらの問いに答えるため，各特徴量の重要度を算出しプロットした（図 4（a））．重要度が高い波数がまとまっている部

分は平均スペクトルにおいて何らかのラマンバンドが存在する部分に対応しており，ランダムフォレストによる識別が分子の観点から「説明可能」な結果であることがわかった．

識別への寄与が大きいラマンバンドは主にタンパク質（1000, 1650 cm^{-1} 付近）および DNA/RNA（780 cm^{-1} 付近）に帰属される．図 4（a）の平均スペクトルから一見してわかる *T. thermophilus* のみが有する 1150, 1520 cm^{-1} 付近のラマンバンドも当然，重要度が高いが，これらはカロテノイドの共鳴ラマンバンドである（*T. thermophilus* は黄色のカロテノイドを持つことが知られている）．2800～3000 cm^{-1} 付近の CH 伸縮振動領域はタンパク質の寄与が支配的であるが，その中でも識別への寄与度が高い 2850 cm^{-1} 付近には脂質の CH_2 対称伸縮振動に特徴的なピークが現れる．細菌の膜脂質が脂肪酸から構成されるのに対して，アーキアの膜脂質はイソプレノイドアルコールからなる骨格を持つため，炭化水素鎖中の CH_2 基の数が相対的に少なくなる．実際，アーキアは細菌に比べて CH_3/CH_2 伸縮振動バンドの強度比が高い傾向にあることが，顕微赤外分光による先行研究で報告されている[12]．また，細菌 3 種の中でも，グラム陰性細菌（*E. coli*, *T. thermophilus*）は外膜を有するためグラム陽性細菌（*B. subtilis*）と比べて膜脂質の量が多いと考えられる．このように，CH 伸縮振動領域は細菌/アーキア，グラム陰性/陽性細菌の細胞構造の違いを反映するラマンスペクトル上の特徴の一つであるが，機械学習による識別結果はその生物学的な差を反映したものであることが確かめられた．

最後に筆者らは，微生物の *in situ* 識別への応用に向けた proof-of-concept として，このランダムフォレスト識別モデルを *B. subtilis*, *T. thermophilus*, *N. viennensis* の 3 種の微生物混合試料における識別に適用した．これらの微生物は図 5（a）の顕微鏡写真からわかるとおり，細胞の形状およびサイズが顕著に異なるので，顕微鏡下で個々の細胞がどの種であるかというラベル付けが可能である．混合試料中の 3 種の微生物からそれぞれ 20 本程度のラマンスペクトルを測定し，6 種識別で構築したランダムフォレストモデルを用いてそれらの識別を行ったところ，この場合も 98.4 % という高い正解率を達成することができた（図 5（b））．



(b)

正解クラス ＼ 予測クラス	*E. coli*	*B. subtilis*	*T. thermophilus*	*T. kodakarensis*	*S. acidocaldarius*	*N. viennensis*
B. subtilis	0	21	0	0	0	0
T. thermophilus	0	0	20	0	0	0
N. viennensis	0	0	0	0	1	20

図 5　ランダムフォレストによる混合試料中の 3 種の微生物の識別

（a）3 種の微生物細胞の位相差顕微鏡像．（b）識別結果．（文献 2 より許可を得て転載）

4　おわりに

本稿で紹介した筆者らの研究では識別した微生物種は 6 種と限定的であったが，最近，ランダムフォレスト同様，決定木学習をベースとした機械学習アルゴリズム LightGBM を用いることで，20 種以上の高精度識別に成功した．さらに，細胞の生理状態や生物ドメイン（細菌/アーキア）の識別にも挑戦している．

この分野の論文数はここ数年で急速に増えている．分子論的な裏付けがなされないまま，精度と速度の向上を目的として様々な深層学習アルゴリズムをラマンスペクトルデータに応用している研究も多く，ラマンスペクトルデータの解析における有意義かつ適切な AI の活用方法をさらに追求していく必要があると筆者は感じている．奇しくも 2024 年のノーベル物理学賞，化学賞はどちらも AI 関連の研究に与えられた．今後も AI の進歩がもたらすデータ駆動型分光学から目が離せない．

文　献

1）尾崎幸洋，宇田明史，赤井俊雄：“化学者のための多変量解析　ケモメトリックス入門”，(2002)，(講談社サイエンティフィク)．

2）N. Kanno, S. Kato, M. Ohkuma, M. Matsui, W. Iwasaki, S. Shigeto : *iScience*, **24**, 102975 (2021)．

3）N. Kanno, S. Kato, M. Ohkuma, M. Matsui, W. Iwasaki, S. Shigeto : *STAR Protoc.*, **3**, 101812 (2022)．

4）K. Tabata, H. Kawagoe, J. N. Taylor, K. Mochizuki, T. Kubo, J.-E. Clement, Y. Kumamoto, Y. Harada, A. Nakamura, K. Fujita, T. Komatsuzaki : *Proc. Natl. Acad. Sci. USA*, **121**, e2304866121 (2024)．

5）C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon, J. Dionne : *Nat. Commun.*, **10**, 4927 (2019)．

6）S. Singh, D. Kumbhar, D. Reghu, S. J. Venugopal, P. T.

Rekha, S. Mohandas, S. Rao, A. Rangaiah, S. K. Chunchanur, D. K. Saini, S. Umapathy : *Anal. Chem.*, **94**, 14745 (2022).
7) K. Wang, L. Chen, X. Ma, L. Ma, K. C. Chou, Y. Cao, I. U. H. Khan, G. Gölz, X. Lu : *Appl. Environ. Microbiol.*, **86**, e00924-20 (2020).
8) Y. Zhang, K. Chang, B. Ogunlade, L. Herndon, L F. Tadess, A. R. Kirane, J. A. Dionne : *ACS Nano*, **18**, 18101 (2024).
9) M. Wu, S. Wang, S. Pan, A. C. Terentis, J. Strasswimmer, X. Zhu : *Sci. Rep.*, **11**, 23842 (2021).
10) L. Breiman : *Mach. Learn.*, **45**, 5 (2001).
11) A. Matsuda, N. Sakaguchi, S. Shigeto : *J. Raman Spectrosc.*, **50**, 768 (2019).
12) M. Igisu, K. Takai, Y. Ueno, M. Nishizawa, T. Nunoura, M. Hirai, M. Kaneko, H. Naraoka, M. Shimojima, K. Hori, S. Nakashima, H. Ohta, S. Maruyama, Y. Isozaki : *Environ. Microbiol. Rep.*, **4**, 42 (2012).

重藤 真介（Shigeto Shinsuke）
関西学院大学大学院理工学研究科化学専攻（〒669−1330 兵庫県三田市学園上ケ原 1）．東京大学大学院理学系研究科化学専攻博士課程修了．博士（理学）．《現在の研究テーマ》ラマン分光の微生物研究への応用．非線形光学イメージング．《趣味》囲碁，野球観戦．
E-mail : shigeto@kwansei.ac.jp

原　稿　募　集

「技術紹介」の原稿を募集しています

対象：以下のような分析機器，分析手法に関する紹介・解説記事
1）分析機器の特徴や性能および機器開発に関わる技術，2）分析手法の特徴および手法開発に関わる技術，3）分析機器および分析手法の応用例，4）分析に必要となる試薬や水および雰囲気などに関する情報・解説，5）前処理や試料の取扱い等に関する情報・解説・注意事項，6）その他，分析機器の性能を十分に引き出すために有用な情報など

新規性：本記事の内容に関しては，新規性は一切問いません．新規の装置や技術である必要はなく，既存の装置や技術に関わるもので構いません．また，社会的要求が高いテーマや関連技術については，データや知見の追加などにより繰り返し紹介していただいても構いません．

お問い合わせ先：
日本分析化学会『ぶんせき』編集委員会
〔E-mail：bunseki@jsac.or.jp〕