

研究課題 (テーマ)		利用可能な全てのウイルスと細菌ゲノムの大規模言語モデル化、及びそれを用いたウイルス遺伝子の機能推定	
研究者	所属学科等	職	氏 名
代表者	生物工学科	助教	杉本竜太
分担者			
研究結果の概要			
現在、公共データベースには膨大な数の細菌およびウイルスのゲノム情報が登録されているが、その多くの遺伝子は機能不明のままである。特にウイルス遺伝子の大半は既知の遺伝子と相同性を示さず、従来の配列比較に基づく方法では機能を推定することが難しい。そこで本研究では、自然言語処理の技術である大規模言語モデルを応用し、ウイルス遺伝子の機能を推定するための機械学習モデルを開発した。中でも RNA ウイルスの複製に必須な酵素、RNA 依存性 RNA ポリメラーゼ (RdRP) に着目し、従来の手法では検出できなかった新規 RdRP を識別可能な分類モデルを構築した。 モデルの学習には、高品質かつ多様な訓練データが必要である。申請者は、ヒト腸内、海水、淡水、下水などの環境由来の RNA シーケンスを公共データベースから取得・アセンブリし、そこから既知の RdRP に類似するアミノ酸配列を抽出した。さらに、既存のウイルスゲノムデータベースからも RdRP 配列を収集し、多様性に富んだ訓練データセットを整備した。最終的に、約 4 万件の RdRP 配列と、RdRP に類似しない約 30 万件のその他のタンパク質配列を訓練用データとして使用した。 これらの配列データは、事前学習済みの言語モデル (ESM-C) を用いてベクトル化し、申請者が設計したニューラルネットワークに入力して分類精度が最大となるよう訓練を行った。データは訓練用、検証用、テスト用に分割され、各ステップで精度をモニタリングした。検証データで最も高精度を示したモデルを最終版として採用し、テストデータによってその汎化性能を評価した結果、未知の配列に対しても高い識別精度を維持できることが確認された。 このモデルを用いて、前述の環境サンプルから得られた約 400 万件のタンパク質配列すべてに対して RdRP である確率を予測した結果、約 2 万 3 千件の新たな RdRP 候補を検出した。中には、既存の分類体系に含まれない新規のグループと考えられるものも含まれていた。系統解析の結果、これらは既知の RdRP とは系統的に大きく異なっており、ウイルスの多様性に関する新たな知見が得られた。以上より、本研究は機械学習を活用することで、従来の配列比較に基づいた方法では見落とされていた RdRP を網羅的に発見できることを示したものであり、他のウイルスや細菌の未知遺伝子の機能推定にも応用可能であることが示唆された。			
今後の展開			
今後はこの手法を拡張し、RdRP 以外のウイルスタンパク質や未知機能タンパク質の分類・同定にも対応できる汎用モデルの開発を目指す。生物工学や創薬などへの応用を視野に入れた基盤技術として発展させていく予定である。			