

生命科学研究を加速する機械学習

小林徹也，舟橋 啓，杉村 薫

近年，生命科学の各領域で，機械学習や深層学習を活用した研究の重要性がさかんに提唱されている．一方で，手法の原理が十分に理解されないままに機械学習，深層学習に対して過剰な期待が寄せられ，研究の実状と乖離してきているのも事実である．本増刊号は，バイオインフォマティクス，画像解析，細胞・組織・個体の統計解析の各領域における，最先端の機械学習の研究事例を紹介することで，こうした期待と実状のギャップを少しでも解消することを目的に企画された．本稿では本書の構成を紹介するとともに，機械学習や深層学習を用いたデータ駆動型研究について，われわれが考える展望と懸念を共有したい．

1. はじめに

ここ数年，生命医科学の分野において機械学習や深層学習を使ったいわゆるデータ駆動型研究への期待が高まっている．実際，深層学習や自然言語処理（natural language processing：NLP），コンピュータービジョンなどの手法を含む生物学の論文数は2017年以降，毎年1.5倍ずつ増えており，2020年は2万件を超える論文の発表が見込まれるという統計もある¹⁾（図1）．

歴史的に，機械学習や深層学習は自然画像認識や機械翻訳などの知見が蓄積され，大量のデータを集めやすい分野を中心に発展し，相対的に利用可能なデータの少ない生命医科学は，臨床診断・創薬など産業に直結した領域を除いて，主要なターゲットではなかった．しかし近年，機械学習が適用可能な生命科学のデータが整備・蓄積されてきたこと^{2) 3)}，画像認識や言語処理の最先端が研究室レベルでは利用不可能な大規模計算機資源を必要とするようになり⁴⁾，飽和した情報系研究者が新たなターゲットを模索しはじめたことなどもあって，遅ればせながら，生命医科学にいわゆるAIのブームが訪れている．特に，過去20年間における次世代シーケンサーとイメージングの2つの計測技術の発展は，生命科学の現場にこれまでにない多次元かつ多量の定量データを提供し，この流

【略語】

AI：artificial intelligence（人工知能）

DNN：deep neural network（ディープニューラルネットワーク）

ML：machine learning（機械学習）

NLP：natural language processing（自然言語処理）

scRNAseq：single cell RNA sequencing（1細胞RNAシーケンシング）

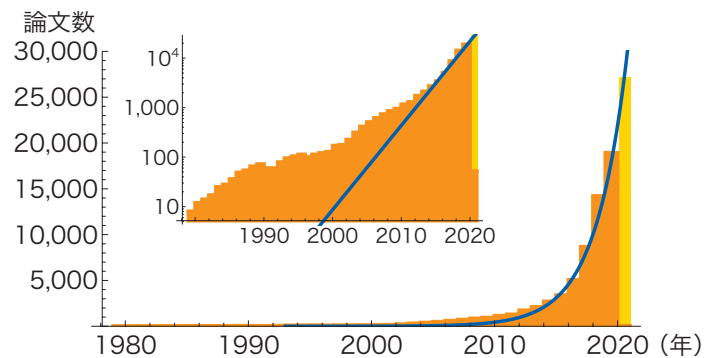


図1 生命科学分野で機械学習、深層学習を含む論文数の急速な増加

文献1を参考に、PubMedよりmachine learning, deep learningなどのキーワードをもつ論文を抽出。黄色は2021年分の直近数年の情報の外挿に基づく予測。青線は $y=a+x^{1.48}$ の曲線。挿入図は同じグラフの対数表示。

れを後押ししている。これが、この原稿執筆時点（2020年11月）の現状であろう。

しかしながら、機械学習や深層学習に対する期待の高まりや「データ駆動」というキーワードの普及とは裏腹に「機械学習や深層学習で実際何ができて、何ができないのか?」「どうすれば研究に利用できるのか?」「どのような分野に活用できるのか?」という分野の根幹が、生命科学の現場ではっきりと目に見える形にはなっていない。これは情報・数理系側から見ても同様で、生命科学のこういった問題にデータサイエンスの手法が有効に使えるのかは、実はまだよくわからない。そしてこの混乱のなか、「仮説によらない生命科学の実現」や「機械学習や深層学習の難しいことはわからなくても何となく使えるツール」、「定量データとコンピューターを使った客観的で透明性の高い研究の実現」といった、現時点での機械学習や深層学習の性質を踏まえると過剰宣伝ともとれるような言説も流布しつつある。

このような背景を受け、本書は機械学習や深層学習を用いた実際の研究例を紹介することで、期待と実際の現場の状況との乖離を少しでも解消することを目的として企画された。生命医科学に機械学習や深層学習がどれくらい有効に使えるかは各研究者による試行錯誤の段階にあり、全体を体系的に俯瞰・学習できるような枠組みは確立していない。そこで、緩やかに関連する複数のテーマを取り上げ、それぞれを専門にする筆者に分野の最先端の流れを紹介してもらい、読者が自身の研究に関係する内容を選択し読んでいただくという形式にした。

以下では、各テーマの間の関係やそれらを取り上げた意図を概説したうえで、最後にデータ駆動型研究の今後とその課題について、われわれがもつ展望と懸念をまとめた。

2. 本書の構成と各記事の紹介

まず最初に断らなければいけないこととして、本書では基礎生命科学への応用に焦点を当て、医科学系の研究はあえて扱わなかった。これは、医科学研究では医療統計学や医療画像解析など古くから続く研究の流れが現在の機械学習や深層学習の応用に接続しており、それらをカバーすることはわれわれの専門と本書の誌面の範囲では難しかったからである。また、系統学、進化生態学においても同様に、フィッシャーから続く連綿とした統計の流れと現代的な統計・学習理論の応用があるが、こちらも今回は割愛させていただいた。われわれの力不足をこの分野の関係者に陳謝するとともに、これらに特化した同様の特集が企画されることを期待したい。

分子・細胞・組織・個体を対象とした生命科学にフォーカスすると、機械学習や深層学習の主要なターゲットは、次世代シーケンサーとイメージングの計測技術から得られるデータである。前者はバイオインフォマティクス研究の流れに、後者はシステム生物学や定量生物学研究の流れに属する。さらに最近では、1細胞RNAシーケンシング(scRNAseq)の台頭やデータサイエンスの応用を背景に、2つの流れは合流しつつある。本書では、この2つの分野をそれぞれ第2章・第3章として大きくくくり、それらから発展した研究や関連する研究を第4章にまとめた。加えて、読者が機械学習や深層学習を用いた研究をはじめのきっかけとなるように、導入の第1章ではGoogle ColaboratoryやImageJを用いたデータ解析の基礎を解説した。

第1章では、舟橋(第1章-1)がgoogleが提供するPythonのプラットフォームGoogle Colaboratoryを用いた深層学習による画像分類を、三浦(第1章-2)がImageJとDeep ImageJを用いた画像解析を具体例やコードを交えて紹介する。基礎的な内容ではあるが、どのようなステップを踏んでデータ解析を進めるのかについて具体的なイメージがつかめると思う。

第2章では、1細胞遺伝子発現解析を中心に免疫・微生物・化学・創薬などバイオインフォマティクスから派生する多様なトピックへの機械学習の応用を紹介する。前半は急速に発展するscRNAseqの解析を扱い、松本(第2章-1)が低次元化やクラスタリングを用いた細胞タイプの同定を、前原(第2章-2)が細胞タイプの遷移動態の推定を、本田(第2章-3)が細胞の空間情報の推定・再構成の最新結果を示す。若本(第2章-4)は、非侵襲的なラマン分光計測から通常は侵襲的にしか得られないトランスクリプトームを機械学習を用いて推定するシーケンス・イメージング・機械学習の融合技術について解説する。後半では、シーケンス技術の免疫細胞群解析への応用を横田(第2章-5)が、微生物メタゲノムへの応用を森(第2章-6)が扱う。そして、岩田(第2章-8)・山西(第2章-7)がそれぞれケモインフォマティクスと薬剤応答推定への機械学習の最

新研究事例を示す。

第3章では、画像解析における機械学習と深層学習の応用を少し体系的に概観する。前半ではまず、山田（第3章-1）が細胞画像の分類問題を紹介し、続いて、高尾（第3章-2）が深層学習を用いて画像から細胞周期を分類・予測できることを示す。備瀬と内田（第3章-3）は、完全な教師データが得られない場合に重要となる弱教師・半教師データや誤りのある学習データの取り扱いを解説する。後半では、徳岡（第3章-4）が深層学習を用いた3D画像からの細胞セグメンテーションの最先端研究を紹介し、備瀬（第3章-5）が細胞トラッキングの機械学習・深層学習手法を示す。最後に、個体行動解析の現場を革新しつつあるDeepLabCutの内容を概説する（第3章-6）。

画像からの分子や細胞、個体の同定と追跡は機械学習応用の一大トピックではあるが、それだけでは解析は完結しない。第4章では、斉藤（第4章-1）と近藤（第4章-2）がそれぞれ画像から細胞形態や応力・物性の情報を統計・機械学習を活用して取り出す最先端の手法を解説する。続いて、画像から同定・追跡された分子もしくは細胞のデータをもとに機械学習や物理モデルのシミュレーションなどを組合わせて、潜在状態の同定と状態遷移の推定を行う手法を、岡本、松永、中島（第4章-3～5）が概説する。そして、個体の意思決定や学習などの高次情報処理の情報を行動データから見出す方法を木村、本田（第4章-6, 7）が示す。

本書で現れる機械学習手法は図2にまとめられている。さまざまな手法が使われていることがわかるが、アドバンスドな内容については用語解説で、基本的な内容については補遺で簡単にではあるが解説しているので、学習の手がかりにいただけたら幸い

特徴量 二乗誤差最小化 ブートストラップ法	統計学・学習理論	ハミング距離 レーベンシュタイン距離 マハラノビス距離 交差エントロピー Jensen-Shannon ダイバージェンス	データ間距離	混合行列 マッシュアップ関係数 サポートベクトルマシン 線形判別解析 (LDA) 決定木 ランダムフォレスト	クラス分類
ベイズの定理・ベイズ推定 EM アルゴリズム 変分ベイズ法 ABC 法	ベイズ理論	主成分分析 (PCA) tSNE 多次元尺度構成法 (MDS) UMAP 正準相関解析 (CCA) オートエンコーダー 非負値行列因子分解 (NMF) テンソル分解 経験モード分解	低次元化	線形回帰 ロジスティック回帰 Lasso 回帰 部分的最小二乗回帰	回帰
Permutation Test フィッシャーの正確確率検定 交差検定・交差検証 AUC	検定・評価・検証	k-means 法 k-medoids 法 スペクトラルクラスタリング 混合ガウスモデル	クラスタリング	マルコフ決定過程 ベルマン方程式 逆強化学習	強化学習
線形計画法・整数計画法 勾配法 最適輸送	最適化法	隠れマルコフモデル カルマンフィルター・スモーカー 粒子フィルター	時系列モデル	ニューラルネットワーク 畳み込みニューラルネットワーク フーリエ畳み込みニューラルネットワーク U-Net Adam Grad-CAM 転移学習	深層学習
疑似逆行列 最大エントロピー法 正則化	逆問題				

図2 本書で紹介されている統計・機械学習、深層学習手法の一覧

である。

3. データ駆動研究の課題と展望

機械学習や深層学習の活用は生命医科学の可能性を確実に大きく広げる。しかし、その未来は必ずしもバラ色な側面だけではない。最後に、機械学習や深層学習を用いたデータ駆動研究の今後について、われわれの展望とそこに付随する懸念を共有したい。

1) ツールとしての機械学習の普及

本書で紹介した研究などから派生し、機械学習や深層学習が生命医科学のツールとして普及してゆくことは確実である。まず、目視で行われることも多い画像データからの細胞追跡などは機械学習手法を活用してさらなる大規模化・自動化が進むだろう。また、より複雑な機械学習手法を用いて多量のシーケンスデータから、詳細な解析に値する遺伝子やゲノム領域、細胞タイプ、化合物の候補を列挙することも堅実な研究の展開である。さらに、イメージング・シーケンスの計測技術と機械学習が相互に結びついた手法の発展と多様化が見込まれる。若本らの手法が代表だが、本書では取り上げられなかった例として、機械学習による処理を前提としたイメージソーティングシステムとして太田らによるゴーストサイトメトリーがある⁵⁾。これらの高度な技術の開発と並行して、初学者にも使いやすい機械学習の解析ツールやプラットフォーム、分野・トピックごとのお決まりの機械学習手法も増えていくはずである。

このように、機械学習を活用した新しい解析手法の普及は確かにデータに基づく新たな研究を切り開くことになるだろうが、他方で、手法の原理を十分に理解しないまま、ツールの表面的な使用方法だけが普及することは、Photoshopを使った画像データ加工と同様の問題も確実に生み出すと、われわれは考えている。本書の記事や補遺を読めば明らかなように、機械学習や深層学習の手法は一見してユーザには可視化されない仮説やパラメータの塊である。また、学習に用いるデータしだいで解析結果が大きくバイアスされることは、基礎科学や産業応用のさまざまな局面で解決すべき問題として認識されている⁶⁾。これらの留意事項を把握せずに、解析の結果のみを指標にして望む結果が得られるように都合よく、モデルを選択したり、パラメータを操作したり、学習データを恣意的に選べば、Photoshopを用いてウエスタンブロットのバンドのコントラストを都合よく強調したように、恣意的なデータのプレゼンテーションに使われてしまう。また、その恣意的な操作を見つけだすことはPhotoshopの場合と比較して、格段に難しくなるだろう。「定量データとコンピューターによる客観的な生命科学」などの標語も見受けられるが「機械学習＝客観的」では全くないのである。

2) 機械学習による新たな科学的発見

ツールとしての普及とあわせた次の展開は、機械学習を人間のデータ解釈や実験計画の補助としてだけ使うのではなく、機械学習の結果を直接的に生物学的な知見・発見と同一視することだろう。実際すでに scRNAseq の解析では、クラスタリングや低次元化の結果をもとに、これまで見つかっていなかった新しい細胞群の存在を示唆する結果が多数報告されている。このようなアプローチは、画像データから細胞形状・動態の新しいモードを発見することなどにも応用が可能である（第4章-1, 3~5も参照のこと）。高次元データと組合わされば人間の認知を超えた多次元データからの科学的発見が可能になる。

しかし、機械学習の結果は前述のようにデータの測り方や手法の原理（仮定、パラメータなど）に依存する。これらの細胞群は本当に生物的に意味のある存在なのだろうか？それらを取り出してきて検証できればいいが、侵襲的なシーケンス技術を用いると、解析後に特定の細胞群を再採取することはできない。もしその細胞群が多数の遺伝子発現の複雑なパターンで規定されている場合は、セルソーターなどで分離することも至難であろう。機械学習で得られた結果を慎重に検証することなく「生物学的な結果・発見」とみなす流れは、有意差が生物学的意味と同一視された仮説検定の乱用やp値ハックと同じような問題をくり返しかねない⁷⁾。このような懸念を解消するためには、機械学習や深層学習も仮説検定と同種の問題をはらんでいることを認識したうえで利用すること、機械学習手法を実験検証手法とセットで発展させることが肝要である。

3) 真のデータ駆動型科学に向けて

ともあれ、現象にかかわる分子や遺伝子・細胞などモノを純粋にデータ科学から発見することを超えて、現象を支配する法則、例えば運動方程式やナビエ・ストークス方程式のような物理法則、を人間の手を介さずに発見する機械学習手法の開発は、生命科学のみならず科学分野全体で挑戦されている⁸⁾。実際それはデータ科学にかかわる研究者の夢でもあり、真の「データ駆動型科学」なのかもしれない。しかし、残念ながら現在の機械学習や深層学習の手法はデータの背後の真理や法則を明らかにするようにはまだデザインされていない。むしろ、そのような目的を手放し、手持ちの教師データから未観測のデータを予測する方向に主眼を移すことで現代的な統計学や学習理論は発展してきた。そこで用いられる確率モデルはデータの背後の法則を取り出そうとするものではなく、ニューラルネットワークのように多様なデータを柔軟に表現できる表現性の高いものである。特殊なものを除いて、データに関する因果的な情報を反映するわけでもない。

この問題はAIや数理・情報学分野でも認識され、データの背後の因果的情報や法則を取り出すことは最重要課題の1つとして取り上げられている⁹⁾。そこでは生命科学でおなじみの現象への介入なども重要な役割を果たす。また、得られる生物的法則は物理や化

学に基づく制約と整合的でなければならない(第4章-2も参照)。そのためにはニュートン力学からの流れをくむ物理的・化学的素過程を数理モデルで表現し実験データと組合わせる手法と統計・情報の流れをくむデータ科学とが何らかの形で融合を果たす必要がある。いまだその形は茫漠としているものの、われわれはデータ科学の流れをくむバイオインフォマティクスと物理モデリングの流れをくむ定量科学研究の合流点がその1つの反応点であると考えている^{10) 11)}。実際、本書でも第2章のscRNAseqの研究と第4章の細胞の定量データに基づく研究に今後の方向性の一致を見出すことができる。本書が、生命科学への機械学習や深層学習の健全な普及のみならず、このような新たな研究を志す若手研究者の一助となれば幸いである。

謝辞

本書および取り上げられている個々の研究は、定量的な生命科学研究や生命科学と情報数理学との融合研究に対する過去10年以上にわたる継続的なサポートに依拠している。最後に、このような研究テーマを志す若手をはじめとした研究者を支えていただいた以下の研究教育への支援制度やコミュニティに改めて感謝をしたい：文部科学省生命動態システム科学推進拠点事業，科学技術振興機構さきがけ・CREST研究（生命システム・生命モデル・生命動態・1細胞・多細胞・情報計測・数学協働・数理情報活用基盤），科研費 新学術領域（細胞コミュニティ・レゾナンスバイオ・生命の情報物理），理化学研究所生命システム研究センター QBiC，東京大学生命普遍性機構，定量生物学の会。

文献

- 1) Benaich N & Hogarth I. “State of AI Report 2020”, <https://www.stateof.ai/>
- 2) RxRx: Datasets released from Recursion’s automated cellular imaging and deep learning platform, <https://RxRx.ai/>
- 3) Broad Bioimage Benchmark Collection (BBBC), https://bbbc.broadinstitute.org/image_sets
- 4) Thompson NC, et al : arXiv:2007.05558, 2020
- 5) Ota S, et al : Science, 360 : 1246–1251, 2018
- 6) Beyer L, et al : arXiv:2006.07159, 2020
- 7) Head ML, et al : PLoS Biol, 13 : e1002106, 2015
- 8) Udrescu SM & Tegmark M : Sci Adv, 6 : eaay2631, 2020
- 9) Schölkopf B : arXiv:1911.10500, 2020
- 10) 実験医学 2013年5月号「生命システムを定量する！—未知なる現象の謎を解け！」(小林徹也/企画)，羊土社，2013
- 11) 小林徹也：数理科学，58：43–49，2020

<著者プロフィール>

小林徹也：東京大学生産技術研究所准教授。2000年東京大学工学部計数工学科卒業。’05年同大学院新領域創成科学研究科博士課程修了。博士（科学）。日本学術振興会特別研究員（DC1, PD），（独）理化学研究所基礎特別研究員を経て，’08年より東京大学生産技術研究所に講師として着任，’10年より同研究所准教授。その間，JST さきがけ研究者・University College London 訪問研究員を兼任。生体の情報処理を数理・情報の立場から定量的な実験と統合させて理解する研究に取り組む。

舟橋 啓：慶應義塾大学理工学部生命情報学科准教授。1995年慶應義塾大学理工学部電気工学科卒業。2000年同大学大学院理工学研究科博士課程計算機科学専攻修了。博士（工学）。日本学術振興

会特別研究員（DC1）、三重大学工学部助手、JST ERATO北野共生システムプロジェクト研究員を経て現在、慶應義塾大学理工学部生命情報学科准教授。システム生物学および定量生物学における計算機基盤の研究に従事。定量生物学の会コアメンバー・世話人。

杉村 薫：京都大学高等研究院iCeMS特定拠点准教授。上村匡京都大学教授の指導のもと、樹状突起のパターンの多様性を生み出すメカニズムの研究で博士（理学）取得。理化学研究所基礎科学特別研究員などを経て現職。機械的な力による多細胞秩序形成原理の研究に取り組む。