特集　人工知能を活用したバイオ産業分野の新潮流

遺伝子解読技術やポストゲノム技術の一般化や低コスト化の加速によりビッグデータ化が進んでいる生物情報から新たな知識を抽出する手段として，人工知能（AI）を活用した技術への期待が高まっている。生物情報のビッグデータとAIの融合により，これまで実験的探索が不可欠であった生物科学関連研究は，AIによるビッグデータ分析を中心としたデータ駆動型の研究手法へとパラダイムシフトを起こしつつある。このような背景の中，バイオ産業分野の研究開発手法も大きく変貌しつつある。AIを活用した従来の探索的研究の効率化や，画像データを活用した培養細胞の品質の診断などが活発に報告され，さらには，細胞やタンパク質などの生体および生体分子の開発にも応用されつつあり，様々な研究プロジェクトが始動している。本特集では，バイオ産業（バイオものつくり）分野における生物情報ビッグデータとAIを融合した最新技術に着目し，その研究事例と展望について紹介する。本特集を通して発展途上であるAIとバイオ技術の融合領域の様々な取組みや事例から，新たな事業化や研究開発戦略のシーズやアイデアのヒントを提供したい。（編集担当：小西正朗・山田真澄・山田剛史・中澤　光）†

深層学習を用いた代謝混合物の解析と重要代謝物の探索

伊達 康博

1. はじめに

昨今の人工知能ブームに伴い，「深層学習」や「ディープラーニング」，「AI」といった言葉を至る所で見かけるようになってきた。実際にこうした技術は，テレビやカメラなどの画像処理技術，インターネットにおける検索エンジン，自動車の自動運転技術などに応用されており，多くの人々が知らず知らずのうちにその恩恵にあずかっている。人工知能関連技術は，このような身近な製品のみならず，農業や医療などのバイオ産業関連分野においても研究開発が進んでおり，社会構造を変える革新的な技術として多くの期待が寄せられている。

人工知能の根底を支える重要な要素技術の一つが機械学習である。機械学習は，読んで字のごとく，コンピュータがたくさんのデータを学習し，学習によって見出した法則性に基づいて分類や予測などをおこなうアルゴリズムであり，深層学習も機械学習の一種である。機械学習の計算アルゴリズムは，昨今の第3次人工知能ブーム以前より，化学や生物学系を含む様々な研究分野で利用されており，筆者の専門分野である核磁気共鳴（nuclear magnetic resonance, NMR）を利用した代謝混合物の解析（いわゆるメタボロミクス研究）においても，2000年代中頃から有用なデータマイニング技術として利用されてきた。本稿では最初に，メタボロミクス研究におけるデータマイニング技術の重要性と機械学習の適用例について簡単に概説する。次に，深層学習のメタボロミクス研究への導入と，深層学習を利用した重要代謝物の探索方法について，自身の研究成果を中心に説明する。最後に，深層学習を適用する上で問題となるサンプル数（n数）に関する課題とその解決策に関する最新の研究成果についても紹介したい。

Deep Learning for Analyses of Metabolic Mixtures and Screening of Important Metabolites
Yasuhiro DATE
2010年　早稲田大学大学院先進理工学研究科生命医科学専攻博士課程修了
現　在　理化学研究所 環境資源科学研究センター 環境代謝分析研究チーム研究員
連絡先：〒230-0045　神奈川県横浜市鶴見区末広町1-7-22
E-mail　yasuhiro.date@riken.jp

2019年11月5日受理

†Konishi, M.　令和元・2年度化工誌編集委員（2号特集主査）北見工業大学工学部
Yamada, M.　同上　千葉大学大学院工学研究科共生応用化学専攻
Yamada, T.　同上　（株）ダイセルイノベーション・パーク生産技術本部シミュレーション技術センター
Nakazawa, H.　同上　東北大学大学院工学研究科

2. メタボロミクスとデータマイニング

メタボロミクスとは，微生物を含む生物の代謝活動によって生じた代謝混合物の網羅的な計測・解析を意味している。メタボロミクス研究におけるサンプル計測では，NMRや質量分析装置が主に用いられているが，共通して言えることは得られたスペクトルの中に多数のピーク（すなわち代謝物）を含んでいることである。これら多数の代謝物ピークの中から，例えば病気のマーカー分子となるような代謝物を探索する際に，機械学習のようなデータマイニング技術が重要となってくる。NMRスペクトルからパターン認識により特徴抽出を試みた初期の報告では，データマイニング技術として主成分分析（principal component analysis，PCA）や階層的クラスター分析（hierarchical cluster analysis，HCA）などの多変量解析が用いられた[1-3]。それ以来今日に至るまで，データの特徴を概観するために利用されるPCAや，二群あるいは三群以上の違いを特徴づける代謝物を探し出すことが可能な部分的最小二乗法（partial least squares，PLS）をベースとした判別分析（discriminant analysis，DA）が，メタボロミクス研究において頻繁に用いられている。筆者らも，微生物処理プロセスにおける代謝動態解析[4]や食品産業におけるプレバイオティクス候補物質のスクリーニング[5]などにメタボロミクス技術とこれらの多変量解析を利用しており，PCAやPLSはメタボロミクス研究における重要な解析ツールである。

PCAやPLSは，多くの有名な統計解析ソフトウェアの中に実装されており，簡便で有用な解析ツールとして広く利用されているため，機械的にこれらの手法を利用するメタボロミクス研究者も少なくない。しかしながら，正しく理解せずに使用すると間違った解釈を導く可能性もはらんでいるため注意が必要であるとGromskiらは警鐘を鳴らしており，同時に代替手法としてサポートベクターマシン（support vector machine，SVM）やランダムフォレスト（random forest，RF）のような機械学習手法の利用も提案している[6]。こうした機械学習手法は，実際に，尿のメタボロミクスデータにおけるSVMとPLSDAの解析性能を比較した研究[7]や山菜を摂食した際の腸内環境変動をRFにより解析した研究[8]などに利用されており，PLSDAに代わる強力なデータマイニング技術として利用されてきた。また，MetaboAnalyst[9]やKODAMA[10]，classyfire[11]などに代表されるような，機械学習を取り入れた解析ツールや統計解析ソフトウェアの開発も盛んにおこなわれており，筆者らも機械学習を利用した重要変数（代謝物）選択法の開発[12, 13]や機械学習と量子化学計算による高精度なNMR化学シフト予測技術の開発[14]などの研究成果を報告している。このように，機械学習はメタボロミクス分野における代謝混合物の解析技術や重要代謝物の探索技術として有用であり，深層学習を導入するための下地は既に構築されていたと言える。

3. 深層学習の導入と重要代謝物探索法

「深層学習」と一言で言っても，画像認識などで使われる畳み込みニューラルネットワークや，自然言語処理などに利用されるリカレントニューラルネットワークなど，そのアルゴリズムは解析するデータセットの性質に応じて多種多様なものがある。それらの最も基本的な枠組みがニューラルネットワーク（neural network，NN）である。**図1**に示されているように，NNの一種である多層パーセプトロン（3層）の基本構造としては，入力層，中間層，出力層の三層からなり，入力層の各ノードから中間層の各ノードへ，中間層の各ノードから出力層の各ノードへと情報の伝達（計算）がおこなわれ，最終的に出力層にて計算の結果が出力される。このようなNNの中間層を多層化したものがディープニューラルネットワーク（deep neural network，DNN）であり，深層学習の最も基幹的なアルゴリズムである。

筆者らがDNNアルゴリズムの適用に関する研究を開始した2016年当初では，DNNなどを含む深層学習を用いたメタボロミクス研究は報告されていなかった。メタボロミクス研究にDNNアルゴリズムを適用する上で当初問題であったのが，分類／回帰モデルの構築に寄与している重要な代謝物を直接的に特定できないことであった。この問題を解決するため，筆者らは基本的なDNNアルゴリズムにパーミュテーション法を組み込んだアルゴリズム作成をおこなった。パーミュテーションとは再配列や並び替えを意味する言葉であるが，この研究では，ある特定の変数（代謝物）に対して，ランダムサンプリングにより取り出してきた値を，各サンプルの値へと一つずつ代入することにより，元の数値から代入値へとランダムに入れ替えた新しい行列を作成する方法である（**図2**）。この方法を用いて作成

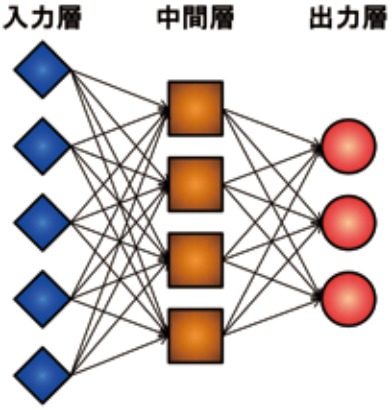

図1　ニューラルネットワークの基本構造

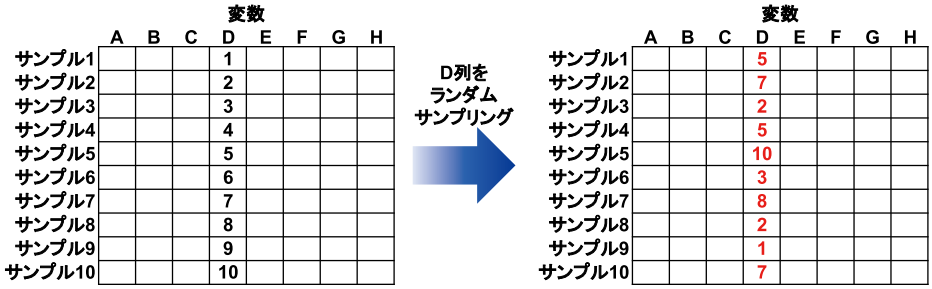

図2　パーミュテーション法による再配列

されたデータセットに対して，DNNアルゴリズムで構築された分類／回帰モデルを適用し，計算された予測値と元のデータセットにおける予測値との差分値を算出する。この工程を数十回（本研究では50回）繰り返し，算出された差分値の平均値を重要度の指標とする，DNN-MDA（mean decrease accuracy，平均減少精度）アルゴリズムを作成した[15]。なお，機械学習の一種であるRFにおいても重要度の算出にMDA法が利用されており，この研究で作成したMDA法は，RFにおけるMDA法をベースに一部改変を加えたものである。

作成したDNN-MDA法を用いて，産地判別問題に対する分類性能の評価をおこなった。この時用いたデータセットは関東地方およびそれ以外の地域（大部分は東北地方）に由来するマハゼ1022個体のNMR計測データである。メタボロミクス分野で頻繁に用いられているPLSおよび機械学習の一種であるRFおよびSVMによる分類性能と比較したところ，PLSでは分類精度の平均値が57.3%であったのに対し，RF，SVMおよびDNN-MDA法ではそれぞれ95.0%，95.8%および97.8%を示し，DNN-MDA法が最も高精度な産地判別能を有することがわかった（**図3**）。また，DNN-MDA法を用いることにより，構築された高精度な分類モデルに寄与している重要な変数としてグリシンやイノシン酸などの代謝物を特定することができ，本手法が重要な代謝物の探索法としても有用であることが示された。なお，筆者らの研究が*Analytical Chemistry*誌に受理された2017年には，DNNを用いた白米の産地判別に関する研究[16]や乳がん患者におけるエストロゲン受容体の状態をDNNにより判別する研究[17]がほぼ同時期に報告されており，深層学習がメタボロミクス研究へと進出し始めた黎明期にあたる研究成果であったと言える。

4. サンプル数に関する課題

深層学習をメタボロミクス研究へと応用する上での最大の課題は，サンプル数の問題である。深層学習はビッグデータの解析において革新的な技術であるが，逆に言うと，ビッグデータが得られない，あるいは得ることが非常に難しいようなデータに関しては適用が難しいため，創意工夫が必要である。バイオ系の研究ではこの問題が特に顕著であり，サンプル自体の入手やその計測データをたくさん集めることは難しい場合が多い。実際に筆者らの研究においても，1000を超えるマハゼサンプルを収集し，計測することに多大な時間と労力を要したが，たかだか1000サンプル程度では，世間一般で言われるビッグデータには遠く及ばない。では，実際にどの程度のサンプル数が得られれば，深層学習を適用できるのであろうか。いわゆるインターネットの検索エンジンや自動運転技術のような人工知能技術と言われるレベルの解析を，バイオ系のサンプルや計測データで実施するために必要なサンプル数についてははっきりとしたことは言えないが，前述したような深層学習の使い方であれば，**図4**に示したようなDNNによる分類精度とサンプル数の関係が得られている[15]。この解析結果から，90%以上の平均分類精度を得るためには，最低200サンプル程度必要であることが読み取れる。これはメタボロ

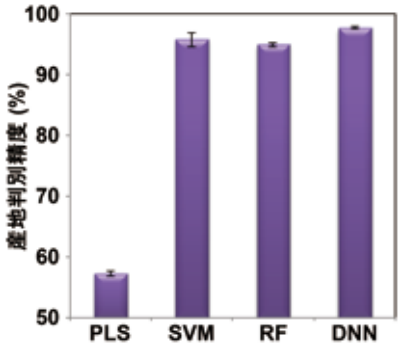

図3　産地判別精度の比較[15]

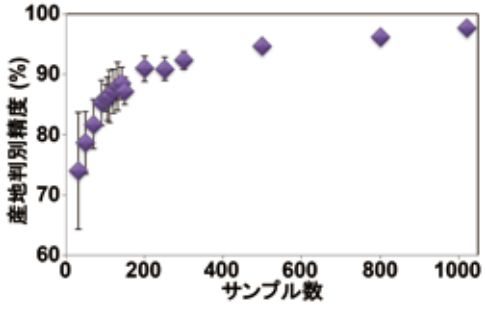

図4　サンプル数と産地判別精度の関係[15]

ミクス分野の研究では一般的なサンプル数であり，DNN-MDA法を用いた判別／回帰分析や重要代謝物の探索という使い方であれば，バイオ関連分野における適用・応用も可能であると考えられる。

一方で筆者らは，より少ないサンプル数のデータセットに対してもDNN-MDA法を適用できるように，アルゴリズムの改良に取り組んでいる。その一端として筆者らは，アンサンブル学習に着目した研究開発をおこなった。アンサンブル学習とは，複数の学習器がそれぞれ別々にモデルを構築し，それらを統合して一つの学習モデルを生成するような機械学習アルゴリズムの一手法であり，代表的なアンサンブル学習としてRFやブースティングなどがある。このアンサンブル学習の概念をDNN-MDA法へと応用し，DNN学習器を複数生成して個別に学習をおこない，最終的に各学習モデルを統合することにより分類／回帰精度を向上させることが可能なアンサンブル型DNN（ensemble DNN，EDNN）アルゴリズムを開発した[18]。本研究では，カタクチイワシやスズキ，ヒラメやブリなど8種類の魚種を対象に，メタボロミクス分野では比較的少なめのサンプル数である18～129サンプルのデータセットを用意し，各魚種に対してEDNN法によるサイズ予測モデルを構築し，その予測精度を二乗平均平方根誤差（root mean square error，RMSE）を用いて評価するとともに，代表的な機械学習手法であるDNN，RF，SVM法の予測精度と比較した。**表1**に示されているように，EDNN法は通常のDNN法と比べて8魚種全てにおいてRMSE値が小さく，予測精度が高いことがわかる。一方で，RFやSVM法と比較すると，8魚種中4魚種で最も高精度な予測を達成しており，通常の機械学習手法と同程度以上の予測性能を有することが示された。従ってEDNN法は，DNN-MDA法におけるサンプル数の問題を改善することに成功し，比較的サンプル数の少ない場合においても利用可能な解析アルゴリズムの開発に成功したと言える。筆者らが開発をおこなってきたDNN-MDA法やEDNN法は，メタボロミクス研究のみならず，将来的にはバイオ産業関連分野における有用な解析ツールとしての利用も期待できる。

表1　RMSEを指標とした回帰精度の比較[18]

魚種名	サンプル数	EDNN	DNN	RF	SVM
カタクチイワシ	30	2.89	3.6	0.56	0.67
スズキ	72	5.83	5.86	4.67	5.2
ヒラメ	18	9.03	10.19	10.32	9.29
ブリ	21	11.93	12.24	9.09	9.84
マアジ	25	3.37	3.94	4.18	5.39
マコガレイ	28	5.17	5.72	6.61	5.81
マサバ	39	3.38	3.7	3.4	3.55
マハゼ	129	3.33	4.01	2.24	2.2

4つの機械学習手法の中で最も誤差の少ない数値を赤字で表記した。

5. おわりに

本稿では，筆者らの研究成果を中心に，メタボロミクス研究における深層学習の適用例について紹介してきた。代謝混合物の解析に深層学習を利用する試みはまだ始まったばかりであるが，今後盛んに研究開発が推進されるものと期待される。特に，筆者の専門とするNMRを用いたメタボロミクス研究では，試料調製が簡単なこと，計測時間が比較的短いこと，さらに機関間の互換性がある[19, 20]（つまり，世界中のどこの研究機関・企業などでサンプルを計測しても比較可能なNMRスペクトルが得られる）ことから，真の意味でのビッグデータとしてスペクトルデータを蓄積してくことが可能である。実際に，オートサンプラーなどの開発・充実化が進んでいるとともに，世界的なネットワークの形成やデータベース化により，2025年にはヘルスケア分野において1000万サンプルの血清のNMR計測データが蓄積されると予測されていることから[21]，深層学習を応用したバイオマーカーの発見や個別化医療への応用など，今後の発展が大いに期待できる。また，医療分野に限らず，食品産業やバイオ産業においても計測データの蓄積・ビッグデータ化は進んできており，計測装置の小型化・低価格化に伴う簡易分析システムの研究開発が進んでいることも考慮すると，将来的には，深層学習とメタボロミクスデータを活用した農畜水産資源の品質管理技術の開発や，食品産業における機能性や付加価値の創造，廃棄物の再資源化技術などへの応用展開が期待できる。

参考文献

1）Gartland, K. P. R. *et al.*：*J. Pharm. Biomed. Anal.*, **8**, 963-968（1990）
2）Gartland, K. P. R. *et al.*：*Mol. Pharmacol.*, **39**, 629-642（1991）
3）Lindon, J. C. *et al.*：*Prog. Nucl. Magn. Reson. Spectrosc.*, **39**, 1-40（2001）
4）Date, Y. *et al.*：*J. Proteome Res.*, **11**, 5602-5610（2012）
5）Date, Y. *et al.*：*Food Chem.*, **152**, 251-260（2014）
6）Gromski, P. S. *et al.*：*Anal. Chim. Acta*, **879**, 10-23（2015）
7）Mahadevan, S. *et al.*：*Anal. Chem.*, **80**, 7562-7570（2008）
8）Shima, H. *et al.*：*Nutrients*, **9**, 1307（2017）
9）Xia, J. *et al.*：*Nucleic Acids Res.*, **37**, W652-660（2009）
10）Cacciatore, S. *et al.*：*Proc. Natl. Acad. Sci. U. S. A.*, **111**, 5117-5122（2014）
11）Chatzimichali, E. A. and C. Bessant：*Metabolomics*, **12**, 16（2016）
12）Asakura, T. *et al.*：*Anal. Methods*, **10**, 2160-2168（2018）
13）Tsutsui, S. *et al.*：*J. Comput. Aided Chem.*, **18**, 31-41（2017）
14）Ito, K. *et al.*：*Chem. Sci.*, **9**, 8213-8220（2018）
15）Date, Y. and J. Kikuchi：*Anal. Chem.*, **90**, 1805-1810（2018）
16）Long, N. P. *et al.*：*Sci. Rep.*, **7**, 8552（2017）
17）Alakwaa, F. M. *et al.*：*J. Proteome Res.*, **17**, 337-347（2018）
18）Asakura, T. *et al.*：*Anal. Chim. Acta*, **1037**, 230-236（2018）
19）Dumas, M. E. *et al.*：*Anal. Chem.*, **78**, 2199-2208（2006）
20）Viant, M. R. *et al.*：*Environ. Sci. Technol.*, **43**, 219-225（2009）
21）Soininen, P. *et al.*：*Circ. Cardiovasc. Genet.*, **8**, 192-206（2015）