FRONTIER SCIENCES

生命を読み解くための新しい情報科学と機械学習

研究室では生命科学に資する機械学習や機械発見の方法をコンピュータサイエンスとして研究しています。機械学習は現在の「AI」の中核技術で、生命の多様な情報が計測できるようになった現在、それらを読み解く新しい情報科学を作るための基礎になります。

コンピュータプログラムは、私が最も確実だと実感できる対象で、数学的形式システムを体現するものです。プログラムが期待した通りに動作しないならば、設計者である私が間違えているからで、プログラミングを通して、自分の認識や思考がいかにいい加減で頼りないものかを日々思い知らされます。

対照的に、もう一つの関心である生命は、不確実で掴みどころのないものです。物質としては代謝によって絶えず入れ替わっているのに、世界は放っておくと乱雑になるというエントロピー増大則に抗って高度な秩序を保っているように見えます。物理学者シュレディンガーはこれを「生命は負のエントロピーを食べている」と表現しました。一方、数学者ゲルファントは「物理学における数学の不合理な有効性よりも、さらに不合理なものが一つだけある。それは生物学における数学の不合理な非有効性だ」と言いました。生物情報科学とは、ふつうには生物学データをコンピュータ解析するという情報科学の一応用分野に過ぎません

生命科学研究系
Division of Biosciences

瀧川 一学　教授

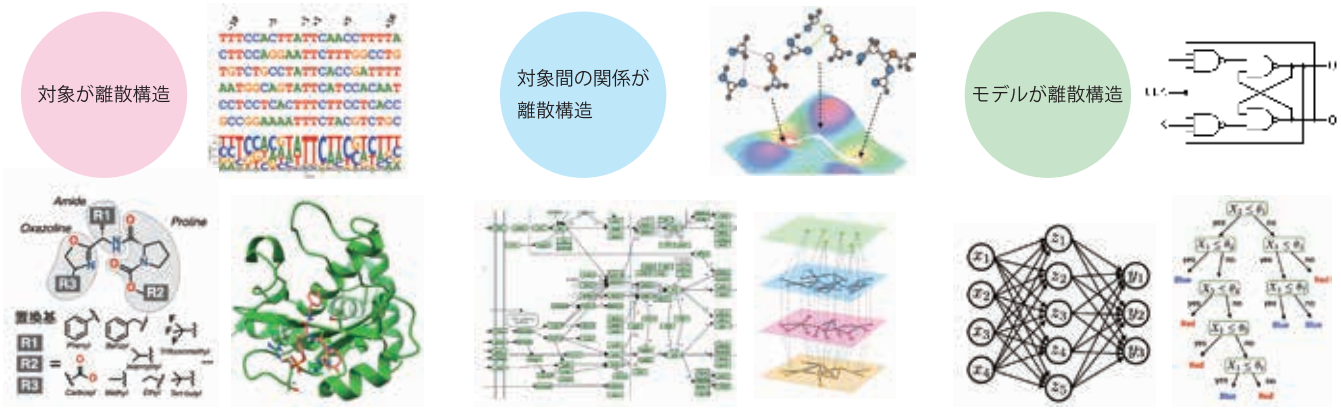
TAKIGAWA Ichigaku

メディカル情報生命専攻 データ駆動知能分野

https://takigawa-lab.tokyo/

離散構造を伴う機械学習

離散構造=有限個の離散要素の「組み合わせ」によって生じる情報。集合、論理、系列、ツリー、グラフ、プログラム、言語など。



が、私にとっては、どう考えても折り合わなさそうなこの両極端な概念（計算と生命）のミッシングピースを希求する新しい情報科学です。

現在の主な関心は「離散構造を伴う機械学習」と「自然科学における機械発見」です。離散構造とは「有限の離散的要素が組み合わされて生じる構造」のことで、集合、論理、シーケンス、ツリー、グラフやネットワークなどの組合せ的な情報構造です。生命を捉えるには構成要素そのものより、「各要素がどのように組み合わされるか」が本質的です。ゲノム配列、分子構造、遺伝子や化学反応のネットワークなど、生命科学は離散構造データに溢れています。機械学習ではふつうは多変量の数値データを対象とし、その確率分布を考えます。一方、離散構造として得られるデータ（例えばゲノム配列）は非数値的であり、その平均値や確率分布とは何なのか（何であるべきなのか）は非自明で、定義から考える必要があります。このように離散構造を統計的に扱うにはさまざまな情報科学の技術課題解決が必要で、離散数学・アルゴリズム・確率統計が融合する、とても楽しい研究トピックです。

もう一つの関心の「自然科学における機械発見」とは、機械学習を上手に使って実際に科学的「発見」を目指すものです。機械学習予測は多数のデータを平均で補間したもので、平均点しか取れません。平均点とは最も平凡であり、科学者が望む「今わかっていないこと・今ないもの」から遠いものです。機械発見は、出来合いの機械学習をただ応用するだけでは不十分であり、「データにないこと」を「データに基づいて」希求する無理設定を技術的に埋める新しいトピックです。

現在、AI技術は画像、音声、言語を処理できるさまざまな実用サービスとして身近な存在になりましたが、科学研究に真に資するためには、まだまだ発展途上の機械学習そのものの情報科学研究も重要だと考えています。

▲機械学習を用いた超強力水中接着剤開発の論文はNature誌の表紙に掲載されました。