TOYOTA CENTRAL R&D LABS NEWS

2019 年（平成 31 年）2 月 28 日
株式会社豊田中央研究所

タンパク質やDNAの構造・機能予測のための新しい機械学習技術を開発

当社の小出智士、河野圭祐、沓名拓郎は、DNAやタンパク質などのバイオデータからその構造や機能を予測するための新たな機械学習技術を開発しました。

タンパク質やDNAはアミノ酸や塩基からなる配列情報を持ち、さらに構造を形成することで生命の基盤となる様々な機能を発揮します。生物進化により形成された、これらのバイオ配列と構造や機能の相関は、科学が進歩した現代においても未だ解明されていない自然科学分野の命題のひとつです。この相関関係を予測することは、医療分野やエネルギー分野（バイオ燃料）の発展にとって極めて重要と考えられます。近年は機械学習技術を利用したデータ駆動型の手法に大きな期待が寄せられていますが、現実のデータに利用するためには予測精度の向上等、まだいくつものハードルがあります。従来の研究では、ニューラルネットワークやサポートベクターマシンなどの機械学習分野における既存技術が利用されてきました。本研究では、二つの配列の類似度計算に用いる「配列アライメント」というバイオインフォマティクス分野の既存アルゴリズムを「畳み込みニューラルネットワーク」 (CNN) に組み込み、新たに「編集不変ニューラルネットワーク」を構築しました。「配列アライメント」を組み込んだのは、この手法がバイオ配列の進化の過程で生じ得る文字の置換・欠損・挿入（編集操作）をモデル化したものであり、バイオ配列の構造・機能予測モデルに適していると期待したためです。

そこで、タンパク質の二次構造予測を、従来法(CNN)と「編集不変ニューラルネットワーク」の両者で行い予測精度を比較したところ、学習データ数が少ないときほど、「編集不変ニューラルネットワーク」の方が従来法 (CNN) よりも精度が高くなることが実証されました。このことは、「編集不変ニューラルネットワーク」が、バイオ配列の構造・機能予測モデルに適した学習効率の高いアーキテクチャであることを示しています。さらに本研究では、上述したような文字列の編集操作とCNNの関係について、いくつかの理論的な考察を行い、ニューラルネットワークのデザインに関する方向性を示しました。

今後は、この技術を用いた予測モデルが、バイオ分野にとどまらず、文字列を扱う自然言語処理やマテリアルズインフォマティクス等の他分野にも応用されることが期待されます。

本成果は国際会議 Neural Information Processing Systems 31 に採択され、発表を行いました (2018/12/4付)。
(https://papers.nips.cc/paper/7744-neural-edit-operations-for-biological-sequences)

＜本件に関するお問い合わせ先＞　株式会社豊田中央研究所　広報室
https://www.tytlabs.co.jp/contact/toiawase.html