

## はじめに

人工知能（AI）という言葉が人口に膾炙して久しい。アミノ酸配列から立体構造を極めて正確に予測できる AlphaFold の登場は言うに及ばず、近年の生命科学系や医歯薬学系のトップジャーナルには毎号のようにコンピューター解析手法をメインにした報告が掲載されており、その多くは広義の AI が使われている。書店に行けば Python の入門書や機械学習・深層学習の書籍が山積みになっているし、インターネットにはさまざまなチュートリアルがあるので、それらを勉強すればすぐにも自分で解析を始められそうな気がしてくる。しかしこういった既存の書籍やチュートリアルはいずれも生命科学とはほど遠い内容を題材としており、特に周りに詳しい人がいない読者にとっては自分自身で機械学習を生命科学研究に使うのは容易ではない。実際、羊土社から 2020 年に発行された実験医学増刊『機械学習を生命科学に使う！』では機械学習の活用事例がオムニバスのように解説されているが、読者からは実践的なコードが欲しいという多数の要望が出版社に寄せられたとのことだった。

そこで企画されたのが本書である。2021 年に羊土社から発行された『独習 Python バイオ情報解析』等の入門書を読み終えたくらいの大学院生～比較的若手の先生方を対象に、**実践的な Python コード例を通して生命科学データを用いた機械学習を学んでいただける書籍**として、画像解析はもちろん、ゲノミクスやシングル解析、創薬などライフサイエンスのさまざまな領域に機械学習をどのように使うのかを体得できることを目指している。今日からでも使えるクラウド計算環境 Google Colaboratory 上で体験しつつ深層学習に不可欠な GPU を用いた解析にも親しむことができるよう配慮した。既存の多くの機械学習入門と異なり、**本書で扱うデータセットは、まさに読者が実際の研究で出会うものばかり**であり、これらを学ぶ中で**生命科学研究における Python の高度な利用法に習熟しつつ、同時に新規課題に対する応用力を身につけられる**ことは間違いない。

本書は**基礎編・実践編・発展編**、そして**終章**に大きく分けられる。まず基礎編においては**第 1 章**で機械学習とは何かを学ぶところから始まり、**第 2 章**で Python を用いたバイオ情報解析を駆け足で復習した後、**第 3 章～第 5 章**にかけて、**がん**と正常のトランスクリプトームパターンを題材に機械学習と深層学習を実際に動かしながら基本となるコンセプトを学んでいく。

続く実践編では、各章読み切りで重要トピックスについて各領域の若手研究者の先生方に執筆いただいた。まず最も AI が活用されているといってもよい画像解析については**第 6 章**で医歯薬学系の 2 つの題材をもとに学び、単なる分類だけではなく画像セグメンテーションや説明可能な AI といった近年のトピックスもあわせて紹介する。**第 7 章**ではゲノムへの AI の応用例として腫瘍ネオ抗原を題材に見ていき、**第 8 章**ではシングルセル解析と生成モデルについて取り扱う。このようなゲノムやトランスクリプトームといった各階層の解析例を第 7・8 章で見た後、**第 9 章**ではエピゲノムも合わせた複数階層の情報を統合しながらがんに向かっていく。**第 10 章**では近年の翻訳 AI 等にも幅広く使われている Transformer という仕組みを使いながらタンパク質の「言語」を読み解き、AI によるタンパク工学の可能性を示す。実践編最後の**第 11 章**では 4 つのプロジェクトに取り組むことで創薬 AI の基本かつ王道を学べるようになっている。

実践編の後ろに用意したのが発展編で、すでに情報解析の経験がある読者にも楽しんでいただけるよう、より advanced な内容を第一線でご活躍されている先生方に執筆いただいた。第12章では近年注目を集める中分子創薬についてAIから迫り、第13章では高度なメタゲノム解析・ゲノムマイニングを学んでいこう。これら2つの章は発展的な内容であるので、本格計算のためにはGoogle Colaboratoryを超える計算リソースが必要になる。このように実践編までの内容の「その先」の高みを見据える読者のために執筆したのが終章で、極めて高度な計算リソースを使う方法を含め、発展的なトピックス・リソースを余すところなく示し、本書を卒業したあとの羅針盤となるよう設計した。

このように本書1冊のみで生命科学における機械学習入門から発展的な内容までを実践的なPythonコードとともに俯瞰でき、本書で学ぶことで読者の研究はさらに大きく加速するだろう。本書の趣旨に賛同くださり、編者からの細かな注文にも快くご対応して下さったすべての執筆者の先生方のご指導の賜物である。また、本書を出版する機会をくださった羊土社の本多正徳氏と早河輝幸氏に心から感謝申し上げる。

次世代シーケンサーやさまざまな網羅解析手法が発展するにつれて、今後ますます生命科学と情報科学の境界はなくなっていくだろう。ぜひ本書で学んだ後は後輩・新人教育の教材の1つとしても使っていただきたい。第2～5章および第8章の筆頭執筆者は編者がオンラインで主催している学生勉強会Biomedical Data Science Clubで学ぶ学部生の方々だが、特に若い世代の方こそ機械学習を次世代のライフサイエンス研究の心強い相棒にさせていただきたいと願っている。読者の生命科学研究のレパートリーに機械学習手法を加えていただき、またその知見を後輩に引き継いでいただくことで、本書が新たな知の創出に貢献できれば執筆者の1人として望外の喜びである。

2023年2月

東京医科歯科大学M&Dデータ科学センター  
AIシステム医科学分野  
清水 秀幸