

研究課題（テーマ）	機械学習モデルを用いた太閤池由来 RNA ウイルスゲノムコレクションの構築		
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

本助成課題では、太閤池由来 RNA ウイルスの完全長ゲノム構築に向け、RNA ウイルスの必須酵素である RNA 依存性 RNA ポリメラーゼ (RdRP) を検出する機械学習プログラム PalmSite を開発した。PalmSite は、タンパク質をベクトル表現に変換した後、それを基に酵素活性中心の位置と、それが RdRP

である確率を出力する (図 1)。
既知の RdRP アミノ酸配列を用いた評価で、PalmSite は高い精度と再現率を達成した (ROC-AUC=0.999、PR-AUC=0.9904) (図 2)。また、活性中心も高精度に認識しており (図 3)、モデルが RdRP の酵素活性中心のパターンを適切に学習したことが示された。
PalmSite は、既存データベースに未登録の未知 RNA ウイルス由来 RdRP も検出できることを確認している。

PalmSite は Github
(<https://github.com/ryota-sugimoto/palmsite>)及び PyPI

(<https://pypi.org/project/palmsite>) で公開済みであり、誰でも自由に使える状態となっている。現在、これらの成果をまとめた論文を執筆中であり、近日中に査読付き学術誌へ投稿する予定である。

PalmSite の開発で得られた知見を活用し、今年度は、単一遺伝子ではなく複数遺伝子の並びから機能を推定する新たなゲノムコンテキストモデルの開発を進めている。今後は、遺伝子の並びを学習したモデルを用いて、未知遺伝子の機能推定に加え、新規ゲノムの生成や実験的検証を目指す。

今後の展開

今後は、PalmSite を太閤池の RNA-seq/メタトランスクリプトームデータに適用し、新規 RNA ウイルス候補を抽出する。分節型 RNA ウイルスについては RNA クロスリンク解析を併用してセグメント対応を推定し、完全長ゲノムコレクションの構築を進める。時系列サンプルを用いて、ゲノム構造やウイルス集団の変動も解析する。

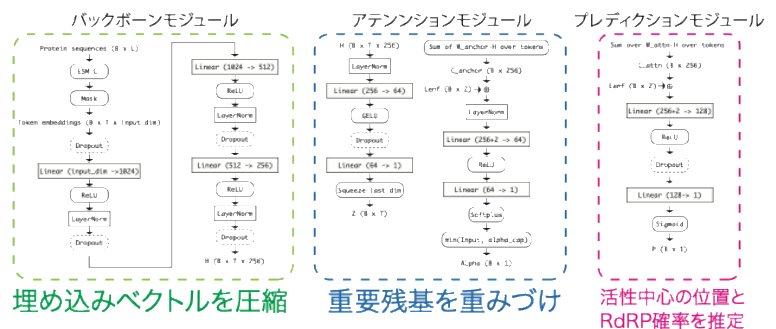


図 1. PalmSite の機械学習レイヤーの概要

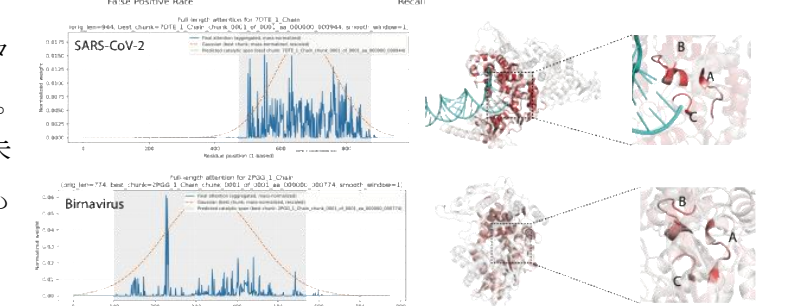
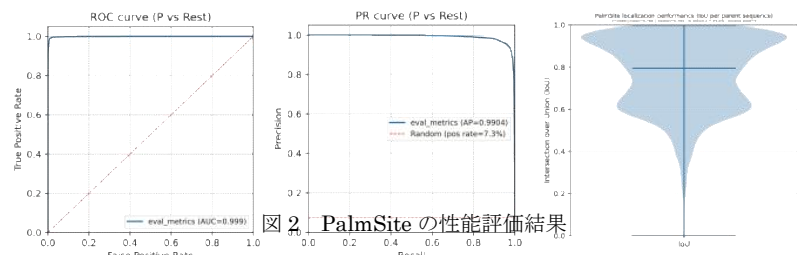


図 3 実際の RNA 依存性 RNA ポリメラーゼを用いた推定結果