2024 年度　利用報告書

課題名（タイトル）: Optimization of crRNA to activate Cas activity by machine learning

利用者氏名:○山崎 大介(1)

理研における所属研究室名:(1) 開拓研究本部 渡邊分子生理研究室

1.　本課題の研究の背景、目的、関係する課題との関係

当研究室では高感度かつ非増幅なウイルス RNA 検出法である SATORI 法を開発し、社会実装に向けてハードウェア、ソフトウェアの両面において実験系の改良を進めている。SATORI 法における RNA 検出は、ウイルス RNA に特異的に結合する crRNA が Cas13 ヌクレアーゼを活性化することにより、ウイルス RNA を切断する機構を利用している。この反応中に核酸蛍光レポーターを共存させておくと、一部のレポーターが切断され蛍光を発するため、特定のウイルス RNA を高感度に検出することができる。SATORI 法の検出感度の重要な因子となるのが crRNA 配列であり、この配列を最適化すると数千倍～数万倍まで Cas13 を活性化できることがこれまでのスクリーニングによりわかっている。crRNA 配列の重要性は SATORI 法のみならず、他の Cas9, Cas12 タンパク質の活性化にも重要であることから、機械学習を利用した crRNA の最適化や規則性の解明がなされているが、教師あり学習が主たるアプローチであり、データセットの不足や適切なアーキテクチャ・計算資源の不足もあり、決定的な解析がなされていないのが現状である。そこで本研究課題では、バクテリア DNA の大規模言語モデルである Evo モデルの転移学習により、これまでのスクリーニングで得られた crRNA 配列と Cas13 活性を教師データとして渡すことで crRNA 配列から Cas13 の活性を予測するモデルを構築する。

2.　具体的な利用内容、計算方法

Evo モデルは長文の文脈理解のために Transformer アーキテクチャをさらに改良した StripedHyena アーキテクチャで配列情報を読み込むため、モデルサイズが 10GB 以上と大きく、1つの 80G GPU 上で訓練を実行することを推奨している。このため、これまで私が利用してきた GoogleColab が提供する 40G GPU では訓練が実行できず、Hokusai の GPU サーバが提供する 80G GPU を利用させていただくこととした。訓練としては、スクリーニングで得られた crRNA 配列とそれに紐づく Cas13 活性を教師データとする教師あり学習を行い、crRNA 配列から Cas13 活性を予測する回帰モデルの構築を目指す。

3.　結果

利用申請が 2 月初旬であり、スーパーコンピュータの利用も初めてであったため、利用方法の理解や環境構築が中心であり、GPU サーバの利用機会も限られていたことから、計算はあまり実行できていない。3 月からようやくジョブ投入できるようになり、訓練が実行できるようになったため、訓練パラメータの調整やモデルアーキテクチャの検討・改良を進めていく。

4.　今後の計画・展望

Evo モデルの転移学習で得られた訓練結果について分析し、ハイパーパラメータの最適化を通じて回帰モデルの構築を目指す。また、データセットの更新や新しくリリースされる DNA 言語モデルについても同様に訓練を実施し、DNA モデルのみならず、Cas13 タンパク質の改良も視野に入れた DNA/タンパク質複合モデルの構築を目指す。